

*AG  
T*

*Algebraic & Geometric  
Topology*

Volume 23 (2023)

Issue 3 (pages 963–1462)



# ALGEBRAIC & GEOMETRIC TOPOLOGY

msp.org/agt

## EDITORS

### PRINCIPAL ACADEMIC EDITORS

John Etnyre  
etnyre@math.gatech.edu  
Georgia Institute of Technology

Kathryn Hess  
kathryn.hess@epfl.ch  
École Polytechnique Fédérale de Lausanne

### BOARD OF EDITORS

Julie Bergner	University of Virginia jeb2md@eservices.virginia.edu	Robert Lipshitz	University of Oregon lipshitz@uoregon.edu
Steven Boyer	Université du Québec à Montréal cohf@math.rochester.edu	Norihiko Minami	Nagoya Institute of Technology nori@nitech.ac.jp
Tara E. Brendle	University of Glasgow tara.brendle@glasgow.ac.uk	Andrés Navas	Universidad de Santiago de Chile andres.navas@usach.cl
Indira Chatterji	CNRS & Université Côte d'Azur (Nice) indira.chatterji@math.cnrs.fr	Thomas Nikolaus	University of Münster nikolaus@uni-muenster.de
Alexander Dranishnikov	University of Florida dranish@math.ufl.edu	Robert Oliver	Université Paris 13 bobol@math.univ-paris13.fr
Corneli Druţu	University of Oxford cornelia.drutu@maths.ox.ac.uk	Birgit Richter	Universität Hamburg birgit.richter@uni-hamburg.de
Tobias Ekholm	Uppsala University, Sweden tobias.ekholm@math.uu.se	Jérôme Scherer	École Polytech. Féd. de Lausanne jerome.scherer@epfl.ch
Mario Eudave-Muñoz	Univ. Nacional Autónoma de México mario@matem.unam.mx	Zoltán Szabó	Princeton University szabo@math.princeton.edu
David Futер	Temple University dfuter@temple.edu	Ulrike Tillmann	Oxford University tillmann@maths.ox.ac.uk
John Greenlees	University of Warwick john.greenlees@warwick.ac.uk	Maggy Tomova	University of Iowa maggy-tomova@uiowa.edu
Ian Hambleton	McMaster University ian@math.mcmaster.ca	Nathalie Wahl	University of Copenhagen wahl@math.ku.dk
Hans-Werner Henn	Université Louis Pasteur henn@math.u-strasbg.fr	Chris Wendl	Humboldt-Universität zu Berlin wendl@math.hu-berlin.de
Daniel Isaksen	Wayne State University isaksen@math.wayne.edu	Daniel T. Wise	McGill University, Canada daniel.wise@mcgill.ca
Christine Lescop	Université Joseph Fourier lescop@ujf-grenoble.fr		

---

See inside back cover or [msp.org/agt](http://msp.org/agt) for submission instructions.


The subscription price for 2023 is US \$650/year for the electronic version, and \$940/year (+ \$70, if shipping outside the US) for print and electronic. Subscriptions, requests for back issues and changes of subscriber address should be sent to MSP. Algebraic & Geometric Topology is indexed by Mathematical Reviews, Zentralblatt MATH, Current Mathematical Publications and the Science Citation Index.

Algebraic & Geometric Topology (ISSN 1472-2747 printed, 1472-2739 electronic) is published 9 times per year and continuously online, by Mathematical Sciences Publishers, c/o Department of Mathematics, University of California, 798 Evans Hall #3840, Berkeley, CA 94720-3840. Periodical rate postage paid at Oakland, CA 94615-9651, and additional mailing offices. POSTMASTER: send address changes to Mathematical Sciences Publishers, c/o Department of Mathematics, University of California, 798 Evans Hall #3840, Berkeley, CA 94720-3840.

---

AGT peer review and production are managed by EditFlow<sup>®</sup> from MSP.

PUBLISHED BY

 **mathematical sciences publishers**  
nonprofit scientific publishing

<http://msp.org/>

© 2023 Mathematical Sciences Publishers

# Projective naturality in Heegaard Floer homology

MICHAEL GARTNER

Let  $\text{Man}_*$  denote the category of closed, connected, oriented and based 3–manifolds, with basepoint preserving diffeomorphisms between them. Juhász, Thurston and Zemke showed that the Heegaard Floer invariants are natural with respect to diffeomorphisms, in the sense that there are functors

$$HF^\circ : \text{Man}_* \rightarrow \mathbb{F}_2[U]\text{-Mod}$$

whose values agree with the invariants defined by Ozsváth and Szabó. The invariant associated to a based 3–manifold comes from a transitive system in  $\mathbb{F}_2[U]\text{-Mod}$  associated to a graph of embedded Heegaard diagrams representing the 3–manifold. We show that the Heegaard Floer invariants yield functors

$$HF^\circ : \text{Man}_* \rightarrow \text{Trans}(P(\mathbb{Z}[U]\text{-Mod}))$$

to the category of transitive systems in a projectivized category of  $\mathbb{Z}[U]$ –modules. In doing so, we will see that the transitive system of modules associated to a 3–manifold actually comes from an underlying transitive system in the projectivized homotopy category of chain complexes over  $\mathbb{Z}[U]\text{-Mod}$ . We discuss an application to involutive Heegaard Floer homology, and potential generalizations of our results.

57M27, 57R58

## 1 Introduction

The Heegaard Floer invariants associated to closed, oriented 3–manifolds were defined in the work of Ozsváth and Szabó [11]. There it was shown that to each such 3–manifold, one can associate an isomorphism class of  $\mathbb{Z}[U]$ –module. Furthermore, cobordisms between 3–manifolds were shown to induce maps between the invariants; see Ozsváth and Szabó [14]. However, there was a gap in the proof of the naturality of these maps. Showing that these invariants are natural with respect even to diffeomorphisms is subtle, and involves detailed consideration of the dependence of the invariants on the choices of Heegaard data, basepoints and embeddings of Heegaard diagrams involved in their construction.

These subtleties were studied extensively by Juhász, Thurston and Zemke in [5]. There they explicated a particular type of loop of Heegaard moves, simple handleswaps, which previous work did not preclude from potentially yielding monodromy in the Heegaard Floer invariants. Moves analogous to these simple handleswap moves were previously studied in detail and suggested as possible candidates for loops with monodromy in the work of Sarkar; eg in [18]. Through a careful analysis of a space of embedded Heegaard diagrams, Juhász, Thurston and Zemke exhausted all possible monodromies and obstructions to the Heegaard Floer assignments being natural with respect to diffeomorphisms, and were then able to provide a minimal set of requirements which could be checked to verify such naturality. They then checked that these requirements are satisfied for all variants of Heegaard Floer homology with coefficients in  $\mathbb{F}_2$ . By building on the work in [14] and [5], Zemke described in [20] the dependence of the cobordism maps defined in [14] on basepoints. Using this dependence, Zemke completed the verification of the fact that the cobordism maps are in fact natural (over  $\mathbb{F}_2$ ) with respect to composition of cobordisms (when the cobordisms are appropriately decorated with graphs).

In this paper we explain the necessary modifications that must be made to obtain naturality with respect to diffeomorphisms of all variants of Heegaard Floer homology, but with coefficients in  $\mathbb{Z}$ . The most immediate goal of our work is simply to fill a gap in the literature. We hope this will be useful both as a resource for nonexperts who aim to understand Heegaard Floer homology itself, and as groundwork which can be used to better understand other invariants associated with Heegaard Floer homology. For example, the contact invariants defined by Ozsváth and Szabó in [13] have proven to be extremely effective in detecting subtle contact properties, and both their definition and many of their applications require the ability to nail down particular elements in the modules  $HF^\circ$ , and the ability to effectively compare two such elements in the same module. We also note that the results in [5] and the analogous integral results presented here are necessary steps for establishing naturality of the integral Heegaard Floer invariants with respect to cobordisms.

## 1.1 Statement of main results

In order to study naturality of many flavors of Heegaard Floer homology and knot Floer homology simultaneously, Juhász, Thurston and Zemke work with sutured 3-manifolds. They consider a graph  $\mathcal{G}$  which encodes the combinatorial structure of a space of sutured Heegaard diagrams related by certain Heegaard moves. Roughly, the

vertices of  $\mathcal{G}$  correspond to isotopy diagrams of sutured manifolds, and between any two such isotopy diagrams there are edges which describe whether they are related by any of the standard Heegaard moves, or additionally whether they are related by a diffeomorphism. The graph  $\mathcal{G}$  contains many sutured isotopy diagrams which are not relevant to the consideration of closed 3-manifolds, so in considering the closed 3-manifold invariants  $HF^\circ$  attention is restricted to a subgraph  $\mathcal{G}(\mathcal{S}_{\text{man}})$ . This is the full subgraph of  $\mathcal{G}$  whose vertices consist only of those isotopy diagrams representing sutured manifolds which can be constructed from a closed 3-manifold in a prescribed way. Since we are only concerned with results regarding closed 3-manifolds in this paper, we will minimize the role of sutured manifolds, and phrase our results in terms of a graph which is isomorphic to  $\mathcal{G}(\mathcal{S}_{\text{man}})$  which we denote by  $\mathcal{G}_{\text{man}}$ . This graph has vertices corresponding to isotopy diagrams of closed, pointed 3-manifolds, where the isotopies are required to be supported away from the basepoint. Edges in  $\mathcal{G}_{\text{man}}$  correspond to sequences of handleslides, stabilizations and diffeomorphisms.

To study naturality using these graphs, we consider the two notions of a *Heegaard invariant* introduced in [5]. The first, a *weak Heegaard invariant* valued in a category  $\mathcal{C}$ , is simply a morphism of graphs from  $\mathcal{G}_{\text{man}}$  to  $\mathcal{C}$  under which all edges in the domain get mapped to isomorphisms. In this language, we can summarize one of the invariance results shown in [11] as stating that the morphisms of graphs

$$HF^\circ: \mathcal{G}_{\text{man}} \rightarrow \mathcal{C}$$

for  $\mathcal{C} = \mathbb{Z}[U]\text{-Mod}$  or  $\mathcal{C} = \mathbb{F}_2[U]\text{-Mod}$  determined by Heegaard Floer homology are weak Heegaard invariants. The second notion, that of a *strong Heegaard invariant*, serves as a minimal set of conditions which are needed to ensure that a weak Heegaard invariant yields a natural invariant of the underlying 3-manifolds; precisely, the authors show that the image of a strong Heegaard invariant  $HF^\circ: \mathcal{G}_{\text{man}} \rightarrow \mathcal{C}$ , when appropriately restricted, forms a transitive system in  $\mathcal{C}$ . This step occupies a majority of the work in the paper, and none of the results in this step depend on the target category  $\mathcal{C}$ . The authors then prove that, in the case when  $\mathcal{C} = \mathbb{F}_2[U]\text{-Mod}$ , such a transitive system yields a functor

$$HF^\circ: \text{Man}_* \rightarrow \mathbb{F}_2[U]\text{-Mod}.$$

Finally, they establish that  $HF^\circ: \mathcal{G}_{\text{man}} \rightarrow \mathbb{F}_2[U]\text{-Mod}$  is in fact a strong Heegaard invariant, completing their proof that the invariants  $HF^\circ$  yield functors from  $\text{Man}_*$  to  $\mathbb{F}_2[U]\text{-Mod}$ .

Our main goal here is to establish similar results for  $\mathcal{C} = P(\mathbb{Z}[U]\text{-Mod})$ , the quotient category obtained from  $\mathbb{Z}[U]\text{-Mod}$  by the relation  $f \sim -f$  for all  $f \in \text{Hom}_{\mathbb{Z}[U]\text{-Mod}}$ . Said simply, we want to show that naturality holds over  $\mathbb{Z}$ , up to a sign. We will consider a category  $\text{Trans}(P(\mathbb{Z}[U]\text{-Mod}))$  of transitive systems in  $P(\mathbb{Z}[U]\text{-Mod})$ , and our main result will be:

**Theorem 1.1** *There are functors*

$$\widehat{HF}, HF^-, HF^+, HF^\infty : \text{Man}_* \rightarrow \text{Trans}(P(\mathbb{Z}[U]\text{-Mod}))$$

whose values on a based 3-manifold  $(Y, z)$  are isomorphic to the modules defined in [11]. Furthermore, isotopic diffeomorphisms have the same image under  $HF^\circ$ .

**Remark 1.2** The finite-rank variant  $HF_{\text{red}}$  of Heegaard Floer homology defined in [11, Definition 4.7] arises as a suitable quotient (or submodule) of  $HF^\pm$ , and Theorem 1.1 implies that this variant also yields a functor  $HF_{\text{red}} : \text{Man}_* \rightarrow \text{Trans}(P(\mathbb{Z}[U]\text{-Mod}))$ .

We will import wholesale the logical structure of [5] used to prove the analog of Theorem 1.1 appearing there. It will therefore suffice to show that

$$HF^\circ : \mathcal{G}_{\text{man}} \rightarrow P(\mathbb{Z}[U]\text{-Mod})$$

is a strong Heegaard invariant. We will in fact show something slightly stronger. Let  $\text{Kom}(\mathbb{Z}[U]\text{-Mod})$  denote the homotopy category of chain complexes over  $\mathbb{Z}[U]\text{-Mod}$ , and, as described above, let  $P(\text{Kom}(\mathbb{Z}[U]\text{-Mod}))$  denote the projectivization of this category. Finally, let  $\text{Trans}(P(\text{Kom}(\mathbb{Z}[U]\text{-Mod})))$  denote the category of transitive systems in  $P(\text{Kom}(\mathbb{Z}[U]\text{-Mod}))$ . We will unpack the precise meaning of these categories in Section 4. A majority of the paper will be occupied with showing:

**Theorem 1.3** *The morphisms*

$$\widehat{CF}, CF^-, CF^+, CF^\infty : \mathcal{G}_{\text{man}} \rightarrow \text{Trans}(P(\text{Kom}(\mathbb{Z}[U]\text{-Mod})))$$

are strong Heegaard invariants.

While proving Theorem 1.3 we will show the analogous result holds on the level of homology:

**Corollary 1.4** *The morphisms*

$$\widehat{HF}, HF^-, HF^+, HF^\infty : \mathcal{G}_{\text{man}} \rightarrow P(\mathbb{Z}[U]\text{-Mod})$$

are strong Heegaard invariants.

We will establish Theorem 1.3 in Sections 7 and 8. We will also obtain from Theorem 1.3 the following statement about the constituent chain complexes.

**Corollary 1.5** *Given a closed, connected, oriented and based 3–manifold  $(Y, z)$  and a  $\text{Spin}^c$ –structure  $\mathfrak{s}$  over  $Y$ , the  $\mathbb{Z}[U]$ –module chain complexes  $CF^\circ(\mathcal{H}, \mathfrak{s})$ , ranging over all strongly  $\mathfrak{s}$ –admissible embedded Heegaard diagrams  $\mathcal{H}$  for  $(Y, z)$ , fit into a transitive system of homotopy equivalences in  $P(\text{Kom}(\mathbb{Z}[U]\text{–Mod}))$  with respect to the maps induced by sequences of pointed handleslides, stabilizations, isotopies, and diffeomorphisms of Heegaard surfaces which are isotopic to the identity in  $Y$ .*

**Remark 1.6** The Heegaard Floer invariants arise as direct sums of invariants

$$HF^\circ(Y, z) = \bigoplus_{\mathfrak{s} \in \text{Spin}^c(Y)} HF^\circ(Y, z, \mathfrak{s})$$

associated to triples  $(Y, z, \mathfrak{s})$  for  $\mathfrak{s} \in \text{Spin}^c(Y)$ . All of the main results have refined statements regarding these invariants of  $(Y, z, \mathfrak{s})$ . Theorem 1.3 and Corollaries 1.4 and 1.5 also depend on choices of coherent orientation systems, which we omit from the statements here. For now, we note that all of the results above hold in particular for the Heegaard Floer chain complexes defined with respect to the canonical coherent orientation systems constructed by Ozsváth and Szabó in [10]. The precise conditions required of the coherent orientation systems implicitly appearing in the results above will be specified in Definition 6.14.

## 1.2 Further directions and applications

We now point out some applications and potential generalizations of our results. Given two based 3–manifolds  $(Y_1, z_1)$  and  $(Y_2, z_2)$ , a cobordism  $W$  between them decorated with a choice of path in  $W$  from  $z_1$  to  $z_2$ , and a choice of  $\mathfrak{t} \in \text{Spin}^c(W)$ , Ozsváth and Szabó constructed in [14] cobordism maps

$$F_{W, \mathfrak{t}}^\circ: HF^\circ(Y_1, z_1, \mathfrak{t}|_{Y_1}) \rightarrow HF^\circ(Y_2, z_2, \mathfrak{t}|_{Y_2}).$$

(The choice of path is not made explicit in [14]). In [20], Zemke extended the results in [5] to show that over  $\mathbb{F}_2$  these maps are well defined and natural with respect to composition of decorated cobordisms. We expect that our results can be used in a similar way to establish such naturality over  $\mathbb{Z}$ , up to an overall sign. Furthermore, in [14], Ozsváth and Szabó showed how naturality of the Heegaard Floer invariants with respect to decorated cobordisms can be used to define the so called mixed invariants of

closed 4–manifolds. Given a closed 4–manifold  $X$  and a choice of  $\mathfrak{t} \in \text{Spin}^c(X)$ , these take the form of maps

$$\Phi_{X,\mathfrak{t}}: \Lambda^*(H_1(X; \mathbb{F}_2)/\text{Tors}) \otimes_{\mathbb{F}_2} \mathbb{F}_2[U] \rightarrow \mathbb{F}_2.$$

These share many of the features of the Seiberg–Witten invariants, and serve as powerful tools in detecting subtle smooth information. If one can establish naturality with respect to cobordisms over  $\mathbb{Z}/\pm$ , we would obtain corresponding mixed invariants

$$\Phi_{X,\mathfrak{t}}: \Lambda^*(H_1(X; \mathbb{Z})/\text{Tors}) \otimes_{\mathbb{Z}} \mathbb{Z}[U] \rightarrow \mathbb{Z}/\pm$$

which we expect would provide fruitful extra information. In fact, before the gap in the literature was noticed, the integral mixed invariants had already been extensively studied in papers including Jabuka and Mark [4], Ozsváth and Szabó [12] and Roberts [16], so establishing naturality with respect to cobordisms over  $\mathbb{Z}$  would immediately prove useful, and would likely also be useful for computations and applications in the future.

A second application of our work comes from involutive Heegaard Floer homology, defined by Hendricks and Manolescu in [3]. To describe it, fix a closed 3–manifold  $Y$  and  $\mathfrak{s} \in \text{Spin}^c(Y)$ . Given a pointed Heegaard diagram  $\mathcal{H} = (\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}, z)$  for  $(Y, z)$ , there is a conjugate diagram  $\bar{\mathcal{H}} = (-\Sigma, \boldsymbol{\beta}, \boldsymbol{\alpha}, z)$  for  $(Y, z)$  given by reversing the orientation on the surface and switching the role of the  $\boldsymbol{\alpha}$  and  $\boldsymbol{\beta}$  curves. Under suitable admissibility hypotheses, there is a chain isomorphism

$$\eta_{\mathcal{H} \rightarrow \bar{\mathcal{H}}}: CF^\circ(\mathcal{H}, \mathfrak{s}) \rightarrow CF^\circ(\bar{\mathcal{H}}, \bar{\mathfrak{s}})$$

given by mapping intersection points to themselves [10, Theorem 2.4]. Note that the role of coherent orientations here is not yet relevant, as Hendricks and Manolescu work over  $\mathbb{F}_2$ . Using the results in [5], Hendricks and Manolescu showed that the  $\mathbb{F}_2$  analog of Corollary 1.5 holds: the modules  $CF^\circ(\mathcal{H}, \mathfrak{s})$  fit into a transitive system in the homotopy category of chain complexes of  $\mathbb{F}_2[U]$ –modules with respect to the maps induced by the Heegaard moves appearing in Corollary 1.5. Thus, since  $\mathcal{H}$  and  $\bar{\mathcal{H}}$  represent the same 3–manifold, there is a chain homotopy equivalence

$$\Phi(\bar{\mathcal{H}}, \mathcal{H}): CF^\circ(\bar{\mathcal{H}}, \bar{\mathfrak{s}}) \rightarrow CF^\circ(\mathcal{H}, \mathfrak{s})$$

of complexes of  $\mathbb{F}_2[U]$ –modules which is well defined up to homotopy. Using these maps, they consider the map  $\iota := \Phi(\bar{\mathcal{H}}, \mathcal{H}) \circ \eta_{\mathcal{H} \rightarrow \bar{\mathcal{H}}}$ , which is well defined up to homotopy, and which is shown to be a homotopy involution in [3, Lemma 2.5]. They then use it to construct an invariant of  $Y$  as follows.



There is a  $\mathbb{Z}/2\mathbb{Z}$  action on  $\text{Spin}^c(Y)$  given by conjugation. Let  $[\text{Spin}^c(Y)]$  denote the set of orbits in  $\text{Spin}^c(Y)$  under this action. Given an orbit  $\bar{\omega} \in [\text{Spin}^c(Y)]$ , let

$$CF^\circ(\mathcal{H}, \bar{\omega}) = \bigoplus_{s \in \bar{\omega}} CF^\circ(\mathcal{H}, s).$$

The authors investigate the map  $(1 + \iota)$ , considered as a chain map between complexes of  $\mathbb{F}_2[U]$ -modules, and consider its cone

$$CFI(\mathcal{H}, \bar{\omega}) := \text{Cone}(1 + \iota) = \left( CF^\circ(\mathcal{H}, \bar{\omega})[-1] \oplus CF^\circ(\mathcal{H}, \bar{\omega}), \partial_{\text{cone}} = \begin{pmatrix} \partial & 0 \\ 1 + \iota & -\partial \end{pmatrix} \right).$$

Here  $CF^\circ(\mathcal{H}, \bar{\omega})[-1]$  indicates the shifted chain complex, whose degree  $n$  piece is given by  $(CF^\circ(\mathcal{H}, \bar{\omega})[-1])_n = CF^\circ(\mathcal{H}, \bar{\omega})_{n-1}$ . They then introduce a formal variable  $Q$  of degree  $-1$  satisfying  $Q^2 = 0$ , and rewrite the map being coned over as

$$CF^\circ(\mathcal{H}, \bar{\omega}) \xrightarrow{Q \cdot (1 + \iota)} Q \cdot CF^\circ(\mathcal{H}, \bar{\omega})[-1].$$

As one can readily check, the cone and its differential can then be rewritten as

$$(1) \quad \text{Cone}(1 + \iota) = (CF^\circ(\mathcal{H}, \bar{\omega})[-1] \otimes \mathbb{F}_2[Q]/(Q^2), \partial + Q(1 + \iota)).$$

Considered in this way, it is a complex of modules over the ring  $\mathcal{R} = \mathbb{F}_2[Q, U]/(Q^2)$ . The authors then show that the quasi-isomorphism class of the complex  $CFI(\mathcal{H}, \bar{\omega})$  of  $\mathcal{R}$ -modules thus defined is an invariant of  $(Y, \bar{\omega})$ .

We now explain how Corollary 1.5 can be used to construct a version of such an invariant defined over  $\mathbb{Z}$ . Before doing so, we make a remark on the reliance of the following discussion on orientation systems.

**Remark 1.7** First we note that the proof establishing that  $\eta$  is an isomorphism given in [10, Theorem 2.4] implicitly proves the statement with respect to an arbitrary coherent orientation system  $\sigma$  over the domain  $\mathcal{H}$  and, ostensibly, the *same* coherent orientation system over the codomain  $\bar{\mathcal{H}}$  (the use of the word *same* makes sense because the underlying diagrams for the domain and codomain of  $\eta$  are the same aside from labeling and orientations). However, to avoid this consideration we will simply focus attention here on the case where both diagrams are equipped with canonical orientation systems, as defined in [10]. We note that the maps  $\Phi$  take canonical orientation systems to canonical orientation systems, since more generally any sequence of maps induced by Heegaard moves takes a canonical orientation system to a canonical orientation system. This follows from the facts that Heegaard moves induce module isomorphisms on the totally twisted module  $\underline{HF}^\infty$  — see [10, Section 8] — and that the canonical

orientation system on a diagram can be characterized by the isomorphism type of  $HF^\infty$ ; see [10, Theorem 10.12]. Similarly, the isomorphism  $\eta$  can be defined with respect to canonical orientation systems on both diagrams. Indeed, since the proof of [10, Theorem 2.4] also shows that the map  $\eta$  yields an isomorphism between the totally twisted module  $HF^\infty$  associated to a diagram equipped with the canonical orientation, and the totally twisted module associated to the reversed diagram equipped with the induced orientation system, the induced orientation system in this case must be the canonical one. With these remarks in mind, we will omit all reference to coherent orientation systems from our notation and description; all remarks in the remainder of the description of this application apply only to the canonical orientation systems.

Fix again a 3-manifold  $Y$ , and diagrams  $\mathcal{H}$  and  $\bar{\mathcal{H}}$  representing  $Y$  as above. Since  $\mathcal{H}$  and  $\bar{\mathcal{H}}$  represent the same 3-manifold, we obtain from Corollary 1.5 (at most) two homotopy classes of chain homotopy equivalences

$$\pm\Psi(\bar{\mathcal{H}}, \mathcal{H}): CF^\circ(\bar{\mathcal{H}}, \bar{\mathfrak{s}}) \rightarrow CF^\circ(\mathcal{H}, \bar{\mathfrak{s}})$$

associated to sequences of Heegaard moves relating the two diagrams. The set  $\{\pm\Psi(\bar{\mathcal{H}}, \mathcal{H})\}$  is well defined up to chain homotopy. We thus obtain two homotopy classes of maps  $\pm\iota := \pm\Psi(\bar{\mathcal{H}}, \mathcal{H}) \circ \eta_{\mathcal{H} \rightarrow \bar{\mathcal{H}}}$ . The same argument used in [3, Lemma 2.5] to show that  $\iota$  is a homotopy involution over  $\mathbb{F}_2$  now shows that  $\pm\iota$  both have order at most 4 (up to homotopy) over  $\mathbb{Z}$ . We define

$$CFI_\pm(\mathcal{H}, \bar{\omega}) := \text{Cone}(1 \pm \iota),$$

where now both complexes are considered as complexes of  $\mathbb{Z}[U]$ -modules. While we can no longer conclude the maps  $\pm\iota$  are homotopy involutions, we still obtain that the collection of the two quasi-isomorphism classes of the complexes of  $\mathbb{Z}[U]$ -modules that we obtain is an invariant of the underlying 3-manifold.

**Theorem 1.8** *With respect to the canonical orientation systems of [10], the unordered pair of quasi-isomorphism classes determined by the complexes*

$$CFI_\pm(\mathcal{H}, \bar{\omega})$$

*(considered as complexes of  $\mathbb{Z}[U]$ -modules) is an invariant of  $(Y, \bar{\omega}, z)$ .*

**Proof** The proof is essentially the same as that in [3], but we include a sketch of it here for the reader’s convenience.

Fix  $(Y, z, \bar{\omega})$ , and consider a diagram  $\mathcal{H}$  and its conjugate  $\bar{\mathcal{H}}$  as above. As we noted earlier, for the fixed diagram  $\mathcal{H}$  the collection of the two chain homotopy equivalences  $\{\pm\Psi(\bar{\mathcal{H}}, \mathcal{H})\}$  is well defined up to chain homotopy by Corollary 1.5. Thus so too is the collection  $\{\pm\iota\}$ . We conclude that the set of the two cones  $\{CFI_{\pm}(\mathcal{H}, \bar{\omega})\}$  associated to  $(\mathcal{H}, \bar{\omega})$  is well defined up to chain homotopy equivalence.

Next, we consider the dependence on the choice of diagram. Consider a different diagram  $\mathcal{H}'$  for  $(Y, z)$  and its conjugate  $\bar{\mathcal{H}}'$ . We obtain corresponding collections  $\{\pm\Psi(\bar{\mathcal{H}}', \mathcal{H}')\}$  and  $\{\pm\iota'\}$  which are both well defined up to homotopy, and  $\{CFI_{\pm}(\mathcal{H}', \bar{\omega})\}$  well defined up to homotopy equivalence. Choose some fixed sequence of Heegaard moves connecting  $\mathcal{H}$  to  $\mathcal{H}'$ , and consider either of the (at most two) corresponding chain homotopy equivalences  $\pm\Psi(\mathcal{H}, \mathcal{H}')$  furnished by Corollary 1.5. We denote our choice by  $\Psi(\mathcal{H}, \mathcal{H}')$ . Consider the diagram, involving the four cone complexes in question,

$$(2) \quad \begin{array}{ccc} CF^{\circ}(\mathcal{H}, \bar{\omega})[-1] & \xrightarrow{1\pm\iota} & CF^{\circ}(\mathcal{H}, \bar{\omega}) \\ \downarrow \Psi(\mathcal{H}, \mathcal{H}') & & \downarrow \Psi(\mathcal{H}, \mathcal{H}') \\ CF^{\circ}(\mathcal{H}', \bar{\omega})[-1] & \xrightarrow{1\pm\iota'} & CF^{\circ}(\mathcal{H}', \bar{\omega}) \end{array}$$

We claim that for a fixed choice in  $\{\pm\iota\}$ , the diagram commutes up to homotopy for at least one of the two choices in  $\{\pm\iota'\}$ . We denote our choice of the fixed homotopy class in the top row by  $\iota$ . To establish the claim, we need to show that

$$\Psi(\mathcal{H}, \mathcal{H}') \circ \Psi(\bar{\mathcal{H}}, \mathcal{H}) \circ \eta_{\mathcal{H} \rightarrow \bar{\mathcal{H}}} \sim \pm\Psi(\bar{\mathcal{H}}', \mathcal{H}') \circ \eta_{\mathcal{H}' \rightarrow \bar{\mathcal{H}}'} \circ \Psi(\mathcal{H}, \mathcal{H}').$$

We note that

$$\eta_{\mathcal{H}' \rightarrow \bar{\mathcal{H}}'} \circ \Psi(\mathcal{H}, \mathcal{H}') \circ \eta_{\bar{\mathcal{H}} \rightarrow \mathcal{H}} \sim \pm\Psi(\bar{\mathcal{H}}, \bar{\mathcal{H}}').$$

To see this, observe that  $\Psi(\mathcal{H}, \mathcal{H}')$  is a map induced by some sequence of Heegaard moves. The map resulting from precomposing and postcomposing this map with the isomorphisms  $\eta$  can be realized as the map induced on  $CF^{\circ}(\bar{\mathcal{H}})$  by the same set of Heegaard moves giving rise to  $\Psi(\mathcal{H}, \mathcal{H}')$  (recall the maps  $\eta$  have no effect on the attaching curves). Thus the conjugated map is homotopic to  $\pm\Psi(\bar{\mathcal{H}}, \bar{\mathcal{H}}')$  by Corollary 1.5. We thus conclude that

$$\begin{aligned} \Psi(\bar{\mathcal{H}}', \mathcal{H}') \circ \eta_{\mathcal{H}' \rightarrow \bar{\mathcal{H}}'} \circ \Psi(\mathcal{H}, \mathcal{H}') &\sim \pm\Psi(\bar{\mathcal{H}}', \mathcal{H}') \circ \Psi(\bar{\mathcal{H}}, \bar{\mathcal{H}}') \circ \eta_{\mathcal{H} \rightarrow \bar{\mathcal{H}}} \\ &\sim \pm\Psi(\mathcal{H}, \mathcal{H}') \circ \Psi(\bar{\mathcal{H}}, \mathcal{H}) \circ \eta_{\mathcal{H} \rightarrow \bar{\mathcal{H}}} \end{aligned}$$

where the last two maps being homotopic up to a sign is also guaranteed by Corollary 1.5. Having established that the diagram with  $\iota$  in the top row commutes up to chain

homotopy for at least one choice of  $\{\pm\iota'\}$  in the bottom row, the argument in [3] now applies directly to establish that  $\text{Cone}(1 + \iota)$  is quasi-isomorphic to at least one of the cones  $\text{Cone}(1 \pm \iota')$ . This concludes the proof.  $\square$

In the case of rational homology three-spheres, the pair of quasi-isomorphism classes in Theorem 1.8 can actually be distinguished from one another to furnish two distinct invariants.

**Corollary 1.9** *Let  $Y$  be a rational homology three sphere. One can specify the maps  $\iota$  so that, with respect to the canonical orientation systems of [10], the quasi-isomorphism classes determined by*

$$CFI_+(\mathcal{H}, \bar{\omega}) \quad \text{and} \quad CFI_-(\mathcal{H}, \bar{\omega})$$

(considered as complexes of  $\mathbb{Z}[U]$ -modules) are each invariants of  $(Y, \bar{\omega}, z)$ .

**Proof** Since  $Y$  is a rational homology three sphere, for each  $\mathfrak{s} \in \text{Spin}^c(Y)$  we have

$$HF^\infty(Y, \mathfrak{s}) \cong \mathbb{Z}[U, U^{-1}]$$

as  $\mathbb{Z}[U]$ -modules by [10, Theorem 10.1].

Consider first the case of a  $\mathbb{Z}/2\mathbb{Z}$ -invariant  $\text{spin}^c$  structure. For each such  $\text{spin}^c$  structure  $\mathfrak{s}$ , the maps  $\pm\iota$  are homotopy equivalences, so induce graded module isomorphisms on  $HF^\infty(Y, \mathfrak{s})$ . Since  $HF^\infty(Y, \mathfrak{s}) \cong \mathbb{Z}[U, U^{-1}]$  there are precisely two such morphisms:  $\pm\text{Id}$ . For each  $\mathbb{Z}/2\mathbb{Z}$ -invariant  $\mathfrak{s}$ , choose  $\iota$  to be the map which induces  $-\text{Id}$  on  $HF^\infty(Y, \mathfrak{s})$ . This can be accomplished for all invariant  $\text{spin}^c$  structures even with a fixed choice of sign on each map  $\Psi(\bar{\mathcal{H}}, \mathcal{H})$ , by altering the signs of the maps  $\eta$  when necessary. Then the proof of Theorem 1.8 carries over directly to show the quasi-isomorphism class determined by

$$CFI_+(\mathcal{H}, \{\mathfrak{s}\})$$

is an invariant of  $(Y, \mathfrak{s}, z)$ . One must only note that the diagram (2) commutes with no sign ambiguity for  $\iota$  and  $\iota'$  specified by our definition. Indeed, the proof of Theorem 1.8 shows that the diagram commutes with  $1 + \iota$  on top for one of  $1 \pm \iota'$  on the bottom, but the diagram could not even commute at the level of homology if  $\iota$  induced  $-\text{Id}$  and  $\iota'$  induced  $\text{Id}$ . By the same argument,

$$CFI_-(\mathcal{H}, \{\mathfrak{s}\})$$

also yields an invariant.

Next consider a  $\mathbb{Z}/2\mathbb{Z}$ -orbit  $\bar{\omega} = \{\mathfrak{s}, \bar{\mathfrak{s}}\}$  coming from a pair of noninvariant  $\text{spin}^c$  structures. We have two homotopy equivalences,  $\iota_{\mathfrak{s} \rightarrow \bar{\mathfrak{s}}}: CF^\circ(\mathcal{H}, \mathfrak{s}) \rightarrow CF^\circ(\mathcal{H}, \bar{\mathfrak{s}})$  and  $\iota_{\bar{\mathfrak{s}} \rightarrow \mathfrak{s}}: CF^\circ(\mathcal{H}, \bar{\mathfrak{s}}) \rightarrow CF^\circ(\mathcal{H}, \mathfrak{s})$ . The total map

$$(1 + \iota): CF^\circ(\mathcal{H}, \mathfrak{s}) \oplus CF^\circ(\mathcal{H}, \bar{\mathfrak{s}}) \rightarrow CF^\circ(\mathcal{H}, \mathfrak{s}) \oplus CF^\circ(\mathcal{H}, \bar{\mathfrak{s}})$$

takes the form

$$(x, y) \mapsto (x + \iota_{\bar{\mathfrak{s}} \rightarrow \mathfrak{s}}(y), y + \iota_{\mathfrak{s} \rightarrow \bar{\mathfrak{s}}}(x)).$$

Define the signs on these maps such that  $\iota_{\bar{\mathfrak{s}} \rightarrow \mathfrak{s}} \circ \iota_{\mathfrak{s} \rightarrow \bar{\mathfrak{s}}}$  induces  $\text{Id}$  on  $HF^\infty(Y, \mathfrak{s})$  and  $\iota_{\mathfrak{s} \rightarrow \bar{\mathfrak{s}}} \circ \iota_{\bar{\mathfrak{s}} \rightarrow \mathfrak{s}}$  induces  $\text{Id}$  on  $HF^\infty(Y, \bar{\mathfrak{s}})$ . As above, the choice of signs can be incorporated into the definition of the maps  $\eta$ . The proof of Theorem 1.8 again shows this gives a well defined invariant

$$CFI_+(\mathcal{H}, \bar{\omega}).$$

Similarly, the choice where  $\iota_{\mathfrak{s} \rightarrow \bar{\mathfrak{s}}} \circ \iota_{\bar{\mathfrak{s}} \rightarrow \mathfrak{s}}$  and  $\iota_{\bar{\mathfrak{s}} \rightarrow \mathfrak{s}} \circ \iota_{\mathfrak{s} \rightarrow \bar{\mathfrak{s}}}$  both induce  $-\text{Id}$  gives a well defined invariant

$$CFI_-(\mathcal{H}, \bar{\omega}). \quad \square$$

**Remark 1.10** The two rational homology sphere invariants given in Corollary 1.9 give rise to distinct involutive Heegaard Floer homologies  $HFI^\infty(Y, \bar{\omega})$ . Namely, one can compute that

$$HFI_+^\infty(Y, \bar{\omega}) \cong \mathbb{Z}[U, U^{-1}, Q]/(Q^2)$$

while

$$HFI_-^\infty(Y, \bar{\omega}) \cong \mathbb{Z}/2\mathbb{Z}[U, U^{-1}].$$

To see this, consider the short exact sequence of chain complexes that results from the definition of  $CFI^\infty(Y, \bar{\omega}) = \text{Cone}(1 + \iota)$ ,

$$0 \rightarrow CF^\infty(\mathcal{H}, \bar{\omega}) \xrightarrow{i} CFI^\infty(Y, \bar{\omega}) \xrightarrow{p} CF^\infty(\mathcal{H}, \bar{\omega}) \rightarrow 0.$$

This gives rise a to a long exact sequence in homology

$$\dots \xrightarrow{p_*} HF^\infty(Y, \bar{\omega}) \xrightarrow{\delta} HF^\infty(Y, \bar{\omega}) \xrightarrow{i_*} HFI^\infty(Y, \bar{\omega}) \xrightarrow{p_*} HF^\infty(Y, \bar{\omega}) \xrightarrow{\delta} HF^\infty(Y, \bar{\omega}) \xrightarrow{i_*} \dots$$

for which the connecting morphism  $\delta$  is precisely the induced map  $(1 + \iota)_*$ .

Consider first the case of invariant  $\text{spin}^c$  structures. When  $\iota$  is chosen such that  $(1 + \iota)_* = 0$ , we get a split short exact sequence, so

$$HFI_+^\infty(Y, \{\mathfrak{s}\}) \cong HF^\infty(Y, \mathfrak{s}) \oplus HF^\infty(Y, \mathfrak{s}) \cong \mathbb{Z}[U, U^{-1}] \oplus \mathbb{Z}[U, U^{-1}].$$

Tracing through the identification analogous to that in (1), this gives

$$HFI_+^\infty(Y, \{\mathfrak{s}\}) \cong \mathbb{Z}[U, U^{-1}, Q]/(Q^2)$$

as a module over  $\mathbb{Z}[Q, U, U^{-1}]/(Q^2)$ . When  $\iota$  is chosen such that  $\delta = (1 + \iota)_* = 2$ , we instead obtain

$$HFI_-^\infty(Y, \{\mathfrak{s}\}) \cong \text{Coker}(\delta) \oplus \text{Ker}(\delta) \cong HF^\infty(Y, \mathfrak{s})/2 \cdot HF^\infty(Y, \mathfrak{s}) \cong \mathbb{Z}/2\mathbb{Z}[U, U^{-1}].$$

Here  $HFI_-^\infty(Y, \mathfrak{s})$  is a  $\mathbb{Z}[Q, U, U^{-1}]/(Q^2)$ -module where  $Q$  acts by zero.

In the case of noninvariant  $\text{spin}^c$  structures, we can use  $\iota_{\mathfrak{s} \rightarrow \bar{\mathfrak{s}}}$  to identify

$$HF^\infty(Y, \mathfrak{s}) \cong HF^\infty(Y, \bar{\mathfrak{s}})$$

and consider the map

$$\phi: HF^\infty(Y, \mathfrak{s}) \oplus HF^\infty(Y, \mathfrak{s}) \rightarrow HF^\infty(Y, \mathfrak{s}) \oplus HF^\infty(Y, \mathfrak{s})$$

defined by the composition  $\phi = (1 \oplus (\iota_{\mathfrak{s} \rightarrow \bar{\mathfrak{s}}})^{-1}) \circ (1 + \iota) \circ (1 \oplus \iota_{\mathfrak{s} \rightarrow \bar{\mathfrak{s}}})$ . More explicitly,

$$\phi(x, y) = (x + \iota_{\bar{\mathfrak{s}} \rightarrow \mathfrak{s}} \circ \iota_{\mathfrak{s} \rightarrow \bar{\mathfrak{s}}}(y), (\iota_{\mathfrak{s} \rightarrow \bar{\mathfrak{s}}})^{-1} \circ \iota_{\mathfrak{s} \rightarrow \bar{\mathfrak{s}}}(x + y)).$$

For  $CFI_+^\infty$ , we defined the constituent maps such that  $\iota_{\bar{\mathfrak{s}} \rightarrow \mathfrak{s}} \circ \iota_{\mathfrak{s} \rightarrow \bar{\mathfrak{s}}}$  induces  $\text{Id}$ , so this becomes

$$\phi(x, y) = (x + y, x + y)$$

and

$$HFI_+^\infty(Y, \{\mathfrak{s}, \bar{\mathfrak{s}}\}) \cong \text{Coker}(\delta) \oplus \text{Ker}(\delta) \cong \mathbb{Z}[U, U^{-1}] \oplus \mathbb{Z}[U, U^{-1}].$$

For  $CFI_-^\infty$ , we defined the constituent maps such that  $\iota_{\bar{\mathfrak{s}} \rightarrow \mathfrak{s}} \circ \iota_{\mathfrak{s} \rightarrow \bar{\mathfrak{s}}}$  induces  $-\text{Id}$ , so this becomes

$$\phi(x, y) = (x - y, x + y)$$

and

$$HFI_-^\infty(Y, \{\mathfrak{s}, \bar{\mathfrak{s}}\}) \cong \text{Coker}(\delta) \oplus \text{Ker}(\delta) \cong \mathbb{Z}/2\mathbb{Z}[U, U^{-1}].$$

The claimed structures as modules over  $\mathbb{Z}[Q, U, U^{-1}]/(Q^2)$  follows as above.

**Remark 1.11** It is plausible that Corollary 1.9 actually extends to the general case of closed, connected, oriented 3-manifolds. To specify an individual invariant in this general case would require a method by which one could naturally make a choice for signs on  $\iota$ . An approach here would be to make an argument like the one in the proof of Corollary 1.9, but by taking advantage of the standard form for the totally twisted module  $\underline{HF}^\infty(Y)$ , rather than the standard form for  $HF^\infty(Y)$  for rational homology spheres. Indeed, by [10, Theorem 10.12] the totally twisted module associated with any  $\text{Spin}^c$  structure is isomorphic to  $\mathbb{Z}[U, U^{-1}]$  (as a  $\mathbb{Z}[U, U^{-1}]$ -module). Using this fact,

one could presumably again pick out particular models for  $CFI_+$  and  $CFI_-$ . What would remain to be shown is that there are analogs to Theorem 1.1 and Corollary 1.5 for the totally twisted complexes, and that these results could be used to carry over the argument used in the proof of Theorem 1.8. We expect that the main results in this paper do carry over to the totally twisted complexes, but we leave investigation of this subtlety to the interested reader.

### 1.3 Organization of the paper

We begin in Section 2 by recalling the notion of sutured 3-manifolds and sutured Heegaard diagrams, as all of the results in [5] are phrased in this setting. We discuss a correspondence between sutured and closed 3-manifolds, and use the correspondence to translate a graph of sutured diagrams central to the setting of [5] into an equivalent graph of closed diagrams which we use throughout the remainder of the paper. In Section 3 we introduce and rephrase the notions of weak and strong Heegaard invariants defined in [5]. Section 4 deals with setting up the algebraic framework in which our main results are phrased, and in particular includes the definitions of the projectivizations and categories of transitive systems appearing in Theorems 1.1 and 1.3. In Section 5, we deduce Theorem 1.1 and Corollary 1.5 from Theorem 1.3 and Corollary 1.4. In Sections 6 and 7 we recall the constructions involved in defining the integral Heegaard Floer chain complexes, and establish that these constructions yield suitably defined weak Heegaard invariants. In Section 7, we check that these weak Heegaard invariants satisfy all but one of the axioms required of a strong Heegaard invariant. In Section 8 we carry out the main work and establish that these weak Heegaard invariants also satisfy the last axiom, known as simple handleswap invariance. Finally, in Section 9 we explain that the construction of the surgery exact triangle works without modification in our setting.

### Acknowledgements

I would like to thank the referee for providing many invaluable and detailed comments and suggestions, and in particular for their help in pointing out errors in the original manuscript and proposing alternative approaches where necessary. It is also my pleasure to thank Robert Lipshitz for his support and encouragement throughout the course of the writing of this paper, for many helpful conversations, and for all of his help as my graduate advisor.

I was partly supported during this work by NSF grant DMS-1810893.

## 2 Background

In order to introduce notation and terminology for the remainder of the paper, we give a quick summary of some relevant background on sutured manifolds and Heegaard diagrams. To unify the approach, the results in [5] are most often phrased in terms of sutured manifolds. Since we are interested here in the closed variants of Heegaard Floer homology, we will set up some background in order to be able to rephrase the results we use from [5] in language more typically used for the closed invariants.

To begin, we will describe how moves on sutured Heegaard diagrams relate to the typical Heegaard moves one considers on Heegaard diagrams for closed 3–manifolds. Next we will recall the definition of the graph of sutured isotopy diagrams  $\mathcal{G}(\mathcal{S}_{\text{man}})$  introduced in Section 1, and describe an isomorphism to a graph  $\mathcal{G}_{\text{man}}$  of closed isotopy diagrams which we will consider instead of  $\mathcal{G}(\mathcal{S}_{\text{man}})$  throughout the remainder of the paper. We refer the reader to [5, Section 2.1] for a more detailed treatment of all of the background in this section.

### 2.1 Background on sutured manifolds

In this paper we will be concerned primarily with closed 3–manifolds, but we will need to refer to numerous results about sutured 3–manifolds along the way. In particular, our results depend on notions of sutured 3–manifolds, sutured diagrams and embedded sutured diagrams for such manifolds, various notions of equivalence of such diagrams, and sutured Heegaard moves. While these notions may be standard, some inequivalent definitions certainly exist, so we explicitly refer the reader to [5] for background on the definitions we will use throughout this paper. We note that the sutured Heegaard moves play a role analogous to that of pointed Heegaard moves on Heegaard diagrams for closed 3–manifolds. There are moves called  $\alpha$  and  $\beta$  equivalences (which correspond to sequences of handleslides), as well as stabilizations and destabilizations, isotopies, and diffeomorphisms. Finally, we note that by restricting attention to the isotopy class of attaching curves on a diagram, one obtains a well-defined notion of a sutured *isotopy diagram*, and one can make sense of sutured Heegaard moves considered as moves on the isotopy diagrams (eg there is a well-defined notion of a diffeomorphism of isotopy diagrams). We again refer the reader to [5] for the relevant definitions of such sutured Heegaard moves; the main relevance here will be their relation to Heegaard moves on diagrams for closed 3–manifolds.



## 2.2 A correspondence between closed and sutured manifolds

Our goal in this paper is to ultimately establish facts about the Heegaard Floer invariants for closed 3–manifolds, so we need a way to translate between sutured and closed manifolds in the cases of interest. Furthermore, certain properties of this correspondence are needed to ensure that the techniques used to obtain functoriality in [5] which we import can be applied to the closed setting of interest here. For our purposes, it will be sufficient to note that there is a correspondence between closed, oriented and based 3–manifolds and sutured manifolds, and that under this correspondence:

- (1) Isotopies of attaching curves in the sutured diagram yield *pointed* isotopies (ie isotopies which do not cross the basepoint) of attaching curves in the closed diagram.
- (2) Diffeomorphisms of sutured isotopy diagrams yield pointed diffeomorphisms of pointed closed isotopy diagrams.
- (3) Stabilizations of sutured isotopy diagrams correspond to stabilizations of pointed isotopy diagrams.
- (4) Two sutured isotopy diagrams  $H_1 = (\Sigma, \alpha_1, \beta_1)$  and  $H_2 = (\Sigma, \alpha_2, \beta_2)$  are  $\alpha$ –equivalent if and only if the curves  $\alpha_1$  and  $\alpha_2$  are related by a sequence of handleslides in the corresponding pointed isotopy diagrams, where the handleslides never cross the basepoint. The analogous statement holds for  $\beta$ –equivalent sutured isotopy diagrams.

Since these last sorts of equivalences will play a prominent role throughout the paper, we use terminology introduced in [14] to describe them:

**Definition 2.1** Given two closed, pointed Heegaard diagrams  $\mathcal{H}_1 = (\Sigma, \alpha_1, \beta_1, z)$  and  $\mathcal{H}_2 = (\Sigma, \alpha_2, \beta_2, z)$  we say they are *strongly equivalent* if they are related by a sequence of isotopies and handleslides which do not cross the basepoint. If the diagrams are related by a sequence of isotopies, and handleslides which occur only among the  $\alpha$  curves, we say the diagrams are *strongly  $\alpha$ –equivalent*. If the diagrams are related by a sequence of isotopies, and handleslides which occur only among the  $\beta$  curves, we say the diagrams are *strongly  $\beta$ –equivalent*.

## 2.3 Graphs of Heegaard diagrams

Following [5, Definition 2.22], construct a directed graph  $\mathcal{G}$  as follows. The class of vertices,  $|\mathcal{G}|$ , of  $\mathcal{G}$  is given by the class of isotopy diagrams of sutured manifolds. Given

two isotopy diagrams  $H_1, H_2 \in |\mathcal{G}|$ , the oriented edges from  $H_1$  to  $H_2$  come in four flavors

$$\mathcal{G}(H_1, H_2) = \mathcal{G}_\alpha(H_1, H_2) \cup \mathcal{G}_\beta(H_1, H_2) \cup \mathcal{G}_{\text{stab}}(H_1, H_2) \cup \mathcal{G}_{\text{diff}}(H_1, H_2).$$

Here

- (1)  $\mathcal{G}_\alpha(H_1, H_2)$  consists of a single edge if the diagrams are  $\alpha$ -equivalent;
- (2)  $\mathcal{G}_\beta(H_1, H_2)$  consists of a single edge if the diagrams are  $\beta$ -equivalent;
- (3)  $\mathcal{G}_{\text{stab}}(H_1, H_2)$  consists of a single edge if the diagrams are related by a stabilization or destabilization;
- (4)  $\mathcal{G}_{\text{diff}}(H_1, H_2)$  consists of a collection of edges, with one edge for each diffeomorphism between the isotopy diagrams.

We denote by  $\mathcal{G}_\alpha, \mathcal{G}_\beta, \mathcal{G}_{\text{stab}}$  and  $\mathcal{G}_{\text{diff}}$  the subgraphs of  $\mathcal{G}$  arising from only considering the corresponding edges on the class of vertices  $|\mathcal{G}|$ .

There is an analog of the Reidemeister–Singer theorem for sutured manifolds (applied to sutured *diagrams*):

**Proposition 2.2** [5, Proposition 2.23] *Two isotopy diagrams  $H_1, H_2 \in |\mathcal{G}|$  can be connected by an oriented path in  $\mathcal{G}$  if and only if they define diffeomorphic sutured manifolds.*

**Remark 2.3** By the definition of  $\mathcal{G}$ , if there is an unoriented path from  $H_1$  to  $H_2$  then there is also an oriented path from  $H_1$  to  $H_2$ .

Let  $S(H)$  denote the sutured manifold associated to the isotopy diagram  $H$ . Given any set  $\mathcal{S}$  of diffeomorphism types of sutured manifolds, denote by  $\mathcal{G}(\mathcal{S})$  the full subgraph of  $\mathcal{G}$  spanned by those isotopy diagrams  $H$  for which  $S(H) \in \mathcal{S}$ . For our purposes, the case of interest will be  $\mathcal{S} = \mathcal{S}_{\text{man}}$ . This is the set of diffeomorphism types of sutured manifolds which arise as the images of closed, oriented, based 3–manifolds under the correspondence discussed above.

Let  $\mathcal{G}_{\text{man}}$  be the oriented graph with vertices given by pointed isotopy Heegaard diagrams of closed, connected 3–manifolds, and with the edges from an isotopy diagram  $H_1$  to an isotopy diagram  $H_2$  given by

$$\mathcal{G}_{\text{man}}(H_1, H_2) = \mathcal{G}_{\text{man}}^\alpha(H_1, H_2) \cup \mathcal{G}_{\text{man}}^\beta(H_1, H_2) \cup \mathcal{G}_{\text{man}}^{\text{stab}}(H_1, H_2) \cup \mathcal{G}_{\text{man}}^{\text{diff}}(H_1, H_2),$$

where:

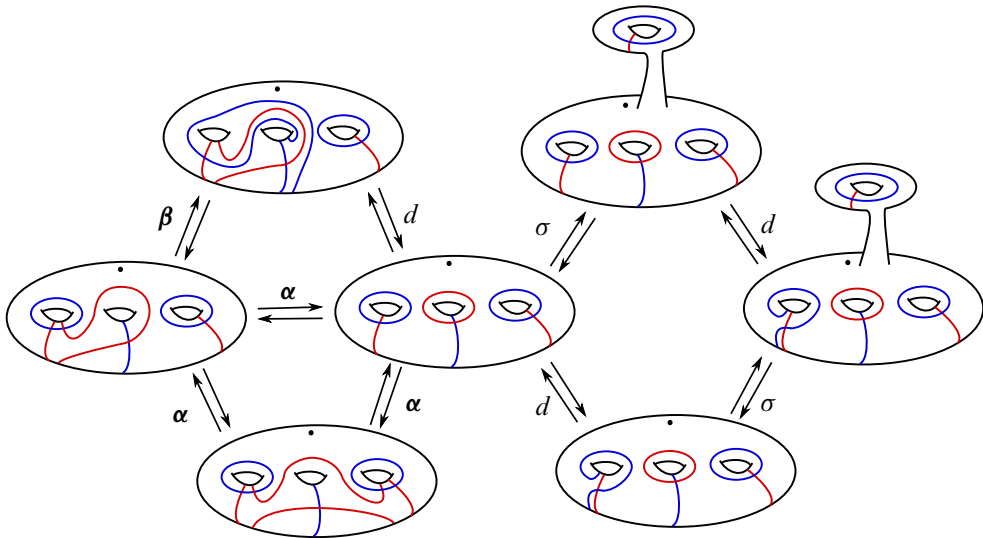


Figure 1: An illustration of a small subgraph in  $\mathcal{G}_{\text{man}}$ . The vertices are isotopy diagrams, which in the picture are depicted by particular Heegaard diagrams representing the isotopy class. We label each pair of edges with  $\alpha$ ,  $\beta$ ,  $\sigma$  or  $d$  according to whether the given pair of edges corresponds to a strong  $\alpha$ -equivalence, a strong  $\beta$ -equivalence, a stabilization/destabilization pair, or a diffeomorphism pair, respectively. We use the convention that on each Heegaard diagram the collection of red attaching curves is denoted  $\alpha$  while the collection of blue attaching curves is denoted  $\beta$ .

- (1)  $\mathcal{G}_{\text{man}}^\alpha(H_1, H_2)$  consists of a single edge if the diagrams are strongly  $\alpha$ -equivalent.
- (2)  $\mathcal{G}_{\text{man}}^\beta(H_1, H_2)$  consists of a single edge if the diagrams are strongly  $\beta$ -equivalent.
- (3)  $\mathcal{G}_{\text{man}}^{\text{stab}}(H_1, H_2)$  consists of a single edge if the diagrams are related by a stabilization or destabilization.
- (4)  $\mathcal{G}_{\text{man}}^{\text{diff}}(H_1, H_2)$  consists of a collection of edges, with one edge for each pointed diffeomorphism between the isotopy diagrams.

We provide a sketch of a piece of the graph  $\mathcal{G}_{\text{man}}$  in Figure 1. The following analog of Proposition 2.2 holds in the closed and pointed setting.

**Proposition 2.4** [11, Proposition 7.1] *Two isotopy diagrams  $H_1, H_2 \in |\mathcal{G}_{\text{man}}|$  can be connected by an oriented path in  $\mathcal{G}_{\text{man}}$  if and only if they define diffeomorphic pointed manifolds.*

Finally we note that the preceding arguments specify an isomorphism of graphs

$$(3) \quad T : \mathcal{G}(\mathcal{S}_{\text{man}}) \rightarrow \mathcal{G}_{\text{man}}$$

which we will use implicitly in the remainder of the paper to rephrase certain results from [5] in terms of  $\mathcal{G}_{\text{man}}$ .

### 3 Heegaard invariants

We now make precise two notions of what one might mean by a Heegaard invariant of closed 3–manifolds. For the interested reader’s convenience, we note that the definitions originally given in [5] apply to sutured manifolds and the graph  $\mathcal{G}(\mathcal{S}_{\text{man}})$ . Instead, we state here the equivalent definitions phrased in terms of closed manifolds and the graph  $\mathcal{G}_{\text{man}}$ .

Suppose we produce some assignment of algebraic objects to Heegaard diagrams (the vertices of the graph  $\mathcal{G}_{\text{man}}$ ), and an assignment of maps between these algebraic objects to each Heegaard move between two diagrams (the edges of  $\mathcal{G}_{\text{man}}$ ). Given Proposition 2.4, the minimal requirement we should ask of such an assignment to obtain an invariant of the underlying 3–manifold is for edges in  $\mathcal{G}_{\text{man}}$  to be assigned isomorphisms. Given any category  $\mathcal{C}$ , we have:

**Definition 3.1** (cf [5, Definition 2.24]) *A weak Heegaard invariant of closed 3–manifolds is a morphism of graphs  $F : \mathcal{G}_{\text{man}} \rightarrow \mathcal{C}$  for which  $F(e)$  is an isomorphism for all edges  $e \in \mathcal{G}_{\text{man}}$ .*

Of course, this level of invariance was established for Heegaard Floer homology at the outset.

**Theorem 3.2** [11] *The morphisms*

$$\widehat{HF}, HF^-, HF^+, HF^\infty : \mathcal{G}_{\text{man}} \rightarrow \mathbb{F}_2[U]\text{-Mod}$$

and

$$\widehat{HF}, HF^-, HF^+, HF^\infty : \mathcal{G}_{\text{man}} \rightarrow \mathbb{Z}[U]\text{-Mod}$$

are weak Heegaard invariants of closed 3–manifolds.

The above results also immediately yield:

**Corollary 3.3** *The morphisms*

$$HF^\circ: \mathcal{G}_{\text{man}} \rightarrow P(\mathbb{Z}[U]\text{-Mod})$$

*are weak Heegaard invariants of closed 3-manifolds.*

In Section 6 we will recall the definition of these morphisms of graphs precisely. In particular, since the vertices of  $\mathcal{G}_{\text{man}}$  are isotopy diagrams, we will need to explain the meaning of  $HF^\circ(H)$  when  $H$  is an isotopy diagram rather than a particular Heegaard diagram representing the isotopy class.

**Remark 3.4** For the reader referencing the corresponding results stated in [5], we note that Theorem 3.2 is instead phrased as “ $HF^\circ: \mathcal{G}(\mathcal{S}_{\text{man}}) \rightarrow \mathbb{F}_2[U]\text{-Mod}$  are weak Heegaard invariants” in [5, Theorem 2.26]. Of course, as they were originally defined,  $HF^\circ$  are invariants assigned to closed, pointed Heegaard diagrams; the meaning of  $HF^\circ(H)$  for  $H$  a sutured isotopy diagram in this statement is interpreted as follows. Recall that vertices of  $\mathcal{G}(\mathcal{S}_{\text{man}})$  correspond to isotopy diagrams  $H$  of sutured manifolds corresponding to closed, oriented 3-manifolds  $Y$ . Given an actual sutured diagram  $\mathcal{H} = (\Sigma, \alpha, \beta)$  (not up to isotopy) for such a 3-manifold, the boundary of the Heegaard surface  $\Sigma$  is  $S^1$ , so it can be capped off with a disk to obtain a closed surface  $\bar{\Sigma}$  and a pointed Heegaard diagram  $\bar{\mathcal{H}} = (\bar{\Sigma}, \alpha, \beta, z)$  for  $Y$ , where the basepoint  $z$  is chosen to lie in the disk. Thus, given a sutured diagram  $\mathcal{H}$  representing the isotopy diagram  $H$ , we define  $CF^\circ(\mathcal{H}) := CF^\circ(\bar{\mathcal{H}})$ . Finally, we will describe how the collection  $\{CF^\circ(\mathcal{H})\}$  gives rise to  $CF^\circ(H)$  in Section 6.5. Equivalently, using the isomorphism of graphs  $T$  specified in (3), the definitions above will amount to defining  $HF^\circ(H) := HF^\circ(T(H))$  for  $H$  a sutured isotopy diagram.

Let  $\text{Man}_*$  be the category whose class of objects consists of closed, connected, oriented and based 3-manifolds, and whose morphisms are basepoint preserving diffeomorphisms. In [11] and [14], significant progress was made towards showing that the weak Heegaard invariants in the theorem above can in fact be assembled into functors from  $\text{Man}_*$  to  $\mathbb{F}_2[U]\text{-Mod}$ . However, there was a gap in the proof. In [5], the authors carefully analyzed the dependence of such a result on the nature of embedded (versus abstract) Heegaard diagrams, and basepoints, and set up a framework which allowed them to finish this program. To do so, they introduced a stronger notion of a Heegaard invariant which we now describe.

To begin, we introduce some terminology for particular subgraphs in  $\mathcal{G}_{\text{man}}$  (or more generally in  $\mathcal{G}$ ) which will serve as minimal data on which this new notion of invariance will rely.

**Definition 3.5** [5, Definition 2.29] A *distinguished rectangle* is a subgraph of  $\mathcal{G}_{\text{man}}$  of the form

$$\begin{array}{ccc} H_1 & \xrightarrow{e} & H_2 \\ \downarrow f & & \downarrow g \\ H_3 & \xrightarrow{h} & H_4 \end{array}$$

which satisfies one of the following conditions.

- (1) The arrows  $e$  and  $h$  are strong  $\alpha$ -equivalences, and the arrows  $f$  and  $g$  are strong  $\beta$ -equivalences.
- (2) The arrows  $e$  and  $h$  are either both strong  $\alpha$ -equivalences or both strong  $\beta$ -equivalences, and the arrows  $f$  and  $g$  are stabilizations.
- (3) The arrows  $e$  and  $h$  are either both strong  $\alpha$ -equivalences or both strong  $\beta$ -equivalences, and the arrows  $f$  and  $g$  are diffeomorphisms. Furthermore,  $f = g$ . (Note in this case  $\Sigma_1 = \Sigma_2$ , and  $\Sigma_3 = \Sigma_4$ , so this requirement makes sense.)
- (4) All of the arrows  $e, f, g$  and  $h$  are stabilizations. Furthermore, there are disjoint disks  $D_1, D_2 \subset \Sigma_1$  and disjoint punctured tori  $T_1, T_2 \subset \Sigma_4$  such that  $\Sigma_1 \setminus (D_1 \cup D_2) = \Sigma_4 \setminus (T_1 \cup T_2)$ ,  $\Sigma_2 = (\Sigma_1 \setminus D_1) \cup T_1$ , and  $\Sigma_3 = (\Sigma_1 \setminus D_2) \cup T_2$ .
- (5) The arrows  $e$  and  $h$  are stabilizations, and the arrows  $f$  and  $g$  are diffeomorphisms. Furthermore, the diffeomorphism  $g$  is an extension of the diffeomorphism  $f$  in the following sense. There are disks  $D_1 \subset \Sigma_1, D_3 \subset \Sigma_3$  and punctured tori  $T_2 \subset \Sigma_2, T_4 \subset \Sigma_4$  such that  $\Sigma_1 \setminus D_1 = \Sigma_2 \setminus T_2, \Sigma_3 \setminus D_3 = \Sigma_4 \setminus T_4, f(D_1) = D_2, g(T_3) = T_4$  and  $f|_{\Sigma_1 \setminus D_1} = g|_{\Sigma_2 \setminus T_2}$ .

We illustrate cases (4) and (5) schematically in Figures 2 and 3.

**Definition 3.6** [5, Definition 2.31] A *simple handleswap* is a subgraph of  $\mathcal{G}_{\text{man}}$  of the form

$$\begin{array}{ccc} H_1 & & \\ g \uparrow & \searrow e & \\ H_3 & \xleftarrow{f} & H_2 \end{array}$$

such that:

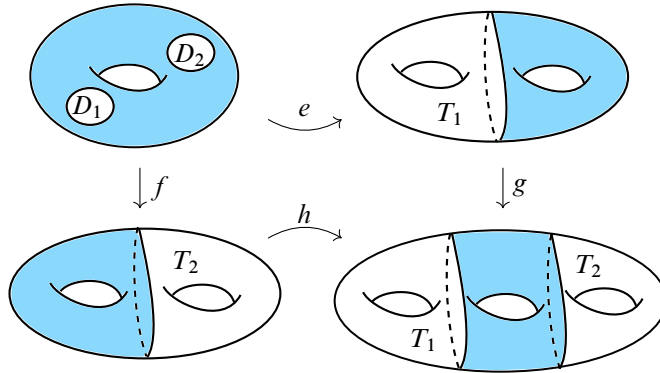


Figure 2: A schematic illustrating case (4) in the definition of a distinguished rectangle. The blue regions indicate the identifications specified in case (4). For ease of visualization, we suppress the attaching curve data in the initial diagram and in the stabilizations.

- (1) The isotopy diagrams  $H_i$  are given by  $H_i = (\Sigma \# \Sigma_0, [\alpha_i], [\beta_i])$ , where  $\Sigma_0$  is a genus two surface.
- (2)  $e$  is a strong  $\alpha$ -equivalence,  $f$  is a strong  $\beta$ -equivalence, and  $g$  is a diffeomorphism.
- (3) In the punctured genus two surface  $P = (\Sigma \# \Sigma_0) \setminus \Sigma$ , the above triangle is equivalent to the triangle in Figure 4 in the following sense. There are

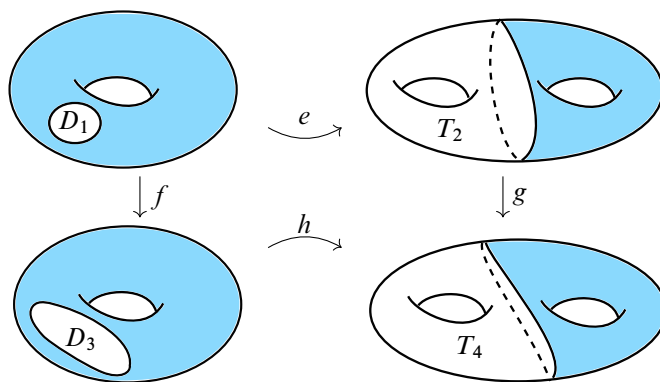


Figure 3: A schematic illustrating case (5) in the definition of a distinguished rectangle. The blue regions indicate the identifications of the regions specified in case (5). For ease of visualization, we suppress the attaching curve data in each diagram.

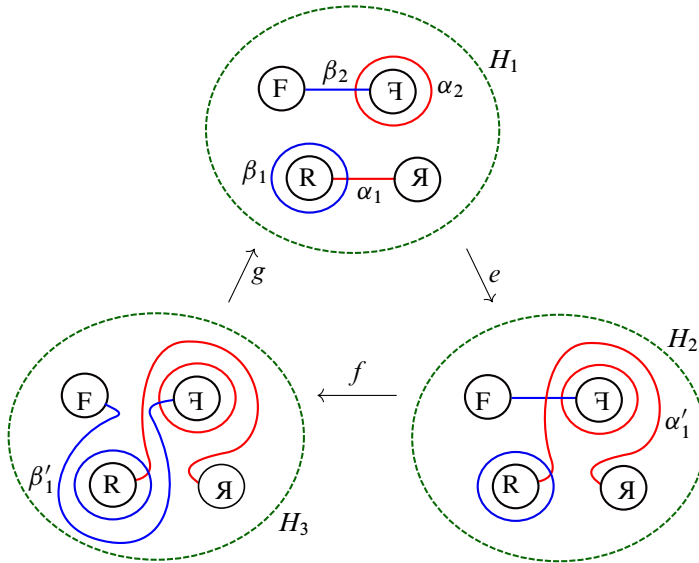


Figure 4: The standard simple handleswap.

diffeomorphisms from  $P \cap H_i$  to the green disks labeled  $H_i$  in the figure, such that the image of the  $\alpha$  curves are the red circles in the figures, and the image of the  $\beta$  curves are the blue circles in the figures.

- (4) The diagrams  $H_1, H_2$  and  $H_3$  are identical when restricted to  $\Sigma$ .

With these notions in hand, the stronger sense of invariance we will ask of our Heegaard invariants is as follows.

**Definition 3.7** [5, Definition 2.32] *A strong Heegaard invariant of closed 3-manifolds is a weak Heegaard invariant  $F: \mathcal{G}_{\text{man}} \rightarrow \mathcal{C}$  that additionally satisfies the following axioms:*

- (1) **Functoriality** The restriction of  $F$  to  $\mathcal{G}_{\text{man}}^\alpha, \mathcal{G}_{\text{man}}^\beta$  and  $\mathcal{G}_{\text{man}}^{\text{diff}}$  are functors to  $\mathcal{C}$ . If  $e: H_1 \rightarrow H_2$  is a stabilization and  $e': H_2 \rightarrow H_1$  is the corresponding destabilization, then  $F(e') = F(e)^{-1}$ .
- (2) **Commutativity** For every distinguished rectangle in  $\mathcal{G}_{\text{man}}$ ,

$$\begin{array}{ccc}
 H_1 & \xrightarrow{e} & H_2 \\
 \downarrow f & & \downarrow g \\
 H_3 & \xrightarrow{h} & H_4
 \end{array}$$

we have  $F(g) \circ F(e) = F(h) \circ F(f)$ .



- (3) **Continuity** If  $H \in |\mathcal{G}_{\text{man}}|$  and  $e \in \mathcal{G}_{\text{man}}^{\text{diff}}(H, H)$  is a diffeomorphism isotopic to  $\text{Id}_{\Sigma}$ , then  $F(e) = \text{Id}_{F(H)}$ .
- (4) **Handleswap invariance** For every simple handleswap in  $\mathcal{G}_{\text{man}}$ ,

$$\begin{array}{ccc}
 H_1 & & \\
 g \uparrow & \searrow e & \\
 H_3 & \xleftarrow{f} & H_2
 \end{array}$$

we have  $F(g) \circ F(f) \circ F(e) = \text{Id}_{F(H_1)}$ .

As we will summarize in Section 5, it was shown in [5] that for any weak Heegaard invariant the axioms required above are sufficient to ensure the images of the invariant, when restricted to a particular subgraph of  $\mathcal{G}_{\text{man}}$  whose vertices represent a fixed 3-manifold, form a transitive system in the given category. For certain categories  $\mathcal{C}$ , this in turn is enough to ensure that the assignments of the invariants can be understood as a functor from an appropriate category of 3-manifolds.

### 4 Transitive systems of chain complexes and projectivization

In this section we describe the algebraic framework which will be necessary to phrase our projective functoriality results. To begin with, we recall the following fundamental notions.

**Definition 4.1** A *directed set*  $(I, \leq)$  is a set  $I$  together with a reflexive and transitive binary relation  $\leq$ , such that for every pair of elements  $a, b \in I$  there is an element  $c \in I$  with  $a \leq c$  and  $b \leq c$ .

**Definition 4.2** Let  $\mathcal{C}$  be a category, and  $(I, \leq)$  be a directed set. Given a collection of objects  $\{O_i\}$  in  $\mathcal{C}$  indexed by  $I$ , and a collection of morphisms  $\{f_{i,j} : O_i \rightarrow O_j\}$  for all  $i, j \in I$  with  $i \leq j$ , we say the collections are a *transitive system in  $\mathcal{C}$  (indexed by  $I$ )* if they satisfy

- (1)  $f_{i,i} = \text{Id}_{O_i}$ ,
- (2)  $f_{i,k} = f_{j,k} \circ f_{i,j}$ .

We also have the following notion of morphisms between transitive systems.

**Definition 4.3** Given two transitive systems

$$T_1 = \{I_1, \leq, \{O_i\}, \{f_{i,j}\}\} \quad \text{and} \quad T_2 = \{I_2, \leq, \{P_i\}, \{g_{i,j}\}\}$$

in a category  $\mathcal{C}$ , a *morphism of transitive systems*  $(M, \{n_i\})$  from  $T_1$  to  $T_2$  consists of a map of directed sets  $M : I_1 \rightarrow I_2$  and a collection of morphisms  $\{n_i : O_i \rightarrow P_{M(i)}\}$  in  $\mathcal{C}$  such that for all  $i, j \in I_1$  with  $i \leq j$  the squares

$$\begin{array}{ccc} O_i & \xrightarrow{n_i} & P_{M(i)} \\ \downarrow f_{i,j} & & \downarrow g_{M(i),M(j)} \\ O_j & \xrightarrow{n_j} & P_{M(j)} \end{array}$$

commute in  $\mathcal{C}$ . We denote the resulting category of transitive systems in  $\mathcal{C}$  by  $\text{Trans}(\mathcal{C})$ .

Finally, given a transitive system in  $\text{Trans}(\mathcal{C})$  indexed by  $J$ , we obtain what one might call a two-dimensional transitive system. Such a two-dimensional transitive system naturally has the structure of a transitive system in  $\mathcal{C}$  indexed by  $I \times J$ , where  $(i, j) \leq (i', j')$  if and only if  $i \leq i'$  and  $j \leq j'$ .

We now explain how these notions will arise in the context of our results. We will begin by considering the category  $\text{Kom}(\mathbb{Z}[U]\text{-Mod})$ , the homotopy category of chain complexes of  $\mathbb{Z}[U]$ -modules. To each pointed isotopy diagram  $H$ , corresponding to a vertex of  $\mathcal{G}_{\text{man}}$ , we will assign a transitive system  $CF^-(H) \in \text{Trans}(\text{Kom}(\mathbb{Z}[U]\text{-Mod}))$ . To be more explicit about the nature of this construction and bridge the gap to the language defined above, given an isotopy diagram  $H$  we consider the directed set  $(I, \leq)$  with  $I$  the set of Heegaard diagrams in the given isotopy class, and  $\leq$  the (in this case trivial, equivalence) relation on the set indicating existence of an isotopy between two elements. Then  $CF^-(H)$  will be a transitive system in  $\text{Kom}(\mathbb{Z}[U]\text{-Mod})$  indexed by  $(I, \leq)$ , with the objects in the transitive system being the Heegaard Floer chain complexes associated to individual diagrams in the fixed isotopy class, and the morphisms in the transitive systems being certain continuation maps between such complexes. The details of precisely how these assignments are made will be specified throughout the course of Section 6. To a diffeomorphism, strong  $\alpha$ -equivalence, strong  $\beta$ -equivalence, or stabilization between two such isotopy diagrams  $H_1$  and  $H_2$  we will then associate a morphism of transitive systems from  $CF^-(H_1)$  to  $CF^-(H_2)$ . Together, these assignments will yield a morphism of graphs

$$CF^- : \mathcal{G}_{\text{man}} \rightarrow \text{Trans}(\text{Kom}(\mathbb{Z}[U]\text{-Mod})).$$

This morphism of graphs may not be a strong Heegaard invariant. We will however be able to establish that this morphism of graphs satisfies the axioms required of a strong Heegaard invariant up to an overall sign in each of the axioms (2), (3) and (4) appearing in Definition 3.7.

Equivalently, we will phrase this result in terms of an appropriate projectivization. Recall that given any category  $\mathcal{C}$ , with an equivalence relation  $\sim$  on every hom set which furthermore respects composition, we may form the quotient category  $\mathcal{C} = \mathcal{C} / \sim$ . This is the category whose objects are those of  $\mathcal{C}$ , and whose morphisms are equivalence classes of morphisms with respect to  $\sim$ . Given an additive category  $\mathcal{C}$ , we define the *projectivization of  $\mathcal{C}$* ,  $P(\mathcal{C})$ , to be the quotient category of  $\mathcal{C}$  with respect to the relation  $f \sim -f$  for all morphisms  $f$ . The last statement in the preceding paragraph is then given precisely by the following statement: considering now the category of transitive systems in the projectivized homotopy category,  $\text{Trans}(P(\text{Kom}(\mathbb{Z}[U]\text{-Mod}))$ ), we will show that the morphism of graphs above yields a strong Heegaard invariant

$$CF^- : \mathcal{G}_{\text{man}} \rightarrow \text{Trans}(P(\text{Kom}(\mathbb{Z}[U]\text{-Mod}))).$$

**Remark 4.4** While the proliferation of transitive systems may seem undesirable, we were unable to produce another framework in which our naturality results could be phrased. There appear to be two issues that arise if one tries to use the same framework developed in [5] to phrase our projective results.

The first issue comes from the fact that the statement in Theorem 1.3 is concerned with the Floer chain complexes. If one wanted to dispense with the category of transitive systems appearing in that statement, one would need to assign a *single* chain complex  $CF^\circ(H)$  of  $\mathbb{Z}[U]$ -modules to an isotopy diagram  $H$ . As we will recall in the next section, what the Heegaard Floer construction actually produces for each isotopy diagram  $H$  is a transitive system of chain homotopy equivalences between chain complexes of  $\mathbb{Z}[U]$ -modules. In general, it is not clear how one should define an object like a colimit of such a transitive system of chain complexes to obtain a single chain complex. We note that it seems likely that this issue is in fact a nonissue, for the following reason. We expect our transitive system of chain homotopy equivalences is homotopy coherent in the sense of [19], which if true would allow one to define a single chain complex  $CF^\circ(H)$  via a homotopy colimit. Indeed, that our transitive systems are homotopy coherent in this sense seems likely to follow from the results in [1].

However, even if one could assign to each isotopy diagram a single chain complex  $CF^\circ(H)$ , there is another key obstruction to phrasing Theorem 1.1 without the use of transitive systems. In the proof of Theorem 1.1, which will be given in Section 5, we will associate to each closed, pointed 3–manifold a transitive system in  $P(\mathbb{Z}[U]\text{–Mod})$ . The author is unaware of a notion of a colimit in  $P(\mathbb{Z}[U]\text{–Mod})$  which would allow Theorem 1.1 to be stated without transitive systems, in such a way that it is also not merely reduced to a statement about the  $\mathbb{F}_2$  invariants.

### 5 Projective naturality from strong Heegaard Floer invariants

In this section we prove Theorem 1.1 assuming Corollary 1.4, which we will prove in turn in Section 7. Our argument will follow the same logical structure as that used to prove the analogous result over  $\mathbb{F}_2$  appearing in [5, Theorem 1.5]. We provide the argument here for the reader’s convenience, but note that the scheme is essentially the same.

In [5] Juhász, Thurston and Zemke show that the images of any strong Heegaard invariant, appropriately restricted, fit into a transitive system. To make this precise, we introduce a few more definitions.

**Definition 5.1** Suppose  $H_1$  and  $H_2$  are embedded isotopy diagrams for a closed, oriented, pointed 3–manifold  $(Y, z)$ , with Heegaard surfaces

$$\iota_1, \iota_2: (\Sigma_1, z), (\Sigma_2, z) \hookrightarrow (Y, z).$$

We say a diffeomorphism of isotopy diagrams  $d: H_1 \rightarrow H_2$  is *isotopic to the identity in  $M$*  if  $\iota_2 \circ d: \Sigma_1 \rightarrow (Y, z)$  is isotopic to  $\iota_1: \Sigma_1 \rightarrow (Y, z)$  relative to the basepoint.

**Definition 5.2** Given  $(Y, z)$ , let  $(\mathcal{G}_{\text{man}})_{(Y,z)}$  be the following subgraph of  $\mathcal{G}_{\text{man}}$  whose vertices are embedded isotopy diagrams for  $(Y, z)$ . The edges  $e \in (\mathcal{G}_{\text{man}})_{(Y,z)}(H_1, H_2)$  between two isotopy diagrams again come in four flavors,

$$(\mathcal{G}_{\text{man}})_{(Y,z)}(H_1, H_2) = \mathcal{G}_{\text{man}}^\alpha(H_1, H_2) \cup \mathcal{G}_{\text{man}}^\beta(H_1, H_2) \cup \mathcal{G}_{\text{man}}^{\text{stab}}(H_1, H_2) \cup (\mathcal{G}_{\text{man}}^{\text{diff}})^0(H_1, H_2).$$

Here  $\mathcal{G}_{\text{man}}^\alpha$ ,  $\mathcal{G}_{\text{man}}^\beta$  and  $\mathcal{G}_{\text{man}}^{\text{stab}}$  are the same collections as in the definition of  $\mathcal{G}_{\text{man}}$ , while  $(\mathcal{G}_{\text{man}}^{\text{diff}})^0(H_1, H_2)$  consists of one edge for each element in the set of diffeomorphisms from  $H_1$  to  $H_2$  which are isotopic to the identity in  $M$ .

With these notions in hand, we have a stronger version of Proposition 2.4 which applies now to embedded diagrams for some fixed  $(Y, z)$ :

**Proposition 5.3** [5, Proposition 2.36] *Given  $(Y, z)$ , any two vertices in the graph  $(\mathcal{G}_{\text{man}})_{(Y,z)}$  can be connected by an oriented path in  $(\mathcal{G}_{\text{man}})_{(Y,z)}$ .*

The salient feature of a strong Heegaard invariant,  $F$ , is that the isomorphisms  $F(e)$  associated to edges  $e$  in  $(\mathcal{G}_{\text{man}})_{(Y,z)}$  fit into a transitive system. This follows from the fact that the isomorphism associated to a path depends only on the endpoints:

**Theorem 5.4** [5, Theorem 2.38] *Let  $F: \mathcal{G}_{\text{man}} \rightarrow \mathcal{C}$  be a strong Heegaard invariant. Given two isotopy diagrams  $H, H' \in |(\mathcal{G}_{\text{man}})_{(Y,z)}|$  and any two oriented paths  $\eta$  and  $\nu$  in  $(\mathcal{G}_{\text{man}})_{(Y,z)}$  from  $H$  to  $H'$ , we have*

$$F(\eta) = F(\nu).$$

Now, for any two isotopy diagrams  $H$  and  $H'$ , and an oriented path  $\eta$  from  $H$  to  $H'$ , we can define the map  $F_{H,H'} = F(\eta)$ .

**Corollary 5.5** [5, Corollary 2.41] *Suppose that  $H, H', H'' \in |(\mathcal{G}_{\text{man}})_{(Y,z)}|$ . Then*

$$F_{H,H''} = F_{H',H''} \circ F_{H,H'}.$$

These results should provide some intuitive justification for the appearance of the notion of a strong Heegaard invariant. At the very least, the notion is enough to ensure such invariants fit into a transitive system. In particular, applying Corollary 5.5 to the strong Heegaard invariants

$$CF^\circ: \mathcal{G}_{\text{man}} \rightarrow \text{Trans}(P(\text{Kom}(\mathbb{Z}[U]\text{-Mod})))$$

of Theorem 1.3 immediately yields Corollary 1.5. We now show that this transitivity is also enough for the functoriality ends we seek in Theorem 1.1.

**Proof of Theorem 1.1** Assuming Corollary 1.4, the Heegaard Floer invariants

$$HF^\circ: \mathcal{G}_{\text{man}} \rightarrow P(\mathbb{Z}[U]\text{-Mod})$$

are strong Heegaard invariants. Let  $\text{Man}_*$  be the category of closed, connected, oriented and based 3-manifolds with based diffeomorphisms. Using the strong Heegaard invariants above, we can obtain functors

$$HF_1^\circ: \text{Man}_* \rightarrow \text{Trans}(P(\mathbb{Z}[U]\text{-Mod}))$$

as follows. Given a manifold  $(Y, z) \in \text{Ob}(\text{Man}_*)$ , Corollary 5.5 ensures that the modules  $HF^\circ(H)$  for isotopy diagrams  $H \in |(\mathcal{G}_{\text{man}})_{(Y,z)}|$ , along with the isomorphisms  $HF_{H,H'}^\circ$ , form a transitive system. We denote this transitive system by

$$HF_1^\circ(Y, z) \in \text{Trans}(P(\mathbb{Z}[U]\text{-Mod})).$$

To a pointed diffeomorphism  $\phi: (Y, z) \rightarrow (Y', z')$ , the functor  $HF_1^\circ$  will assign a morphism of transitive systems

$$HF_1^\circ(\phi): HF_1^\circ(Y, z) \rightarrow HF_1^\circ(Y', z')$$

defined as follows. Given any isotopy diagram  $H = (\Sigma, A, B, z)$  for  $(Y, z)$ , let  $\phi_H = \phi|_\Sigma$  and  $H'$  be the isotopy diagram  $\phi(H)$  for  $(Y', z')$ . By virtue of being a strong Heegaard invariant,  $HF^\circ$  associates a morphism  $HF^\circ(\phi_H): HF^\circ(H) \rightarrow HF^\circ(H')$  in  $P(\mathbb{Z}[U]\text{-Mod})$  to any such diffeomorphism of isotopy diagrams  $\phi_H$ . The collection of morphisms  $\{\phi_H\}$  for  $H \in |(\mathcal{G}_{\text{man}})_{(Y,z)}|$  will thus yield a collection of morphisms  $\{HF^\circ(\phi_H)\}$ . We claim that this collection of morphisms is in fact a morphism of transitive systems

$$HF_1^\circ(\phi): HF_1^\circ(Y, z) \rightarrow HF_1^\circ(Y', z')$$

as desired. According to Definition 4.3, we must check that for any path of edges  $\gamma$  in  $(\mathcal{G}_{\text{man}})_{(Y,z)}$  from  $H_1$  to  $H_2$ , we have  $HF^\circ(\phi_{H_2}) \circ HF^\circ(\gamma) = HF^\circ(\gamma') \circ HF^\circ(\phi_{H_1})$  for some path  $\gamma'$  in  $(\mathcal{G}_{\text{man}})_{(Y',z')}$  from  $H'_1$  to  $H'_2$ . If  $\gamma$  is given by the path of edges

$$D_0 \xrightarrow{e_1} D_1 \xrightarrow{e_2} \dots \xrightarrow{e_{n-1}} D_{n-1} \xrightarrow{e_n} D_n$$

in  $(\mathcal{G}_{\text{man}})_{(Y,z)}$  from  $D_0 = H_1$  to  $D_n = H_2$ , we pick out a path  $\gamma'$  in  $(\mathcal{G}_{\text{man}})_{(Y',z')}$  from  $H'_1$  to  $H'_2$  given by

$$D'_0 \xrightarrow{e'_1} D'_1 \xrightarrow{e'_2} \dots \xrightarrow{e'_{n-1}} D'_{n-1} \xrightarrow{e'_n} D'_n$$

as follows. We define the intermediate isotopy diagrams in the path  $\gamma'$  by  $D'_i = \phi(D_i)$ . If the edge  $e_i$  is given by a strong  $\alpha$ -equivalence, a strong  $\beta$ -equivalence, or a (de)stabilization, we let  $e'_i$  denote the corresponding strong  $\alpha$ -equivalence, strong  $\beta$ -equivalence, or (de)stabilization. If  $e_i$  corresponds to a diffeomorphism  $e_i: D_{i-1} \rightarrow D_i$  isotopic to the identity, we set  $e'_i = \phi_{D_i} \circ e_i \circ \phi_{D_{i-1}}^{-1}$ . We then have a subgraph in  $\mathcal{G}_{\text{man}}$  given by

$$\begin{array}{ccccccc} D_0 & \xrightarrow{e_1} & D_1 & \xrightarrow{e_2} & \dots & \xrightarrow{e_{n-1}} & D_{n-1} & \xrightarrow{e_n} & D_n \\ \downarrow \phi_{H_1} & & \downarrow \phi_{D_1} & & & & \downarrow \phi_{D_{n-1}} & & \downarrow \phi_{H_2} \\ D'_0 & \xrightarrow{e'_1} & D'_1 & \xrightarrow{e'_2} & \dots & \xrightarrow{e'_{n-1}} & D'_{n-1} & \xrightarrow{e'_n} & D'_n \end{array}$$

The condition that needs to be verified is that the image under  $HF^\circ$  of the outer rectangle in this subgraph commutes. By construction of the path  $\gamma'$ , each small square in the diagram is either a distinguished rectangle (recall Definition 3.7) or a commuting square of diffeomorphisms. Commutativity of the large rectangle now follows by virtue of  $HF^\circ$  being a strong Heegaard invariant. Since the restriction of  $HF^\circ$  to  $\mathcal{G}_{\text{man}}^{\text{diff}}$  is a functor, the image under  $HF^\circ$  of the commuting square of diffeomorphisms also commutes. Since the image under  $HF^\circ$  of any distinguished rectangle also commutes, we thus see that the morphism of transitive systems

$$HF_1^\circ(\phi): HF_1^\circ(Y, z) \rightarrow HF_1^\circ(Y', z')$$

associated to a pointed diffeomorphism  $\phi$  is well defined.

The assignments above thus define the functor  $HF_1^\circ$ ; we note that composition of morphisms in  $\text{Man}_*$  are respected under  $HF_1^\circ$  because  $HF^\circ$  is a strong Heegaard invariant, and in particular must be a functor when restricted to  $\mathcal{G}_{\text{man}}^{\text{diff}}$  (see axiom (1) in Definition 3.7).

Finally, we note that isotopic diffeomorphisms in  $\text{Man}_*$  induce identical maps under  $HF_1^\circ$ . To see this, suppose  $\phi: (Y, z) \rightarrow (Y, z)$  is isotopic to  $\text{Id}_{(Y, z)}$ , and fix an isotopy diagram  $H = (\Sigma, A, B, z)$  for  $(Y, z)$ . Then  $\phi_H = \phi|_H$  is isotopic to  $\text{Id}_H$  and  $H' = \phi(H) = H$ , so by virtue of  $HF^\circ$  being a strong Heegaard invariant we must have  $HF^\circ(\phi_H) = \text{Id}_{HF^\circ(H)}$ . Thus  $HF_1^\circ(\phi)$  is the map of transitive systems defined by the data  $\{HF^\circ(\phi_H) = \text{Id}_{HF^\circ(H)}\}$  for  $H \in (\mathcal{G}_{\text{man}})_{(Y, z)}$ , and is thus an identity morphism in  $\text{Trans}(P(\mathbb{Z}[U]\text{-Mod}))$ .  $\square$

## 6 Heegaard Floer homology as a weak Heegaard invariant

In this section we very briefly recall numerous maps defined on the Heegaard Floer chain complexes, and then use these maps to define the underlying morphisms of graphs of the strong Heegaard invariants appearing in Theorem 1.3. For the most part we just seek to establish notation in Sections 6.1–6.4, and refer the reader to [5], [6] and [11] for detailed descriptions of the constructions involved in the definitions appearing there.

For concreteness and ease of notation, we will phrase the results in this section in terms of  $CF^-$ ; however we note that the definitions vary in a cosmetic way, and analogous results hold, for all of the variants  $CF^\circ$ . In particular, the proof of Theorem 1.3 for  $CF^\circ$  will follow by the same arguments given here for  $CF^-$ . In fact, one could also

obtain the results for the other variants directly from those we prove, as  $\widehat{CF}$ ,  $CF^+$  and  $CF^\infty$  can all be obtained by taking suitable tensor products with  $CF^-$  and quotients thereof.

Finally, we note at the outset that we will use  $\sim$  to indicate homotopic chain maps.

### 6.1 $\text{Spin}^c$ structures and strong admissibility

We must first address the fact that while the graph  $\mathcal{G}_{\text{man}}$  that we have been considering thus far contains arbitrary Heegaard diagrams, the Heegaard Floer chain complexes defined in [11] are defined only with respect to certain *admissible* diagrams. Since we will focus on the case of  $CF^-$  in this section, the admissibility we will need is given by the notion of *strong admissibility*, which we now summarize.

We begin by recalling the setting of Heegaard Floer homology, and the role of  $\text{Spin}^c$  structures in the construction of the Heegaard Floer chain complexes. Given a genus  $g$  based Heegaard diagram

$$\mathcal{H} = (\Sigma, \boldsymbol{\alpha} = (\alpha_1, \alpha_2, \dots, \alpha_g), \boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_g), z)$$

for a closed, connected, oriented and based 3-manifold  $(Y, z)$ , one considers the tori

$$\mathbb{T}_{\boldsymbol{\alpha}} = \alpha_1 \times \alpha_2 \times \dots \times \alpha_g, \quad \mathbb{T}_{\boldsymbol{\beta}} = \beta_1 \times \beta_2 \times \dots \times \beta_g$$

in the symmetric product  $\text{Sym}^g(\Sigma) := (\Sigma \times \dots \times \Sigma) / S_g$ . A choice of complex structure on  $\Sigma$  induces an almost complex structure on  $\text{Sym}^g(\Sigma)$ , and with respect to such an induced structure the tori  $\mathbb{T}_{\boldsymbol{\alpha}}$  and  $\mathbb{T}_{\boldsymbol{\beta}}$  are totally real. The Heegaard Floer homology is then defined as a variation of Lagrangian intersection Floer homology applied to these tori. To define the chain complexes one must fix a complex structure  $j$  on  $\Sigma$ , and a choice of generic path  $J_s$  of almost complex structures on  $\text{Sym}^g(\Sigma)$  through  $\text{Sym}^g(j)$ ; see [11].

The basepoint  $z$  induces a map

$$s_z: \mathbb{T}_{\boldsymbol{\alpha}} \cap \mathbb{T}_{\boldsymbol{\beta}} \rightarrow \text{Spin}^c(Y)$$

which associates to each intersection point a  $\text{Spin}^c$ -structure. One first defines a chain complex

$$CF^-(\mathcal{H}, \mathfrak{s})$$

which is freely generated as an abelian group by  $[x, i]$ , for  $x \in \mathbb{T}_{\boldsymbol{\alpha}} \cap \mathbb{T}_{\boldsymbol{\beta}}$  with  $s_z(x) = \mathfrak{s}$  and for  $i \in \mathbb{Z}$  with  $i < 0$ . Given two intersection points  $x, y \in \mathbb{T}_{\boldsymbol{\alpha}} \cap \mathbb{T}_{\boldsymbol{\beta}}$ , we let



$\pi_2(\mathbf{x}, \mathbf{y})$  denote the set of homotopy classes of Whitney disks connecting  $\mathbf{x}$  to  $\mathbf{y}$  in  $\text{Sym}^g(\Sigma)$ , with the usual boundary conditions. Given a homotopy class  $\phi \in \pi_2(\mathbf{x}, \mathbf{y})$ , we denote by  $\mathcal{M}_{J_s}(\phi)$  the moduli space of  $J_s$ -holomorphic disks in the class  $\phi$ , and write  $\widehat{\mathcal{M}}_{J_s}(\phi) = \mathcal{M}_{J_s}(\phi)/\mathbb{R}$  for the quotient with respect to the  $\mathbb{R}$ -action coming from the translation action on the disks. We let  $\mu(\phi)$  denote the Maslov index of the class  $\phi$ , and let  $n_z(\phi)$  denote the algebraic intersection number of  $\phi$  with  $z \times \text{Sym}^{g-1}(\Sigma)$ . We then have a well-defined relative (in general cyclic) grading on the generators defined above, given by the formula

$$\text{gr}([\mathbf{x}, i], [\mathbf{y}, j]) = \mu(\phi) - 2n_z(\phi) + 2i - 2j,$$

where  $\phi$  is any class  $\phi \in \pi_2(\mathbf{x}, \mathbf{y})$ . This grading is only integral if  $c_1(\mathfrak{s}) = 0$ . Finally, the differential

$$\partial: CF^-(\mathcal{H}, \mathfrak{s}) \rightarrow CF^-(\mathcal{H}, \mathfrak{s})$$

is defined by the formula

$$\partial([\mathbf{x}, i]) = \sum_{\{\mathbf{y} \in \mathbb{T}_\alpha \cap \mathbb{T}_\beta \mid s_z(\mathbf{y}) = \mathfrak{s}\}} \sum_{\{\phi \in \pi_2(\mathbf{x}, \mathbf{y}) \mid \mu(\phi) = 1\}} \#\widehat{\mathcal{M}}_{J_s}(\phi) \cdot [\mathbf{y}, i - n_z(\phi)].$$

There is an action of the polynomial ring  $\mathbb{Z}[U]$  on the complex  $CF^-(\mathcal{H}, \mathfrak{s})$ , where

$$U \cdot [\mathbf{x}, i] = [\mathbf{x}, i - 1]$$

decreases the relative grading by 2. We will always consider  $CF^-(\mathcal{H}, \mathfrak{s})$  as a complex of  $\mathbb{Z}[U]$ -modules. Finally, the total chain complex associated to  $\mathcal{H}$  then splits by definition as

$$CF^-(\mathcal{H}) = \bigoplus_{\mathfrak{s} \in \text{Spin}^c(Y)} CF^-(\mathcal{H}, \mathfrak{s}).$$

Given a  $\text{Spin}^c$  structure  $\mathfrak{s}$ , we call a pointed Heegaard diagram  $\mathfrak{s}$ -realized if there is an intersection point  $\mathbf{x} \in \mathbb{T}_\alpha \cap \mathbb{T}_\beta$  with  $s_z(\mathbf{x}) = \mathfrak{s}$ . We note that for any  $\mathfrak{s} \in \text{Spin}^c(Y, z)$  there is an  $\mathfrak{s}$ -realized pointed Heegaard diagram for  $(Y, z)$  by [11, Lemma 5.2].

The chain complex  $CF^-(\mathcal{H}, \mathfrak{s})$  can in fact only be defined for Heegaard diagrams  $\mathcal{H} = (\Sigma, \alpha, \beta, z)$  which satisfy an admissibility hypothesis. Given  $\mathfrak{s} \in \text{Spin}^c(Y)$ , we say the diagram  $\mathcal{H}$  is *strongly  $\mathfrak{s}$ -admissible* if every nontrivial periodic domain  $D$  on  $\mathcal{H}$  satisfying  $\langle c_1(\mathfrak{s}), H(D) \rangle = 2n \geq 0$  has some coefficient that is greater than  $n$ . Here  $H(D) \in H_2(Y; \mathbb{Z})$  is the homology class naturally associated to the periodic domain  $D$ . It turns out that this notion of admissibility is enough to ensure that differential  $\partial$  given above consists of a finite sum and is well defined on  $CF^-(\mathcal{H}, \mathfrak{s})$ , and to ensure that it in

fact yields a chain complex. It is shown in [11, Lemma 5.4] that given any  $\mathfrak{s} \in \text{Spin}^c(Y)$ , there is an  $\mathfrak{s}$ -realized, strongly  $\mathfrak{s}$ -admissible pointed diagram for  $(Y, z)$ .

To define triangle maps on the Floer chain complexes, we will need an analogous notion of admissibility for Heegaard triple diagrams. A pointed triple diagram

$$\mathcal{T} = (\Sigma, \alpha, \beta, \gamma, z)$$

specifies a 4-manifold with boundary, which we denote by  $X_{\alpha, \beta, \gamma}$ . Given now a  $\text{Spin}^c$ -structure  $\mathfrak{s}$  on  $X_{\alpha, \beta, \gamma}$ , denote by  $\mathfrak{s}_{\alpha, \beta}$  the restriction of  $\mathfrak{s}$  to the boundary component  $Y_{\alpha, \beta}$ . We will say the triple diagram  $\mathcal{T}$  is *strongly  $\mathfrak{s}$ -admissible* if any triply periodic domain  $D$  which is the sum of doubly periodic domains

$$D = D_{\alpha, \beta} + D_{\beta, \gamma} + D_{\alpha, \gamma}$$

and which furthermore satisfies

$$\langle c_1(\mathfrak{s}_{\alpha, \beta}), H(D_{\alpha, \beta}) \rangle + \langle c_1(\mathfrak{s}_{\beta, \gamma}), H(D_{\beta, \gamma}) \rangle + \langle c_1(\mathfrak{s}_{\alpha, \gamma}), H(D_{\alpha, \gamma}) \rangle = 2n \geq 0$$

has some coefficient greater than  $n$ . It is shown in [11, Lemma 8.11] that given any pointed triple diagram  $\mathcal{T}$  and a  $\text{Spin}^c$  structure  $\mathfrak{s}$  on  $X_{\alpha, \beta, \gamma}$ , there is a pointed triple diagram isotopic to  $\mathcal{T}$  which is strongly  $\mathfrak{s}$ -admissible.

## 6.2 Orientation systems

**6.2.1 Coherent orientation systems of disks** We recall that to define the differential on the Heegaard Floer chain complexes with coefficients in  $\mathbb{Z}$ , one must perform signed counts of the points in certain moduli spaces of pseudoholomorphic disks. To do so, one must ensure that on a pointed Heegaard diagram  $\mathcal{H} = (\Sigma, \alpha, \beta, z)$  the moduli spaces of holomorphic disks in a homotopy class  $A \in \pi_2(x, y)$ , which we denote by  $\mathcal{M}^A$  or  $\mathcal{M}(A)$ , are orientable. By [11, Proposition 3.10] (or [6, Proposition 6.3] for the reader more comfortable in the cylindrical setting), these moduli spaces are orientable whenever they are smoothly cut out. There this is shown by trivializing the determinant line bundle  $\mathcal{L}$  of the virtual index bundle of the linearized  $\bar{\partial}$ -equation defining the moduli space in question, so when necessary we will specify our orientations by specifying sections of these determinant line bundles.

In order for these orientations to allow for the structure of a chain complex on the Heegaard Floer chain modules, we actually need somewhat more: we want the moduli spaces for different homotopy classes of disks to be oriented coherently. To make this precise, Ozsváth and Szabó used the notion of a *coherent orientation system* for

the moduli spaces of holomorphic disks in a Heegaard diagram  $\mathcal{H} = (\Sigma, \alpha, \beta, z)$ . Such an orientation system consists of a collection  $\sigma_{\mathcal{H}} = \sigma_{\alpha, \beta} := \{\sigma_{\alpha, \beta}^A\}$  of sections  $\sigma_{\alpha, \beta}^A$  of the determinant line bundle  $\mathcal{L}$  over all possible homotopy classes of disks  $A \in \pi_2(x, y)$  (ranging over all  $x, y \in \mathbb{T}_{\alpha} \cap \mathbb{T}_{\beta}$ ). Roughly, the coherence condition amounts to requiring that these sections are compatible with a process of gluing holomorphic disks together. We refer the reader to [11] for the precise definition of the coherence condition, or to Section 8.2 where we will formulate a precise version of the notion in the cylindrical setting. For our purposes in this section, we mainly just want to recall the fact that every pointed Heegaard diagram equipped with complex structure data achieving transversality admits a coherent orientation system by the remarks following [11, Definition 3.12]. We also want to make explicit the following equivalence relation on orientation systems.

**Definition 6.1** Fix two coherent orientation systems  $\sigma_{\alpha, \beta}$  and  $\sigma'_{\alpha, \beta}$  on a diagram  $\mathcal{H} = (\Sigma, \alpha, \beta, z)$ . We say the orientation systems are *equivalent* if there is a function

$$\epsilon: \mathbb{T}_{\alpha} \cap \mathbb{T}_{\beta} \rightarrow \{\pm 1\}$$

such that for each  $x, y \in \mathbb{T}_{\alpha} \cap \mathbb{T}_{\beta}$ ,

$$\sigma_{\alpha, \beta}^A = \epsilon(x) \cdot \epsilon(y) \cdot \sigma'_{\alpha, \beta}{}^A$$

for all  $A \in \mathcal{M}(x, y)$ .

It follows directly from the definition of the differential on  $CF^-$  that equivalent orientation systems give rise to isomorphic Heegaard Floer chain complexes. In what follows, we will often be concerned with specifying orientation systems which are unique up to equivalence. For these discussions, it will be useful to explicitly recall one more definition from the literature.

**Definition 6.2** [11, Definition 3.12] Given a  $\text{Spin}^c$  structure  $\mathfrak{s}$ , a strongly  $\mathfrak{s}$ -admissible diagram  $\mathcal{H} = (\Sigma, \alpha, \beta, z)$ , and an intersection point  $\mathbf{x}_0 \in \mathbb{T}_{\alpha} \cap \mathbb{T}_{\beta}$ , we will say a collection of classes  $\{A_{\mathbf{y}}\}$  where  $A_{\mathbf{y}} \in \pi_2(\mathbf{x}_0, \mathbf{y})$  and  $\mathbf{y}$  ranges over the intersection points in  $(\mathbb{T}_{\alpha} \cap \mathbb{T}_{\beta}) \setminus \{\mathbf{x}_0\}$  which represent  $\mathfrak{s}$ , is a *complete set of paths (based at  $\mathbf{x}_0$ )* for  $(\mathcal{H}, \mathfrak{s})$ .

**6.2.2 Coherent orientation systems of triangles** Given a pointed Heegaard triple diagram  $\mathcal{T} = (\Sigma, \alpha, \beta, \gamma, z)$ , we also note that moduli spaces of holomorphic triangles in a homotopy class  $\psi$ , which we denote by  $\mathcal{M}^{\psi}$  or  $\mathcal{M}(\psi)$ , are also orientable when

they are smoothly cut out, by [11, Section 8.2] (or [6, Proposition 10.3]). Given a collection  $\mathfrak{o}_{\mathcal{T}} := \{\mathfrak{o}_{\alpha,\beta,\gamma}, \mathfrak{o}_{\alpha,\beta}, \mathfrak{o}_{\beta,\gamma}, \mathfrak{o}_{\alpha,\gamma}\}$ , where  $\mathfrak{o}_{\alpha,\beta,\gamma}$  is a collection of sections of the determinant line bundle over all homotopy classes of triangles, and  $\mathfrak{o}_{\alpha,\beta}$ ,  $\mathfrak{o}_{\beta,\gamma}$  and  $\mathfrak{o}_{\alpha,\gamma}$  are collections of sections of the determinant line bundle over all homotopy classes of disks in the respective double diagrams, we will consider a related notion of coherence; see [11, Definition 8.6]. Roughly, the coherence condition here will amount to the requirement that each collection of orientations of the moduli spaces of strips on the respective double diagrams are coherent, and that all possible pregluings of triangles with strips satisfy the analogous gluing condition (this coherence condition will also be spelled out precisely in Section 8.2). The existence of such coherent orientation systems is guaranteed by the following result.

**Lemma 6.3** [11, Lemma 8.7] *Fix a pointed Heegaard triple diagram  $(\Sigma, \alpha, \beta, \gamma, z)$ , and let  $\mathfrak{s}$  be a  $\text{Spin}^c$  structure on  $X_{\alpha,\beta,\gamma}$  whose restriction to each boundary component is realized by an intersection point in the corresponding Heegaard diagram. Then for any coherent orientation systems  $\mathfrak{o}_{\alpha,\beta}$  and  $\mathfrak{o}_{\beta,\gamma}$  for two of the boundary components, there exists at least one coherent orientation system  $\mathfrak{o}_{\alpha,\gamma}$  for the remaining boundary component and a coherent orientation system  $\mathfrak{o}_{\alpha,\beta,\gamma}$  such that the entire collection of orientations is coherent.*

**Remark 6.4** We note here that this lemma does not guarantee that the orientation systems  $\mathfrak{o}_{\alpha,\gamma}$  and  $\mathfrak{o}_{\alpha,\beta,\gamma}$  are unique, as can be seen from inspection of the proof provided in [11, Lemma 8.7]. We mainly provide the reference to this lemma as it is stated for background context on the existence of coherent orientation systems. In what follows we will actually be interested in using a strengthened version of this lemma that applies in a particular situation to produce a unique induced coherent orientation system, which we will specify more precisely when the time comes. We note in particular that we only cite Lemma 6.3 in two places in this paper (in Sections 6.4 and 6.8), and in both cases an additional argument is used to explain why the induced orientation system is unique in the context under consideration.

It will be useful later to have a clear understanding of the indeterminacy in the orientation systems furnished by this lemma, and to have terminology with which we can refer to the sources of indeterminacy. To do so, we will now describe a high level outline of the proof of the above lemma, and point out explicitly where in the proof the indeterminacies arise. For details of the proof, we just point to the original source, since we have no new perspectives or value to add in reproducing them.

Assume we have fixed  $\circ_{\alpha,\beta}$  and  $\circ_{\beta,\gamma}$  as in the statement of the lemma. The way to produce  $\circ_{\alpha,\gamma}$  and the coherent orientation system  $\circ_{\alpha,\beta,\gamma}$  on the triple diagram guaranteed by the lemma can be summarized as follows.

- (1) Choose an arbitrary orientation over a single class of triangle  $\psi_0 \in \pi_2(x_0, y_0, z_0)$  connecting intersection points  $x_0, y_0$  and  $z_0$ .
- (2) Next, fix orientations over all periodic classes  $\phi_{\alpha,\gamma} \in \Pi_{z_0} \subset \pi_2(z_0, z_0)$  as follows:

- (a) Define a subgroup  $K \subset \Pi_{z_0}$  by

$$K = \{ \phi_{\alpha,\gamma} \in \pi_2(z_0, z_0) \mid \exists \phi_{\alpha,\beta} \in \pi_2(x_0, x_0), \phi_{\beta,\gamma} \in \pi_2(y_0, y_0) \text{ such that } \psi_0 + \phi_{\alpha,\gamma} = \psi_0 + \phi_{\alpha,\beta} + \phi_{\beta,\gamma} \}.$$

- (b) Show that the periodic classes split as

$$\Pi_{z_0} = K \oplus Q$$

for some free abelian group  $Q$ .

- (c) Using the defining property of  $K$  and a small lemma, extend  $\circ_{\alpha,\gamma}$  over all periodic classes in  $K$  such that the resulting orientations are consistent with  $\circ_{\alpha,\beta}$  and  $\circ_{\beta,\gamma}$ .
  - (d) Choose the orientations  $\circ_{\alpha,\gamma}$  arbitrarily over a basis for  $Q$ . We will call this collection of choices the *indeterminacy over  $Q$* .
  - (e) Obtain orientations over all classes of triangles  $\psi \in \pi_2(x_0, y_0, z_0)$  by bootstrapping from the above.
- (3) Next, choose a complete set of paths for  $Y_{\alpha,\gamma}$ , and choose orientations for  $\circ_{\alpha,\gamma}$  over the classes defining the complete set of paths. We will call this collection of choices the *indeterminacy over a complete set of paths*.
  - (4) The previously defined orientations together uniquely determine a coherent orientation system for the triple diagram. We see that, up to a sign, the indeterminacy in the orientation systems  $\circ_{\alpha,\gamma}$  and  $\circ_{\alpha,\beta,\gamma}$  furnished by the lemma is due to the indeterminacy over  $Q$  and the indeterminacy over a complete set of paths.

Finally, we note that the indeterminacy over a complete set of paths mentioned above does in fact vanish in general, so long as we consider orientation systems up to equivalence. For given a complete set of paths and two orientation systems  $\circ_{\alpha,\gamma}$  and  $\circ'_{\alpha,\gamma}$  which differ on the complete set of paths, it is straightforward to construct a third orientation system  $\circ''_{\alpha,\gamma}$  which is equivalent to  $\circ'_{\alpha,\gamma}$  and which agrees with

$\circ_{\alpha, \gamma}$  on the complete set of paths. Indeed, if the complete set of paths is denoted by  $\{A_{\mathbf{y}}\}$ , construct an equivalence function  $\epsilon$  by declaring for each  $\mathbf{y}$  that  $\epsilon(\mathbf{y}) = -1$  if  $\circ_{\alpha, \gamma}(A_{\mathbf{y}}) \neq \circ'_{\alpha, \gamma}(A_{\mathbf{y}})$ . Altering  $\circ'_{\alpha, \gamma}$  by any such equivalence function will yield an orientation system  $\circ''_{\alpha, \gamma}$  as desired. Thus up to equivalence and sign, we see that the indeterminacy in the orientation systems  $\circ_{\alpha, \gamma}$  and  $\circ_{\alpha, \beta, \gamma}$  furnished by the lemma is solely due to the indeterminacy over  $Q$  (coming from the indeterminacy in the orientations over the periodic classes).

### 6.3 Change of almost complex structures

Next, we recall the dependence of the construction of the Heegaard Floer invariants on the choices of almost complex structures involved. The definition of the Heegaard Floer chain complex associated to a pointed Heegaard diagram  $(\Sigma, \alpha, \beta, z)$  in fact requires a choice of complex structure  $j$  on  $\Sigma$ , and a generic path of almost complex structures  $J_s \subset \mathcal{U}$  on  $\text{Sym}^g(\Sigma)$  going through the structure  $\text{Sym}^g(j)$  induced by  $j$ . Here  $g$  is the genus of  $\Sigma$  and  $\mathcal{U}$  is a particular contractible set of almost complex structures specified by Ozsváth and Szabó in [11, Theorem 3.15 and Section 4.1]. Given a strongly  $\mathfrak{s}$ -admissible pointed Heegaard diagram  $\mathcal{H} = (\Sigma, \alpha, \beta, z)$ , a coherent orientation  $\circ$  on  $\mathcal{H}$ , and two choices of such almost complex structure data  $(j, J_s)$  and  $(j', J'_s)$ , there is a chain homotopy equivalence

$$\Phi_{J_s \rightarrow J'_s} : CF^-_{J_s}(\Sigma, \alpha, \beta, z, \mathfrak{s}, \circ) \rightarrow CF^-_{J'_s}(\Sigma, \alpha, \beta, z, \mathfrak{s}, \circ').$$

Here  $\circ'$  is an orientation system uniquely determined by  $\circ$ , as described in the beginning of [6, Section 9]. These equivalences fit into a transitive system in the homotopy category of chain complexes of  $\mathbb{Z}[U]$ -modules, in the sense that  $\Phi_{J_s \rightarrow J_s} \sim \text{id}_{CF^-(\Sigma, \alpha, \beta)}$  and  $\Phi_{J'_s \rightarrow J''_s} \circ \Phi_{J_s \rightarrow J'_s} \sim \Phi_{J_s \rightarrow J''_s}$ . This is shown in [14, Lemma 2.11]. We denote this transitive system in the homotopy category of complexes of  $\mathbb{Z}[U]$ -modules by

$$CF^-(\Sigma, \alpha, \beta, z, \mathfrak{s}, \circ).$$

Of course we also obtain from the maps  $\Phi_{J_s \rightarrow J'_s}$  a transitive system of isomorphisms on homology. We will denote the colimit of the  $\mathbb{Z}[U]$ -modules  $HF^-_{J_s}(\Sigma, \alpha, \beta, z, \mathfrak{s}, \circ)$  with respect to this transitive system by

$$HF^-(\Sigma, \alpha, \beta, z, \mathfrak{s}, \circ).$$

### 6.4 Triangle maps and continuation maps

Given a pointed Heegaard triple diagram  $\mathcal{T} = (\Sigma, \alpha, \beta, \gamma, z)$  which is strongly  $\mathfrak{s}$ -admissible for a  $\text{Spin}^c$  structure  $\mathfrak{s}$  on  $X_{\alpha, \beta, \gamma}$ , as well as a coherent orientation system

$\circ_{\alpha, \beta, \gamma}$  compatible with coherent orientation systems  $\circ_{\alpha, \beta}$ ,  $\circ_{\beta, \gamma}$  and  $\circ_{\alpha, \gamma}$ , there are  $\mathbb{Z}[U]$ -module chain maps  $\mathcal{F}_{\alpha, \beta, \gamma}(\cdot, \mathfrak{s}, \circ_{\alpha, \beta, \gamma})$  of the form

$$CF_{J_s}^-(\Sigma, \alpha, \beta, \mathfrak{s}_{\alpha, \beta}, \circ_{\alpha, \beta}) \otimes_{\mathbb{Z}[U]} CF_{J_s}^-(\Sigma, \beta, \gamma, \mathfrak{s}_{\beta, \gamma}, \circ_{\beta, \gamma}) \rightarrow CF_{J_s}^-(\Sigma, \alpha, \gamma, \mathfrak{s}_{\alpha, \gamma}, \circ_{\alpha, \gamma})$$

defined in [11, Theorem 8.12]. Here, and throughout this section, we sometimes suppress the basepoint  $z$  from the chain complex notation for brevity, but the dependence is always implied. Put simply, these chain maps count pseudoholomorphic triangles on the triple diagram. In fact, the homotopy class of the chain map  $F_{\alpha, \beta, \gamma}$  does not depend on the choice of almost complex structure data. More precisely, for two choices of almost complex structure data the maps above commute up to homotopy with the change of almost complex structure maps by [11, Proposition 8.13]. Thus with respect to the transitive systems  $CF^-(\Sigma, \alpha, \beta, z, \mathfrak{s}, \circ)$ , the map  $F_{\alpha, \beta, \gamma}$  is a morphism in  $\text{Trans}(\text{Kom}(\mathbb{Z}[U]\text{-Mod}))$ , ie a morphism between two transitive systems in the homotopy category of  $\mathbb{Z}[U]$  modules. We denote this morphism by  $\mathcal{F}_{\alpha, \beta, \gamma}(\cdot, \mathfrak{s}, \circ_{\alpha, \beta, \gamma})$  and it takes the form

$$CF^-(\Sigma, \alpha, \beta, \mathfrak{s}_{\alpha, \beta}, \circ_{\alpha, \beta}) \otimes_{\mathbb{Z}[U]} CF^-(\Sigma, \beta, \gamma, \mathfrak{s}_{\beta, \gamma}, \circ_{\beta, \gamma}) \rightarrow CF^-(\Sigma, \alpha, \gamma, \mathfrak{s}_{\alpha, \gamma}, \circ_{\alpha, \gamma}).$$

We also obtain induced maps of  $\mathbb{Z}[U]$ -modules  $\mathcal{F}_{\alpha, \beta, \gamma}(\cdot, \mathfrak{s}, \circ_{\alpha, \beta, \gamma})$  of the form

$$HF^-(\Sigma, \alpha, \beta, \mathfrak{s}_{\alpha, \beta}, \circ_{\alpha, \beta}) \otimes_{\mathbb{Z}[U]} HF^-(\Sigma, \beta, \gamma, \mathfrak{s}_{\beta, \gamma}, \circ_{\beta, \gamma}) \rightarrow HF^-(\Sigma, \alpha, \gamma, \mathfrak{s}_{\alpha, \gamma}, \circ_{\alpha, \gamma}).$$

The triangle maps above allow one to define maps associated to handleslides. To describe the handleslide maps, we first recall the following fact.

**Lemma 6.5** [11, Lemma 9.4, Remark 9.2 and Section 9.1; 5, Lemma 9.2] *Let  $(\Sigma, \beta, \gamma', z)$  be a pointed genus  $g$  Heegaard diagram such that  $\gamma'$  can be obtained from  $\beta$  by performing a sequence of handleslides among the curves in  $\beta$ . Then the diagram represents  $\#^g(S^1 \times S^2)$ . There is a unique  $\text{Spin}^c$  structure  $\mathfrak{s}_0 \in \text{Spin}^c(\#^g(S^1 \times S^2))$  such that  $c_1(\mathfrak{s}_0) = 0$ , and upon performing a particular small Hamiltonian isotopy of  $\gamma'$  — specified in [11] — to obtain  $(\Sigma, \beta, \gamma, z)$ , one can ensure this new diagram is strongly  $\mathfrak{s}_0$ -admissible. Furthermore, there is a choice of coherent orientation system  $\circ_{\beta, \gamma}$  on this diagram such that*

$$\widehat{HF}(\Sigma, \beta, \gamma, z, \mathfrak{s}_0, \circ_{\beta, \gamma}) \cong H_*(T^g; \mathbb{Z}),$$

$$HF^-(\Sigma, \beta, \gamma, z, \mathfrak{s}_0, \circ_{\beta, \gamma}) \cong \mathbb{Z}[U] \otimes H_*(T^g; \mathbb{Z}).$$

*In this case it follows that in the highest nontrivial relative homological grading  $HF^-(\Sigma, \beta, \gamma, z, \mathfrak{s}_0, \circ_{\beta, \gamma})$  is isomorphic to  $\mathbb{Z} =: \langle \theta_{\beta, \gamma} \rangle$ , for a generator we denote*

by  $\theta_{\beta, \gamma}$ . Finally, there is only one equivalence class of orientation system with these properties.

**Remark 6.6** For such a diagram, we can also identify a particular intersection point  $\theta_{\beta, \gamma} \in CF^-(\Sigma, \beta, \gamma, z, \mathfrak{s}_0, \mathfrak{o}_{\beta, \gamma})$  representing this element of homology. Indeed, the strongly admissible diagram referred to in the lemma statement yields a chain complex whose rank is the same as that of its homology, and which has a unique intersection point realizing  $\mathfrak{s}_0$  in the maximal relative grading.

**Remark 6.7** All of the statements in the lemma other than the last sentence are explicitly proved in the cited references. The last sentence is also contained implicitly in the references cited, but since it is particularly relevant to our arguments we provide a sketch of the proof below.

The last sentence in Lemma 6.5 follows from the next result.

**Lemma 6.8** *Equivalence classes of coherent orientation systems over the diagram  $(\Sigma, \beta, \gamma, z)$  for  $(S^1 \times S^2)^{\#g}$  from Lemma 6.5 are in bijection with morphisms*

$$\pi_1(T^g) \rightarrow \text{Aut}(\mathbb{Z}),$$

where  $T^g$  is a torus. Furthermore, for a corresponding orientation system  $\mathfrak{o}$  and morphism  $\mathcal{L}$ ,

$$\widehat{HF}((S^1 \times S^2)^{\#g}, \mathfrak{o}) \cong H_*(T^g; \mathcal{L}).$$

**Proof sketch** Fix a diagram  $(\Sigma, \beta, \gamma, z)$  for  $(S^1 \times S^2)^{\#g}$  as described in Lemma 6.5, an intersection point  $\mathbf{x}_0 \in \mathbb{T}_{\beta} \cap \mathbb{T}_{\gamma}$ , and a complete set of paths based at  $\mathbf{x}_0$ . As described in Remark 6.4, all coherent orientation systems on the diagram agree on the complete set of paths up to equivalence. Thus equivalence classes of coherent orientation systems are determined by their values on a basis for the periodic domains based at  $\mathbf{x}_0$ . Note that specifying values in  $\{\pm 1\}$  for each class in a basis for the periodic domains based at  $\mathbf{x}_0$  is the same as specifying a morphism  $\pi_1(T^g) \rightarrow \text{Aut}(\mathbb{Z})$ , since the group of periodic classes is identified with  $H^1((S^1 \times S^2)^{\#g})$ . This establishes the first sentence in the lemma.

The second statement in the lemma follows from a direct comparison of the contributions to homology (Heegaard Floer or singular) in the diagrams in question for a given choice of values over a basis for the periodic domains based at  $\mathbf{x}_0$ . For example, assigning 1



to each periodic domain corresponds to the isomorphism class of local system over  $T^g$  specified by the trivial homomorphism  $\mathcal{L}: \pi_1(T^g) \rightarrow \mathbb{Z}/2\mathbb{Z}$ , and to some equivalence class of coherent orientation system on  $(\Sigma, \beta, \gamma, z)$ . Using the local picture and calculations developed in [11, Lemma 9.4], one can establish an identification between generators of  $\pi_1$  and generators of  $\widehat{HF}$ .  $\square$

Now to establish the last sentence in Lemma 6.5, just note that there is a single local system  $\mathcal{L}$  over the torus  $T^g$  for which the singular homology is  $H_*(T^g; \mathbb{Z})$  (namely the trivial local system).

Given a strongly  $\mathfrak{s}$ -admissible triple diagram  $(\Sigma, \alpha, \beta, \gamma, z)$  with  $\gamma$  related to  $\beta$  as in the statement of Lemma 6.5, we will write

$$\Psi_{\beta \rightarrow \gamma}^\alpha(\cdot, \mathfrak{s}, \mathfrak{o}_{\alpha, \beta, \gamma}) := F_{\alpha, \beta, \gamma}(\cdot \otimes \theta_{\beta, \gamma}, \mathfrak{s}, \mathfrak{o}_{\alpha, \beta, \gamma}),$$

where

$$F_{\alpha, \beta, \gamma}(\cdot \otimes \theta_{\beta, \gamma}, \mathfrak{s}, \mathfrak{o}_{\alpha, \beta, \gamma}): CF^-(\Sigma, \alpha, \beta, z, \mathfrak{s}_{\alpha, \beta}, \mathfrak{o}_{\alpha, \beta}) \rightarrow CF^-(\Sigma, \alpha, \gamma, z, \mathfrak{s}_{\alpha, \gamma}, \mathfrak{o}_{\alpha, \gamma}).$$

Here we have used an arbitrary coherent orientation system  $\mathfrak{o}_{\alpha, \beta}$  and the coherent orientation system  $\mathfrak{o}_{\beta, \gamma}$  of Lemma 6.5, and enlarged them to a coherent orientation system  $\mathfrak{o}_{\alpha, \beta, \gamma}$ . That this can be done in some way is ensured by Lemma 6.3; in fact, though, this enlargement is unique up to equivalence in this particular case, as we now explain. Recall we have seen in Remark 6.4 that the indeterminacy in the equivalence classes of the orientation systems furnished by Lemma 6.3 is due solely to the indeterminacy over the group  $Q$ . It is shown in the proof of [11, Lemma 8.7] that this group  $Q$  is the image of the composition  $q \circ i$ ,

$$H_2(Y_{\alpha, \gamma}) \xrightarrow{i} H_2(X_{\alpha, \beta, \gamma}) \xrightarrow{q} H_2(X_{\alpha, \beta, \gamma}, Y_{\alpha, \beta} \cup Y_{\beta, \gamma})$$

where  $i$  is induced by inclusion and  $q$  comes from the relative long exact sequence for the relevant pair. In the case at hand, we have  $Y_{\alpha, \beta} \cong Y_{\alpha, \gamma}$  are arbitrary 3-manifolds, and  $X_{\alpha, \beta, \gamma}$  is  $Y_{\alpha, \beta} \times I$  with a neighborhood of a bouquet of  $g$  circles removed. Thus we have  $i(H_2(Y_{\alpha, \beta})) = i(H_2(Y_{\alpha, \gamma}))$ , and  $Q = 0$ . This establishes that the coherent orientation systems used in our definition of the map  $\Psi_{\beta \rightarrow \gamma}^\alpha$  above are well defined. Similarly, if instead  $\beta$  is related to  $\alpha$  as in the statement of Lemma 6.5, we will write

$$\Psi_{\gamma}^{\alpha \rightarrow \beta}(\cdot, \mathfrak{s}, \mathfrak{o}_{\beta, \alpha, \gamma}) := F_{\beta, \alpha, \gamma}(\theta_{\beta, \alpha} \otimes \cdot, \mathfrak{s}, \mathfrak{o}_{\beta, \alpha, \gamma}),$$

where

$$F_{\beta, \alpha, \gamma}(\theta_{\beta, \alpha} \otimes \cdot, \mathfrak{s}, \mathfrak{o}_{\beta, \alpha, \gamma}): CF^-(\Sigma, \alpha, \gamma, z, \mathfrak{s}_{\alpha, \gamma}, \mathfrak{o}_{\alpha, \gamma}) \rightarrow CF^-(\Sigma, \beta, \gamma, z, \mathfrak{s}_{\beta, \gamma}, \mathfrak{o}_{\beta, \gamma}).$$

These can be thought of as maps on the Floer invariants associated to (small variations of) sequences of handleslides on diagrams. These maps are in fact homotopy equivalences according to the following result.

**Lemma 6.9** [11, Theorem 9.5 and Section 9.1] (1) *If  $(\Sigma, \alpha, \beta, \gamma, z)$  is a strongly  $s$ -admissible triple diagram and  $\beta$  is related to  $\gamma$  as in the statement of Lemma 6.5, then  $\Psi_{\beta \rightarrow \gamma}^\alpha$  is a chain homotopy equivalence.*

(2) *Furthermore, such equivalences are transitive: for two triples satisfying the conditions above we have*

$$\Psi_{\beta \rightarrow \gamma}^\alpha \sim \Psi_{\delta \rightarrow \gamma}^\alpha \circ \Psi_{\beta \rightarrow \delta}^\alpha.$$

(3) *The analogous results hold for the maps induced by changing the  $\alpha$  curves.*

There are also maps associated to special Hamiltonian isotopies of diagrams [11, Proof of Theorem 7.3]. Given strongly  $s$ -admissible diagrams  $(\Sigma, \alpha, \beta, z)$  and  $(\Sigma, \alpha', \beta', z)$  and an exact Hamiltonian isotopy on  $(\Sigma, \omega)$  taking  $\alpha$  to  $\alpha'$  and  $\beta$  to  $\beta'$ , which furthermore never crosses the basepoint, we claim that each coherent orientation system  $\circ_{\alpha, \beta}$  for the first diagram determines a unique equivalence class of coherent orientation system  $\circ_{\alpha', \beta'}$  for the second. This is part of the statement of [11, Theorem 7.3], and can be understood as follows. First note that it will suffice to show that there is a correspondence  $\pi_2(\mathbf{x}, \mathbf{x})_{\mathcal{H}_1} \cong \pi_2(\mathbf{y}, \mathbf{y})_{\mathcal{H}_2}$  between homotopy classes of periodic disks based at some intersection point  $\mathbf{x}$  on  $\mathcal{H}_1$  and homotopy classes of periodic disks based at some intersection point  $\mathbf{y}$  on  $\mathcal{H}_2$ . With this fact established, a coherent orientation system on the first diagram uniquely determines an equivalence class on the second diagram, since as we have already observed equivalence classes of orientation systems on  $\mathcal{H}_2$  are determined by their values on the periodic domains based at a single intersection point. The correspondence  $\pi_2(\mathbf{x}, \mathbf{x})_{\mathcal{H}_1} \cong \pi_2(\mathbf{y}, \mathbf{y})_{\mathcal{H}_2}$  is realized by a certain concatenation with a homotopy class with varying boundary conditions, as we now explain.

Following [11, Proof of Theorem 7.3], let us denote our isotopy by  $\Psi_t: \Sigma \rightarrow \Sigma$  and set  $\alpha_t = \Psi_t(\alpha)$  and  $\beta_t = \Psi_t(\beta)$ . Define  $\pi_2^{\Psi_t}(\mathbf{x}, \mathbf{y})$  to be the set of homotopy classes of Whitney disks which connect  $\mathbf{x} \in \mathbb{T}_{\alpha} \cap \mathbb{T}_{\beta}$  to  $\mathbf{y} \in \mathbb{T}_{\alpha'} \cap \mathbb{T}_{\beta'}$  and have boundary conditions  $u(0, t) \in \alpha_t$ ,  $u(1, t) \in \beta_t$ . We now explain how a single class  $\phi \in \pi_2^{\Psi_t}(\mathbf{x}, \mathbf{y})$  establishes the desired correspondence  $\pi_2(\mathbf{x}, \mathbf{x})_{\mathcal{H}_1} \cong_{\phi} \pi_2(\mathbf{y}, \mathbf{y})_{\mathcal{H}_2}$  via a certain conjugation. Given  $u$  representing  $A \in \pi_2(\mathbf{x}, \mathbf{x})_{\mathcal{H}_1}$  and a disk  $v$  representing the class  $\phi \in \pi_2^{\Psi_t}(\mathbf{x}, \mathbf{y})$ , we can construct a disk  $\bar{v} \natural u \natural v$  by concatenation. Such

a disk lies in  $\pi_2^{\Psi_{1-t} * \text{Id} * \Psi_t}(\mathbf{y}, \mathbf{y})$ , which is the set of homotopy classes of Whitney disks which connect  $\mathbf{y} \in \mathbb{T}_{\alpha'} \cap \mathbb{T}_{\beta'}$  to itself, and have boundary conditions matching  $\bar{\alpha} * \alpha_0 * \alpha$  and  $\bar{\beta} * \beta_0 * \beta$  on its two sides, where  $\bar{\alpha}$  and  $\bar{\beta}$  are the curves traversed in the opposite direction. We now claim two things:

- (1) This correspondence establishes a bijection

$$\pi_2(\mathbf{x}, \mathbf{x}) \cong \pi_2^{\Psi_{1-t} * \text{Id} * \Psi_t}(\mathbf{y}, \mathbf{y}).$$

- (2) There is also a bijection

$$\pi_2^{\Psi_{1-t} * \text{Id} * \Psi_t}(\mathbf{y}, \mathbf{y}) \cong \pi_2(\mathbf{y}, \mathbf{y}).$$

We omit the proofs of these facts, but note that both can be understood by thinking of the space of periodic domains at  $\mathbf{x}$  as a subspace of the fundamental group of the path space between the Heegaard curves, based at the constant path  $\mathbf{x}$ . In this context, one can show that an isotopy of the Heegaard curves gives rise to an identification between path spaces, and that the class  $\phi$  yields an identification between the corresponding loop spaces. This line of reasoning can be used to establish both bijections. For the interested reader, a precise argument explaining related facts in a more general setting can be found in [2, Section 3.3]. Finally, one should note that a class  $\phi \in \pi_2^{\Psi_t}(\mathbf{x}, \mathbf{y})$  does in fact exist for  $\mathbf{y} = \Psi_1(\mathbf{x})$ , because given an intersection point  $\mathbf{x} \in \mathbb{T}_{\alpha} \cap \mathbb{T}_{\beta}$ , we may just follow it with the isotopy to obtain a disk  $u(s, t) = \Psi_t(\mathbf{x})$  which satisfies the requirements for a disk with varying boundary conditions between  $\mathbf{x}$  and  $\mathbf{y} = \Psi_1(\mathbf{x})$ . This completes the explanation of the identification between equivalence classes of coherent orientation systems on  $(\Sigma, \alpha, \beta, z)$  and  $(\Sigma, \alpha', \beta', z)$ .

With respect to the aforementioned orientation systems there is an induced chain homotopy equivalence

$$\Gamma_{\beta \rightarrow \beta'}^{\alpha \rightarrow \alpha'} : CF^-(\Sigma, \alpha, \beta, z, s, o_{\alpha, \beta}) \rightarrow CF^-(\Sigma, \alpha', \beta', z, s, o_{\alpha', \beta'}),$$

which we call a continuation map associated to the Hamiltonian isotopy  $\phi_t$ . We will also use the notation

$$\Gamma_{\beta}^{\alpha \rightarrow \alpha'} = \Gamma_{\beta \rightarrow \beta}^{\alpha \rightarrow \alpha'} \quad \text{and} \quad \Gamma_{\beta \rightarrow \beta'}^{\alpha} = \Gamma_{\beta \rightarrow \beta'}^{\alpha \rightarrow \alpha}.$$

By [14, Lemma 2.12], these equivalences compose naturally under concatenation of isotopies in the sense that

$$\Gamma_{\beta}^{\alpha \rightarrow \alpha''} \sim \Gamma_{\beta}^{\alpha' \rightarrow \alpha''} \circ \Gamma_{\beta}^{\alpha \rightarrow \alpha'} \quad \text{and} \quad \Gamma_{\beta \rightarrow \beta'}^{\alpha \rightarrow \alpha'} \sim \Gamma_{\beta'}^{\alpha \rightarrow \alpha'} \circ \Gamma_{\beta \rightarrow \beta'}^{\alpha} \sim \Gamma_{\beta \rightarrow \beta'}^{\alpha'} \circ \Gamma_{\beta}^{\alpha \rightarrow \alpha'}.$$

Furthermore, by their definition in [11, Proof of Theorem 7.3], they satisfy  $\Gamma_{\beta \rightarrow \beta}^{\alpha \rightarrow \alpha} = \text{id}_{CF^-(\Sigma, \alpha, \beta, z, \mathfrak{s}, \mathfrak{o}_{\alpha, \beta})}$ .

As suggested by the notation, we note that while the continuation map is a priori associated to a Hamiltonian isotopy between the isotopic attaching curves, in the cases of interest for us its chain homotopy class will actually be independent of the choice of isotopy. To see this, we recall:

**Lemma 6.10** [11, Lemma 9.1 and Section 9.1] *Let  $(\Sigma, \beta, \beta', z)$  be a pointed diagram such that each curve  $\beta'_i$  in  $\beta'$  is obtained from the curve  $\beta_i$  in  $\beta$  by performing a small Hamiltonian isotopy which introduces two transverse intersection points between  $\beta_i$  and  $\beta'_i$ , and no intersection points between  $\beta'_i$  and  $\beta_j$  for  $j \neq i$ . Then the diagram represents  $\#^g(S^1 \times S^2)$ . There is a unique  $\text{Spin}^c$  structure  $\mathfrak{s}_0 \in \text{Spin}^c(\#^g(S^1 \times S^2))$  such that  $c_1(\mathfrak{s}_0) = 0$ , and the diagram  $(\Sigma, \beta, \beta', z)$  is strongly  $\mathfrak{s}_0$ -admissible. Furthermore, there is a choice of coherent orientation system  $\mathfrak{o}_{\beta, \beta'}$  on this diagram such that in the highest nontrivial relative homological grading  $HF^-(\Sigma, \beta, \beta', z, \mathfrak{s}_0, \mathfrak{o}_{\beta, \beta'})$  is isomorphic to  $\mathbb{Z} =: \langle \theta_{\beta, \beta'} \rangle$  for a generator we denote by  $\theta_{\beta, \beta'}$ .*

Using the generator  $\theta_{\beta, \beta'}$  we have an analogous triangle map to that defined above, which is also shown to be an equivalence:

**Lemma 6.11** [11, Theorem 9.8 and Section 9.1] *If  $(\Sigma, \alpha, \beta, \beta', z)$  is a strongly  $\mathfrak{s}$ -admissible triple diagram and  $\beta'$  is related to  $\beta$  as in the statement of Lemma 6.10 by a sufficiently small isotopy, then*

$$F_{\alpha, \beta, \beta'}(\cdot \otimes \theta_{\beta, \beta'}) : CF^-(\Sigma, \alpha, \beta, z, \mathfrak{s}_{\alpha, \beta}, \mathfrak{o}_{\alpha, \beta}) \rightarrow CF^-(\Sigma, \alpha, \beta', z, \mathfrak{s}_{\alpha, \beta'}, \mathfrak{o}_{\alpha, \beta'})$$

*is a chain homotopy equivalence.*

Furthermore, we have:

**Lemma 6.12** [6, Proposition 11.4] *If the triple diagram  $(\Sigma, \alpha, \beta, \beta', z)$  is strongly  $\mathfrak{s}$ -admissible and  $\beta'$  is related to  $\beta$  as in the statement of Lemma 6.10 by a sufficiently small isotopy, then the continuation map associated to any Hamiltonian isotopy  $\phi_t$  between  $\beta$  and  $\beta'$  satisfies*

$$\Gamma_{\beta \rightarrow \beta'}^{\alpha} \sim F_{\alpha, \beta, \beta'}(\cdot \otimes \theta_{\beta, \beta'}).$$

We thus see that the continuation maps associated to small Hamiltonian isotopies of the attaching curves are independent of the choice of isotopy.

Finally, we introduce notation for a composition of triangle maps and continuation maps associated to strong  $\alpha$ -equivalences and strong  $\beta$ -equivalences.

**Definition 6.13** [14, Section 2 and Lemma 2.13] Given two strongly  $\mathfrak{s}$ -admissible diagrams  $(\Sigma, \alpha_1, \beta_1, z)$  and  $(\Sigma, \alpha_2, \beta_2, z)$  which are strongly equivalent, one can construct another pointed diagram  $(\Sigma, \alpha'_1, \beta'_1, z)$  such that:

- (1)  $\alpha'_1$  and  $\beta'_1$  are obtained respectively from  $\alpha_1$  and  $\beta_1$  by special isotopies.
- (2)  $\alpha_2$  and  $\beta_2$  are obtained respectively from  $\alpha'_1$  and  $\beta'_1$  by (small variations of) sequences of handleslides as in Lemma 6.5.
- (3) The quadruple diagram  $(\Sigma, \alpha'_1, \beta'_1, \alpha_2, \beta_2)$  is strongly  $\mathfrak{s}$ -admissible for the unique  $\text{Spin}^c$ -structure on  $X_{\alpha'_1, \beta'_1, \alpha_2, \beta_2}$  which restricts to  $\mathfrak{s}$  on  $Y_{\alpha'_1, \beta_2}$  and  $\mathfrak{s}_0$  on  $Y_{\alpha'_1, \alpha_2}$  and  $Y_{\beta'_1, \beta_2}$ .

We define a map,

$$\Phi_{\beta_1 \rightarrow \beta_2}^{\alpha_1 \rightarrow \alpha_2}(\cdot, \mathfrak{s}) : CF^-(\Sigma, \alpha_1, \beta_1, z, \mathfrak{s}) \rightarrow CF^-(\Sigma, \alpha_2, \beta_2, z, \mathfrak{s})$$

associated to two such strongly equivalent diagrams by the formula

$$\Phi_{\beta_1 \rightarrow \beta_2}^{\alpha_1 \rightarrow \alpha_2}(\cdot, \mathfrak{s}) = \Psi_{\beta'_1 \rightarrow \beta_2}^{\alpha_2} \circ \Psi_{\beta'_1}^{\alpha'_1 \rightarrow \alpha_2} \circ \Gamma_{\beta_1 \rightarrow \beta'_1}^{\alpha_1 \rightarrow \alpha'_1}.$$

We will sometimes use the notation

$$\Phi_{\beta \rightarrow \beta'}^{\alpha} = \Phi_{\beta \rightarrow \beta'}^{\alpha \rightarrow \alpha} \quad \text{and} \quad \Phi_{\beta}^{\alpha \rightarrow \alpha'} = \Phi_{\beta \rightarrow \beta}^{\alpha \rightarrow \alpha'}.$$

### 6.5 The weak Heegaard Floer invariants

Using the previous two subsections, we are now in position to define the value on vertices of the morphism of graphs

$$CF^- : \mathcal{G}_{\text{man}} \rightarrow \text{Trans}(P(\text{Kom}(\mathbb{Z}[U]\text{-Mod})))$$

which will partially define the weak invariants underlying the maps in Theorem 1.3. In doing so, we will also define the value on vertices of the morphism of graphs

$$HF^- : \mathcal{G}_{\text{man}} \rightarrow P(\mathbb{Z}[U]\text{-Mod})$$

appearing in Corollary 1.4.

**Definition 6.14** Fix some pointed isotopy diagram  $H = (\Sigma, A, B, z)$  (corresponding to a vertex in  $\mathcal{G}_{\text{man}}$ ) representing the pointed 3-manifold  $(Y, z)$ . For  $\mathfrak{s} \in \text{Spin}^c(Y)$ , let

$$\text{Admiss}_{(\Sigma, A, B, z)}(\mathfrak{s}) = \{\text{strongly } \mathfrak{s}\text{-admissible diagrams } (\Sigma, \alpha, \beta, z) \mid [\alpha] = A, [\beta] = B\}$$

be the set of strongly  $\mathfrak{s}$ -admissible diagrams representing  $H$ . By [11, Proofs of Lemma 5.2 and Lemma 5.4], this is nonempty for all  $\mathfrak{s} \in \text{Spin}^c(Y)$ . Choose any diagram  $\mathcal{H} = (\Sigma, \alpha, \beta, z) \in \text{Admiss}_{(\Sigma, A, B, z)}(\mathfrak{s})$ , and fix a coherent orientation system  $\mathfrak{o}_{\alpha, \beta}$  on it. By [11, Lemma 7.3], the transitive system  $CF^-(\Sigma, \alpha, \beta, z, \mathfrak{s}, \mathfrak{o}_{\alpha, \beta})$  can be used along with the continuation maps  $\Gamma$  to induce coherent orientation systems for all strongly  $\mathfrak{s}$ -admissible diagrams representing the isotopy diagram  $H$ . Then by [14, Lemma 2.12], the transitive systems  $CF^-(\Sigma, \alpha, \beta, z, \mathfrak{s}, \mathfrak{o}_{\alpha, \beta})$  ranging over all  $(\Sigma, \alpha, \beta, z) \in \text{Admiss}_{(\Sigma, A, B, z)}(\mathfrak{s})$  fit into a transitive system (of morphisms between transitive systems) with respect to the continuation maps  $\Gamma_{\beta \rightarrow \beta'}^{\alpha \rightarrow \alpha'}$ . We can therefore define a single transitive system (see Section 4) in  $\text{Kom}(\mathbb{Z}[U]\text{-Mod})$ , which we denote by

$$CF^-(H, \mathfrak{s}).$$

Finally, we define the value of the weak Heegaard invariant  $CF^-$  on the isotopy diagram  $H$  by

$$CF^-(H) = \bigoplus_{\mathfrak{s} \in \text{Spin}^c(Y)} CF^-(H, \mathfrak{s}).$$

Passing to homology, we obtain instead that the  $\mathbb{Z}[U]$ -modules  $HF^-(\Sigma, \alpha, \beta, z, \mathfrak{s}, \mathfrak{o}_{\alpha, \beta})$  for  $(\Sigma, \beta, \alpha, z) \in \text{Admiss}_{(\Sigma, A, B, z)}(\mathfrak{s})$  fit into a transitive system of isomorphisms with respect to the continuation maps. We denote the colimit of this transitive system by

$$HF^-(H, \mathfrak{s})$$

and define

$$HF^-(H) = \bigoplus_{\mathfrak{s} \in \text{Spin}^c(Y)} HF^-(H, \mathfrak{s}).$$

We now proceed to fix the data of the underlying coherent orientation systems we will use to define  $CF^-(H')$  for all other isotopy diagrams  $H'$  in  $\mathcal{G}_{\text{man}}$ . First consider the path component of  $\mathcal{G}_{\text{man}}$  containing the fixed isotopy diagram  $H$  chosen above. We note that by Proposition 2.4, the collection of vertices in this path component corresponds to the collection of all isotopy diagrams representing the fixed 3-manifold  $(Y, z)$ . Given another isotopy diagram  $H'$  in this path component, choose a sequence of edges  $\gamma$  in  $(\mathcal{G}_{\text{man}})_{(Y, z)}$  from  $H$  to  $H'$ . For any diagrams  $\mathcal{H} \in H$  and  $\mathcal{H}'$  in  $H'$ , the constructions

described in the previous subsections yield a composition of maps associated to  $\gamma$  on the underlying chain complexes,

$$CF^-(\gamma): CF^-(\mathcal{H}) \rightarrow CF^-(\mathcal{H}').$$

Here the sequence of maps  $CF^-(\gamma)$  of course depends on our previously fixed choice of coherent orientation system for  $\mathcal{H}$ ; we described in the previous subsections how each of the possible constituent maps in the composition  $CF^-(\gamma)$  induces a coherent orientation system on the target given a coherent orientation system on the domain, and it is this induced orientation system that we fix on  $\mathcal{H}'$ . One can check that this induced orientation on  $\mathcal{H}'$  is independent of the choice of path  $\gamma$ , by verifying the commutativity of the induced orientations occurring in each of the five types of distinguished rectangle, and in a simple handleswap. We will verify this commutativity in Section 6.8. We thus see that our specification of the coherent orientation systems  $\mathfrak{o}_{\alpha,\beta}$  on all diagrams  $\mathcal{H}$  representing  $H$  actually yields a choice of coherent orientation systems for all diagrams in the same path component as  $H$ . Repeating this entire procedure for all path components in  $\mathcal{G}_{\text{man}}$ , we have thus defined

$$CF^-(H) = \bigoplus_{\mathfrak{s} \in \text{Spin}^c(Y)} CF^-(H, \mathfrak{s}) \quad \text{and} \quad HF^-(H) = \bigoplus_{\mathfrak{s} \in \text{Spin}^c(Y)} HF^-(H, \mathfrak{s})$$

for all isotopy diagrams  $H$  in  $\mathcal{G}_{\text{man}}$ .

**Remark 6.15** We interpret the role of coherent orientations in the definition above loosely as follows. If one fixes any Heegaard diagram for a 3–manifold, there are numerous inequivalent choices of coherent orientation system — in fact there are  $2^{b_1(Y)}$  such choices; see [11, Lemma 4.16]. The above definition just says one should fix whichever choice they prefer, and then take care to use the maps induced by the standard Heegaard moves (or diffeomorphisms isotopic to the identity) to carry this choice around when considering different Heegaard diagrams for the same 3–manifold.

To finish defining the weak Heegaard invariants, we need to associate isomorphisms to all edges in  $\mathcal{G}_{\text{man}}$ . We begin by assigning maps to edges corresponding to strong  $\alpha$ –equivalences and strong  $\beta$ –equivalences.

**Definition 6.16** Given two strongly  $\alpha$ –equivalent isotopy diagrams

$$H_1 = (\Sigma, A, B, z), H_2 = (\Sigma, A', B, z) \in |\mathcal{G}_{\text{man}}|$$

representing  $(Y, z)$ , and  $\mathfrak{s} \in \text{Spin}^c(Y)$ , fix strongly  $\mathfrak{s}$ -admissible diagrams  $(\Sigma, \alpha, \beta, z)$  and  $(\Sigma, \alpha', \beta', z)$  representing them. As above, this is possible by [11, Section 5]. Then by [14, Theorem 2.3 and Lemma 2.13], the chain homotopy equivalences  $\Phi_{\beta}^{\alpha \rightarrow \alpha'}$  fit into a morphism of transitive systems between the transitive systems  $CF^-(H, \mathfrak{s})$  appearing in Definition 6.14. Thus for the edge  $e \in \mathcal{G}_{\text{man}}^{\alpha}(H_1, H_2)$  corresponding to the strong  $\alpha$ -equivalence, we can associate this collection of chain homotopy equivalences (or equivalently, this collection of isomorphisms in  $\text{Kom}(\mathbb{Z}[U]\text{-Mod})$ ) to obtain a morphism

$$\Phi_e := \Phi_B^{A \rightarrow A'} : CF^-(H_1) \rightarrow CF^-(H_2).$$

We note that such a collection of chain homotopy equivalences is precisely the notion of an isomorphism in  $\text{Trans}(\text{Kom}(\mathbb{Z}[U]\text{-Mod}))$ . We define the chain homotopy equivalences associated to a strong  $\beta$ -equivalence analogously.

To finish defining the weak Heegaard invariants, we assign isomorphisms to stabilizations and diffeomorphisms in the next two subsections.

## 6.6 Stabilization maps

We recall maps on the Heegaard Floer chain complexes which can be associated to stabilizations. Given a strongly  $\mathfrak{s}$ -admissible diagram  $\mathcal{H} = (\Sigma, \alpha, \beta, z)$  and a stabilization thereof,  $\mathcal{H}' = (\Sigma \# \Sigma_0, \alpha', \beta', z)$ , each coherent orientation system  $\mathfrak{o}$  on  $\mathcal{H}$  induces a coherent orientation system  $\mathfrak{o}'$  on  $\mathcal{H}'$ . With respect to these orientation systems, there is a  $\mathbb{Z}[U]$ -equivariant chain isomorphism

$$\sigma_{\mathcal{H} \rightarrow \mathcal{H}'} : CF_{J_s}^-(\Sigma, \alpha, \beta, z, \mathfrak{s}, \mathfrak{o}) \rightarrow CF_{J'_s(T)}^-(\Sigma \# \Sigma_0, \alpha', \beta', z, \mathfrak{s}, \mathfrak{o}')$$

defined for sufficiently large values of a parameter  $T$ . This is established in [11, Theorems 10.1 and 10.2].

The curves  $\alpha' \cup \beta'$  are obtained as the disjoint union of  $\alpha \cup \beta$  along with a pair of closed curves  $\alpha'$  and  $\beta'$  contained in  $\Sigma_0$  which intersect transversally in a single point we will denote by  $c$ . We can identify the intersection points in the two diagrams above by assigning to an intersection point  $x \in \mathbb{T}_{\alpha} \cap \mathbb{T}_{\beta}$  the intersection point

$$\sigma_{\mathcal{H} \rightarrow \mathcal{H}'}(x) = x \times c \in \mathbb{T}_{\alpha'} \cap \mathbb{T}_{\beta'}.$$

Fix complex structures  $j_{\Sigma}$  on  $\Sigma$  and  $j_{\Sigma_0}$  on  $\Sigma_0$ , and let  $j'(T)$  denote the complex structure on  $\Sigma \# \Sigma_0$  defined by inserting a neck of length  $T$  between  $(\Sigma, j_{\Sigma})$  and  $(\Sigma_0, j_{\Sigma_0})$ . Then one can associate to a perturbation  $J_s$  of  $\text{Sym}^g(j_{\Sigma})$  on  $\text{Sym}^g(\Sigma)$  and



a perturbation  $J_s^0$  of  $j_{\Sigma_0}$ , a perturbation  $J'_s(T)$  of  $\text{Sym}^{g+1}(j'(T))$  on  $\text{Sym}^{g+1}(\Sigma \# \Sigma_0)$ . The key argument needed to establish the above chain isomorphism then comes in the form of a neck stretching argument which yields the following gluing result: for sufficiently large values of  $T$ , a homotopy class of a Whitney disk  $\phi \in \pi_2(\mathbf{x}, \mathbf{y})$  on  $\Sigma$  with Maslov index 1, and the corresponding homotopy class  $\phi' \in \pi_2(\mathbf{x} \times \mathbf{c}, \mathbf{y} \times \mathbf{c})$  on  $\Sigma \# \Sigma_0$  with Maslov index 1, there is an identification of moduli spaces  $\mathcal{M}_{J_s}(\phi) \cong \mathcal{M}_{J'_s(T)}(\phi')$ . From this it follows readily that the above map is a  $\mathbb{Z}[U]$ -equivariant chain isomorphism.

**Definition 6.17** Given isotopy diagrams  $H$  and  $H'$ , with  $H'$  obtained from  $H$  via a stabilization, we can associate a morphism of transitive systems

$$\sigma_{H \rightarrow H'}: CF^-(H) \rightarrow CF^-(H')$$

as follows. Fixing any  $\text{Spin}^c$ -structure  $\mathfrak{s}$ , strongly  $\mathfrak{s}$ -admissible representatives  $\mathcal{H}$  and  $\mathcal{H}'$  which realize the stabilization, and almost complex structure data on  $\mathcal{H}$ , there is some choice of almost complex structure data on  $\mathcal{H}'$  for which the stabilization isomorphism is defined. As described in [14, Lemma 2.15], the stabilization maps  $\sigma_{\mathcal{H} \rightarrow \mathcal{H}'}$  commute with the change of almost complex structure maps, and with the strong equivalence maps. This implies that the chain isomorphisms  $\{\sigma_{\mathcal{H} \rightarrow \mathcal{H}'}\}$ , when the complex structures are chosen so that they are defined, satisfy the commutativity requirements required of a morphism of transitive systems as in Definition 4.3. We can complete this partially defined morphism of transitive systems for other choices of complex structure data by declaring the stabilization map  $\sigma_{\mathcal{H} \rightarrow \mathcal{H}'}$  to be computed for allowable complex structure data, followed by the appropriate change of almost complex structure homotopy equivalence  $\Phi_{J_s \rightarrow J'_s}$ . We define the morphism of transitive systems associated to the corresponding destabilization to be the inverse of  $\sigma_{H \rightarrow H'}$ .

On the level of homology, we obtain via the colimit construction in Definition 6.14 canonical isomorphisms  $i_{\mathcal{H}}: HF^-(\mathcal{H}) \rightarrow HF^-(H)$  and  $i_{\mathcal{H}'}: HF^-(\mathcal{H}') \rightarrow HF^-(H')$ . We set  $\sigma_{H \rightarrow H'} = i_{\mathcal{H}'} \circ \sigma_{\mathcal{H} \rightarrow \mathcal{H}'} \circ i_{\mathcal{H}}^{-1}$  for any choice of such  $\mathcal{H}$  and  $\mathcal{H}'$ . This is independent of the choice of diagrams  $\mathcal{H}$  and  $\mathcal{H}'$  by the aforementioned result [14, Lemma 2.15].

## 6.7 Diffeomorphism maps

Finally, we need to discuss how diffeomorphisms of Heegaard surfaces lead to maps on the associated chain complexes. We use the following definition.

**Definition 6.18** [5, Definition 9.23] Fix a strongly  $\mathfrak{s}$ -admissible diagram  $(\Sigma, \alpha, \beta, z)$ , with  $|\alpha| = |\beta| = k$ . Let  $j$  be an almost complex structure on  $\Sigma$ , and  $J_s$  be a perturbation of the almost complex structure  $\text{Sym}^k(j)$  on  $\text{Sym}^k(\Sigma)$ . Let  $\mathfrak{o}$  be a coherent orientation system on the diagram. Fix a diffeomorphism  $d: \Sigma \rightarrow \Sigma'$ , and set  $d(\alpha) = \alpha'$  and  $d(\beta) = \beta'$ . We define an associated map as follows. First, the almost complex structure  $j$  and perturbation  $J_s$  can be conjugated via the differential of  $d$  to obtain  $j' = d_*(j)$  on  $\Sigma$  and  $J'_s = d_*(J_s)$  a perturbation of  $d_*(j)$  on  $\text{Sym}^k(\Sigma')$ . The diffeomorphism  $d$  provides an identification between periodic classes  $\pi_2(x, x) \cong \pi_2(x', x')$  for  $x \in \mathbb{T}_\alpha \cap \mathbb{T}_\beta$  and  $x' \in \mathbb{T}_{\alpha'} \cap \mathbb{T}_{\beta'}$ . We use this identification to push forward the coherent orientation system  $\mathfrak{o}$  to obtain an induced orientation system  $\mathfrak{o}'$ . This yields a chain isomorphism

$$d_{J_s, J'_s}: CF_{J_s}^-(\Sigma, \alpha, \beta, z, \mathfrak{s}, \mathfrak{o}) \rightarrow CF_{J'_s}^-(\Sigma', \alpha', \beta', z', d(\mathfrak{s}), \mathfrak{o}')$$

as can be seen easily by a direct argument pushing forward all intersection points, and holomorphic disks connecting two such, via  $d$ . We note that the change of complex structure maps commute with the maps  $d_{J_s, J'_s}$  (by a direct check), so there is also an induced map of transitive systems

$$d_*: CF^-(\Sigma, \alpha, \beta, z, \mathfrak{s}) \rightarrow CF^-(\Sigma', \alpha', \beta', z', d(\mathfrak{s})).$$

Finally, by Lemma 6.12 and [5, Lemma 9.24], the maps  $d_*$  commute with the maps  $\Gamma_{\beta \rightarrow \beta'}^{\alpha \rightarrow \alpha'}$  appearing in Definition 6.14. Thus by using the continuation maps the maps  $d_*$  can be extended to a morphism of the transitive systems in Definition 6.14,

$$d_*: CF^-(H, \mathfrak{s}) \rightarrow CF^-(H', d(\mathfrak{s})),$$

where  $H = (\Sigma, [\alpha], [\beta], z)$  and  $H' = (\Sigma', [\alpha'], [\beta'], z')$ .

On the level of homology, the above definitions give a well defined map of the  $\mathbb{Z}[U]$ -modules in Definition 6.14,

$$d_*: HF^-(H, \mathfrak{s}) \rightarrow HF^-(H', d(\mathfrak{s})).$$

## 6.8 Monodromy of orientation systems

We now establish the claim made in Definition 6.14 that there is no monodromy of induced orientations systems around loops of diagrams. This will finish the proof that the Heegaard Floer invariants are weak Heegaard invariants, and will also establish

in particular that there is no monodromy of induced orientation systems around the special loops relevant to strong Heegaard invariance. We will show:

**Lemma 6.19** *There is no monodromy of coherent orientation systems around loops composed of isomorphisms associated to isotopies, handleslides, stabilizations and diffeomorphisms of diagrams.*

**Corollary 6.20** *There is no monodromy of coherent orientation systems around the loops determined by simple handleswaps and distinguished rectangles.*

**Remark 6.21** Note that we have already described how each type of Heegaard move induces a map on the Heegaard Floer chain complex which is defined with respect to an orientation system induced on the codomain from one on the domain. See Section 6.4, Definition 6.17 and Definition 6.18 for the relevant definitions and references.

To prove Lemma 6.19, it will be useful to think first about the canonical orientation systems, introduced in [10], in particular. We first note the following fact about those orientation systems.

**Lemma 6.22** *The maps associated to isotopies, handleslides, stabilizations and diffeomorphisms take canonical orientations to canonical orientations.*

**Proof** By [10, Section 8], each such map induces an isomorphism on the totally twisted module  $\underline{HF}^\infty$ . By [10, Theorem 10.1], the canonical orientation system is characterized by the resulting isomorphism type of  $\underline{HF}^\infty$ .  $\square$

**Corollary 6.23** *The maps associated to the loops defining simple handleswaps and distinguished rectangles take canonical orientations to canonical orientations.*

Having established the statement of Lemma 6.19 for canonical orientation systems, we now turn to proving it for general orientation systems. Fix a Heegaard diagram  $\mathcal{H}$  for  $Y$ , and let

$$\mathcal{O}_{\mathcal{H}} = \{\text{equivalence classes of coherent orientation systems on } \mathcal{H}\}$$

and  $\sigma_c \in \mathcal{O}_{\mathcal{H}}$  be the canonical orientation system. There is a map

$$\text{diff}_{\sigma_c} : \mathcal{O}_{\mathcal{H}} \rightarrow \text{Hom}(\Pi_{\mathbf{x}}^{\mathcal{H}}, \mathbb{Z}/2\mathbb{Z}),$$

where  $\Pi_{\mathbf{x}}^{\mathcal{H}}$  is the group of periodic domains based at  $\mathbf{x}$  in  $\mathcal{H}$ , defined by measuring the difference between an orientation system and  $\sigma_c$  on each periodic domain. In symbols, for an orientation system  $\sigma$ ,  $\text{diff}_{\sigma_c}(\sigma)$  is a map

$$\text{diff}_{\sigma_c}(\sigma): \Pi_{\mathbf{x}}^{\mathcal{H}} \rightarrow \mathbb{Z}/2\mathbb{Z}$$

satisfying

$$\text{diff}_{\sigma_c}(\sigma)[D] = \begin{cases} 0 & \text{if } \sigma|_D = \sigma_c|_D, \\ 1 & \text{if } \sigma|_D \neq \sigma_c|_D. \end{cases}$$

We note that the analogous map  $\text{diff}_{\sigma}$  can be defined for any coherent orientation system  $\sigma$  on  $\mathcal{H}$ .

**Lemma 6.24**  $\text{diff}_{\sigma_c}: \mathcal{O}_{\mathcal{H}} \rightarrow \text{Hom}(\Pi_{\mathbf{x}}^{\mathcal{H}}, \mathbb{Z}/2\mathbb{Z})$  is a well-defined bijection.

**Proof** We've already seen in the proof of Lemma 6.8 that equivalence classes of orientation systems are determined by their values on a basis for the periodic domains based at  $\mathbf{x}$ . Thus if  $\text{diff}_{\sigma_c}(\sigma) = \text{diff}_{\sigma_c}(\sigma')$ , then  $\sigma$  and  $\sigma'$  agree with  $\sigma_c$  on the same set of domains, and  $\sigma = \sigma'$ , so  $\text{diff}_{\sigma_c}$  is injective.

Given a morphism  $\phi: \Pi_{\mathbf{x}}^{\mathcal{H}} \rightarrow \mathbb{Z}/2\mathbb{Z}$ , define  $\sigma$  to satisfy

$$\sigma|_D = \begin{cases} \sigma_c|_D & \text{if } \phi(D) = 0, \\ -\sigma_c|_D & \text{if } \phi(D) = 1. \end{cases}$$

By the comments in Remark 6.4,  $\sigma$  can then be extended over a complete set of paths to obtain a coherent orientation system satisfying  $\text{diff}_{\sigma_c}(\sigma) = \phi$ .  $\square$

Note that by Remark 6.21, any loop  $L$  composed of Heegaard moves induces a map

$$L: \mathcal{O}_{\mathcal{H}} \rightarrow \mathcal{O}_{\mathcal{H}}.$$

With this notation, proving Lemma 6.19 amounts to showing that  $L(\sigma) = \sigma$  for each diagram  $\mathcal{H}$  and each coherent orientation system  $\sigma \in \mathcal{O}_{\mathcal{H}}$ , while Corollary 6.23 says  $L(\sigma_c) = \sigma_c$ . To prove Lemma 6.19, we will show that the maps on orientation systems induced by Heegaard moves commute with the diff maps in the following sense. Given a Heegaard move from a diagram  $\mathcal{H}_1$  to a diagram  $\mathcal{H}_2$ , let  $f: \mathcal{O}_{\mathcal{H}_1} \rightarrow \mathcal{O}_{\mathcal{H}_2}$  be the induced map of coherent orientation systems. Similarly, let

$$\tilde{f}: \text{Hom}(\Pi_{\mathbf{x}}^{\mathcal{H}_1}, \mathbb{Z}/2\mathbb{Z}) \rightarrow \text{Hom}(\Pi_{\mathbf{x}'}^{\mathcal{H}_2}, \mathbb{Z}/2\mathbb{Z})$$

be the map induced from precomposition with the identifications  $\Pi_{\mathbf{x}'}^{\mathcal{H}_2} \cong H_2(Y) \cong \Pi_{\mathbf{x}}^{\mathcal{H}_1}$  described in [11, Proposition 2.15 and Lemma 2.17]. We then have the result:

**Lemma 6.25** For each of the maps  $f$  on coherent orientation systems induced by Heegaard moves,

$$\tilde{f} \circ \text{diff}_o = \text{diff}_{f(o)} \circ f$$

for all coherent orientation systems  $o$ .

**Proof** For each type of Heegaard move, the definition of the map  $f$  (which specifies how a coherent orientation on the starting diagram determines one on the target diagram) can be described by an identification

$$\phi: \Pi_{\mathbf{x}'}^{\mathcal{H}_2} \rightarrow \Pi_{\mathbf{x}}^{\mathcal{H}_1}.$$

If we let  $\tilde{\phi}$  be the map

$$\tilde{\phi}: \text{Hom}(\Pi_{\mathbf{x}}^{\mathcal{H}_1}, \mathbb{Z}/2\mathbb{Z}) \rightarrow \text{Hom}(\Pi_{\mathbf{x}'}^{\mathcal{H}_2}, \mathbb{Z}/2\mathbb{Z})$$

induced by precomposition with  $\phi$ , then for each Heegaard move one can show that

$$(4) \quad \tilde{\phi} \circ \text{diff}_o = \text{diff}_{f(o)} \circ f.$$

For example, in the case of a handleslide this follows from an inspection of the proof of [11, Lemma 8.7] (or see for comparison Remark 6.4), as we now explain. Let  $(\Sigma, \alpha, \beta, \gamma, z)$  be a triple diagram determining a handleslide, as in Lemma 6.9, and  $\mathcal{H}_{\alpha, \beta}$  and  $\mathcal{H}_{\alpha, \gamma}$  be the initial and final diagrams for the handleslide. Fix a homotopy class of triangle  $\psi_0 \in \pi_2(x, y, z)$  in the triple diagram. The map  $f: \mathcal{O}_{\mathcal{H}_{\alpha, \beta}} \rightarrow \mathcal{O}_{\mathcal{H}_{\alpha, \gamma}}$  on orientation systems induced by the handleslide is defined by applying Lemma 6.3 with the orientation  $o_{\beta, \gamma}$  on the intermediary diagram  $\mathcal{H}_{\beta, \gamma}$  chosen to be that of Lemma 6.5. Recall that in this case the indeterminacy of Lemma 6.3 disappears, as the group  $Q$  from Remark 6.4 is zero, so an orientation  $o_{\alpha, \gamma}$  is uniquely determined by an orientation  $o_{\alpha, \beta}$ . The key property we will need to recall from this particular application of the construction of Lemma 6.3 is that for each periodic class,  $A_{\alpha, \gamma} \in \Pi_{\mathbf{x}}^{\mathcal{H}_{\alpha, \gamma}}$ , there is a unique pair of classes  $A_{\beta, \gamma}$  and  $A_{\alpha, \beta}$  such that

$$(5) \quad \psi_0 + A_{\alpha, \gamma} = \psi_0 + A_{\beta, \gamma} + A_{\alpha, \beta},$$

and furthermore the induced orientation is constructed such that this relation is respected by the orientations over these domains; see [11, Proof of Lemma 8.7]. With this in mind, we proceed to establish (4) as follows. Fix a periodic domain  $P_{\alpha, \gamma} \in \Pi_{\mathbf{x}}^{\mathcal{H}_{\alpha, \gamma}}$ , and two orientation systems  $o_{\alpha, \beta}$  and  $o'_{\alpha, \beta}$  on  $\mathcal{H}_{\alpha, \beta}$ . Let  $o_{\alpha, \gamma} = f(o_{\alpha, \beta})$  and  $o'_{\alpha, \gamma} = f(o'_{\alpha, \beta})$  be the corresponding orientations induced by the handleslide, as described above.

Finally, let  $\phi: \Pi_x^{\mathcal{H}_{\alpha,\beta}} \rightarrow \Pi_x^{\mathcal{H}_{\alpha,\gamma}}$  be the map which sends a domain  $P_{\alpha,\beta}$  to the unique class  $P_{\alpha,\gamma}$  specified by (5). We then compare the two sides of (4), and find

$$(\tilde{\phi} \circ \text{diff}_{\sigma_{\alpha,\beta}}(\sigma'_{\alpha,\beta}))[P_{\alpha,\gamma}] = (\sigma_{\alpha,\beta} - \sigma'_{\alpha,\beta})[P_{\alpha,\beta}]$$

while

$$\begin{aligned} (\text{diff}_{f(\sigma_{\alpha,\beta})} \circ f(\sigma'_{\alpha,\beta}))[P_{\alpha,\gamma}] &= (\sigma_{\alpha,\gamma} - \sigma'_{\alpha,\gamma})[P_{\alpha,\gamma}] \\ &= (\sigma_{\alpha,\beta} - \sigma'_{\alpha,\beta})[P_{\alpha,\beta}] + (\sigma_{\beta,\gamma} - \sigma'_{\beta,\gamma})[P_{\beta,\gamma}] \\ &= (\sigma_{\alpha,\beta} - \sigma'_{\alpha,\beta})[P_{\alpha,\beta}], \end{aligned}$$

where the second equality uses the fact that the induced orientations respect (5) by construction, and the third equality uses the fact that  $\sigma_{\beta,\gamma} = \sigma'_{\beta,\gamma}$ . This completes the proof of (4) for the case of handleslides.

With (4) understood, we further claim that for each Heegaard move the identification

$$\phi: \Pi_{x'}^{\mathcal{H}_2} \rightarrow \Pi_x^{\mathcal{H}_1}$$

used to define  $f$  and  $\tilde{\phi}$  agrees with the identification

$$\Pi_{x'}^{\mathcal{H}_2} \cong H_2(Y) \cong \Pi_x^{\mathcal{H}_1}$$

used to define  $\tilde{f}$ . In particular, this implies that  $\tilde{f} = \tilde{\phi}$ , which together with (4) proves the lemma.

In the case of a handleslide, the aforementioned claim follows from the following observations:

- (1) Each periodic domain is uniquely determined by the part of its boundary that lies on one set of attaching curves.
- (2) In the case at hand, the identification  $\phi$  preserves the part of the boundary of domains that lies on one set of attaching circles, and can be characterized as the unique identification of periodic domains with that property.
- (3) The identifications

$$\Pi_x^{\mathcal{H}_1} \cong H_2(Y) \quad \text{and} \quad \Pi_{x'}^{\mathcal{H}_2} \cong H_2(Y)$$

(described in [11, Lemma 2.17 and subsequent remarks]) are determined by the part of the boundary of each periodic domain that lies on one set of attaching curves.

The first two facts imply that for each periodic domain  $D \in \Pi_x^{H_2}$ ,  $D$  and  $\phi(D)$  share the part of their boundary that lies on one set of attaching curves, and the third fact then ensures that  $D$  and  $\phi(D)$  have the same image in  $H_2(Y)$ , which establishes the claim.

The preceding argument proves that the lemma holds for handleslides. For the other Heegaard moves, (4) once again follows directly from the definition of the maps  $f$  and  $\phi$  (although in these cases the definitions are themselves simpler), and the second claim follows from considerations analogous to those listed above. We leave the details of these cases to the reader.  $\square$

**Corollary 6.26** *For any loop  $L$  of Heegaard moves,  $\tilde{L} \circ \text{diff}_\circ = \text{diff}_{L(\circ)} \circ L$  for all coherent orientation systems  $\circ$ .*

**Proof of Lemma 6.19** Corollary 6.26 applied to  $\circ_c$  yields

$$\tilde{L} \circ \text{diff}_\circ(\circ_c) = \text{diff}_{L(\circ)} \circ L(\circ_c) = \text{diff}_{L(\circ)}(\circ_c)$$

where the last equality comes from Corollary 6.23. Since we are considering here a loop of diagrams, the map  $\tilde{L}$  is the identity, and the previous equation yields

$$\text{diff}_\circ(\circ_c) = \text{diff}_{L(\circ)}(\circ_c)$$

or, equivalently,

$$\circ = L(\circ),$$

as desired  $\square$

## 7 Heegaard Floer homology as a strong Heegaard invariant

In the previous section we recalled the definition of the weak Heegaard invariants

$$CF^- : \mathcal{G}_{\text{man}} \rightarrow \text{Trans}(P(\text{Kom}(\mathbb{Z}[U]\text{-Mod})))$$

and

$$HF^- : \mathcal{G}_{\text{man}} \rightarrow P(\mathbb{Z}[U]\text{-Mod})$$

underlying the strong Heegaard invariants appearing in Theorem 1.3 and Corollary 1.4, respectively. To establish Theorem 1.3 we need to check the four axioms required of a strong Heegaard invariant in Definition 3.7.

The proofs of axioms (1) and (2) given in [5, Section 9.2, page 131] for  $\mathbb{F}_2[U]\text{-Mod}$  apply almost directly to establish axioms (1) and (2) for  $CF^-$  and  $HF^-$  as Heegaard invariants valued in  $\text{Trans}(P(\text{Kom}(\mathbb{Z}[U]\text{-Mod})))$  and  $P(\mathbb{Z}[U]\text{-Mod})$ , respectively, as we now summarize for  $CF^-$ .

For axiom (1), the functoriality of  $CF^-$  restricted to  $\mathcal{G}_{\text{man}}^\alpha$  and  $\mathcal{G}_{\text{man}}^\beta$  follows from Lemma 6.9 and [14, Theorem 2.3]. The functoriality of  $CF^-$  restricted to  $\mathcal{G}_{\text{man}}^{\text{diff}}$  is immediate from Definition 6.18. Finally, for a stabilization  $e$  and the corresponding destabilization  $e'$ ,  $CF^-(e') = CF^-(e)^{-1}$  by Definition 6.17.

For axiom (2), we need to establish that the images under  $CF^-$  of distinguished rectangles in  $\mathcal{G}_{\text{man}}$  (recall Definition 3.5) form commuting rectangles. For a rectangle of type (1), commutativity follows from Lemma 6.9 and [14, Theorem 2.3]. For a rectangle of type (2), commutativity follows from [14, Lemma 2.15]. For a rectangle of type (3), commutativity follows from [5, Lemma 9.24]. Finally, rectangles of type (4) and (5) can be seen to commute by directly applying the arguments in [5, page 131].

We now investigate axiom (3). Let  $H = (\Sigma, A, B, z) \in |\mathcal{G}_{\text{man}}|$  be an isotopy diagram,  $d: H \rightarrow H$  a diffeomorphism of isotopy diagrams which is isotopic to  $\text{Id}_\Sigma$ , and  $d_* := CF^-(e)$  where  $e \in \mathcal{G}_{\text{man}}^{\text{diff}}(H, H)$  is the edge corresponding to  $d$ . We need to show  $d_* = \text{Id}_{CF^-(H)}$  as morphisms of transitive systems in  $P(\text{Kom}(\mathbb{Z}[U]\text{-Mod}))$ . We adapt and restate the argument given in [5, Proposition 9.27] in order to explain why it can be applied to the case of (projective) integral coefficients. We show the following result.

**Theorem 7.1** *Let  $(\Sigma, \alpha, \beta, z)$  be a strongly  $\mathfrak{s}$ -admissible diagram with  $|\alpha| = |\beta| = g$ . Suppose that  $d: \Sigma \rightarrow \Sigma$  is a diffeomorphism isotopic to  $\text{Id}_\Sigma$ , and let  $\alpha' = d(\alpha)$  and  $\beta' = d(\beta)$ . Let  $\mathfrak{o}_{\alpha, \beta}$  be a coherent orientation system on  $(\Sigma, \alpha, \beta, z)$  and  $\mathfrak{o}_{\alpha', \beta'}$  be the coherent orientation system on  $(\Sigma, \alpha', \beta', z)$  induced by  $d$ . Then, with respect to these orientation systems,*

$$d_* = \pm \Gamma_{\beta \rightarrow \beta'}^{\alpha \rightarrow \alpha'}: HF^-(\Sigma, \alpha, \beta, z, \mathfrak{s}, \mathfrak{o}_{\alpha, \beta}) \rightarrow HF^-(\Sigma, \alpha', \beta', z', \mathfrak{s}, \mathfrak{o}_{\alpha', \beta'}).$$

Furthermore, as maps

$$d_*, \pm \Gamma_{\beta \rightarrow \beta'}^{\alpha \rightarrow \alpha'}: CF^-(\Sigma, \alpha, \beta, z, \mathfrak{s}, \mathfrak{o}_{\alpha, \beta}) \rightarrow CF^-(\Sigma, \alpha', \beta', z', \mathfrak{s}, \mathfrak{o}_{\alpha', \beta'}),$$

$d_*$  is chain homotopic to one of  $\pm \Gamma_{\beta \rightarrow \beta'}^{\alpha \rightarrow \alpha'}$ .

In fact, this theorem will establish axiom (3) in Definition 3.7 for the weak Heegaard invariants  $CF^-$  and  $HF^-$  above. Since  $d$  is isotopic to  $\text{Id}_\Sigma$  by hypothesis, we have  $\alpha'$  is isotopic to  $\alpha$  and  $\beta'$  is isotopic to  $\beta$ , so  $H := (\Sigma, [\alpha], [\beta], z) = (\Sigma, [\alpha'], [\beta'], z')$ . The induced map of transitive systems  $d_*: CF^-(H) \rightarrow CF^-(H)$  defined in Definition 6.18 is then computed by extending the following map by conjugation with the continuation maps:

$$CF^-(\Sigma, \alpha, \beta, z, \mathfrak{o}_{\alpha, \beta}) \xrightarrow{d_*} CF^-(\Sigma, \alpha', \beta', z, \mathfrak{o}_{\alpha', \beta'}) \xrightarrow{\Gamma_{\alpha' \rightarrow \alpha}^{\beta' \rightarrow \beta}} CF^-(\Sigma, \alpha, \beta, z, \mathfrak{o}_{\alpha, \beta}).$$



Since  $\Gamma_{\alpha' \rightarrow \alpha}^{\beta' \rightarrow \beta} \sim (\Gamma_{\alpha \rightarrow \alpha'}^{\beta \rightarrow \beta'})^{-1}$  and  $d_* \sim \pm \Gamma_{\alpha \rightarrow \alpha'}^{\beta \rightarrow \beta'}$  by Theorem 7.1, we see that

$$d_*: CF^-(H) \rightarrow CF^-(H)$$

is the extension of a map  $CF^-(\Sigma, \alpha, \beta, z, \circ_{\alpha, \beta}) \rightarrow CF^-(\Sigma, \alpha, \beta, z, \circ_{\alpha, \beta})$  which is homotopic to plus or minus the identity. Thus we see that  $d_* = \text{Id}_{CF^-(H)}$  as morphisms in  $\text{Trans}(P(\text{Kom}(\mathbb{Z}[U]\text{-Mod}))$ .

**Proof of Theorem 7.1** Since  $d$  is isotopic to  $\text{id}_\Sigma$ , we may decompose it into a composition of diffeomorphisms  $d_i$  on some diagrams  $\mathcal{H}_i = (\Sigma, \alpha_i, \beta_i)$ , such that each  $d_i$  is Hamiltonian isotopic to  $\text{id}_\Sigma$  for some symplectic form  $\omega_i$  on  $\Sigma$ , and the diagrams satisfy the intersection properties  $|\alpha \cap d_i(\alpha)| = |\beta \cap d_i(\beta)| = 2$  for all  $\alpha \in \alpha_{i-1}$  and  $\beta \in \beta_{i-1}$ . As described in [5, Proposition 9.27], it will suffice to prove the result for such a  $d_i$ . So let  $d_t$  for  $t \in \mathbb{R}$  be a Hamiltonian isotopy which is independent of  $t$  for  $t \in (-\infty, 0]$  and  $t \in [1, \infty)$ , and which connects  $\text{id}_\Sigma$  to a diffeomorphism  $d$  of  $\mathcal{H} = (\Sigma, \alpha, \beta)$ . Throughout the proof, we will use the notation  $d_t(\alpha) = \alpha_t, d_t(\beta) = \beta_t$ , and use primes to indicate the values of various quantities at  $t = 1$ .

Fix the data of a complex structure  $j$  on  $\Sigma$  and a perturbation  $J_s$  of  $\text{Sym}^g(j)$  on  $\text{Sym}^g(\Sigma)$ , and for  $t \in \mathbb{R}$  let  $j_t = (d_t)_*(j)$  and  $J_{s,t} = (\text{Sym}^g(d_t))_*(J_s)$ . As described in the sections above, there are numerous chain maps on the Heegaard Floer chain complexes we can associate with the isotopy  $d_t$  and this induced almost complex structure data. We will be concerned here with the following three.

- (1) We can change the almost complex structure on  $\text{Sym}^g(\Sigma)$  from  $J_s = J_{s,0}$  to  $J'_s = J_{s,1}$ , while leaving the attaching curves unchanged, and consider the induced map

$$\Phi_{J_s \rightarrow J'_s}: CF_{J_s}^-(\Sigma, \alpha, \beta, z, \circ_{\alpha, \beta}) \rightarrow CF_{J'_s}^-(\Sigma, \alpha, \beta, z, \circ_{\alpha, \beta}).$$

We recall here that this map is defined (in [11]) by counting Maslov index 0 disks  $u: [0, 1] \times \mathbb{R} \rightarrow \text{Sym}^g(\Sigma)$  connecting some  $x \in \mathbb{T}_\alpha \cap \mathbb{T}_\beta$  to some  $y \in \mathbb{T}_\alpha \cap \mathbb{T}_\beta$ , which satisfy  $u(0, t) \in \alpha, u(1, t) \in \beta$  and  $du/ds + J_{s,t}(du/dt) = 0$ .

- (2) We can leave the almost complex structures  $(j, J_s)$  fixed, and consider the effect on the Floer complex of altering only the attaching curves via the map

$$\Gamma_{\alpha \rightarrow \alpha'}^{\beta \rightarrow \beta'}: CF_{J_s}^-(\Sigma, \alpha, \beta, z, \circ_{\alpha, \beta}) \rightarrow CF_{J_s}^-(\Sigma, \alpha', \beta', z, \circ_{\alpha', \beta'})$$

associated to the Hamiltonian isotopy  $d_t$ . In this case, the map is defined by counting Maslov index 0 disks  $u$  connecting some  $x \in \mathbb{T}_\alpha \cap \mathbb{T}_\beta$  to some  $y \in \mathbb{T}_{\alpha'} \cap \mathbb{T}_{\beta'}$  as above, but with dynamic boundary conditions  $u(0, t) \in \alpha_t, u(1, t) \in \beta_t$ , and which satisfy  $du/ds + J_s(du/dt) = 0$ .

(3) We define a new sort of continuation map associated with  $d_t$ ,

$$\Gamma_{d_t} : CF_{J_s}^-(\Sigma, \alpha, \beta, z, \sigma_{\alpha, \beta}) \rightarrow CF_{J'_s}^-(\Sigma, \alpha', \beta', z, \sigma_{\alpha', \beta'})$$

which combines the ideas from the previous two. This map is defined to count Maslov index 0 disks  $u$  which connect some  $\mathbf{x} \in \mathbb{T}_\alpha \cap \mathbb{T}_\beta$  to some  $\mathbf{x}' \in \mathbb{T}_{\alpha'} \cap \mathbb{T}_{\beta'}$ , have dynamic boundary conditions  $u(0, t) \in \alpha_t$ ,  $u(1, t) \in \beta_t$ , and satisfy  $du/ds + J_{s,t}(du/dt) = 0$ . We will denote the set of homotopy classes of Whitney disks (not necessarily  $J_{s,t}$ -holomorphic) satisfying these boundary conditions by  $\pi_2^{d_t}(\mathbf{x}, \mathbf{x}')$ , and for  $\phi \in \pi_2^{d_t}(\mathbf{x}, \mathbf{x}')$  we will denote the moduli space of  $J_{s,t}$ -holomorphic maps representing  $\phi$  by  $\mathcal{M}^{d_t}(\phi)$ .

We claim that the third map in the list above is in fact chain homotopic to the map  $d_{J_s, J'_s}$  from Definition 6.18. To see this, we first explain that if a diffeomorphism (which we also indicate by  $d$ , as an abuse of notation)  $d: \Sigma \rightarrow \Sigma$  isotopic to the identity (via an isotopy  $d_t$ ) is sufficiently close to  $\text{Id}_\Sigma$ , then the map defined in case (3) above satisfies  $\Gamma_{d_t} = d_{J_s, J'_s}$  as chain maps. Indeed, by taking  $d$  to be a sufficiently small perturbation of  $\text{Id}_\Sigma$ , we may ensure the isotopy  $d_t$  is arbitrarily close to being constant in  $t$ . For an isotopy which is constant in  $t$ , the definition of the continuation map in (3) above counts Maslov index 0 disks with fixed boundary conditions which are  $J_s$ -holomorphic. The only such maps are constant maps. Thus, by Gromov compactness, if the isotopy  $d_t$  is sufficiently close to being constant, the Maslov index 0 solutions to the equation appearing in the definition of  $\Gamma_{d_t}$  will be close enough to constant disks to ensure that  $\Gamma_{d_t}$  will be a nearest-point map.

Next we note that the definition of  $\Gamma_{d_t}$  depends on a choice of coherent orientation system for the moduli spaces  $\mathcal{M}^{d_t}(\phi)$ . As explained in [11, Proof of Proposition 7.3], when  $\pi_2^{d_t}(\mathbf{x}, \mathbf{x}') \neq 0$  a single homotopy class  $\phi \in \pi_2^{d_t}(\mathbf{x}, \mathbf{x}') \cong \mathbb{Z}$  yields via gluing an identification between periodic classes  $\pi_2(\mathbf{x}, \mathbf{x}) \cong_\phi \pi_2(\mathbf{x}', \mathbf{x}')$  on the two diagrams, and a choice of orientation for  $\mathcal{M}^{d_t}(\phi)$  then yields an identification between coherent orientation systems on the two diagrams. Thus, given a coherent orientation system  $\sigma_{\alpha, \beta}$  on  $(\Sigma, \alpha, \beta)$ , and an orientation on  $\mathcal{M}^{d_t}(\phi)$ , we obtain an induced orientation  $\sigma_{\alpha', \beta'}$  on  $(\Sigma, \alpha', \beta')$  with respect to which the map is defined. We claim that we may arrange for this induced orientation to agree with that induced by  $d_{J_s, J'_s}$ . Indeed, fix for each  $\mathbf{x} \in \mathbb{T}_\alpha \cap \mathbb{T}_\beta$  a homotopy class  $\phi_{\mathbf{x}} \in \pi_2^{d_t}(\mathbf{x}, \mathbf{x}')$ . We can choose orientations on all such  $\mathcal{M}^{d_t}(\phi_{\mathbf{x}})$  freely such that  $\Gamma_{d_t}$  is the positive nearest-point map (with the generator corresponding to an intersection point being taken to the positive generator corresponding to the nearest intersection point after the isotopy is performed), and then extend these choices to a coherent system. The coherent orientation  $\sigma_{\alpha', \beta'}$  on  $(\Sigma, \alpha', \beta', z')$

induced by  $\Gamma_{d_t}$  that results will then be the same as that induced by  $d_{J_s, J'_s}$ , as we now explain. Fix  $\mathbf{x}, \mathbf{y} \in \mathbb{T}_\alpha \cap \mathbb{T}_\beta$  and let  $\mathbf{x}' = d(\mathbf{x})$  and  $\mathbf{y}' = d(\mathbf{y})$  be the corresponding intersection points in  $\mathbb{T}_{\alpha'} \cap \mathbb{T}_{\beta'}$ . Given a homotopy class  $\psi \in \pi_2(\mathbf{x}, \mathbf{y})$  and a positively oriented Whitney disk  $u$  from  $\mathbf{x}$  to  $\mathbf{y}$  in the class  $\psi$ , the orientation system induced by  $d_{J_s, J'_s}$  will positively orient the corresponding disk  $d(u)$  representing the class  $d(\psi) \in \pi_2(\mathbf{x}', \mathbf{y}')$ ; see Definition 6.18. We need to show that the disk  $d(u)$  is also positively oriented in the orientation system induced by  $\Gamma_{d_t}$ . As described above, the orientation on  $d(u)$  induced by  $\Gamma_{d_t}$  is specified as follows. We consider representative disks  $v_1$  and  $v_2$  for the classes  $\phi_x \in \pi_2^{d_t}(\mathbf{x}, \mathbf{x}')$  and  $\phi_y \in \pi_2^{d_t}(\mathbf{y}, \mathbf{y}')$ , which we may assume are both positively oriented by the choice we made for orientations on  $\mathcal{M}^{d_t}(\phi_x)$  and  $\mathcal{M}^{d_t}(\phi_y)$ . We then consider the glued disk  $v_2 \natural u \natural \bar{v}_1$ . Since an orientation has been specified on each constituent disk and our system is coherent, this glued disk also has a specified orientation, which is positive given our choices. Finally, we note that this disk is identified with  $d(u)$  under the identification between coherent orientation systems in the two diagrams, and thus  $d(u)$  must also be oriented positively. We thus see that both maps induce the same coherent orientation system on the target and both take the form of the positive nearest-point map, so  $\Gamma_{\phi_t} = \phi_{J_s, J'_s}$ .

Finally, we can decompose our original diffeomorphism  $d : (\Sigma, \alpha_0, \beta_0) \rightarrow (\Sigma, \alpha_1, \beta_1)$  into a sequence of diffeomorphisms  $d^1, d^2, \dots, d^N$ , where

$$d^i : (\Sigma, \alpha_{(i-1)/N}, \beta_{(i-1)/N}) \rightarrow (\Sigma, \alpha_{i/N}, \beta_{i/N})$$

and each  $d^i$  is isotopic to  $\text{Id}_\Sigma$  via isotopies  $d_t^i$ . For sufficiently large  $N$ , we can ensure that the continuation map  $\Gamma_{d_t^i}$  associated to each constituent isotopy satisfies

$$\Gamma_{d_t^i} = (d^i)_{J_s, (i-1)/N, J_s, i/N}$$

by the argument in the preceding paragraphs. Furthermore, by inserting long necks one can see that the composition of the corresponding continuation maps is homotopic to the original continuation map

$$\Gamma_{d_t} \sim (\Gamma_{d_t^N} \circ \dots \circ \Gamma_{d_t^1}).$$

Since

$$d_{J_s, J'_s} = d_{J_s, (N-1)/N, J_s, 1}^N \circ \dots \circ d_{J_s, 0, J_s, 1/N}^1,$$

we thus see that  $d_{J_s, J'_s} \sim \Gamma_{d_t}$ , which establishes the claim.

Using Definition 6.18 we have  $d_* = \Phi_{J'_s \rightarrow J_s} \circ d_{J_s, J'_s}$ . Thus to complete the proof it will in fact suffice to show that  $\Phi_{J'_s \rightarrow J_s} \circ d_{J_s, J'_s} \sim \pm \Gamma_{\beta \rightarrow \beta'}$ , or, since  $d_{J_s, J'_s} \sim \Gamma_{d_t}$

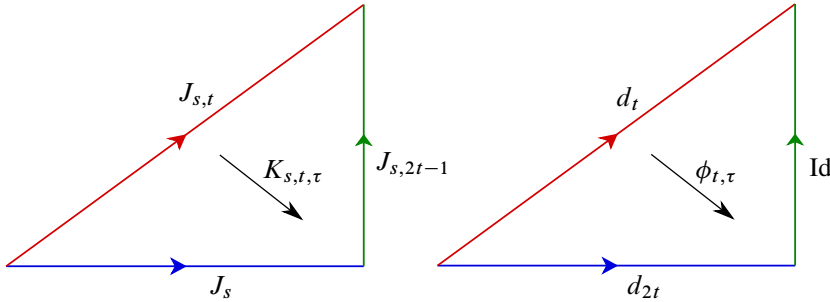


Figure 5: A schematic of the complex structure and isotopy data defining the continuation maps  $\Gamma_{d_t}$  and (a continuation map homotopic to)  $\Phi_{J_s \rightarrow J'_s} \circ \Gamma_{\beta \rightarrow \beta'}^{\alpha \rightarrow \alpha'}$ , and the homotopies between the two sets of data. The data defining  $\Gamma_{d_t}$  is represented by the top edges of the two triangles, while the data defining  $\Phi_{J_s \rightarrow J'_s} \circ \Gamma_{\beta \rightarrow \beta'}^{\alpha \rightarrow \alpha'}$  is represented by the bottom edges followed by the vertical edges.

and  $\Phi_{J'_s \rightarrow J_s}^{-1} \sim \Phi_{J_s \rightarrow J'_s}$ , to show that

$$(6) \quad \Gamma_{d_t} \sim \pm \Phi_{J_s \rightarrow J'_s} \circ \Gamma_{\beta \rightarrow \beta'}^{\alpha \rightarrow \alpha'}$$

To see that (6) is true, we consider the following generalized notion of a continuation map, of which each of the three maps involved are a special case. Consider a Hamiltonian isotopy  $\phi_t$  and a generic two parameter family of almost complex structures  $K_{s,t}$  on  $\text{Sym}^g(\Sigma)$  which are perturbations of  $\text{Sym}^g(k_t)$  where  $k_t$  is a one parameter family of complex structures on  $\Sigma$ . Here we assume for convenience as above that this data is independent of  $t$  for  $t \in (-\infty, 0]$  and  $t \in [1, \infty)$ . We set  $\alpha_t = \phi_t(\alpha)$  and  $\beta_t = \phi_t(\beta)$ . Given such data we can associate the *continuation map with respect to*  $(\phi_t, K_{s,t})$ ,

$$(7) \quad \Gamma_{(\phi_t, K_{s,t})} : CF_{K_{s,0}}^-(\Sigma, \alpha_0, \beta_0) \rightarrow CF_{K_{s,1}}^-(\Sigma, \alpha_1, \beta_1),$$

by counting Maslov index 0 disks  $u$  connecting some  $x \in \mathbb{T}_{\alpha_0} \cap \mathbb{T}_{\beta_0}$  to some  $y \in \mathbb{T}_{\alpha_1} \cap \mathbb{T}_{\beta_1}$ , with dynamic boundary conditions  $u(0, t) \in \alpha_t$ ,  $u(1, t) \in \beta_t$ , and which satisfy

$$\frac{du}{ds} + K_{s,t} \left( \frac{du}{dt} \right) = 0.$$

The maps  $\Gamma_{d_t}$ ,  $\Phi_{J_s \rightarrow J'_s}$  and  $\Gamma_{\beta \rightarrow \beta'}^{\alpha \rightarrow \alpha'}$  above are then the continuation maps with respect to the data  $(d_t, J_{s,t})$ ,  $(\text{id}_{\Sigma}, J_{s,t})$  and  $(d_t, J_{s,0})$  respectively. Furthermore, since the homotopy classes of such continuation maps are natural under concatenation and rescaling of the  $\phi_t$  and  $K_{s,t}$  by [14, Lemma 2.12] (see also the argument below), the

composite  $\Phi_{J_s \rightarrow J'_s} \circ \Gamma_{\beta \rightarrow \beta'}^{\alpha \rightarrow \alpha'}$  is homotopic to the continuation map defined with respect to the data

$$(d_{t,1}, J_{s,t,1}) := \begin{cases} (d_{2t}, J_{s,0}) & \text{if } t \in [0, \frac{1}{2}], \\ (\text{id}_\Sigma, J_{s,2t-1}) & \text{if } t \in [\frac{1}{2}, 1]. \end{cases}$$

Consider now two Hamiltonian isotopies  $\phi_{t,0}$  and  $\phi_{t,1}$  with  $\phi_{0,0} = \phi_{0,1} = \text{id}_\Sigma$  and  $\phi_{1,0} = \phi_{1,1}$ , and two generic two parameter families  $K_{s,t,0}$  and  $K_{s,t,1}$  with  $K_{s,0,0} = K_{s,0,1}$  and  $K_{s,1,0} = K_{s,1,1}$ . We will complete the proof by showing that a generic homotopy  $h = (\phi_{t,\tau}, K_{s,t,\tau})$  between  $(\phi_{t,0}, K_{s,t,0})$  and  $(\phi_{t,1}, K_{s,t,1})$  induces a chain homotopy between  $\Gamma_{(\phi_{t,0}, K_{s,t,0})}$  and  $\pm \Gamma_{(\phi_{t,1}, K_{s,t,1})}$ . In particular, (6) will follow, as the data  $(d_t, J_{s,t})$  used to define  $\Gamma_{d_t, J_{s,t}} =: \Gamma_{d_t}$  is homotopic to the data  $(d_{t,1}, J_{s,t,1})$  used to define  $\Gamma_{d_{t,1}, J_{s,t,1}} \sim \Phi_{J_s \rightarrow J'_s} \circ \Gamma_{\beta \rightarrow \beta'}^{\alpha \rightarrow \alpha'}$ .

Fixing  $\tau$ , let  $\pi_2^\tau(x, y)$  denote the homotopy classes of disks  $u$  which connect  $x$  to  $y$ , and which satisfy the boundary conditions  $u(0, t) \in \phi_{t,\tau}(\alpha)$  and  $u(1, t) \in \phi_{t,\tau}(\beta)$ . Given a homotopy class  $\phi \in \pi_2^\tau(x, y)$ , we denote by  $\mathcal{M}_\tau(\phi)$  the moduli space of disks in the class  $\phi$  satisfying

$$\frac{du}{ds} + K_{s,t,\tau} \left( \frac{du}{dt} \right) = 0.$$

We note that for fixed  $\tau$ , the definition of the continuation map with respect to  $(\phi_{t,\tau}, K_{s,t,\tau})$  given above can be restated succinctly as counting Maslov index 0 disks in the moduli spaces  $\mathcal{M}_\tau(\phi)$ . For any  $\tau$ , the homotopy  $h$  induces an identification between homotopy classes of disks  $\pi_2^0(x, y) \cong \pi_2^\tau(x, y)$ . Using this identification, we may define for each  $\phi \in \pi_2^0(x, y)$  the moduli space

$$(8) \quad \mathcal{M}^h(\phi) = \bigcup_{\tau \in I} \mathcal{M}_\tau(\phi) \times \{\tau\}.$$

For a generic choice of homotopy  $h$ , this is a manifold of dimension  $\mu(\phi) + 1$ . We use this moduli space to define a chain homotopy

$$H^h : CF_{K_{s,0}}^-(\Sigma, \alpha_0, \beta_0) \rightarrow CF_{K_{s,1}}^-(\Sigma, \alpha_1, \beta_1)$$

between  $\Gamma_{(\phi_{t,0}, K_{s,t,0})}$  and  $\Gamma_{(\phi_{t,1}, K_{s,t,1})}$  associated with the homotopy  $h$ . For  $x \in \mathbb{T}_\alpha \cap \mathbb{T}_\beta$  we set

$$H^h([x, i]) = \sum_{y \in \mathbb{T}_{\alpha_1} \cap \mathbb{T}_{\beta_1}} \sum_{\substack{\phi \in \pi_2^0(x, y) \\ \mu(\phi) = -1}} \#(\mathcal{M}^h(\phi))[y, i - n_p(\phi)].$$

To see that this is a chain homotopy, we will consider the ends of the moduli spaces  $\mathcal{M}^h(\psi)$  for  $\psi$  with Maslov index  $\mu(\psi) = 0$ . Since such spaces  $\mathcal{M}^h(\psi)$  are smooth 1-dimensional manifolds for generic choices of almost complex structure data, and since they are orientable, the signed count of the ends is zero for any choice of orientation.

The ends can be partitioned into three types: those corresponding to  $\tau = 0$ , those corresponding to  $\tau = 1$ , and those corresponding to strips breaking off for values  $0 < \tau < 1$ . For the ends corresponding to  $\tau = 0$ , the contribution to the count of the ends is given by the count of the zero-dimensional moduli space  $\#\mathcal{M}_{\tau=0}(\psi)$ . Modulo signs, this is precisely the count occurring in the definition of  $\Gamma_{(\phi_{t,0}, K_{s,t,0})}$ . For  $\tau = 1$ , the contribution to the count of the ends is similarly given by  $\#\mathcal{M}_{\tau=1}(\psi)$ , which is the count occurring in the definition of  $\Gamma_{(\phi_{t,1}, K_{s,t,1})}$ , modulo signs. We will discuss the signed contributions below. Finally, the ends corresponding to strip breaking come from the space

$$\left( \coprod_{\substack{\phi * \phi' = \psi \\ \mu(\phi)=0, \mu(\phi')=1}} \mathcal{M}^h(\phi) \times \widehat{\mathcal{M}}(\phi') \right) \amalg \left( \coprod_{\substack{\phi' * \phi = \psi \\ \mu(\phi)=0, \mu(\phi')=1}} \widehat{\mathcal{M}}(\phi') \times \mathcal{M}^h(\phi) \right).$$

Supposing the orientations on the moduli spaces  $\mathcal{M}^h$  are chosen to be coherent with respect to pregluings of strips, the count of the terms in the first parentheses is precisely the count occurring in the composition  $\partial_0^- \circ H^h$ , while the count of the terms in the second parentheses is precisely the count occurring in  $H^h \circ (\partial_1)^-$ . Here  $\partial_0^-$  indicates the differential on  $CF_{K_{s,0}}^-(\Sigma, \alpha_0, \beta_0)$  and  $(\partial_1)^-$  indicates the differential on  $CF_{K_{s,1}}^-(\Sigma, \alpha_1, \beta_1)$ .

Finally, we note that we may arrange for the spaces  $\mathcal{M}^h(\phi)$  to be coherently oriented such that the total signed count of the ends of  $\mathcal{M}^h(\psi)$  is given by

$$0 = \Gamma_{(\phi_{t,0}, K_{s,t,0})} - \Gamma_{(\phi_{t,1}, K_{s,t,1})} - ((\partial_1)^- \circ H^h + H^h \circ \partial_0^-).$$

Indeed, we have

$$(9) \quad \mathcal{M}^h(\psi) = \bigcup_{\tau \in I} \mathcal{M}_\tau(\psi) \times \{\tau\} = \{(u, \tau) \in C^\infty(I \times \mathbb{R}, \text{Sym}^g(\Sigma)) \times I \mid u \in \mathcal{M}_\tau(\psi)\},$$

so for each homotopy class  $\psi$  we may choose orientations on  $\mathcal{M}_{\tau=0}(\psi)$  fitting together coherently, and obtain induced orientations on the spaces  $\mathcal{M}^h(\psi)$  via the product structure in (9). Such an induced orientation will enjoy the property that the restrictions to the ends at  $\tau = 0$  and  $\tau = 1$  yield the counts  $-\#\mathcal{M}_{\tau=0}(\psi)$  and  $+\#\mathcal{M}_{\tau=1}(\psi)$  respectively. We omit the technical details of this argument, and refer the interested reader to the proof of Lemma 8.13, where an analogous argument dealing with holomorphic triangles is spelled out in detail. We have thus shown that a generic homotopy  $h = (\phi_{t,\tau}, K_{s,t,\tau})$  between  $(\phi_{t,0}, K_{s,t,0})$  and  $(\phi_{t,1}, K_{s,t,1})$  induces a chain homotopy between  $\Gamma_{(\phi_{t,0}, K_{s,t,0})}$  and  $\pm \Gamma_{(\phi_{t,1}, K_{s,t,1})}$ .

Finally, we note that since the homotopy  $h$  is constant in  $\tau$  for  $t = 0$  and  $t = 1$ , the chain homotopy  $H^h$ , defined with respect to the orientations on  $\mathcal{M}^h(\phi)$  specified above, is a chain homotopy between the continuation maps  $\Gamma_{(\phi_{t,0}, K_{s,t,0})}$  and  $\Gamma_{(\phi_{t,1}, K_{s,t,1})}$ , which both take the form

$$CF_{K_{s,0,0}=K_{s,0,1}}^-(\Sigma, \alpha_0, \beta_0, z, \mathfrak{o}_{\alpha_0, \beta_0}) \rightarrow CF_{K_{s,1,0}=K_{s,1,1}}^-(\Sigma, \alpha_1, \beta_1, z, \mathfrak{o}_{\alpha_1, \beta_1})$$

and are defined with respect to the same coherent orientation systems on their domains, and the same coherent orientation systems on their targets. In particular, in the case of interest—ie (6)—we may choose orientations on  $\mathcal{M}_{\tau=0} = \mathcal{M}^{d_t}$  so that  $d_{J_s, J'_s} \sim \Gamma_{d_t}$  (which we established is possible earlier), which together with the above remarks establishes (6). This completes the proof of the theorem.  $\square$

Finally, we relegate the proof of axiom (4), simple handleswap invariance, to Section 8. Given a simple handleswap in  $\mathcal{G}_{\text{man}}$ ,

$$\begin{array}{ccc} H_1 & & \\ g \uparrow & \searrow e & \\ H_3 & \xleftarrow{f} & H_2 \end{array}$$

we will show that the composition of the induced maps in the category of transitive systems in the projectivized homotopy category yields the identity. We recall from Definition 3.6 that here  $H_i = (\Sigma \# \Sigma_0, \alpha_i, \beta_i)$  are isotopy diagrams,  $e$  is a strong  $\alpha$ -equivalence,  $f$  is a strong  $\beta$ -equivalence, and  $g$  is a diffeomorphism of isotopy diagrams.

**Theorem 7.2** (cf [5, Theorem 9.30]) *Let  $(\{H_i\}, e, f, g)$  be data defining a simple handleswap as above. For the weak Heegaard invariants  $CF^\circ$  defined in Definition 6.14, the induced maps  $g_* := CF^\circ(g)$ ,  $\Phi_e := CF^\circ(e)$ , and  $\Phi_f := CF^\circ(f)$  satisfy*

$$g_* \circ \Phi_f \circ \Phi_e = \text{Id}_{CF^-(H_1)}.$$

*Thus the weak Heegaard invariants  $CF^\circ : \mathcal{G}_{\text{man}} \rightarrow \text{Trans}(P(\text{Kom}(\mathbb{Z}[U]\text{-Mod}))$  satisfy simple handleswap invariance.*

**Corollary 7.3** *The weak Heegaard invariants  $HF^- : \mathcal{G}_{\text{man}} \rightarrow P(\mathbb{Z}[U]\text{-Mod})$  satisfy simple handleswap invariance.*

Theorem 7.2 and Corollary 7.3 will establish Theorem 1.3 and Corollary 1.4, which by Section 5 also establishes Theorem 1.1.

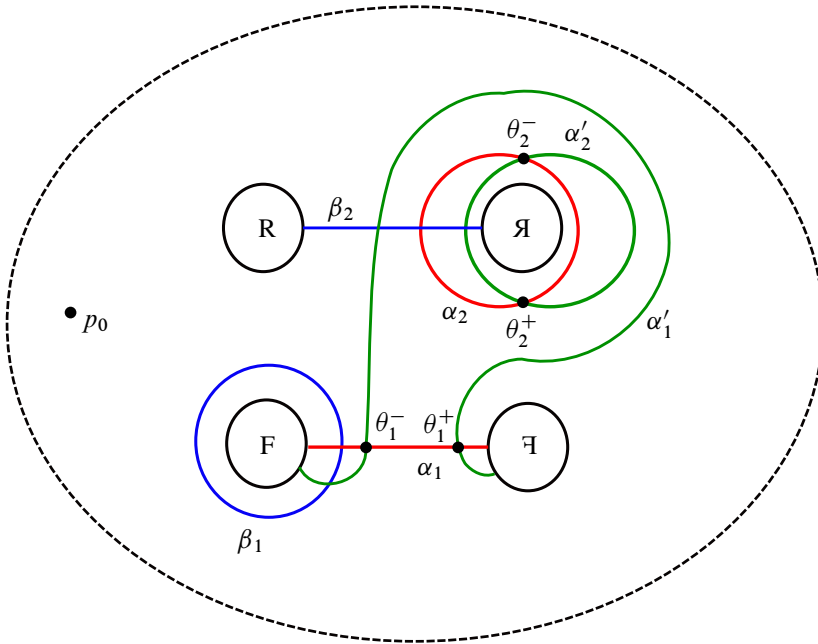


Figure 6: The pointed triple diagram  $\mathcal{T}_0$ , with the curves  $\alpha'_0 = (\alpha'_1, \alpha'_2)$ ,  $\alpha_0 = (\alpha_1, \alpha_2)$ ,  $\beta_0 = (\beta_1, \beta_2)$ , and the  $\theta$  intersection points, labeled.

### 8 Simple handleswap invariance

In this section we prove Theorem 7.2. The key result which will need to be established is the integral analog of a triangle count proved in [5, Proposition 9.31]. We will consider the pointed genus two Heegaard triple diagram  $\mathcal{T}_0$  shown in Figure 6 (compare the diagrams in Figure 4). Given any triple diagram  $\mathcal{T}$  we will show that triangle maps on the stabilized diagram  $\mathcal{T} \# \mathcal{T}_0$ , endowed with a sufficiently stretched neck, are determined by triangle maps on the unstabilized diagram  $\mathcal{T}$ .

We now fix some notation regarding the intersection points in the triple diagram  $\mathcal{T}_0 = (\Sigma, \alpha'_0, \alpha_0, \beta_0, p_0)$ . We write  $\mathbb{T}_{\alpha_0} \cap \mathbb{T}_{\beta_0} = \{\mathbf{a}\}$ ,  $\mathbb{T}_{\alpha'_0} \cap \mathbb{T}_{\beta_0} = \{\mathbf{b}\}$ , and

$$\mathbb{T}_{\alpha'_0} \cap \mathbb{T}_{\alpha_0} = \{\theta_1^+ \theta_2^+, \theta_1^+ \theta_2^-, \theta_1^- \theta_2^+, \theta_1^- \theta_2^-\}.$$

Here the intersection points  $\theta_1^\pm \in \alpha'_1 \cap \alpha_1$  and  $\theta_2^\pm \in \alpha'_2 \cap \alpha_2$  are those labeled in Figure 6. We write  $\Theta := \theta_1^+ \theta_2^+$ . We will show:

**Proposition 8.1** (cf [5, Proposition 9.31]) *Fix a strongly  $\mathfrak{s}$ -admissible Heegaard triple  $\mathcal{T} = (\Sigma, \alpha', \alpha, \beta, p)$ , and consider the diagram  $\mathcal{T} \# \mathcal{T}_0$ , where  $\mathcal{T}_0 = (\Sigma, \alpha'_0, \alpha_0, \beta_0, p_0)$*



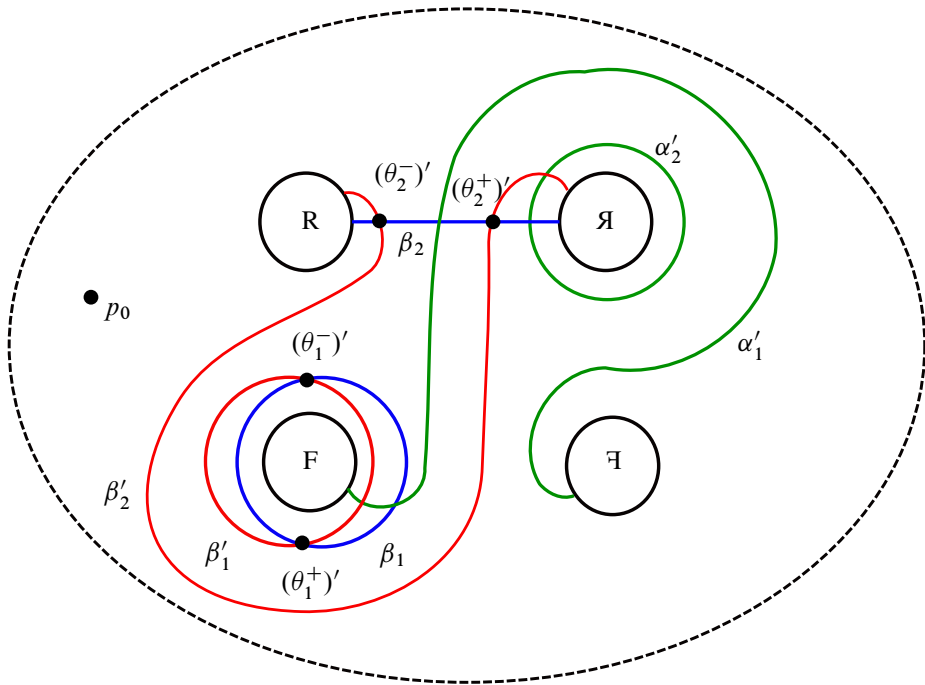


Figure 7: The pointed triple diagram  $\mathcal{T}'_0$ , with the curves  $\alpha'_0 = (\alpha'_1, \alpha'_2)$ ,  $\beta_0 = (\beta_1, \beta_2)$ , and  $\beta'_0 = (\beta'_1, \beta'_2)$ , and the  $\theta'$  intersection points, labeled.

is the diagram in Figure 6 and the connect sum is taken at the basepoints  $p$  and  $p_0$ . Then for a generic and sufficiently stretched almost complex structure there is a coherent orientation system  $\sigma_{\mathcal{T}_0}$  on  $\mathcal{T}_0$ , which together with any coherent orientation system  $\sigma_{\mathcal{T}}$  on  $\mathcal{T}$  induces a coherent orientation system  $\sigma_{\mathcal{T} \# \mathcal{T}_0}$  on  $\mathcal{T} \# \mathcal{T}_0$ . Furthermore, with respect to these orientations,

$$\mathcal{F}_{\mathcal{T} \# \mathcal{T}_0}((x \times \Theta) \otimes (y \times a), s) = \pm \mathcal{F}_{\mathcal{T}}(x \otimes y, s) \times b$$

for any  $x \in \mathbb{T}_{\alpha'} \cap \mathbb{T}_{\alpha}$  and  $y \in \mathbb{T}_{\alpha} \cap \mathbb{T}_{\beta}$ .

In fact when we prove handleswap invariance the diagram  $\mathcal{T}_0$  and the triangle count just stated will be relevant only to the consideration of the strong  $\alpha$ -equivalence involved in the statement. We will need an analogous result which pertains to the strong  $\beta$ -equivalence map occurring in the statement. We now state the precise result we will need for this. Let  $\mathcal{T}'_0 = (\Sigma_0, \alpha'_0, \beta_0, \beta'_0, p_0)$  denote the pointed genus two triple diagram shown in Figure 7, where  $\alpha'_0 = \{\alpha'_1, \alpha'_2\}$ ,  $\beta_0 = \{\beta_1, \beta_2\}$  and  $\beta'_0 = \{\beta'_1, \beta'_2\}$  (again compare the diagrams in Figure 4).

We further fix the following notation for intersection points in the diagram: we let  $\mathbb{T}_{\alpha'_0} \cap \mathbb{T}_{\beta_0} = \{\mathbf{b}\}$ ,  $\mathbb{T}_{\alpha'_0} \cap \mathbb{T}_{\beta'_0} = \{\mathbf{c}\}$ , and  $\Theta'$  denote the generator in  $\mathbb{T}_{\beta_0} \cap \mathbb{T}_{\beta'_0}$  with the highest relative grading. Let  $\mathcal{T}' = (\Sigma, \alpha', \beta, \beta', p)$  be another pointed Heegaard triple, and consider the diagram  $\mathcal{T}' \# \mathcal{T}'_0$ , where the connect sum is taken at the basepoints  $p$  and  $p_0$ . Then we will have an analogous triangle count:

**Proposition 8.2** (cf [5, Proposition 9.32]) *Fix a strongly  $s$ -admissible Heegaard triple  $\mathcal{T}' = (\Sigma, \alpha', \beta, \beta', p)$ , and consider the diagram  $\mathcal{T}' \# \mathcal{T}'_0$  as above. Then for a generic and sufficiently stretched almost complex structure there is a coherent orientation system  $\circ_{\mathcal{T}'_0}$  on  $\mathcal{T}'_0$ , which together with any coherent orientation system  $\circ_{\mathcal{T}'}$  on  $\mathcal{T}'$  induces a coherent orientation system  $\circ_{\mathcal{T}' \# \mathcal{T}'_0}$  on  $\mathcal{T}' \# \mathcal{T}'_0$ . Furthermore, with respect to these orientations,*

$$\mathcal{F}_{\mathcal{T}' \# \mathcal{T}'_0}((\mathbf{x} \times \mathbf{b}) \otimes (\mathbf{y} \times \Theta'), \mathfrak{s}) = \pm \mathcal{F}_{\mathcal{T}'}(\mathbf{x} \otimes \mathbf{y}, \mathfrak{s}) \times \mathbf{c}$$

for any  $\mathbf{x} \in \mathbb{T}_{\alpha'} \cap \mathbb{T}_{\beta}$  and  $\mathbf{y} \in \mathbb{T}_{\beta} \cap \mathbb{T}_{\beta'}$ .

We will prove Proposition 8.1 in the following subsection. Since a nearly identical proof can be used to establish Proposition 8.2, we omit the proof of that result. We now assume Propositions 8.1 and 8.2 and use them to establish Theorem 7.2.

**Proof of Theorem 7.2** We consider a simple handleswap  $(H_1, H_2, H_3, e, f, g)$  as in Definition 3.6. We first note that to prove the statement about transitive systems appearing in Theorem 7.2, it will suffice to find representatives  $\mathcal{H}_1, \mathcal{H}_2$ , and  $\mathcal{H}_3$  for the isotopy diagrams, and show that for these representatives,

$$g_* \circ \Phi_f \circ \Phi_e = \pm \text{Id}_{CF^-(\mathcal{H}_1)}$$

in  $\text{Kom}(\mathbb{Z}[U]\text{-Mod})$ , or equivalently

$$g_* \circ \Phi_f \circ \Phi_e = \text{Id}_{CF^-(\mathcal{H}_1)}$$

in  $P(\text{Kom}(\mathbb{Z}[U]\text{-Mod}))$ . Indeed, since each of the maps  $\Phi_e, \Phi_f$ , and  $g_*$  above are contained in the morphisms  $\Phi_e, \Phi_f$  and  $g_*$  of the transitive systems  $CF^-(H)$ , by the results in Sections 6 and 7, this monodromy relation will automatically yield corresponding monodromy relation for all such triangles.

Let  $\mathcal{H}_1 = (\Sigma \# \Sigma_0, \alpha_1, \beta_2)$  be a representative for the first isotopy diagram in the collection of data specifying the simple handleswap. By definition,  $\mathcal{H}_1$  decomposes as  $\mathcal{H} \# \mathcal{H}_0$ , where  $\mathcal{H} = (\Sigma, \alpha, \beta)$  and  $\mathcal{H}_0 = (\Sigma_0, \alpha_0, \beta_0)$  are as in Figure 4 ( $\mathcal{H}_0$  here is what we were denoting by  $P \cap \mathcal{H}_1$  in Definition 3.6).

Fix two new curves  $\alpha'_0$  on  $\Sigma_0$  which are related to  $\alpha_0$  as in the diagram  $\mathcal{T}_0$  in the statement of Proposition 8.1. Fix also a collection of curves  $\alpha'$  on  $\Sigma$  which are obtained by performing a small Hamiltonian isotopy on the curves in  $\alpha$ . The second isotopy diagram  $H_2$  can then be represented as  $H_2 = (\Sigma \# \Sigma_0, \alpha' \cup \alpha'_0, \beta \cup \beta_0)$ , and the morphism associated to the strong  $\alpha$ -equivalence  $e$  is given by the triangle map  $\Phi_e := \Psi_{\beta \cup \beta_0}^{\alpha \cup \alpha_0 \rightarrow \alpha' \cup \alpha'_0}$ . We note that our choices of representatives for the isotopy diagrams  $H_1$  and  $H_2$  ensure that the strong equivalence map of Definition 6.13 applied to these representatives is computed using only a single triangle map, as opposed to a composition of triangle maps and continuation maps. As in the notation of Proposition 8.1, we set  $\mathbb{T}_{\alpha_0} \cap \mathbb{T}_{\beta_0} = \{a\}$  and  $\mathbb{T}_{\alpha'_0} \cap \mathbb{T}_{\beta_0} = \{b\}$ . We then have for any  $y \times a \in \mathbb{T}_{\alpha \cup \alpha_0} \cap \mathbb{T}_{\beta \cup \beta_0}$ ,

$$\begin{aligned} \Phi_e(y \times a) &= \Psi_{\beta \cup \beta_0}^{\alpha \cup \alpha_0 \rightarrow \alpha' \cup \alpha'_0}(y \times a) \\ &= \mathcal{F}_{\alpha' \cup \alpha'_0, \alpha \cup \alpha_0, \beta \cup \beta_0}(\Theta_{\alpha' \cup \alpha'_0, \alpha \cup \alpha_0} \otimes (y \times a)) \\ &= \mathcal{F}_{\alpha' \cup \alpha'_0, \alpha \cup \alpha_0, \beta \cup \beta_0}((\Theta_{\alpha', \alpha} \times \Theta) \otimes (y \times a)) \\ &= \pm \mathcal{F}_{\alpha', \alpha, \beta}(\Theta_{\alpha', \alpha} \times y) \times b \\ &= \pm \Gamma_{\beta}^{\alpha \rightarrow \alpha'}(y) \times b. \end{aligned}$$

Here we have used Proposition 8.1 in the second to last equality, and Lemma 6.12 in the last equality.

We perform the analogous calculation for the strong  $\beta$ -equivalence. Fix two new curves  $\beta'_0$  on  $\Sigma_0$  which are related to  $\beta_0$  as in the diagram  $\mathcal{T}'_0$  in the statement of Proposition 8.2. Fix also a collection of curves  $\beta'$  on  $\Sigma$  which are obtained by performing a small Hamiltonian isotopy on the curves in  $\beta$ . The third isotopy diagram  $H_3$  can then be represented as  $H_3 = (\Sigma \# \Sigma_0, \alpha' \cup \alpha'_0, \beta' \cup \beta'_0)$ , and the morphism associated to the strong  $\beta$ -equivalence  $f$  is given by the triangle map  $\Phi_f := \Psi_{\beta' \cup \beta'_0}^{\alpha' \cup \alpha'_0}$ . As in the notation of Proposition 8.2, we set  $\mathbb{T}_{\alpha'_0} \cap \mathbb{T}_{\beta'_0} = \{c\}$ . By the same sequence of computations as in the previous case we then have for any  $x \times b \in \mathbb{T}_{\alpha' \cup \alpha'_0} \cap \mathbb{T}_{\beta \cup \beta_0}$ ,

$$\begin{aligned} \Phi_f(x \times b) &= \Psi_{\beta' \cup \beta'_0}^{\alpha' \cup \alpha'_0}(x \times b) \\ &= \mathcal{F}_{\alpha' \cup \alpha'_0, \beta \cup \beta_0, \beta' \cup \beta'_0}((x \times b) \otimes \Theta_{\beta \cup \beta_0, \beta' \cup \beta'_0}) \\ &= \mathcal{F}_{\alpha' \cup \alpha'_0, \beta \cup \beta_0, \beta' \cup \beta'_0}((x \times b) \otimes (\Theta_{\beta, \beta'} \times \Theta)) \\ &= \pm \mathcal{F}_{\alpha', \beta, \beta'}(x \times \Theta_{\beta, \beta'}) \times c \\ &= \pm \Gamma_{\beta'}^{\alpha' \rightarrow \beta'}(x) \times c. \end{aligned}$$

This time we have used Proposition 8.2 in the second-to-last equality, and again used Lemma 6.12 in the last equality.

We note that in the collection of representatives for the isotopy diagrams in a simple handleswap one could leave the  $\alpha$  and  $\beta$  curves unchanged throughout the handleswap, which would necessitate the diffeomorphism  $g$  restricting to the identity on  $\Sigma$ . Here we have altered  $\alpha$  and  $\beta$  slightly, so that the strong  $\alpha$ -equivalence and strong  $\beta$ -equivalence maps could each be computed via a single triangle map  $\Psi$ . Since our alteration of the curves  $\alpha$  and  $\beta$  on  $\Sigma$  came from small Hamiltonian isotopies, we can however still ensure that for our representatives for the handleswap the diffeomorphism  $g$  is isotopic to the identity when restricted to  $\Sigma$ . Furthermore, since  $g$  is part of a simple handleswap it must satisfy  $g(\alpha') = g(\alpha)$  and  $g(\beta') = g(\beta)$ . Thus, by definition of the maps induced by diffeomorphisms of diagrams, we have

$$g_*(z \times c) = (g|_\Sigma)_*(z) \times a$$

for all  $(z \times c) \in \mathbb{T}_{\alpha' \cup \alpha'_0} \cap \mathbb{T}_{\beta' \cup \beta'_0}$ .

Putting these formulas for each of the induced maps together, we find that

$$\begin{aligned} g_* \circ \Phi_f \circ \Phi_e(y \times a) &= (g_* \circ \Psi_{\beta \cup \beta_0 \rightarrow \beta' \cup \beta'_0}^{\alpha' \cup \alpha'_0} \circ \Psi_{\beta \cup \beta_0}^{\alpha \cup \alpha_0 \rightarrow \alpha' \cup \alpha'_0})(y \times a) \\ &= \pm((g|_\Sigma)_* \circ \Gamma_{\beta \rightarrow \beta'}^{\alpha' \rightarrow \alpha'})_*(y) \times a. \end{aligned}$$

Since the restriction of  $g$  to  $\Sigma$  is isotopic to the identity, Theorem 7.1 ensures

$$(g|_\Sigma)_* \circ \Gamma_{\beta \rightarrow \beta'}^{\alpha' \rightarrow \alpha'} \sim \pm \text{Id}_{CF^-(\mathcal{H})}.$$

We thus have

$$\begin{aligned} g_* \circ \Phi_f \circ \Phi_e &= \pm((g|_\Sigma)_* \circ \Gamma_{\beta \rightarrow \beta'}^{\alpha' \rightarrow \alpha'}) \otimes \text{Id}_{CF^-(\mathcal{H}_0)} \\ &\sim \pm \text{Id}_{CF^-(\mathcal{H})} \otimes \text{Id}_{CF^-(\mathcal{H}_0)} \\ &\sim \pm \text{Id}_{CF^-(\mathcal{H}_1)}, \end{aligned}$$

which by the remarks at the beginning of the proof completes the argument. □

Having established that Propositions 8.1 and 8.2 together imply Theorem 7.2, we now turn towards proving Proposition 8.1.

We employ the strategy used in [5] for proving the analog of Proposition 8.1 appearing there. We import many results exactly as they are stated there, while in a few cases we make small modifications in order to be able to apply their results. For the reader's convenience we provide statements of some results from [5], and provide proofs of any

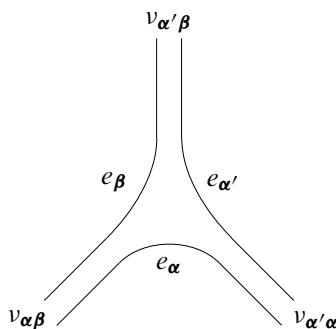


Figure 8: The region  $\Delta$ .

imported results which must be modified slightly for our purposes. We also provide sketches of proofs of certain statements from [5] which we do not need to modify, but whose exposition we hope will aid in the readability of this paper.

In the remainder of this section we work in the cylindrical formulation of Heegaard Floer homology introduced by Lipshitz in [6].

### 8.1 Moduli spaces of triangles

We begin by recalling some notation and terminology regarding holomorphic triangles in the cylindrical setting of Heegaard Floer homology; see [6]. We denote by  $\Delta$  the subset of  $\mathbb{C}$  shown in Figure 8, which has three cylindrical ends modeled on  $[0, 1] \times [0, \infty)$ . We will think of this region as a triangle with its vertices removed. We also introduce in the figure notation we will use to indicate the boundary components and ends of this region.

We will consider almost complex structures  $J$  on  $\Sigma \times \Delta$  which satisfy the following conditions:

- (J'1')  $J$  is tamed by the split symplectic form on  $\Sigma \times \Delta$ .
- (J'2') On each component of  $\Sigma \setminus (\alpha' \cup \alpha \cup \beta)$  there is at least one point at which  $J = j_{\Sigma} \times j_{\Delta}$ .
- (J'3') On each cylindrical end  $\Sigma \times [0, 1] \times \mathbb{R}$  of  $\Sigma \times \Delta$ , there is a 2-plane distribution  $\eta$  on  $\Sigma \times [0, 1] \times \{0\}$  such that the restriction of  $\omega$  to  $\eta$  is nondegenerate,  $J$  preserves  $\eta$ , and the restriction of  $J$  to  $\eta$  is compatible with  $\omega$ . Furthermore,  $\eta$  is tangent to  $\Sigma$  near  $(\Sigma \times \{0, 1\} \times \{0\}) \cup (\Sigma \times [0, 1] \times \{0\})$ .
- (J'4') The planes  $T_d(\{p\} \times \Delta)$  are complex lines of  $J$  for all  $(p, d) \in \Sigma \times \Delta$ .

(J'5') There is an open set  $U \subset \Delta$  containing  $\partial\Delta \setminus \{v_{\alpha'\alpha}, v_{\alpha\beta}, v_{\alpha'\beta}\}$  such that the planes  $T_p(\Sigma \times \{d\})$  are complex lines of  $J$  for all  $(p, d)$  near  $(\alpha' \cup \alpha \cup \beta) \times \Delta$  and for all  $(p, d) \in \Sigma \times U$ .

$J$ -holomorphic curves in  $\Sigma \times \Delta$  for almost complex structures  $J$  of this sort enjoy the following property.

**Lemma 8.3** [6, Lemma 3.1] *Let  $J$  be an almost complex structure on  $\Sigma \times \Delta$  that satisfies the axioms (J'1')–(J'5'). If  $u: S \rightarrow \Sigma \times \Delta$  is  $J$ -holomorphic and  $\pi_\Sigma \circ u$  is nonconstant on a component  $S_0$  of  $S$ , then  $\pi_\Sigma \circ u|_{S_0}$  is an open map. Furthermore, there are coordinates near any critical point of  $\pi_\Sigma \circ u|_{S_0}$  where  $\pi_\Sigma \circ u$  takes the form  $z \mapsto z^k$  for some  $k > 0$ .*

In fact, this result follows immediately from [9, Theorem 7.1].

To understand Proposition 8.1, we will need to investigate the nature of triangle maps on the diagram  $\mathcal{T} \# \mathcal{T}_0$ . In the cylindrical setting, the notion of a holomorphic triangle in a Heegaard triple diagram takes the following form.

**Definition 8.4** Let  $\mathcal{T} = (\Sigma, \alpha', \alpha, \beta)$  be a triple diagram, and set  $d = |\alpha'| = |\alpha| = |\beta|$ . By a *holomorphic triangle in the triple diagram  $\mathcal{T}$*  we will mean a  $(j, J)$ -holomorphic map  $u: S \rightarrow \Sigma \times \Delta$  satisfying:

- (M1)  $(S, j)$  is a (possibly nodal) Riemann surface with boundary and  $3d$  punctures on  $\partial S$ .
- (M2)  $u$  is locally nonconstant.
- (M3)  $u(\partial S) \subset (\alpha' \times e_{\alpha'}) \cup (\alpha \times e_\alpha) \cup (\beta \times e_\beta)$ .
- (M4)  $u$  has finite energy.
- (M5) For each  $i \in \{1, \dots, d\}$  and  $\sigma \in \{\alpha', \alpha, \beta\}$ , the preimage  $u^{-1}(\sigma_i \times e_\sigma)$  consists of exactly one component of the punctured boundary of  $S$ .
- (M6) As one approaches the punctures of  $\partial S$ , the map  $u$  converges to a collection of intersection points on the Heegaard triple in the cylindrical ends of  $\Sigma \times \Delta$ .

We will often ask holomorphic triangles to satisfy the following additional two requirements:

- (M7)  $\pi_\Delta \circ u$  is nonconstant on each component of  $S$ .
- (M8)  $S$  is smooth, and  $u$  is an embedding.

Unless otherwise specified, we will use the term holomorphic triangle to refer to maps satisfying axioms (M1)–(M6), and explicitly note when we are considering curves satisfying the additional axioms (M7) and (M8).

For any homology class  $\psi$  of triangles on a Heegaard triple diagram  $\mathcal{T}$ , we will denote by  $\mathcal{M}(\psi)$  the moduli space of holomorphic triangles on  $\mathcal{T}$  in the homology class  $\psi$ . Given a Riemann surface  $S$ , we will indicate by  $\mathcal{M}(\psi, S)$  the subspace of  $\mathcal{M}(\psi)$  consisting of holomorphic triangles with source  $S$ .

To obtain the triangle count we are after on a sufficiently stretched copy of  $\mathcal{T} \# \mathcal{T}_0$ , we will need to understand compactifications of these moduli spaces of triangles. These compactifications allow for a weaker notion of triangle which we refer to as broken:

**Definition 8.5** Let  $\mathcal{T} = (\Sigma, \alpha', \alpha, \beta)$  and  $d$  be as above. We say that a collection of  $(j, J)$ -holomorphic curves  $BT = (u_1, v_1, \dots, v_n, w_1, \dots, w_m)$  is a *broken holomorphic triangle on  $\mathcal{T}$  representing the homology class  $\psi$*  if

- (BT1)  $u_1$  is a curve mapping to  $\Sigma \times \Delta$  satisfying (M1) and (M3)–(M6).
- (BT2)  $v_i$  are curves mapping to  $\Sigma \times I \times \mathbb{R}$  which satisfy the analogs of (M1) and (M3)–(M6), each representing some homology class of strips in one of the diagrams  $(\Sigma, \alpha, \alpha')$ ,  $(\Sigma, \alpha', \beta)$  or  $(\Sigma, \alpha, \beta)$ .
- (BT3) The  $w_i$  are curves from Riemann surfaces with  $d$  boundary components and a single puncture on each boundary component, and which map to

$$(\Sigma \times I \times \mathbb{R}) \amalg (\Sigma \times \Delta).$$

For each  $i$ , the boundary components of the curve  $w_i$  all map to a single set of attaching curves.

- (BT4) The total homology class of the curves in  $BT$  is equal to  $\psi$ .

With this notion in hand, we can state the following compactness result which describes the behavior of triangles on  $\mathcal{T} \# \mathcal{T}_0$  as we stretch the neck:

**Proposition 8.6** [5, Proposition 9.40] *Let  $\psi \# \psi_0$  be a homology class of triangles on  $(\Sigma \# \Sigma_0) \times \Delta$ , and  $u_{T_i}$  be a sequence of holomorphic triangle representatives for  $\psi \# \psi_0$  on  $(\Sigma \# \Sigma_0) \times \Delta$ , with respect to almost complex structures  $J(T_i)$  for neck lengths  $T_i \rightarrow \infty$ . Then there is a subsequence which converges to a triple  $(U, V, U_0)$  where  $U$  and  $U_0$  are broken holomorphic triangles on  $\Sigma \times \Delta$  and  $\Sigma_0 \times \Delta$  representing*

$\psi$  and  $\psi_0$  respectively, and  $V$  is a collection of holomorphic curves on the neck regions  $S^1 \times \mathbb{R} \times \Delta$  or  $S^1 \times \mathbb{R} \times [0, 1] \times \mathbb{R}$  which are asymptotic to (possibly multiply covered) Reeb orbits  $S^1 \times \{d\}$  for  $d \in \Delta$  or  $d \in [0, 1] \times \mathbb{R}$ .

**Remark 8.7** More precisely, the asymptotic condition on the curves appearing in  $V$  in Proposition 8.6 above has the following meaning. By a “Reeb orbit” in this context, we mean a periodic orbit  $\gamma$  of the vector field  $d/d\theta$  on  $S^1 \times \mathbb{R} \times \Delta$  or  $S^1 \times \mathbb{R} \times I \times \mathbb{R}$ , where  $\theta$  is the coordinate on  $S^1$ . The curves  $v$  in  $V$  have as sources punctured Riemann surfaces. Let  $S$  be a connected component of such a source,  $q$  a puncture of  $S$ , and  $v: S \rightarrow S^1 \times \mathbb{R} \times \Delta$ . Write  $(\theta, r, z)$  for coordinates on the target. Then  $v$  is asymptotic to  $\gamma$  at  $q$  if:

- (1) There is a neighborhood  $U$  of  $q$  in  $S$  and a biholomorphic diffeomorphism  $\phi: U \cong S^1 \times (0, \infty)$ . Write  $(x, y)$  for coordinates on  $S^1 \times (0, \infty)$ .
- (2)  $r \circ v \circ \phi^{-1} \rightarrow \infty$  as  $y \rightarrow \infty$ .
- (3)  $(\theta, z) \circ v \circ \phi^{-1}(x, y) \rightarrow \gamma(x)$  as  $y \rightarrow \infty$  as maps  $S^1 \rightarrow S^1 \times \Delta$  in  $C_{\text{loc}}^\infty$ .

## 8.2 Matched moduli spaces and orientations

Fix a triple diagram  $\mathcal{T} = (\Sigma, \alpha', \alpha, \beta)$  and a point  $p \in \Sigma \setminus (\alpha' \cup \alpha \cup \beta)$ . Let  $u: S \rightarrow \Sigma \times \Delta$  be a  $J$ -holomorphic curve satisfying (M1)–(M6), for some almost complex structure  $J$  on  $\Sigma \times \Delta$  satisfying (J'1')–(J'5'). Then  $u$  is locally nonconstant by condition (M2), so, by Lemma 8.3,  $\pi_\Sigma \circ u$  is an open map on each component of  $S$ , and takes the form  $z \mapsto z^k$  near any critical point. Thus  $(\pi_\Sigma \circ u)^{-1}(p)$  is a finite set of points. Furthermore, using property (J'4') of the almost complex structure  $J$ , positivity of complex intersections for  $J$ -holomorphic curves — see eg [9] or [8] — ensures that all intersections between  $p \times \Delta$  and the image of  $u$  are positive.

We will write  $(\pi_\Sigma \circ u)^{-1}(p) = \{x_1, \dots, x_{n_p(u)}\} \in \text{Sym}^{n_p(u)}(S)$ , and define

$$\rho^p(u) := \{\pi_\Delta \circ u(x_1), \dots, \pi_\Delta \circ u(x_{n_p(u)})\} \in \text{Sym}^{n_p(u)}(\Delta).$$

We remark that our notation involving set braces is somewhat misleading, as there may of course be repetitions among the points  $x_i$  in the symmetric product, corresponding to intersection points occurring with positive multiplicity greater than 1.

To understand the triangle count, we will be concerned with holomorphic triangles  $u$  for which  $\rho^p(u)$  takes prescribed values. As a first step towards understanding the moduli spaces of such triangles, Juhász, Thurston and Zemke show that, for any prescribed value outside the fat diagonal, such a triangle is somewhere injective.



**Lemma 8.8** [5, Lemma 9.45] *Let  $(\Sigma, \alpha', \alpha, \beta, p)$  be a triple diagram, and*

$$\mathbf{d} \in \text{Sym}^k(\Delta) \setminus \text{Diag}(\Delta).$$

*If  $u: S \rightarrow \Sigma \times \Delta$  is a  $J$ -holomorphic curve satisfying (M1)–(M6) for an almost complex structure satisfying (J'1')–(J'5'), which furthermore has  $\rho^p(u) = \mathbf{d}$ , then every component of  $u$  is somewhere injective.*

Fix a Heegaard triple diagram  $\mathcal{T} = (\Sigma, \alpha', \alpha, \beta, p)$  and a homology class of triangle  $\psi$ , with  $n_p(\psi) = k$ . Given a subset  $X \subset \text{Sym}^k(\Delta)$ , we let

$$\mathcal{M}(\psi, S, X) = \{u \in \mathcal{M}(\psi, S) \mid \rho^p(u) \in X\}$$

and

$$\mathcal{M}(\psi, X) = \{u \in \mathcal{M}(\psi) \mid \rho^p(u) \in X\}.$$

Using techniques similar to those used in the standard setting, Juhász, Thurston and Zemke prove the following result, which shows that generically these matched moduli spaces are smooth manifolds.

**Proposition 8.9** [5, Proposition 9.47] *Let  $(\Sigma, \alpha', \alpha, \beta)$  be a triple diagram, and fix a point  $p \in \Sigma \setminus (\alpha' \cup \alpha \cup \beta)$ . Suppose  $X \subset \text{Sym}^k(\Delta)$  for some  $k \in \mathbb{N}$  is a nonempty submanifold that does not intersect the fat diagonal. Furthermore, suppose that for every  $x \in X$ , the  $k$ -tuple  $x$  has no coordinate in the open set  $U \subset \Delta$  from (J'5'). Then, for a generic choice of almost complex structure  $J$ , the set  $\mathcal{M}(\psi, S, X)$  is a smooth manifold of dimension*

$$\text{ind}(\psi, S) - \text{codim}(X)$$

*where  $\text{ind}(\psi, S)$  denotes the Fredholm index of the linearized  $\bar{\partial}$  operator at any representative  $u: S \rightarrow \Sigma \times \Delta$  for  $\psi$ . For  $X = \text{Sym}^k(\Delta)$ , the same statement holds near any curve  $u$  that has no component  $T$  on which  $\pi_\Delta \circ u|_T$  is constant and has image in  $U$ , and such that all components of  $u$  are somewhere injective.*

It will be important for our purposes to note that these moduli spaces are also orientable when they are smoothly cut out, which follows in a straightforward manner from the framework in which the proof of the previous proposition is carried out. We now provide a sketch of the argument.

**Lemma 8.10** *For  $J$  and  $X$  as in Proposition 8.9, with  $X \subset \text{Sym}^k(\Delta)$  furthermore assumed to be an orientable submanifold,  $\mathcal{M}(\psi, S, X)$  is orientable.*

**Proof** Forgetting the matching condition — ie taking  $X = \text{Sym}^k(\Delta)$  — we consider  $\mathcal{M}(\psi, S, \text{Sym}^k(\Delta)) = \mathcal{M}(\psi, S)$ . By [6, Proposition 6.3 and Section 10.3], whenever this space is transversely cut out it is an orientable smooth manifold.

For the case when  $X \neq \text{Sym}^k(\Delta)$ , we briefly recall how one can establish the existence of a smooth manifold structure on  $\mathcal{M}(\psi, S, X)$ , as in the proof of [5, Proposition 9.47]. Consider the map  $\rho^p: \mathcal{M}(\psi, S) \rightarrow \text{Sym}^k(\Delta)$ . To obtain the smooth manifold structure on  $\mathcal{M}(\psi, S, X)$ , one considers the universal moduli space  $\mathcal{M}_{\text{univ}}^\ell(\psi, S)$ . This consists of triples  $(u, j, J)$ , where  $j$  is a  $C^\ell$  complex structure on  $S$ ,  $J$  is a  $C^\ell$  almost complex structure on  $\Sigma \times \Delta$  satisfying conditions  $(J'1')$ – $(J'5')$ , and  $u$  is a  $(j, J)$ –holomorphic map  $u: S \rightarrow \Sigma \times \Delta$  in the homology class  $\psi$ , which furthermore satisfies certain regularity conditions; see [6, page 968]. It is shown in the proof of Proposition 8.9, using the technique of [6, Proposition 3.7], that the universal moduli space  $\mathcal{M}_{\text{univ}}^\ell(\psi, S)$  is a Banach manifold and the evaluation map  $\rho^p: \mathcal{M}_{\text{univ}}^\ell(\psi, S) \rightarrow \text{Sym}^k(\Delta)$  is a submersion at all triples  $(u, j, J)$  for which  $\rho^p(u)$  is not in the fat diagonal. Thus for  $X$  missing the fat diagonal, the universal matched moduli space  $\mathcal{M}_{\text{univ}}^\ell(\psi, S, X) := (\rho^p)^{-1}(X)$  is a Banach manifold. One can then apply the Sard–Smale theorem to the Fredholm map  $\pi: \mathcal{M}_{\text{univ}}^\ell(\psi, S, X) \rightarrow \mathcal{J}^\ell$  to obtain a regular value  $J \in \mathcal{J}^\ell$  so that

$$\mathcal{M}^\ell(\psi, S, X) = \pi^{-1}(J)$$

is a smooth manifold. Finally, one uses an approximating bootstrapping argument to obtain the same result for  $C^\infty$  complex structures. More precisely, one obtains that for a generic choice of  $J$  the space  $\mathcal{M}(\psi, S)$  is a smooth manifold and the map

$$\rho^p: \mathcal{M}(\psi, S) \rightarrow \text{Sym}^k(\Delta)$$

is transverse to  $X$ . Thus, for  $X$  missing the fat diagonal,  $\mathcal{M}(\psi, S, X) := (\rho^p)^{-1}(X)$  is a smooth manifold.

Fixing  $u \in \mathcal{M}(\psi, S, X)$ ,

$$T_u \mathcal{M}(\psi, S) \cong T_u \mathcal{M}(\psi, S, X) \oplus N_u$$

where  $N$  is any choice of orthogonal complement. Since  $\mathcal{M}(\psi, S)$  is orientable, it will suffice to show  $N$  is orientable to establish that  $\mathcal{M}(\psi, S, X)$  is orientable. Since  $\rho^p$  is transverse to  $X$ ,

$$d\rho^p(T_u \mathcal{M}(\psi, S)) + T_{\rho^p(u)} X = T_{\rho^p(u)} \text{Sym}^k(\Delta).$$

Since  $(d\rho^p)^{-1}(TX) = T\mathcal{M}(\psi, S, X)$ , the two equations above yield a direct sum decomposition

$$d\rho^p(N_u) \oplus T_{\rho^p(u)}X \cong T_{\rho^p(u)}\text{Sym}^k(\Delta).$$

Finally, since  $X$  and  $\text{Sym}^k(\Delta)$  are orientable, and  $d\rho^p|_N$  is an isomorphism on each fiber, the last equation establishes orientability of the complement  $N$ . Thus  $\mathcal{M}(\psi, S, X)$  is orientable, as desired.  $\square$

We now turn to an investigation of the behavior of orientations on these moduli spaces. We recall again the notion of coherent orientation systems, and now provide the precise definitions in the cylindrical setting, as we will need them in some of our computations. We begin with the moduli space of holomorphic *strips* in a homology class  $A \in \pi_2(\mathbf{x}, \mathbf{y})$ , denoted by  $\mathcal{M}^A$ , on some Heegaard (double) diagram  $\mathcal{H} = (\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta})$ . We set  $\widehat{\mathcal{M}}^A = \mathcal{M}^A/\mathbb{R}$ . As noted above, these moduli spaces are orientable whenever they are smoothly cut out by [6, Proposition 6.3]. There this is shown by trivializing the determinant line bundle of the virtual index bundle of the linearized  $\bar{\partial}$ -equation. In fact, this line bundle is trivialized over a larger auxiliary space of curves which are not necessarily holomorphic, which we denote by  $\mathcal{B}^A$ , rather than over  $\mathcal{M}^A$ . We ask for the trivializations of these determinant lines  $\mathcal{L}$  over  $\mathcal{B}^A$  to satisfy the following compatibility under gluing.

**Definition 8.11** Given a Heegaard diagram  $\mathcal{H}$ , homology classes of strips  $A$  and  $A'$  which are adjacent on the diagram — ie  $A \in \pi_2(\mathbf{x}, \mathbf{y})$  and  $A' \in \pi_2(\mathbf{y}, \mathbf{z})$  — and maps  $u: S \rightarrow \Sigma \times I \times \mathbb{R}$  and  $u': S' \rightarrow \Sigma \times I \times \mathbb{R}$  representing  $A$  and  $A'$  respectively, one can preglue the positive corners of  $u$  to the negative corners of  $u'$ ; see [6, Appendix A] for one such construction. In fact, there is a 1-parameter family of such pregluings  $(u \natural_r u': S \natural_r S' \rightarrow \Sigma \times I \times \mathbb{R})$  in the class  $A + A'$ , defined for sufficiently large values of the parameter  $r$ . One can show that this map preserves the analogs of (M1), (M3) and (M4) for strips, and the asymptotic conditions one asks of the strips. Denote the collection of maps of the form  $S \rightarrow \Sigma \times I \times \mathbb{R}$  in a given homology class  $A$  which furthermore satisfy (M1), (M3), (M4) and the asymptotic conditions by  $\mathcal{B}^A(S)$ . We say a choice of orientations for all  $\widehat{\mathcal{M}}^A$ , specified by a collection of nonvanishing sections  $\sigma_{\mathcal{H}} = \sigma_{\boldsymbol{\alpha}, \boldsymbol{\beta}} = \{\sigma^A\}$  of  $\mathcal{L}$  over all of the  $\widehat{\mathcal{M}}^A$ , is a *coherent orientation system on  $\mathcal{H}$*  if the induced map of determinant lines covering the map  $\natural_r: \mathcal{B}^A(S) \times \mathcal{B}^{A'}(S') \times (R, \infty) \rightarrow \mathcal{B}^{A+A'}(S \natural_r S')$  satisfies  $(\natural_r)_*(\sigma^A \times \sigma^{A'}) = +\sigma^{A+A'}$ .

That such coherent orientation systems exist is shown in numerous places. One construction sufficient for our purposes can be found in [6, Section 6].

In the case of holomorphic triangles, the moduli spaces  $\mathcal{M}(\psi)$  are also orientable. For a collection of orientations on  $\mathcal{M}(\psi)$  for all homology classes  $\psi$  of triangles in a triple diagram, we will consider a related notion of coherence.

**Definition 8.12** Given a Heegaard triple diagram  $\mathcal{T}$ , we will say a choice of orientations for  $\mathcal{M}^{\psi_{\alpha,\beta}}$ ,  $\mathcal{M}^{\psi_{\beta,\gamma}}$ ,  $\mathcal{M}^{\psi_{\alpha,\gamma}}$  and  $\mathcal{M}(\psi)$  (for  $\psi_{\alpha,\beta}$ ,  $\psi_{\beta,\gamma}$  and  $\psi_{\alpha,\gamma}$  ranging over all classes of strips in the respective double diagrams, and  $\psi$  ranging over all classes of triangles in the triple diagram) specified by a collection of sections  $\sigma_{\mathcal{T}} = \{\sigma_{\alpha,\beta,\gamma}, \sigma_{\alpha,\beta}, \sigma_{\beta,\gamma}, \sigma_{\alpha,\gamma}\}$  is a *coherent orientation system of triangles* if each collection of orientations of the moduli spaces of strips on the respective double diagrams are coherent, and all possible pregluings of triangles with strips satisfy the analogous gluing condition.

Following [6, Section 6], given a homology class of triangles  $\psi$  on the triple diagram  $\mathcal{T}$ , let  $T(\psi)$  denote the space of pairs  $(u, j)$ , where  $u: S \rightarrow \Sigma \times \Delta$  is a curve in the class  $\psi$  satisfying (M1), (M3) and (M4), and  $j$  is a complex structure on  $S$ . We declare two such pairs  $(u: S \rightarrow \Sigma \times \Delta, j)$  and  $(u': S' \rightarrow \Sigma \times \Delta, j')$  to be equivalent if there is a biholomorphism  $\phi: (S, j) \rightarrow (S', j')$  such that the diagram

$$(10) \quad \begin{array}{ccc} S & \xrightarrow{\phi} & S' \\ & \searrow u & \swarrow u' \\ & \Sigma \times \Delta & \end{array}$$

commutes. We denote the quotient of  $T(\psi)$  by this equivalence relation by  $\mathcal{B}(\psi)$ .

Let  $p: I \rightarrow \text{Sym}^k(\Delta)$  be an embedded path missing the fat diagonal. We consider the following moduli spaces of holomorphic triangles associated to homology classes  $\psi_0 \in \pi_2(\Theta, \mathbf{a}, \mathbf{b})$  in the triple diagram  $\mathcal{T}_0$  from Proposition 8.1,

$$(11) \quad \begin{aligned} \mathcal{M}_I^{\psi_0} &= \mathcal{M}(\psi_0, p(I)) \\ &= \{(u, t) \mid u \in \mathcal{M}(\psi_0) \text{ such that } \rho^p(u) \in p(t) \text{ for some } t \in I\} \end{aligned}$$

and

$$(12) \quad \mathcal{M}_t^{\psi_0} = \mathcal{M}(\psi_0, p(t)) = \{u \in \mathcal{M}(\psi_0) \text{ such that } \rho^p(u) \in p(t)\}.$$

By Proposition 8.9, for a generic choice of almost complex structure on  $\Sigma_0 \times \Delta$  the moduli spaces  $\mathcal{M}_I^{\psi_0}$  are smooth manifolds of dimension  $\mu(\psi_0) - \text{codim}(p(I))$ . By Lemma 8.15, we have  $\mu(\psi_0) = 2n_{p_0}(\psi_0)$ , so the expected dimension becomes  $2n_{p_0}(\psi_0) - (2k - 1)$ . In particular, when  $k = n_{p_0}(\psi_0)$  the moduli space  $\mathcal{M}_I^{\psi_0}$  is a

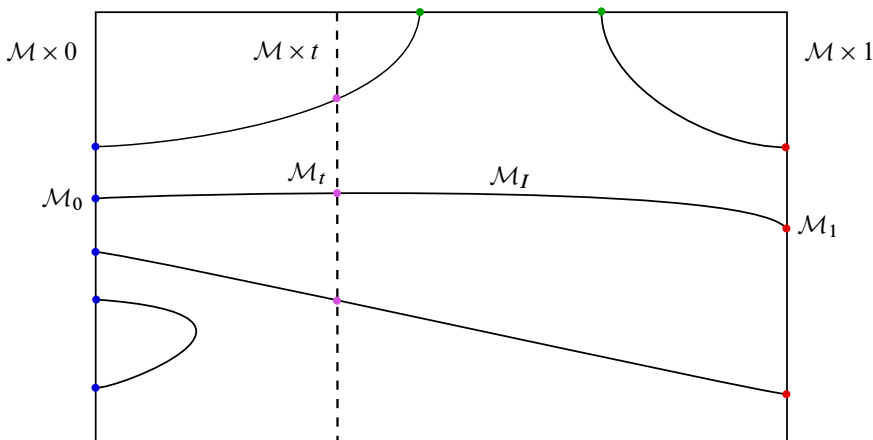


Figure 9: A schematic of the space  $\mathcal{M} \times I$  with  $\mathcal{M}_I$  inside it. Vertical slices of the picture such as the vertical dashed line represent the spaces  $\mathcal{M} \times t$ , while the solid curves collectively represent the smooth moduli space  $\mathcal{M}_I$ . The left and right endpoints on  $\mathcal{M}_I$  represent  $\mathcal{M}_0 \subset \mathcal{M} \times 0$  and  $\mathcal{M}_1 \subset \mathcal{M} \times 1$  respectively, while the endpoints of  $\mathcal{M}_I$  on the top and bottom of the figure represent degenerations of triangles into broken triangles in the compactification.

smooth 1-manifold when it is transversely cut out. Similarly, the expected dimension of  $\mathcal{M}_t^{\psi_0}$  is 0 when  $k = n_{p_0}(\psi_0)$ . Finally, we define the spaces

$$\mathcal{M} = \coprod_{\substack{\psi_0 \in \pi_2(\Theta, \mathbf{a}, \mathbf{b}) \\ n_{p_0}(\psi_0) = k}} \mathcal{M}^{\psi_0}, \quad \mathcal{M}_I = \coprod_{\substack{\psi_0 \in \pi_2(\Theta, \mathbf{a}, \mathbf{b}) \\ n_{p_0}(\psi_0) = k}} \mathcal{M}_I^{\psi_0}, \quad \mathcal{M}_t = \coprod_{\substack{\psi_0 \in \pi_2(\Theta, \mathbf{a}, \mathbf{b}) \\ n_{p_0}(\psi_0) = k}} \mathcal{M}_t^{\psi_0}.$$

We provide a schematic of these spaces and their relationships in Figure 9.

We note for the following arguments that by the remarks above  $\mathcal{M}_I$  is a smooth manifold of dimension 1 for a generic choice of almost complex structure, and for each  $t$  a (potentially different) generic choice of almost complex structure will ensure  $\mathcal{M}_t$  is a smooth manifold of dimension 0. We will denote by  $\mathfrak{o}_{\mathcal{M}_I}$  and  $\mathfrak{o}_{\mathcal{M}_t}$  nowhere zero sections of the bundles  $\mathcal{L}_I$  and  $\mathcal{L}_t$  respectively, which are the determinant line bundles of the virtual index bundles of the linearized equations defining these moduli spaces. We recall that such sections determine orientations of the moduli spaces.

For arguments appearing later, we want to ensure we can achieve the following intuitively achievable constraints on our orientations.

**Lemma 8.13** *Let  $\mathcal{M}_I$  and  $\mathcal{M}_t$  be as above. Then there are coherent orientation systems  $\mathfrak{o}_{\mathcal{M}_0}$  on  $\mathcal{M}_0$ ,  $\mathfrak{o}_{\mathcal{M}_1}$  on  $\mathcal{M}_1$ , and  $\mathfrak{o}_{\mathcal{M}_I}$  on  $\mathcal{M}_I$  such that  $(\mathfrak{o}_{\mathcal{M}_I})|_{\mathcal{M}_0} \cong -\mathfrak{o}_{\mathcal{M}_0}$  and  $(\mathfrak{o}_{\mathcal{M}_I})|_{\mathcal{M}_1} \cong \mathfrak{o}_{\mathcal{M}_1}$ .*

**Proof** The proof is an elaboration on that of Lemma 8.10. Consider again the universal moduli space of holomorphic maps  $\mathcal{M}_{\text{univ}}$  consisting of triples  $(u, j, J)$  satisfying the conditions as in the proof of Lemma 8.10. We consider the map

$$\rho^p \times \text{id}: \mathcal{M}_{\text{univ}} \times I \rightarrow \text{Sym}^k(\Delta) \times I$$

given by  $(\rho^p \times \text{id})(u, j, J, t) = (\rho^p(u), t)$ . This map is again a submersion when  $\rho^p(u)$  is not in the fat diagonal, by [5, Proposition 9.47]. Let

$$P = \{(p(t), t) \mid t \in I\}$$

and note that  $(\rho^p \times \text{id})^{-1}(P) = (\mathcal{M}_{\text{univ}})_I$ , where here we have used the notation  $(\mathcal{M}_{\text{univ}})_I$  to indicate the universal moduli space matched to  $p(I)$  (as the notation is used in (11)). Since we are working with a path  $p$  missing the fat diagonal,  $\rho^p \times \text{id}$  is a submersion, and we have, as a consequence of the Sard–Smale theorem, parametric transversality: denoting by  $\mathcal{M}_J$  the moduli space of holomorphic curves with respect to the almost complex structure  $J$ ,  $\rho^p \times \text{id}: \mathcal{M}_J \times I \rightarrow \text{Sym}^k(\Delta) \times I$  is transverse to  $P$  for generic  $J$ . By [6, Section 6] we may orient such  $\mathcal{M}_J$ , which in turn specifies a product orientation on  $\mathcal{M}_J \times I$ . Since  $P$  is also oriented (by a fixed orientation for  $I$ ), we then have for such  $J$  that  $(\rho^p \times \text{id})^{-1}(P) = \mathcal{M}_I$  inherits an orientation  $\mathfrak{o}_{\mathcal{M}_I}$ ; furthermore, this orientation satisfies the boundary conditions

$$(13) \quad (\mathfrak{o}_{\mathcal{M}_I})|_{\mathcal{M}_0} = -\mathfrak{o}_{\mathcal{M}_0} \quad \text{and} \quad (\mathfrak{o}_{\mathcal{M}_I})|_{\mathcal{M}_1} = \mathfrak{o}_{\mathcal{M}_1},$$

where  $\mathfrak{o}_{\mathcal{M}_0}$  and  $\mathfrak{o}_{\mathcal{M}_1}$  are the orientations coming from the previously fixed choice of orientation on  $\mathcal{M}_J$ , as desired. Finally, we note that by the same argument used to prove [6, Lemma 10.10], we may arrange for the orientation systems  $\mathfrak{o}_{\mathcal{M}_I}$ ,  $\mathfrak{o}_{\mathcal{M}_0}$  and  $\mathfrak{o}_{\mathcal{M}_1}$  in the preceding paragraph to be enlarged to coherent systems in the sense of Definition 8.12.  $\square$

Having discussed the smooth manifold structure and a particular construction of coherent orientations on the matched moduli spaces of triangles on a triple diagram, we now state a gluing result from [5] which will allow us to relate these matched moduli spaces of triangles on the diagram  $\mathcal{T}_0$  to the triangles on  $\mathcal{T} \# \mathcal{T}_0$  we seek to count. We consider homology classes of triangles  $\psi$  on an arbitrary pointed triple diagram

$\mathcal{T} = (\Sigma, \alpha', \alpha, \beta, p)$  and  $\psi_0$  on the pointed diagram  $\mathcal{T}_0 = (\Sigma_0, \alpha'_0, \alpha_0, \beta_0, p_0)$ . We form the connected sum of the diagrams at the points  $p$  and  $p_0$ , and consider the resulting homology class  $\psi \# \psi_0$ :

**Proposition 8.14** [5, Proposition 9.49] *Let  $u$  and  $u_0$  be holomorphic triangles representing homology classes  $\psi$  and  $\psi_0$  in  $\Sigma \times \Delta$  and  $\Sigma_0 \times \Delta$  respectively. Let  $k = n_p(\psi) = n_{p_0}(\psi_0)$ , and suppose  $\mu(u) = 0$ ,  $\mu(u_0) = 2k$ , and*

$$\rho^p(u) = \rho^{p_0}(u_0) \in \text{Sym}^k(\Delta) \setminus \text{Diag}^k(\Delta).$$

*Suppose further that  $\mathcal{M}(\psi)$  and  $\mathcal{M}(\psi_0, \rho^p(u))$  are transversely cut out near  $u$  and  $u_0$ . Then there is a homeomorphism  $h$  between  $[0, 1)$  and a neighborhood of  $(u, u_0)$  in the compactified 1-dimensional moduli space*

$$\overline{\bigcup_T \mathcal{M}_{J(T)}(\psi \# \psi_0)}$$

*such that  $h(u, u_0) = \{0\}$ .*

Finally, the following three facts will also be useful in the proof of the triangle count of Proposition 8.1, so we state them here as lemmas for convenience in referencing.

**Lemma 8.15** [5, Lemma 9.50] *Consider the triple diagram  $\mathcal{T}_0 = (\Sigma_0, \alpha'_0, \alpha_0, \beta_0)$ . If  $\mathbf{x} \in \mathbb{T}_{\alpha'_0} \cap \mathbb{T}_{\alpha_0}$  and  $\psi_0 \in \pi_2(\mathbf{x}, \mathbf{a}, \mathbf{b})$ , then*

$$(14) \quad \mu(\psi_0) = 2n_{p_0}(\psi_0) + \mu(\mathbf{x}, \Theta).$$

**Lemma 8.16** *The differential on  $\widehat{CF}(\Sigma_0, \alpha'_0, \alpha_0, p_0, \circ_{\alpha'_0, \alpha_0})$ , defined with respect to the coherent orientation system  $\circ_{\alpha'_0, \alpha_0}$  specified in Lemma 6.5, vanishes.*

**Proof** By [11, Lemma 9.4],  $\text{rank}_{\mathbb{Z}}(\widehat{HF}(\Sigma_0, \alpha'_0, \alpha_0, p_0, \circ_{\alpha'_0, \alpha_0})) = 4$ . By inspection  $\text{rank}_{\mathbb{Z}}(\widehat{CF}) = 4$ , so the differential must vanish. □

**Lemma 8.17** *The map*

$$\Psi_{\beta_0}^{\alpha_0 \rightarrow \alpha'_0} : \widehat{CF}(\Sigma_0, \alpha_0, \beta_0, p_0) \rightarrow \widehat{CF}(\Sigma_0, \alpha'_0, \beta_0, p_0)$$

*satisfies  $\Psi_{\beta_0}^{\alpha_0 \rightarrow \alpha'_0}(\mathbf{a}) = \pm \mathbf{b}$ .*

**Proof** By Lemma 6.9,  $\Psi_{\beta_0}^{\alpha_0 \rightarrow \alpha'_0}$  is a quasi-isomorphism. Since the two complexes in question are trivial of rank one over  $\mathbb{Z}$ , the quasi-isomorphism must be an isomorphism between trivial, rank one complexes over  $\mathbb{Z}$ , of which there are precisely two. □

### 8.3 Counting triangles

We are now in position to prove the main triangle count, and conclude the proof of handleswap invariance.

**Proof of Proposition 8.1** As we did in Sections 6 and 7, we will consider the case of the chain complexes  $CF^-$  in what follows in order to fix definitions; however we note that the proof carries over equally well for all variants  $CF^\circ$ .

For an almost complex structure  $J$  which achieves transversality we have, by definition,

$$\mathcal{F}_{\mathcal{T}\#\mathcal{T}_0}((\mathbf{x} \times \Theta) \otimes (\mathbf{y} \times \mathbf{a})) = \sum_z \sum_{\substack{A \in \pi_2(\mathbf{x} \times \Theta, \mathbf{y} \times \mathbf{a}, \mathbf{z} \times \mathbf{b}) \\ \mu(A)=0}} (\#\mathcal{M}_J(A))U^{n_p(A)} \cdot \mathbf{z} \times \mathbf{b}$$

and

$$\mathcal{F}_{\mathcal{T}}(\mathbf{x} \otimes \mathbf{y}) \times \mathbf{b} = \left( \sum_z \sum_{\substack{A \in \pi_2(\mathbf{x}, \mathbf{y}, \mathbf{z}) \\ \mu(A)=0}} (\#\mathcal{M}_J(A))U^{n_p(A)} \cdot \mathbf{z} \right) \times \mathbf{b}.$$

To obtain the result we will count Maslov index 0 holomorphic triangles in the homology class  $A$ , for each generator  $\mathbf{z} \in \mathbb{T}_{\alpha'} \cap \mathbb{T}_{\beta}$  and class  $A \in \pi_2(\mathbf{x} \times \Theta, \mathbf{y} \times \mathbf{a}, \mathbf{z} \times \mathbf{b})$ .

Consider two homology classes of triangles  $\psi \in \pi_2(\mathbf{x}, \mathbf{y}, \mathbf{z})$  on  $\mathcal{T} = (\Sigma, \alpha', \alpha, \beta, p)$  and  $\psi_0 \in \pi_2(\Theta, \mathbf{a}, \mathbf{b})$  on  $\mathcal{T}_0 = (\Sigma_0, \alpha'_0, \alpha_0, \beta_0, p_0)$ . If  $n_p(\psi) = n_{p_0}(\psi_0)$ , so the classes match across the connect sum point, then the homology classes can be combined to give a class  $\psi \# \psi_0 \in \pi_2(\mathbf{x} \times \Theta, \mathbf{y} \times \mathbf{a}, \mathbf{z} \times \mathbf{b})$ . Conversely, it is clear that any class  $A \in \pi_2(\mathbf{x} \times \Theta, \mathbf{y} \times \mathbf{a}, \mathbf{z} \times \mathbf{b})$  can be written uniquely as a connect sum of suitable classes with this matching condition.

So for any such homology class  $A = \psi \# \psi_0$  with  $\mu(A) = 0$ , we aim to count Maslov index zero holomorphic representatives as we stretch the neck, ie to count  $\#\mathcal{M}_{J(T_i)}(\psi \# \psi_0)$ , where  $J(T_i)$  is a sequence of almost complex structures being stretched along the neck. To do so, suppose  $u_{T_i}$  is a sequence of  $J(T_i)$ -holomorphic triangles representing  $\psi \# \psi_0$ , where  $\mu(\psi \# \psi_0) = 0$ . We note here that by [17, Theorem 4.1] and Lemma 8.15 we have

$$\mu(\psi \# \psi_0) = \mu(\psi) + \mu(\psi_0) - 2n_p(\psi_0) = \mu(\psi) + \mu(\theta, \theta) = \mu(\psi).$$

Hence  $\mu(\psi) = 0$  and  $\mu(\psi_0) = 2n_{p_0}(\psi_0)$ .

By Proposition 8.6, there is a subsequence of  $u_{T_i}$  which converges to a triple  $(U, V, U_0)$  where  $U$  is a broken holomorphic triangle in  $\Sigma \times \Delta$  representing  $\psi$ ,  $U_0$  is a broken



holomorphic triangle in  $\Sigma_0 \times \Delta$  representing  $\psi_0$ , and  $V$  is a collection of holomorphic curves mapping into the neck regions that are asymptotic to (possibly multiply covered) Reeb orbits of the form  $S^1 \times \{d\}$ .

The proof will now proceed in steps as follows:

- (1) We will show  $U$  consists of a single holomorphic triangle  $u$  with Maslov index zero, with  $u$  satisfying (M1)–(M8), and potentially some number of constant holomorphic curves.
- (2) We then show that  $U_0$  consists of a single Maslov index  $2n_{p_0}(\psi_0)$  triangle  $u'_0$ , with  $u'_0$  satisfying (M1)–(M8) and  $\rho^p(u) = \rho^{p_0}(u_0)$ , and potentially some number of constant holomorphic curves.
- (3) We rule out the possibility of constant curves occurring in steps (1) and (2), and show that  $V$  consists of a collection of trivial holomorphic cylinders.
- (4) Using this knowledge of  $(U, V, U_0)$  and the gluing result, we reduce the proof to showing Lemma 8.18 below.

In fact, the proofs of steps (1)–(3) given in [5] carry over exactly as they are stated there, so we will only carry out step (4).

**Step (4)** By steps (1)–(3), a sequence  $u_{T_i}$  of  $J(T_i)$ –holomorphic triangles representing  $\psi \# \psi_0$  converges to a broken holomorphic triangle  $(U, V, U_0)$ , where  $U = u$  is a single holomorphic triangle satisfying  $\mu(u) = 0$ ,  $V$  is a collection of trivial holomorphic cylinders,  $U_0$  is a single holomorphic triangle  $u_0$  satisfying  $\mu(u_0) = 2n_p(\psi)$ , and  $\rho^p(u) = \rho^{p_0}(u_0)$ . By Proposition 8.14, there is therefore a homeomorphic identification  $h$  between a neighborhood of  $(u, u_0)$  in the compactified 1–dimensional moduli space

$$\overline{\bigcup_{T_i} \mathcal{M}_{J(T_i)}(\psi \# \psi_0)}$$

and the interval  $[0, 1)$ , such that  $h(u, u_0) = \{0\}$ . This yields an identification

$$\mathcal{M}_{J(T_i)}(\psi \# \psi_0) \cong \{(u, u_0) \in \mathcal{M}(\psi) \times \mathcal{M}(\psi_0) \mid \rho^p(u) = \rho^{p_0}(u_0)\}$$

for sufficiently large  $T_i$ . We now fix  $J_{T_i}$  for such a sufficiently large value of  $T_i$ , and drop this choice of almost complex structure from our notation.

Given coherent orientation systems  $\sigma_{\mathcal{T}}$  over  $\mathcal{T}$  and  $\sigma_{\mathcal{T}_0}$  over  $\mathcal{T}_0$ , there is a coherent orientation system  $\sigma_{\mathcal{T} \# \mathcal{T}_0}$  with respect to which the signed count of the 0–dimensional

moduli space  $\mathcal{M}(\psi \# \psi_0)$  is given by

$$\#\mathcal{M}(\psi \# \psi_0) = \#\{(u, u_0) \in \mathcal{M}(\psi) \times \mathcal{M}(\psi_0) \mid \rho^P(u) = \rho^P(u_0)\}.$$

Indeed, given two homology classes of triangles  $\psi$  on  $\mathcal{T}$  and  $\psi_0$  on  $\mathcal{T}_0$ , the gluing map  $\natural$  (see [6, Appendix A, page 1082] for the definition) used to identify the two moduli spaces is covered by a map of determinant lines  $(\natural)_{\#}$  which can be used to produce an orientation  $\sigma_{\mathcal{T}\#\mathcal{T}_0}^{\psi\#\psi_0}$  over  $\mathcal{M}(\psi \# \psi_0)$  from orientations  $\sigma_{\mathcal{T}}^{\psi}$  over  $\mathcal{M}(\psi)$  and  $\sigma_{\mathcal{T}_0}^{\psi_0}$  over  $\mathcal{M}(\psi_0)$ . Similarly, for two homology classes of strips  $A$  on  $\mathcal{T}$  and  $A_0$  on  $\mathcal{T}_0$ , the same procedure can be used to determine an orientation  $\sigma_{\mathcal{T}\#\mathcal{T}_0}^{A\#A_0}$  from  $\sigma_{\mathcal{T}}^A$  and  $\sigma_{\mathcal{T}_0}^{A_0}$ . The fact that homology classes of strips and triangles on  $\mathcal{T} \# \mathcal{T}_0$  are in bijective correspondence to pairs of homology classes of strips on  $\mathcal{T}$  and  $\mathcal{T}_0$  ensures that the coherent orientation systems  $\sigma_{\mathcal{T}}$  and  $\sigma_{\mathcal{T}_0}$  thus determine a single orientation system  $\sigma_{\mathcal{T}\#\mathcal{T}_0}$  over all classes of strips and triangles in the connect summed diagram (ie the determinations for a particular class of triangle or strip on the summed diagram are not overspecified). That this induced orientation is coherent follows from the coherence of the two constituent orientations, along with the fact that gluing map  $(\natural)_{\#}$  above commutes with the map  $(\natural)_*$  appearing in Definition 8.12. More precisely, the coherence follows from these facts as

$$\begin{aligned} \sigma_{\mathcal{T}\#\mathcal{T}_0}^{(\psi+A)\#(\psi_0+A_0)} &:= (\natural)_{\#}(\sigma_{\mathcal{T}}^{\psi+A} \times \sigma_{\mathcal{T}_0}^{\psi_0+A_0}) \\ &= (\natural)_{\#}((\natural)_*(\sigma_{\mathcal{T}}^{\psi} \times \sigma_{\mathcal{T}}^A) \times (\natural)_*(\sigma_{\mathcal{T}_0}^{\psi_0} \times \sigma_{\mathcal{T}_0}^{A_0})) \\ &= (\natural)_*((\natural)_{\#}(\sigma_{\mathcal{T}}^{\psi} \times \sigma_{\mathcal{T}_0}^{\psi_0}) \times (\natural)_{\#}(\sigma_{\mathcal{T}}^A \times \sigma_{\mathcal{T}_0}^{A_0})) \\ &=: (\natural)_*(\sigma_{\mathcal{T}\#\mathcal{T}_0}^{\psi\#\psi_0} \times \sigma_{\mathcal{T}\#\mathcal{T}_0}^{A\#A_0}) \end{aligned}$$

where the second equality is the definition of coherence for the orientation systems  $\sigma_{\mathcal{T}}$  and  $\sigma_{\mathcal{T}_0}$ , and the third equality is the statement of the commutativity of the two induced gluing maps referenced above. This commutativity follows from the fact that the two gluing maps can be viewed as taking place in a small neighborhood of the curves being glued, and can thus be performed in either order, or simultaneously, via the construction in [6, Appendix A]. This establishes coherence of the system  $\sigma_{\mathcal{T}\#\mathcal{T}_0}$ .

For  $u \in \mathcal{M}(\psi)$  let

$$\mathcal{M}_{(\Theta, \mathbf{a}, \mathbf{b})}(\rho^P(u)) = \coprod_{\substack{\psi_0 \in \pi_2(\Theta, \mathbf{a}, \mathbf{b}) \\ \mu(\psi_0) = 2n_p(\psi)}} \mathcal{M}(\psi_0, \rho^P(u)).$$

With respect to a coherent orientation system  $\sigma_{\mathcal{T}\#\mathcal{T}_0}$  on  $\mathcal{T} \# \mathcal{T}_0$  determined from any coherent systems  $\sigma_{\mathcal{T}}$  and  $\sigma_{\mathcal{T}_0}$  as above, the triangle map in question can then be written

as

$$\begin{aligned}
 \mathcal{F} &= \mathcal{F}_{\mathcal{T}\#\mathcal{T}_0}((\mathbf{x} \times \Theta) \otimes (\mathbf{y} \times \mathbf{a})) \\
 &= \sum_z \sum_{\substack{\psi \in \pi_2(\mathbf{x}, \mathbf{y}, z) \\ \psi_0 \in \pi_2(\Theta, \mathbf{a}, \mathbf{b}) \\ \mu(\psi \# \psi_0) = 0}} \#\{(u, u_0) \in \mathcal{M}(\psi) \times \mathcal{M}(\psi_0) \mid \rho^p(u) = \rho^p(u_0)\} U^{n_p(\psi \# \psi_0)} \cdot \mathbf{z} \times \mathbf{b} \\
 &= \sum_z \sum_{\substack{\mu(\psi) = 0 \\ \psi \in \pi_2(\mathbf{x}, \mathbf{y}, z)}} \sum_{\substack{\psi_0 \in \pi_2(\Theta, \mathbf{a}, \mathbf{b}) \\ \mu(\psi_0) = 2n_p(\psi)}} \#\{(u, u_0) \in \mathcal{M}(\psi) \times \mathcal{M}(\psi_0) \mid \rho^p(u) = \rho^p(u_0)\} U^{n_p(\psi \# \psi_0)} \cdot \mathbf{z} \times \mathbf{b} \\
 &= \sum_z \sum_{\substack{\psi \in \pi_2(\mathbf{x}, \mathbf{y}, z) \\ \mu(\psi) = 0}} \sum_{\substack{\psi_0 \in \pi_2(\Theta, \mathbf{a}, \mathbf{b}) \\ \mu(\psi_0) = 2n_p(\psi)}} \sum_{u \in \mathcal{M}(\psi)} \#(u \times \mathcal{M}(\psi_0, \rho^p(u))) U^{n_p(\psi \# \psi_0)} \cdot \mathbf{z} \times \mathbf{b} \\
 &= \sum_z \sum_{\substack{\psi \in \pi_2(\mathbf{x}, \mathbf{y}, z) \\ \mu(\psi) = 0}} \sum_{u \in \mathcal{M}(\psi)} \#(u \times \mathcal{M}(\Theta, \mathbf{a}, \mathbf{b})(\rho^p(u))) U^{n_p(\psi \# \psi_0)} \cdot \mathbf{z} \times \mathbf{b}.
 \end{aligned}$$

We will show in Lemma 8.18 below that there is a coherent orientation system  $\sigma_{\mathcal{T}_0}$  on  $\mathcal{T}_0$  for which either

$$\#\mathcal{M}_{(\Theta, \mathbf{a}, \mathbf{b})}(\rho^p(u)) = 1$$

for all  $\psi$  with  $\mu(\psi) = 0$  and all  $u \in \mathcal{M}(\psi)$ , or

$$\#\mathcal{M}_{(\Theta, \mathbf{a}, \mathbf{b})}(\rho^p(u)) = -1$$

for all  $\psi$  with  $\mu(\psi) = 0$  and all  $u \in \mathcal{M}(\psi)$ . Then we will have

$$\begin{aligned}
 \mathcal{F} &= \mathcal{F}_{\mathcal{T}\#\mathcal{T}_0}^-(\mathbf{x} \times \Theta) \otimes (\mathbf{y} \times \mathbf{a}) \\
 &= \sum_z \sum_{\substack{\psi \in \pi_2(\mathbf{x}, \mathbf{y}, z) \\ \mu(\psi) = 0}} \sum_{u \in \mathcal{M}(\psi)} \#(u \times \mathcal{M}_{(\Theta, \mathbf{a}, \mathbf{b})}(\rho^p(u))) U^{n_p(\psi \# \psi_0)} \cdot \mathbf{z} \times \mathbf{b} \\
 &= \pm \sum_z \sum_{\substack{\psi \in \pi_2(\mathbf{x}, \mathbf{y}, z) \\ \mu(\psi) = 0}} \#\mathcal{M}(\psi) U^{n_p(\psi \# \psi_0)} \cdot \mathbf{z} \times \mathbf{b} \\
 &= \pm \left( \sum_z \sum_{\substack{\psi \in \pi_2(\mathbf{x}, \mathbf{y}, z) \\ \mu(\psi) = 0}} (\#\mathcal{M}(\psi)) U^{n_p(\psi)} \cdot \mathbf{z} \right) \times \mathbf{b} \\
 &= \pm \mathcal{F}_{\mathcal{T}}^-(\mathbf{x} \otimes \mathbf{y}) \times \mathbf{b}.
 \end{aligned}$$

This completes the proof of the proposition, modulo Lemma 8.18. □

**Lemma 8.18** For  $\mathbf{d} \in \text{Sym}^k(\Delta) \setminus \text{Diag}(\Delta)$  and a generic choice of almost complex structure  $J$ , the moduli space  $\mathcal{M}_{(\Theta, \mathbf{a}, \mathbf{b})}(\mathbf{d})$  is a smoothly cut out 0–manifold. For

such  $J$ , there is a coherent orientation system  $\sigma_{\tau_0}$  on  $\mathcal{T}_0$  for which the signed count of points in the moduli space is

$$\#\mathcal{M}_{(\Theta, a, b)}(\mathbf{d}) = \pm 1$$

where the constant is independent of  $\mathbf{d}$ .

**Proof** The proof is again carried out in steps:

- (1) We show the moduli space is transversely cut out for generic  $J$ .
- (2) We show that for generic  $\mathbf{d} \in \text{Sym}^k(\Delta) \setminus \text{Diag}(\Delta)$ , the signed count  $\#\mathcal{M}_{(\Theta, a, b)}(\mathbf{d})$  is independent of  $\mathbf{d}$ .
- (3) We find one choice of  $\mathbf{d}$  giving the desired count.

In fact, the proof of step (1) given in [5] carries over exactly as it is stated there, so we will only prove steps (2) and (3).

**Step (2)** Let  $p: I \rightarrow \text{Sym}^k(\Delta)$  be a path from  $\mathbf{d}_0$  to  $\mathbf{d}_1$ , where the image of  $p$  satisfies the conditions of Proposition 8.9. We consider the moduli space

$$\bigcup_{t \in I} \mathcal{M}_{(\Theta, a, b)}(p(t))$$

which by Proposition 8.9 and Lemma 8.10 is a smooth, orientable 1–manifold. From orientability, we know that the signed count of the ends of the moduli space above is zero. We now describe all contributions to the count of the ends. We begin by making considerations which will hold for any choice of coherent orientation system satisfying the property appearing in Lemma 8.13.

The ends of  $\bigcup_{t \in I} \mathcal{M}_{(\Theta, a, b)}(p(t))$  fall into three classes. They arise from  $\mathcal{M}_{(\Theta, a, b)}(\mathbf{d}_0)$ ,  $\mathcal{M}_{(\Theta, a, b)}(\mathbf{d}_1)$ , and degenerations of holomorphic triangles to broken holomorphic triangles in the compactification. Let  $u_i: S_0 \rightarrow \Sigma_0 \times \Delta$  be a sequence of holomorphic triangles in  $\bigcup_{t \in I} \mathcal{M}_{(\Theta, a, b)}(p(t))$ . As shown in [5, Lemma 9.58], the only degenerations that can occur correspond to “strip breaking”. In particular, if  $u_i$  converges to a broken holomorphic triangle

$$U = (u_1, v_1, \dots, v_n, w_1, \dots, w_m)$$

(in the sense of Definition 8.5), then in fact  $U = (u_1, v_1, \dots, v_n)$  where the  $v_i$  are holomorphic strips. We note that the argument used to rule out other types of degenerations has nothing to do with orientations. Furthermore, we will see presently that among

degenerations corresponding to strip breaking, the only ones which can occur yield broken triangles  $U$  consisting of a triangle  $u_1$  of index  $2k - 1$  which matches a divisor  $p(t)$  for some  $t \in I$ , as well as a single curve  $v_1: S \rightarrow \Sigma_0 \times I \times \mathbb{R}$  with index 1.

To see this, note that if  $U$  is genuinely broken then  $U = (u_1, v_1, \dots, v_n)$  with  $u_1$  a holomorphic triangle representing a class in  $\pi_2(x, \mathbf{a}, \mathbf{b})$  and  $v_i$  holomorphic curves in  $\pi_2(\mathbf{y}_i, \mathbf{z}_i)$  for some  $\mathbf{y}_i, \mathbf{z}_i \in \mathbb{T}_{\alpha'} \cap \mathbb{T}_{\alpha}$ . We now analyze what contributions to the ends can occur for the four possible intersection points  $\mathbf{x} \in \mathbb{T}_{\alpha'} \cap \mathbb{T}_{\alpha}$ .

Suppose  $\mathbf{x} = \Theta$ . Then by applying Lemma 8.15 to  $u_1$  we obtain  $\mu(u_1) = 2n_{p_0}(u_1)$ . Since  $u_1$  satisfies a matching condition with  $p(t)$  for some  $t \in I$ ,

$$2n_{p_0}(u_1) = |\rho^P(p(t))| = k = 2n_{p_0}(\psi_0) = \mu(\psi_0).$$

Thus  $\mu(u_1) = \mu(\psi_0)$ . Since the total homology class of  $U$  must be  $\psi_0$ , we therefore must have  $\mu(v_i) = 0$  and  $n_{p_0}(v_i) = 0$  for all  $i$ . Since the  $v_i$  satisfy (M1) and (M3)–(M6), the only possibility for such curves is that each is a collection of constant components. Indeed, if any  $v_i$  were locally nonconstant, it would satisfy (M2); hence, by [5, Corollary 7.2], the dimension of the relevant moduli space containing it would be negative. Thus  $U = (u_1)$  (plus potentially some constant curves) is in the interior of  $\bigcup_{t \in I} \mathcal{M}_{(\Theta, \mathbf{a}, \mathbf{b})}(p(t))$ , and so contributes nothing to the signed count of the ends.

Next we consider the cases  $\mathbf{x} = \theta_1^+ \theta_2^-, \theta_1^- \theta_2^+$ . In these cases Lemma 8.15 yields that the index of the triangle must be  $\mu(u_1) = 2n_{p_0}(u_1) - 1 = 2n_{p_0}(\psi_0) - 1$ , so the remaining curves must have indices which sum to 1. Similarly,

$$0 = n_{p_0}(\psi_0) - n_{p_0}(u_1) = \sum_i n_{p_0}(v_i),$$

so  $v_i$  must have multiplicity 0 at the basepoint for each  $i$ . The only possibility in this case is that there is a single Maslov index 1 strip  $v_1$ . Thus in this case, we have additional contributions to the ends coming from

$$\bigcup_{\substack{t \in I \\ \mathbf{x} \in \{\theta_1^+ \theta_2^-, \theta_1^- \theta_2^+\}}} \bigcup_{\substack{\phi \in \pi_2(\Theta, \mathbf{x}) \\ n_{p_0}(\phi) = 0}} \mathcal{M}_{(\mathbf{x}, \mathbf{a}, \mathbf{b})}(p(t)) \times \widehat{\mathcal{M}}(\phi).$$

Fix  $\mathbf{x} \in \{\theta_1^+ \theta_2^-, \theta_1^- \theta_2^+\}$ . Then by Lemma 8.16 we know that

$$\sum_{\substack{\phi \in \pi_2(\Theta, \mathbf{x}) \\ n_{p_0}(\phi) = 0}} \#\widehat{\mathcal{M}}(\phi) = 0.$$

Thus

$$\begin{aligned}
 & \# \left( \bigcup_{\substack{t \in I \\ \mathbf{x} \in \{\theta_1^+ \theta_2^-, \theta_1^- \theta_2^+\}}} \bigcup_{\substack{\phi \in \pi_2(\Theta, \mathbf{x}) \\ n_{p_0}(\phi) = 0}} \mathcal{M}_{(\mathbf{x}, \mathbf{a}, \mathbf{b})}(p(t)) \times \widehat{\mathcal{M}}(\phi) \right) \\
 &= \sum_{\substack{t \in I \\ \mathbf{x} \in \{\theta_1^+ \theta_2^-, \theta_1^- \theta_2^+\}}} \sum_{\substack{\phi \in \pi_2(\Theta, \mathbf{x}) \\ n_{p_0}(\phi) = 0}} \#(\mathcal{M}_{(\mathbf{x}, \mathbf{a}, \mathbf{b})}(p(t)) \times \widehat{\mathcal{M}}(\phi)) \\
 &= \sum_{\substack{t \in I \\ \mathbf{x} \in \{\theta_1^+ \theta_2^-, \theta_1^- \theta_2^+\}}} \sum_{\substack{\phi \in \pi_2(\Theta, \mathbf{x}) \\ n_{p_0}(\phi) = 0}} (\#\mathcal{M}_{(\mathbf{x}, \mathbf{a}, \mathbf{b})}(p(t))) \cdot (\#\widehat{\mathcal{M}}(\phi)) \\
 &= 0.
 \end{aligned}$$

Here we have used in the last equality the fact that we have endowed the orientable manifold  $\bigcup_{t \in I} \mathcal{M}_{(\Theta, \mathbf{a}, \mathbf{b})}(p(t))$  with some coherent orientation system. This implies in particular that the orientation induced on the compactification agrees with the product orientation at ends such as those above. So we see these cases also contribute nothing to the count of signed ends of the moduli space.

Lastly, we consider the case  $\mathbf{x} = \theta_1^- \theta_2^-$ . For any  $\psi_0 \in \pi_2(\theta_1^- \theta_2^-, \mathbf{a}, \mathbf{b})$  we have by Lemma 8.15  $\mu(\psi_0) = 2n_{p_0}(\psi_0) - 2 = 2k - 2$ . By Proposition 8.9, for a generic choice of almost complex structure  $J$ , and a fixed source  $S$ , the matched moduli space  $\mathcal{M}(\psi_0, S, p(I))$  is a smooth manifold of dimension

$$\text{ind}(\psi_0, S) - \text{codim}(p(I)) = \text{ind}(\psi_0, S) - (2k - 1) \leq \mu(\psi_0) - (2k - 1) = -1.$$

Here the fact being used to establish the inequality is that for any holomorphic triangle  $u$  in the homology class  $A$  (not necessarily embedded), the index of the linearized  $\bar{\partial}$  operator at  $u$  satisfies  $\text{ind}(A, S) = \mu(A) - 2\text{sing}(u)$ , and in particular  $\text{ind}(A, S) \leq \mu(A)$ . This is [5, equation 9.46], which comes from adapting [7, Proposition 5.69]. This shows that for a generic choice of  $J$ , the broken triangle  $U$  can not in fact contain a triangle  $u_1$  in such a class  $\psi_0$ .

To summarize, we have shown that the ends of  $\bigcup_{t \in I} \mathcal{M}_{(\Theta, \mathbf{a}, \mathbf{b})}(p(t))$  correspond to  $\mathcal{M}_{(\Theta, \mathbf{a}, \mathbf{b})}(\mathbf{d}_0)$ ,  $\mathcal{M}_{(\Theta, \mathbf{a}, \mathbf{b})}(\mathbf{d}_1)$ , and to degenerations of triangles into broken triangles containing one triangle and one strip, and that the last types of ends contribute nothing to the total signed count of the ends. Since we have chosen a collection of orientation systems satisfying the conclusion of Lemma 8.13, we see that the signed count of the ends of  $\bigcup_{t \in I} \mathcal{M}_{(\Theta, \mathbf{a}, \mathbf{b})}(p(t))$  is given by

$$\#\mathcal{M}_{(\Theta, \mathbf{a}, \mathbf{b})}(\mathbf{d}_1) - \#\mathcal{M}_{(\Theta, \mathbf{a}, \mathbf{b})}(\mathbf{d}_0) = 0.$$

This concludes step (2).

We note that by Lemma 8.13, a coherent orientation system on  $\mathcal{M}_{(\Theta, \mathbf{a}, \mathbf{b})}(p(0))$  induces a coherent orientation system over  $\bigcup_{t \in I} \mathcal{M}_{(\Theta, \mathbf{a}, \mathbf{b})}(p(t))$  and  $\mathcal{M}_{(\Theta, \mathbf{a}, \mathbf{b})}(p(1))$  satisfying the conclusion of the lemma. We thus see that if we can find a single divisor  $\mathbf{d}$  and a coherent orientation system  $\sigma$  over  $\mathcal{M}_{(\Theta, \mathbf{a}, \mathbf{b})}(\mathbf{d})$  giving the desired count, then the argument of step (2) shows that there are induced coherent orientations over all divisors  $\mathbf{d}'$  in the same path component as  $\mathbf{d}$  for which the counts are the same. We will construct such a divisor in step (3) below.

**Step (3)** To construct a divisor  $\mathbf{d} \in \text{Sym}^k(\Delta) \setminus \text{Diag}(\Delta)$  giving the desired count, we consider a path of divisors subject to constraints, and evaluate the asymptotics of the moduli spaces of triangles matched to divisors in this path. Our argument is an explication of that in [5], which is in turn based on an analogous argument in [15, page 653] which deals with holomorphic strips. Our goal in summarizing these proofs is to make explicit the dependence of all statements on signs and orientations.

We consider any path  $p: [1, \infty) \rightarrow \text{Sym}^k(\Delta) \setminus \text{Diag}(\Delta)$  for which each point in  $p(t)$  is at least a distance of  $t$  away from all other points in  $p(t)$ , with respect to a metric on  $\Delta$  for which the corners are infinite strips in  $\mathbb{C}$ ; see Figure 8. We further require that the points in  $p(t)$  smoothly approach the vertex  $v_{\alpha_0 \beta_0}$  of  $\Delta$  as  $t \rightarrow \infty$ . For such a path of divisors, we have as before a matched moduli space

$$\mathcal{M}_{(\Theta, \mathbf{a}, \mathbf{b})}(p) = \bigcup_{t \in [1, \infty]} \mathcal{M}_{(\Theta, \mathbf{a}, \mathbf{b})}(p(t)).$$

By the same arguments used in step 2, the ends of this moduli space corresponding to degenerations of triangles at finite values of  $t$ , with  $t \neq 1$ , will contribute nothing to the signed count of the ends, for any choice of coherent orientation system. Consider any coherent orientation system  $\sigma$  satisfying the properties of that furnished by Lemma 8.13; then with respect to such an orientation system the signed count  $\#\mathcal{M}_{(\Theta, \mathbf{a}, \mathbf{b})}(p(1))$  must agree with the signed count of the ends of  $\mathcal{M}_{(\Theta, \mathbf{a}, \mathbf{b})}(p)$  coming from degenerations of triangles as  $t \rightarrow \infty$ . So we now count these ends.

We claim that as  $t \rightarrow \infty$ , the only broken triangles which can occur in the limit consist of a single genuine triangle  $\tau$  of index 0 on  $(\Sigma_0, \alpha'_0, \alpha_0, \beta_0)$ , along with  $k$  index 2 curves on  $(\Sigma_0, \alpha_0, \beta_0)$  which satisfy matching conditions with some collection of divisors  $c_i \in [0, 1] \times \mathbb{R}$ . To see this, we note that each point in the path  $p$  consists of  $k$  distinct points in  $\Delta$ , and the fact that these  $k$  points separate and approach the vertex  $v_{\alpha_0 \beta_0}$  in the limit necessitates that the limiting broken triangle must contain  $k$  strips

satisfying matching conditions. To see the index of each of these curves must be 2, we make some simple observations about the diagram  $(\Sigma_0, \alpha_0, \beta_0)$  for  $S^3$ .

First, note that the only homology classes of disks supporting holomorphic representatives are  $\{e_a + s[\Sigma_0]\}$  for nonnegative integers  $s$ , where  $e_a$  is the constant disk at  $a$ . The Maslov indices for such classes are  $\mu(e_a + s[\Sigma_0]) = 2s$ . The fact that each strip satisfies a matching condition implies we must have  $s \geq 1$  for each homology class. Since the total index of each holomorphic triangle in the moduli space  $\mathcal{M}_{(\Theta, a, b)}(p)$  is  $2k$ , the limiting broken holomorphic triangle must have index  $2k$ , so the only possibility is that each of the  $k$  curves has index 2 (ie has  $s = 1$ ), and the triangle  $\tau$  has index 0. By counting multiplicities and noting positivity of intersections, we see that the triangle  $\tau$  must satisfy  $n_{p_0}(\tau) = 0$ . Using the same arguments as in the preceding proposition, we have that all of the curves in the broken triangle must satisfy (M1)–(M8).

Applying the gluing result of Lipshitz [6, Proposition A.1], we see that we can obtain the signed count of the ends occurring as degenerations as  $t \rightarrow \infty$ , or equivalently the count  $\#\mathcal{M}_{(\Theta, a, b)}(p(1))$ , as

$$\#\mathcal{M}_{(\Theta, a, b)}(p(1)) = (\#\mathcal{M}_{(a, a)}(c))^k \cdot \sum_{\substack{\psi \in \pi_2(\Theta, a, b) \\ n_{p_0}(\psi) = 0}} \#\mathcal{M}(\psi),$$

where  $c$  is a divisor in  $[0, 1] \times \mathbb{R}$  and  $\mathcal{M}_{(a, a)}(c)$  is the moduli space of index 2 strips on  $(\Sigma_0, \alpha_0, \beta_0)$  with  $\rho^p(u) = c$ . Here the counts are occurring with respect to any coherent orientation system  $\sigma_{\mathcal{T}_0} = \{\sigma_{\alpha'_0, \alpha_0, \beta_0}, \sigma_{\alpha_0, \beta_0}, \sigma_{\alpha'_0, \alpha_0}, \sigma_{\alpha'_0, \beta_0}\}$  on  $\mathcal{T}_0$  and the compatible orientation system  $\sigma_{\alpha_0, \beta_0}$  included in the data  $\sigma_{\mathcal{T}_0}$ . The sum on the right hand side is precisely the count occurring in the triangle map in Lemma 8.17, and is thus  $\pm 1$ . Thus to finish this step it suffices to show that there is a coherent orientation system  $\sigma_{\mathcal{T}_0}$  for which

$$\#\mathcal{M}_{(a, a)}(c) = \pm 1.$$

Consider the standard diagram  $\mathcal{H}_{S^1 \times S^2}$  for  $S^1 \times S^2$ , twice stabilized via the diagram  $(\Sigma_0, \alpha_0, \beta_0)$  as shown in Figure 10. The figure depicts this genus 3 diagram for  $S^1 \times S^2$ , along with a choice of basepoint  $z$ . Both bigons in  $\mathcal{H}_{S^1 \times S^2}$  for  $S^1 \times S^2$  admit a single holomorphic representative. We consider a choice of coherent orientation system on  $\mathcal{H}_{S^1 \times S^2}$  for which the bigons cancel, and the resulting Floer homology is  $\widehat{HF} \cong \mathbb{Z}^2$ . By invariance of  $\widehat{HF}$ , the twice stabilized bigon in the twice stabilized diagram must also have a single holomorphic representative. As in the proof of stabilization invariance in [6], this implies via a neck stretching argument that there is a coherent orientation



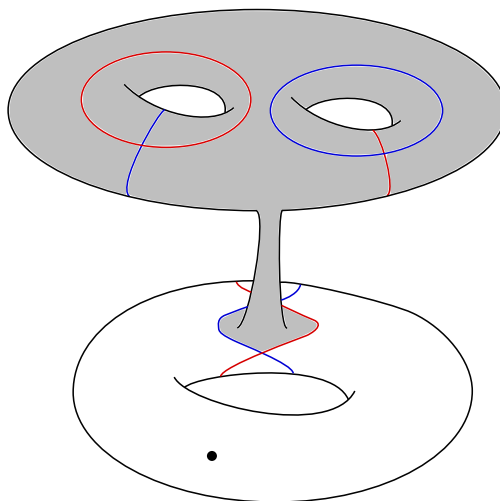


Figure 10: The diagram  $\mathcal{H}_{S^1 \times S^2}$  on the bottom of the figure is twice stabilized via a connect sum with  $(\Sigma_0, \alpha_0, \beta_0)$ . Shaded in gray is a domain on the genus 3 diagram, the “twice stabilized bigon”, which arises from one of the bigons in  $\mathcal{H}_{S^1 \times S^2}$ .

system  $\sigma_{\alpha_0, \beta_0}$  on  $(\Sigma_0, \alpha_0, \beta_0)$  for which

$$\#\mathcal{M}_{(a, a)}(c) = \pm 1.$$

By [11, Lemma 8.7], this coherent orientation system can be extended to a coherent orientation system  $\sigma_{\mathcal{T}_0}$  for which the same condition holds. This completes step (3), and the proof of the lemma.  $\square$

## 9 The surgery exact triangle

In this section, we provide a brief explanation of how our results fit into the construction of the surgery exact sequence defined by Ozsváth and Szabó in [10, Section 9]. The fact that these constructions are compatible with ours turns out to be a matter of bookkeeping. We provide a sketch of the argument here with the hope that it will be useful in extending our naturality results to results about general cobordisms.

First, we recall one version of the construction of the surgery exact triangle and its relation to our naturality results. The relation between other versions of the statement of the exact triangle and our naturality results follows analogously. Let  $Y$  be a closed oriented 3-manifold, and  $K \hookrightarrow Y$  be a knot with a longitude  $\lambda$  and meridian  $\mu$ . We

denote by  $Y_0$  the 3–manifold obtained by performing  $\lambda$ –surgery on  $Y$ , and by  $Y_1$  the 3–manifold obtained by performing  $(\lambda + \mu)$ –surgery on  $Y$ . Call any such triple  $(Y, Y_0, Y_1)$  of 3–manifolds a *triad*. Ozsváth and Szabó showed:

**Theorem 9.1** [10, Theorem 9.12] *For any triad  $(Y, Y_0, Y_1)$  there are long exact sequences of  $\mathbb{Z}[U]$ –modules*

$$\begin{array}{ccc}
 HF^+(Y) & \xrightarrow{F} & HF^+(Y_0) \\
 & \swarrow F_1 & \searrow F_0 \\
 & HF^+(Y_1) &
 \end{array}
 \qquad
 \begin{array}{ccc}
 \widehat{HF}(Y) & \xrightarrow{\widehat{F}} & \widehat{HF}(Y_0) \\
 & \swarrow \widehat{F}_1 & \searrow \widehat{F}_0 \\
 & \widehat{HF}(Y_1) &
 \end{array}$$

The statement above is established via a corresponding statement made at the level of diagrams. To describe it, we recall a particular class of diagrams representing the manifolds in such a triad  $(Y, Y_0, Y_1)$ . Let  $(\mathcal{H}, \mathcal{H}_0, \mathcal{H}_1)$  be a tuple of diagrams for the 3–manifolds  $(Y, Y_0, Y_1)$  respectively. We will say the tuple of diagrams is *subordinate to the surgery triad* if there is a pointed genus  $g$  Heegaard quadruple diagram  $(\Sigma, \alpha, \beta, \gamma, \delta, z)$  satisfying the following properties:

- The diagrams  $(\Sigma, \alpha, \beta)$ ,  $(\Sigma, \alpha, \gamma)$  and  $(\Sigma, \alpha, \delta)$  represent  $Y$ ,  $Y_0$  and  $Y_1$  respectively.
- For  $i \neq g$ ,  $\beta_i$ ,  $\gamma_i$  and  $\delta_i$  are isotopic translates of one another, each intersecting transversally in two points.
- $\gamma_g$  is isotopic to the juxtaposition of  $\delta_g$  and  $\beta_g$  (see [10, Figure 9] for a depiction of juxtaposition).
- Every multiperiodic domain on the quadruple diagram has positive and negative coefficients.

Existence of such subordinate diagrams was established by Ozsváth and Szabó:

**Lemma 9.2** [10, Lemma 9.2] *Given a triad  $(Y, Y_0, Y_1)$ , there is a tuple of Heegaard diagrams  $(\mathcal{H}, \mathcal{H}_0, \mathcal{H}_1)$  subordinate to the triad.*

Theorem 9.1 is then to be interpreted as a compact way of phrasing the following statement at the level of diagrams.

**Theorem 9.3** *Let  $(\mathcal{H}, \mathcal{H}_0, \mathcal{H}_1)$  be a tuple subordinate to the triad  $(Y, Y_0, Y_1)$ , and fix a coherent orientation system  $\sigma_{\mathcal{H}_0}$  on  $\mathcal{H}_0$ . Then there are coherent orientation systems on  $\mathcal{H}$  and  $\mathcal{H}_1$ , and maps  $F, F_0$  and  $F_1$  induced by triangle counts as above, such that,*

with respect to the chosen coherent orientation systems, there are exact triangles

$$\begin{array}{ccc}
 HF^+(\mathcal{H}) & \xrightarrow{F} & HF^+(\mathcal{H}_0) \\
 \swarrow F_1 & & \swarrow F_0 \\
 & HF^+(\mathcal{H}_1) &
 \end{array}
 \qquad
 \begin{array}{ccc}
 \widehat{HF}(\mathcal{H}) & \xrightarrow{\widehat{F}} & \widehat{HF}(\mathcal{H}_0) \\
 \swarrow \widehat{F}_1 & & \swarrow \widehat{F}_0 \\
 & \widehat{HF}(\mathcal{H}_1) &
 \end{array}$$

We now use this restatement of the theorem at the level of diagrams to show that the surgery exact triangle is also well defined with respect to the transitive systems specified by Definition 6.14 and Theorem 1.1.

Recall that Definition 6.14 and Theorem 1.1 describe the four variants of Heegaard Floer homology as functors

$$HF^\circ : \text{Man}_* \rightarrow \text{Trans}(P(\mathbb{Z}[U] - \text{Mod})).$$

The precise restatement of the surgery exact sequence that we wish to establish in this context is just that the exact sequence defined by Ozsváth and Szabó extends to an exact sequence at the level of the transitive systems associated to a triad  $(Y, Y_0, Y_1)$ . Given transitive systems  $T, T_0$  and  $T_1$  and morphisms of transitive systems  $F : T \rightarrow T_0$ ,  $F_0 : T_0 \rightarrow T_1$  and  $F_1 : T_1 \rightarrow T$ , we will say the morphisms form an *exact sequence of transitive systems* if the morphisms restricted to constituent objects form exact sequences. We then have:

**Corollary 9.4** *For any triad  $(Y, Y_0, Y_1)$ , there are exact sequences of transitive systems*

$$\begin{array}{ccc}
 HF^+(Y) & \xrightarrow{F} & HF^+(Y_0) \\
 \swarrow F_1 & & \swarrow F_0 \\
 & HF^+(Y_1) &
 \end{array}
 \qquad
 \begin{array}{ccc}
 \widehat{HF}(Y) & \xrightarrow{\widehat{F}} & \widehat{HF}(Y_0) \\
 \swarrow \widehat{F}_1 & & \swarrow \widehat{F}_0 \\
 & \widehat{HF}(Y_1) &
 \end{array}$$

**Proof** Fix a triad  $(Y, Y_0, Y_1)$ , and a tuple  $(\mathcal{H}, \mathcal{H}_0, \mathcal{H}_1)$  of diagrams subordinate to the triad; such a subordinate tuple exists by Lemma 9.2. By Theorem 9.3, applying  $HF^+$  and  $\widehat{HF}$  to the diagrams in this tuple yields long exact triangles relating the  $\mathbb{Z}[U]$ -modules associated to the diagrams. Note that Theorem 9.3 ensures this statement is true with respect to any choice of coherent orientation system over  $\mathcal{H}_0$ , and the coherent orientations it induces on  $\mathcal{H}$  and  $\mathcal{H}_1$  via the triangle maps  $F_0$  and  $F_1$ . For the remainder of the proof, we fix coherent orientations  $(\sigma_{\mathcal{H}}, \sigma_{\mathcal{H}_0}, \sigma_{\mathcal{H}_1})$  on  $(\mathcal{H}, \mathcal{H}_0, \mathcal{H}_1)$  which are related in this way.

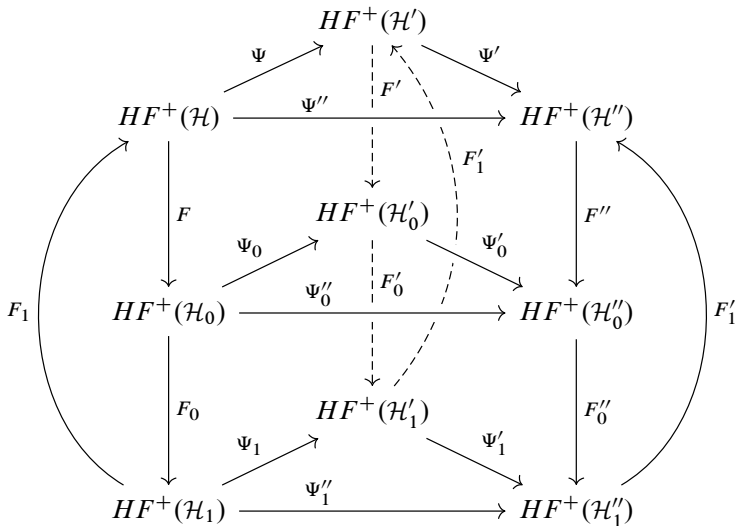


Figure 11: A depiction of the diagrams involved in the proof of Corollary 9.4.

Notice that by their definition, the transitive systems  $HF^\circ(Y)$ ,  $HF^\circ(Y_0)$  and  $HF^\circ(Y_1)$  contain as constituent objects the modules  $HF^\circ(\mathcal{H})$ ,  $HF^\circ(\mathcal{H}_0)$  and  $HF^\circ(\mathcal{H}_1)$ . Thus the exact triangle associated to this tuple of diagrams in Theorem 9.3 begins to partially define an exact sequence between the transitive systems. The situation is depicted in the leftmost column of Figure 11.

It is easy to extend this partially defined triangle of maps between transitive systems to a (more) partially defined triangle, defined now on all objects which correspond to triples of diagrams subordinate to  $(Y, Y_0, Y_1)$ . Given two tuples of diagrams  $(\mathcal{H}, \mathcal{H}_0, \mathcal{H}_1)$  and  $(\mathcal{H}'', \mathcal{H}''_0, \mathcal{H}''_1)$  subordinate to  $(Y, Y_0, Y_1)$ , and equivalences  $(\Psi'', \Psi''_0, \Psi''_1)$  induced by Heegaard moves relating the two tuples of diagrams, the maps  $F_i$  and  $F''_i$  appearing in the respective exact triangles commute (up to sign) with the equivalence maps by [14, Theorem 4.4]. See Figure 11 for a depiction of the situation. In other words, the surgery triangle immediately extends by this result to a partially defined triangle of transitive systems, which is now defined on all diagrams occurring in a subordinate tuple. This can be described in the diagram of Figure 11 by saying that the front square faces in the diagram commute (up to sign).

The final thing that remains to be shown is that these partially defined morphisms of transitive systems can be extended to maps defined on the Heegaard Floer modules associated to any admissible diagram, while preserving the consistency required of a morphism of transitive systems. With respect to the notation in Figure 11, this can be

phrased as asking for maps

$$\begin{aligned} F' &: HF^+(\mathcal{H}') \rightarrow HF^+(\mathcal{H}'_0), \\ F'_0 &: HF^+(\mathcal{H}'_0) \rightarrow HF^+(\mathcal{H}'_1), \\ F'_1 &: HF^+(\mathcal{H}'_1) \rightarrow HF^+(\mathcal{H}') \end{aligned}$$

defined with respect to a tuple  $(\mathcal{H}', \mathcal{H}'_0, \mathcal{H}'_1)$  which is not subordinate to  $(Y, Y_0, Y_1)$ , such that all of the faces in Figure 11 commute.

This is again straightforward: let  $(\Psi, \Psi_0, \Psi_1)$  be a tuple of equivalences induced by Heegaard moves relating  $(\mathcal{H}, \mathcal{H}_0, \mathcal{H}_1)$  and  $(\mathcal{H}', \mathcal{H}'_0, \mathcal{H}'_1)$ , and  $(\Psi', \Psi'_0, \Psi'_1)$  be a tuple of equivalences induced by Heegaard moves relating  $(\mathcal{H}', \mathcal{H}'_0, \mathcal{H}'_1)$  and  $(\mathcal{H}'', \mathcal{H}''_0, \mathcal{H}''_1)$ . Again, we refer to Figure 11 to help recall the meaning of the notation. Define the map  $F' : HF^+(\mathcal{H}') \rightarrow HF^+(\mathcal{H}'_0)$  by  $F' := \Psi_0 \circ F \circ \Psi^{-1}$  where  $\Psi^{-1}$  is a homotopy inverse for the equivalence  $\Psi$ . Similarly, define  $F'_0 := \Psi_1 \circ F_0 \circ \Psi_0^{-1}$  and  $F'_1 := \Psi \circ F_1 \circ \Psi_1^{-1}$ . Note that

$$\begin{aligned} F' &= \Psi_0 \circ F \circ \Psi^{-1} \\ &= \Psi_0 \circ (\Psi''_0)^{-1} \circ F'' \circ \Psi'' \circ \Psi^{-1} \quad (\text{by [14, Theorem 4.4]}) \\ &= \pm(\Psi'_0)^{-1} \circ F'' \circ \Psi' \quad (\text{by Theorem 1.1}). \end{aligned}$$

This shows that we can provide maps on all of the objects of our transitive systems, and furthermore by the computation above that this gives a well defined morphism of transitive systems. Exactness of all “columns” follows by construction as well. This completes the proof.  $\square$

## References

- [1] **K Hendricks, R Lipshitz, S Sarkar**, *A flexible construction of equivariant Floer homology and applications*, J. Topol. 9 (2016) 1153–1236 MR Zbl
- [2] **K Hendricks, R Lipshitz, S Sarkar**, *A simplicial construction of  $G$ -equivariant Floer homology*, Proc. Lond. Math. Soc. 121 (2020) 1798–1866 MR Zbl
- [3] **K Hendricks, C Manolescu**, *Involutive Heegaard Floer homology*, Duke Math. J. 166 (2017) 1211–1299 MR Zbl
- [4] **S Jabuka, T E Mark**, *Product formulae for Ozsváth–Szabó 4-manifold invariants*, Geom. Topol. 12 (2008) 1557–1651 MR Zbl
- [5] **A Juhász, D Thurston, I Zemke**, *Naturality and mapping class groups in Heegaard Floer homology*, Mem. Amer. Math. Soc. 1338, Amer. Math. Soc., Providence, RI (2021) MR Zbl

- [6] **R Lipshitz**, *A cylindrical reformulation of Heegaard Floer homology*, *Geom. Topol.* 10 (2006) 955–1096 MR Zbl
- [7] **R Lipshitz, P S Ozsvath, D P Thurston**, *Bordered Heegaard Floer homology*, *Mem. Amer. Math. Soc.* 1216, Amer. Math. Soc., Providence, RI (2018) MR Zbl
- [8] **D McDuff, D Salamon**, *J–holomorphic curves and symplectic topology*, 2nd edition, American Mathematical Society Colloquium Publications 52, Amer. Math. Soc., Providence, RI (2012) MR Zbl
- [9] **M J Micallef, B White**, *The structure of branch points in minimal surfaces and in pseudoholomorphic curves*, *Ann. of Math.* 141 (1995) 35–85 MR Zbl
- [10] **P Ozsváth, Z Szabó**, *Holomorphic disks and three-manifold invariants: properties and applications*, *Ann. of Math.* 159 (2004) 1159–1245 MR Zbl
- [11] **P Ozsváth, Z Szabó**, *Holomorphic disks and topological invariants for closed three-manifolds*, *Ann. of Math.* 159 (2004) 1027–1158 MR Zbl
- [12] **P Ozsváth, Z Szabó**, *Holomorphic triangle invariants and the topology of symplectic four-manifolds*, *Duke Math. J.* 121 (2004) 1–34 MR Zbl
- [13] **P Ozsváth, Z Szabó**, *Heegaard Floer homology and contact structures*, *Duke Math. J.* 129 (2005) 39–61 MR Zbl
- [14] **P Ozsváth, Z Szabó**, *Holomorphic triangles and invariants for smooth four-manifolds*, *Adv. Math.* 202 (2006) 326–400 MR Zbl
- [15] **P Ozsváth, Z Szabó**, *Holomorphic disks, link invariants and the multi-variable Alexander polynomial*, *Algebr. Geom. Topol.* 8 (2008) 615–692 MR Zbl
- [16] **L P Roberts**, *Rational blow-downs in Heegaard–Floer homology*, *Commun. Contemp. Math.* 10 (2008) 491–522 MR Zbl
- [17] **S Sarkar**, *Maslov index formulas for Whitney  $n$ -gons*, *J. Symplectic Geom.* 9 (2011) 251–270 MR Zbl
- [18] **S Sarkar**, *Moving basepoints and the induced automorphisms of link Floer homology*, *Algebr. Geom. Topol.* 15 (2015) 2479–2515 MR Zbl
- [19] **R M Vogt**, *Homotopy limits and colimits*, from “Proceedings of the International Symposium on Topology and its Applications”, *Savez Društava Mat. Fiz. i Astronom.*, Belgrade (1973) 235–241 MR Zbl
- [20] **I Zemke**, *Graph cobordisms and Heegaard Floer homology*, preprint (2015) arXiv 1512.01184

Philadelphia, PA, United States

mgartner112@gmail.com

Received: 8 January 2020      Revised: 26 July 2021

# Geometrically bounding 3–manifolds, volume and Betti numbers

JIMING MA

FANGTING ZHENG

A hyperbolic 3–manifold is geometrically bounding if it is the only boundary of a totally geodesic hyperbolic 4–manifold. According to previous results of Long and Reid (2000) and Meyerhoff and Neumann (1992), geometrically bounding closed hyperbolic 3–manifolds are very rare. Assume the value  $v \approx 4.3062\dots$  for the volume of the regular right-angled hyperbolic dodecahedron  $P$  in  $\mathbb{H}^3$ . For each positive integer  $n$  and each odd integer  $k$  in  $[1, 5n + 3]$ , we construct a closed hyperbolic 3–manifold  $M$  with  $\beta^1(M) = k$  and  $\text{vol}(M) = 16nv$  which bounds a totally geodesic hyperbolic 4–manifold. In particular, for every positive odd integer  $k$ , there are infinitely many geometrically bounding 3–manifolds whose first Betti numbers are  $k$ . The proof exploits the real toric manifold theory over a sequence of stacking dodecahedra, together with some results obtained by Kolpakov, Martelli and Tschantz (2015).

57R90, 57M50, 57S25

## 1 Introduction

### 1.1 Geometrically bounding 3–manifolds

There is a well-known result given by Rohlin in 1951, saying that any closed orientable 3–manifold is null-cobordant (see, for example, Corollary 2.5 of [18]), whereas for higher dimensions, it remains an open problem to say which closed  $n$ –manifolds can bound  $(n+1)$ –manifolds. Farrell and Zdravkovska [7] conjectured that every almost flat  $n$ –manifold bounds an  $(n+1)$ –manifold; see also Davis and Fang [5]. This conjecture is far from being solved. Farrell and Zdravkovska also conjectured in the same paper that every flat  $n$ –manifold  $M$  is the cusp section of a one-cusped hyperbolic  $(n+1)$ –manifold. However, Long and Reid [11] refuted this stronger conjecture by showing that

if  $M$  is the cusp section of a one-cusped hyperbolic  $4n$ -manifold, its  $\eta$ -invariant  $\eta(M)$  must be an integer.

If a hyperbolic  $n$ -manifold  $M$  is the unique totally geodesic boundary of a hyperbolic  $(n+1)$ -manifold  $N$ , we say that  $M$  *bounds geometrically* or  $M$  is a *geometrically bounding*  $n$ -manifold. In this context, Long and Reid [11] studied what kinds of 3-manifolds bound geometrically; Ratcliffe and Tschantz [16] provided some cosmological motivations for studying geometrically bounding 3-manifolds. In general, it is not a trivial task to look for geometrically bounding 3-manifolds, since only few explicit hyperbolic 4-manifolds are known. Moreover, Long and Reid showed in [11] that if a closed hyperbolic 3-manifold  $M$  is geometrically bounding, its  $\eta$ -invariant  $\eta(M)$  must be an integer. This, together with the result of Meyerhoff and Neumann [13] that the set of  $\eta$ -invariants of all hyperbolic 3-manifolds is dense in  $\mathbb{R}$ , shows that geometrically bounding 3-manifolds are very rare in the set of hyperbolic 3-manifolds. To the best of our knowledge, the following question remains open:

**Question 1.1** Given a closed hyperbolic 3-manifold  $M$  with  $\eta$ -invariant  $\eta(M) \in \mathbb{Z}$ , is there a totally geodesic hyperbolic 4-manifold  $N$  with  $\partial N = M$ ?

By Jorgensen–Thurston’s Dehn surgery theory [23], we know that there are only finitely many (possibly zero) hyperbolic 3-manifolds with a given volume  $x$ . More precisely, if we consider the function

$$f(x) = \sup\{n \mid \text{there are } n \text{ different hyperbolic 3-manifolds with volume } v \leq x\},$$

then Jorgensen–Thurston theory implies that  $f(x)$  is finite. Furthermore, Millichap [14] showed that  $f(x)$  grows at least factorially.

In this paper, we consider instead the number of geometrically bounding 3-manifolds with a given volume. That is, we focus on the function

$$f_b(x) = \sup\{n \mid \text{there are } n \text{ different geometrically bounding 3-manifolds with volume } v \leq x\}.$$

Building on Kolpakov, Martelli and Tschantz [9] and real toric manifold theory, we prove the following:

**Theorem 1.2** Assume that  $v \approx 4.3062\dots$  is the volume of the regular right-angled hyperbolic dodecahedron in  $\mathbb{H}^3$ . Then, for each positive integer  $n$  and each odd integer  $k$  in  $[1, 5n + 3]$ , there is a closed hyperbolic 3-manifold  $M$  with  $\beta^1(M) = k$  and  $\text{vol}(M) = 16nv$  that bounds a totally geodesic hyperbolic 4-manifold.



Therefore, we construct some families  $\mathcal{F}_n$ ,  $n \geq 1$ , of closed hyperbolic 3-manifolds having the following special features:

- They all *bound geometrically*, ie for any  $n$ , each manifold in  $\mathcal{F}_n$  is the connected geodesic boundary of a compact hyperbolic 4-manifold.
- Each manifold in  $\mathcal{F}_n$  can be decomposed into  $16n$  right-angled dodecahedra. The set  $\mathcal{F}_n$  contains manifolds with first Betti numbers  $1, 3, 5, \dots, 5n + 3$ . In particular,  $\mathcal{F}_n$  contains at least  $n$  elements.

This implies that the above-defined function  $f_b(x)$  grows at least linearly. Moreover, we have a corollary of Theorem 1.2 as follows.

**Corollary 1.3** *For every positive odd number  $k$ , there are infinitely many geometrically bounding 3-manifolds whose first Betti numbers are  $k$ .*

We refer to the paper of Ratcliffe and Tschantz [17] for counting the number of totally geodesic hyperbolic 4-manifolds with the same 3-manifold  $M$  as boundary, and to Chu and Kolpakov [4] and Slavich [19; 20] for other topics regarding geometrically bounding hyperbolic manifolds. Also see the recent paper by Kolpakov, Reid and Slavich [10] for problems related to geodesically embedding hyperbolic manifolds. However, we emphasize that being geometrically bounding is a more subtle property than being geodesically embedding.

## 1.2 Real toric manifolds

Small covers, also known as Coxeter orbifold coverings, have been studied by Davis and Januszkiewicz [6], see also Vesnin [24]. They are a class of  $n$ -manifolds which admit locally standard  $\mathbb{Z}_2^n$ -actions, such that the orbit spaces are  $n$ -dimensional simple polytopes. The algebraic and topological properties of a small cover are closely related to the combinatorics of the orbit polytope and to the coloring on the codimension-one faces of that polytope. For example, the mod 2 Betti numbers  $\beta_i^{(2)}$  of a small cover  $M$  over the polytope  $L$  is equal to  $h_i$ , where  $h = (h_0, h_1, \dots, h_n)$  is the  $h$ -vector of the polytope  $L$ ; see [6].

Those manifolds admitting locally standard  $\mathbb{Z}_2^k$ -actions are usually referred to as *real toric manifolds* and form a wider class. Given an  $n$ -dimensional simple polytope  $L$ , we can define a map  $\lambda: \mathcal{F} \rightarrow \mathbb{Z}_2^k$  that satisfies certain conditions, where  $\mathcal{F}$  is the set of codimension-one faces of  $L$ . Furthermore, by the equivalence relation determined by the map  $\lambda$ , we can construct a smooth closed manifold  $M(L, \lambda)$ . See Section 2.1 for more details.

For instance, we may color the four codimension-one faces of a tetrahedron by  $e_1, e_2, e_3$  and  $e_1 + e_2 + e_3$ , where  $e_1, e_2$  and  $e_3$  are the standard basis of  $\mathbb{Z}_2^3$ . From the construction mentioned in the previous paragraph, we construct the closed orientable 3-manifold  $\mathbb{R}\mathbb{P}^3$ . Note that a tetrahedron admits a unique right-angled spherical structure. We thus naturally obtain a unique spherical structure on  $\mathbb{R}\mathbb{P}^3$  by inheriting spherical structures from the four tetrahedral copies.

In the rest of this section, we assume that  $P$  is the regular right-angled hyperbolic dodecahedron in  $\mathbb{H}^3$  with twelve 2-dimensional facets, and  $nP$  is the polytope obtained by stacking  $n$  copies of  $P$ . It is obvious that  $nP$  has 12 pentagonal facets and  $5n - 5$  hexagonal facets. See Section 2.3 for more details.

Given a  $\mathbb{Z}_2^3$ -coloring  $\lambda$  over the polytope  $nP$ , we generate the natural  $\mathbb{Z}_2^4$ -coloring  $\delta$  on  $nP$  in the following manner. Suppose  $\{e_1, e_2, e_3, e_4\}$  is the standard basis of  $\mathbb{Z}_2^4$ . For each facet  $F$  of  $nP$ , if  $\lambda(F) = \sum_{i=1}^3 x_i e_i$  with  $x_i = 1$  or  $0$ , we take  $\delta(F) = \sum_{i=1}^4 x_i e_i$ , where  $x_4 = 1 + \sum_{i=1}^3 x_i \pmod{2}$ . A  $\mathbb{Z}_2^3$ -coloring  $\lambda$  is called *nonorientable* if the corresponding 3-manifold  $M(nP, \lambda)$  is nonorientable. Furthermore, if the 3-manifold  $M(nP, \lambda)$  is nonorientable, then its natural  $\mathbb{Z}_2^4$ -coloring  $\delta$  is called the *natural  $\mathbb{Z}_2^4$ -extension* of  $\lambda$ . It can be shown that  $M(nP, \delta)$  is the orientable double cover of  $M(nP, \lambda)$  when  $M(nP, \lambda)$  is nonorientable. Our main technical theorem is the following.

**Theorem 1.4** *For each positive integer  $n$  and each odd integer  $k$  in  $[1, 5n + 3]$ , there is a nonorientable  $\mathbb{Z}_2^3$ -coloring  $\lambda$  on the polytope  $nP$  such that the first Betti number of the orientable 3-manifold  $M(nP, \delta)$  is  $k$ , where  $\delta$  is the natural  $\mathbb{Z}_2^4$ -extension of  $\lambda$ .*

From Theorem 1.4, given a positive integer  $n$  and an odd integer  $k$  in  $[1, 5n + 3]$ , there exists an orientable 3-manifold  $M(nP, \delta)$  whose first Betti number is exactly  $k$ . Moreover, we conjecture that there is no coloring on  $nP$  leading to an orientable manifold  $M(nP, \delta)$  with first Betti number not an odd integer  $k \leq 5n + 3$ . The converse has been checked numerically, but has not been proved rigorously yet.

**Proof of Theorem 1.2** For a nonorientable  $\mathbb{Z}_2^3$ -coloring  $\lambda$  on the polytope  $nP$ , there is a natural  $\mathbb{Z}_2^4$ -extension  $\delta$  on  $nP$ . Both  $M(nP, \delta)$  and  $M(nP, \lambda)$  are 3-manifolds and  $M(nP, \delta)$  is the orientable double cover of  $M(nP, \lambda)$ . See Proposition 2.11 in Section 2.4 for more details.

Next, we want to show that  $M(nP, \delta)$  is geometrically bounding. First, we use Proposition 2.9 in [9] to extend the  $\mathbb{Z}_2^4$ -coloring  $\delta$  on the 3-dimensional polytope  $nP$  to

a  $\mathbb{Z}_2^5$ -coloring  $\varepsilon$  on the 4-dimensional polytope  $nE$ . Here,  $nE$  is a 4-dimensional polytope obtained by stacking  $n$  copies of the hyperbolic right-angled 120-cell  $E$ . Then  $M(nE, \varepsilon)$  is an orientable hyperbolic 4-manifold in which  $M(nP, \lambda)$  can be embedded. Second, since  $M(nP, \delta)$  is the orientable double cover of  $M(nP, \lambda)$ , it admits a fixed-point-free orientation-reversing involution. We may thus apply Corollary 9 of [12]. By cutting  $M(nE, \varepsilon)$  along the hypersurface  $M(nP, \lambda)$  and applying completion, we can obtain a totally geodesic hyperbolic 4-manifold with boundary  $M(nP, \delta)$ . Now, Theorem 1.2 follows from Theorem 1.4.  $\square$

## Outline of the paper

In Section 2, we provide some preliminaries on the algebraic theory of real toric manifolds. In Section 3, we prove Lemma 3.1, which is the key element of the main theorem. In Sections 4 and 5, we prove Theorem 1.4 for the cases of even and odd  $n$ , respectively.

## Acknowledgements

Jiming Ma was partially supported by NSFC 11771088 and 12171092. Fangting Zheng was supported by NSFC 12101504 and XJTLU Research Development Fund RDF-19-01-29. The authors appreciate greatly the referees and Alastair Darby for their valuable and constructive comments on improving the text, which made our paper much more precise and readable.

## 2 Preliminaries

In this section, we list some facts concerning real toric manifolds and introduce the 3-dimensional right-angled hyperbolic polytope  $nP$ . Proofs, details, and definitions can be found in [1]. For the sake of brevity, we write  $n$ -polytope instead of  $n$ -dimensional polytope, and by *facet* we mean a face of codimension one. An  $n$ -polytope is called simple if every  $r$ -face belongs to exactly  $n - r$  facets.

### 2.1 Real toric manifolds

Given a simple  $n$ -polytope  $L$ , let  $\mathcal{F}(L) = \{F_1, F_2, \dots, F_m\}$  be its set of facets. Let us define the  $\mathbb{Z}_2^k$ -coloring characteristic function,  $n \leq k \leq m$ , as a function

$$\lambda: \mathcal{F}(L) = \{F_1, F_2, \dots, F_m\} \rightarrow \mathbb{Z}_2^k$$

that satisfies the *nonsingularity condition*. That is,  $\lambda(F_{i_1}), \lambda(F_{i_2}), \dots, \lambda(F_{i_n})$  generate a subgroup of  $\mathbb{Z}_2^k$  which is isomorphic to  $\mathbb{Z}_2^n$  when the  $n$  facets  $F_{i_1}, F_{i_2}, \dots, F_{i_n}$  share a common vertex. The binary matrix  $\Lambda_{(n \times m)} = (\lambda(F_1), \lambda(F_2), \dots, \lambda(F_m))$  is called the *characteristic matrix* of  $\lambda$ .

Then, we can construct a smooth manifold  $M(L, \lambda) := L \times \mathbb{Z}_2^k / \sim$ , called a *real toric manifold over the polytope  $P$* , by the equivalence relation

$$(x, g_1) \sim (y, g_2) \iff \begin{cases} x = y \text{ and } g_1 = g_2 & \text{if } x \in \text{Int } L, \\ x = y \text{ and } g_1^{-1}g_2 \in G_f & \text{if } x \in \partial L, \end{cases}$$

where  $f = F_{i_1} \cap \dots \cap F_{i_{n-r}}$  is the unique face of codimension  $n - r$  that contains  $x$  as an interior point, and  $G_f$  is the subgroup generated by  $\lambda(F_{i_1}), \lambda(F_{i_2}), \dots, \lambda(F_{i_{n-r}})$ . The notation  $M(L, \lambda)$  also highlights that each real toric manifold corresponds to a pair  $\{(L, \lambda)\}$  that is made of a polytope and a characteristic function. For brevity, we refer to the colorings when the polytope is given instead of talking about both colorings and manifolds. When  $k = m$ ,  $M(L, \lambda)$  is known as the *real moment-angle manifold* over the polytope  $L$ , which admits a natural  $\mathbb{Z}_2^m$ -action. If  $k = n$ , then the corresponding manifold is called a *small cover*. By the four color theorem, we know that small covers can always be realized over any 3-dimensional simple polytope.

**Example 2.1** Define a  $\mathbb{Z}_2^3$ -coloring characteristic function  $\lambda$  on the right-angled spherical triangle  $\Delta^2$  as shown in Figure 1. Namely, the characteristic function is

$$\begin{aligned} \lambda : \{\{a, b\}, \{b, c\}, \{a, c\}\} &\rightarrow \{(1, 0, 0), (0, 1, 0), (0, 0, 1)\}, \\ (a, b) &\mapsto (1, 0, 0), \\ (b, c) &\mapsto (0, 1, 0), \\ (a, c) &\mapsto (0, 0, 1), \end{aligned}$$

where  $(1, 0, 0) = e_1, (0, 1, 0) = e_2$  and  $(0, 0, 1) = e_3$  are the standard basis vectors of  $\mathbb{Z}_2^3$ .

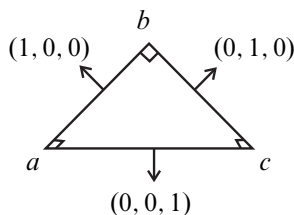


Figure 1: The coloring in Example 2.1.

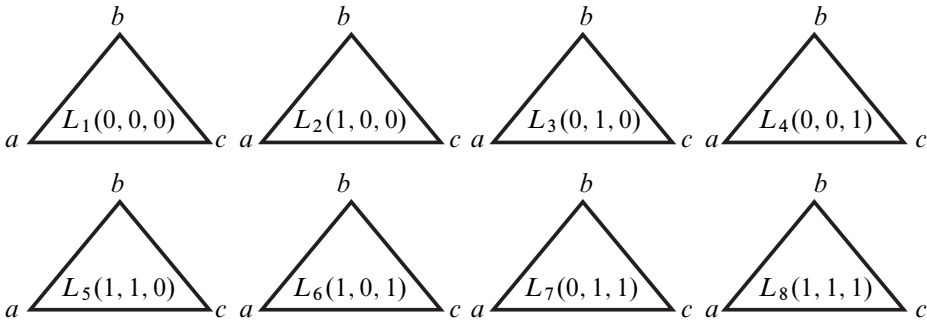


Figure 2: The eight polytopes  $\Delta^2 \times \mathbb{Z}_3^2$  of Example 2.1.

Now, we have eight copies of the polytope, namely  $\Delta^2 \times \mathbb{Z}_2^3$ , as shown in Figure 2.

By the equivalence relation

$$(p, g_1) \sim (q, g_2) \iff \begin{cases} p = q, \\ g_1 - g_2 \in \{(1, 0, 0), (0, 1, 0), (0, 0, 1)\}, \end{cases}$$

we can finally obtain the manifold  $M(\Delta^2, \lambda) \approx \mathbb{S}^2$  as shown in Figure 3, which inherits a spherical structure from the eight copies of right-angled triangles.  $\square$

In order to keep notation concise, we regard every  $\mathbb{Z}_2^*$ -color as a binary number and encode it with an integer. For example in the  $\mathbb{Z}_2^3$ -coloring case, we can use 1, 2, 3, 4, 5, 6 and 7 to represent the seven colors (1, 0, 0), (0, 1, 0), (1, 1, 0), (0, 0, 1), (1, 0, 1), (0, 1, 1) and (1, 1, 1), respectively. Then, a characteristic matrix can also be viewed

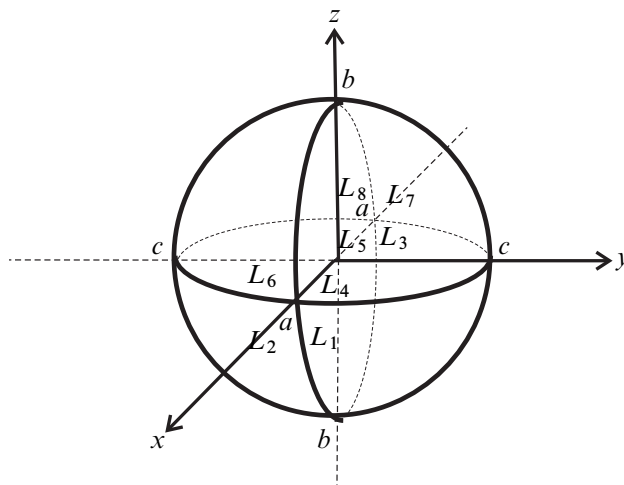


Figure 3: The real toric manifold  $M(\Delta^2, \lambda)$ .

as a characteristic vector. For example, the characteristic matrix of the  $\mathbb{Z}_2^3$ -coloring characteristic function in Example 2.1 is

$$\Lambda_{(3 \times 3)} = (\lambda(\mathcal{F}_1), \lambda(\mathcal{F}_2), \lambda(\mathcal{F}_3)) = (\lambda(a, b), \lambda(b, c), \lambda(a, c)) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Then the corresponding characteristic vector  $C$  is  $(1, 2, 4)$ . The characteristic function  $\lambda$ , characteristic matrix  $\Lambda$ , and the characteristic vector  $C$  can be constructed from each other easily; the characteristic vector  $C$  represents the most concise form.

## 2.2 Cohomology of real toric manifolds

Davis and Januszkiewicz [6] formulated how to calculate the  $\mathbb{Z}_2$ -coefficient cohomology groups of a small cover from the polytope and characteristic function. In 2013, Cai [2] suggested a method to calculate the  $\mathbb{Z}$ -coefficient cohomology groups of a real moment-angle manifold. Based on the results of Cai, Suciu and Trevisanon [21; 22] on rational homology groups of real toric manifolds, Choi and Park [3] obtained a formula for the cohomology groups of real toric manifolds. This can also be viewed as a combinatorial version of the Hochster theorem [8].

Since the dual of the boundary of a simple polytope  $L$  is a simplicial complex  $K$  (see eg [1]), the definition of real toric manifolds introduced above has a dual version. By substituting the facet set  $\mathcal{F}(L)$  with the vertex set  $\mathcal{V}$  of the simplicial complex  $K$ , we can define the *characteristic function*  $\lambda$  on  $K$ , namely

$$\lambda: \mathcal{V}(K) = \{v_1, v_2, \dots, v_m\} \rightarrow \mathbb{Z}_2^k.$$

The nonsingularity condition changes as follows: if for  $n$  vertices  $v_{i_1}, v_{i_2}, \dots, v_{i_n}$  the convex hull  $\text{conv}\{v_{i_1}, v_{i_2}, \dots, v_{i_n}\}$  is a facet of  $K$ , the images  $\lambda(v_{i_1}), \lambda(v_{i_2}), \dots, \lambda(v_{i_n})$  shall generate a subgroup isomorphic to  $\mathbb{Z}_2^n$ . For the sake of brevity, we denote the linear space  $\mathbb{Z}_2^{|\mathcal{V}|}$  by  $\mathbb{Z}_2^\mathcal{V}$ . In addition, we can identify  $\mathbb{Z}_2^\mathcal{V}$  with the power set  $2^\mathcal{V}$  in the canonical way, where  $\emptyset$  corresponds to the identity element and multiplication to the symmetric difference. Namely, we have a map  $\varphi: \mathbb{Z}_2^\mathcal{V} \rightarrow 2^\mathcal{V}$ . Denote by  $K_\omega$  the full subcomplex of  $K = (\partial L)^*$  obtained by restricting to  $\omega \subseteq \mathcal{V}$ . Then every full subcomplex  $K_\omega$  of  $K$ , where  $\omega \subseteq \mathcal{V}$ , is identified with an element of  $\mathbb{Z}_2^\mathcal{V}$ .

Let  $\lambda$  be a  $\mathbb{Z}_2^k$ -coloring characteristic function. Denote by row  $\Lambda$  the *row space* of the characteristic matrix  $\Lambda$ . The following Choi–Park theorem shows that the cohomology group of a real toric manifold  $M(L, \lambda)$  is the direct sum of the cohomology groups of

some full subcomplexes of the dual polytope  $K = (\partial L)^*$ . The full subcomplexes are determined by the characteristic function.

**Theorem 2.2** (Choi–Park [3]) *Assume  $G$  is the coefficient ring  $\mathbb{Q}$  or  $\mathbb{Z}_q$  for a positive odd integer  $q$ . There is an additive isomorphism*

$$H^p(M(L, \lambda); G) \cong \bigoplus_{\varphi^{-1}(\omega) \in \text{row } \Lambda} \tilde{H}^{p-1}(K_\omega; G),$$

where  $\Lambda$  is the characteristic matrix of  $\lambda$ .

We use  $\beta^i$  to denote the rank of  $H^i(M(L, \lambda); \mathbb{Q})$ , called the  $i^{\text{th}}$  Betti number of  $M(L, \lambda)$ ; and use  $\tilde{\beta}^0$  to denote the rank of  $\tilde{H}^0(K_\omega; \mathbb{Q})$ , called the reduced zeroth Betti number of  $K_\omega$ . For the purpose of this paper, we only need the following result.

**Corollary 2.3** *For a simple polytope  $L$ ,*

$$\beta^1(M(L, \lambda); \mathbb{Q}) = \sum_{\varphi^{-1}(\omega) \in \text{row } \Lambda} \tilde{\beta}^0(K_\omega; \mathbb{Q}),$$

where  $\Lambda$  is the characteristic matrix of  $\lambda$ .

By means of Corollary 2.3, we can calculate the first Betti number of a real toric manifold using the combinatorial information of the orbit polytope and the row space of its characteristic matrix. In the following, we show a simple example.

**Example 2.4** Calculate the first Betti number of the Klein bottle  $S = M(L, \lambda)$ .

Figure 4, left, is a colored 2-dimensional square  $L$ , whereas Figure 4, right, is its dual  $K = (\partial L)^*$ , with its vertices colored accordingly.

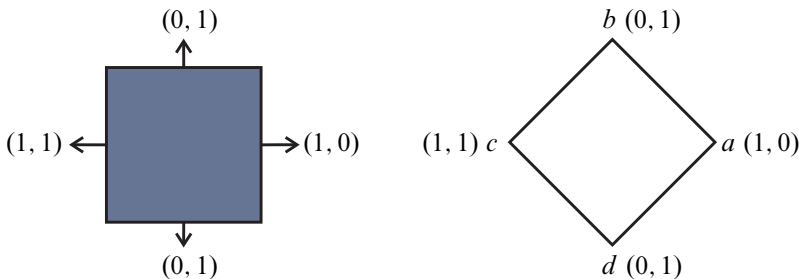


Figure 4: The colored square for Example 2.4.

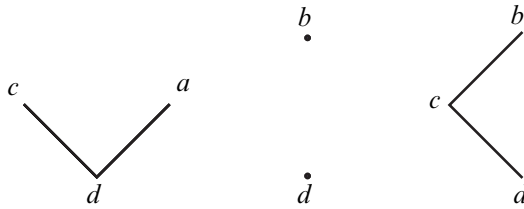


Figure 5: Left to right, the subcomplexes  $K_{\omega_i}$ ,  $2 \leq i \leq 4$ .

Thus, the row space is

$$\text{row } \Lambda = \langle (1, 0, 1, 0), (0, 1, 1, 1) \rangle = \{(0, 0, 0, 0), (1, 0, 1, 1), (0, 1, 0, 1), (1, 1, 1, 0)\}.$$

For  $\omega_1 = (0, 0, 0, 0)$ ,  $K_{\omega_1} = \emptyset$ .

For  $\omega_2 = (1, 0, 1, 1)$ , then  $K_{\omega_2}$  is as shown in Figure 5, left. So  $\tilde{\beta}^0(K_{\omega_2}) = 0$ .

For  $\omega_3 = (0, 1, 0, 1)$ , then  $K_{\omega_3}$  is as shown in Figure 5, center. So  $\tilde{\beta}^0(K_{\omega_3}) = 1$ .

For  $\omega_4 = (1, 1, 1, 0)$ , then  $K_{\omega_4}$  is as shown in Figure 5, right. So  $\tilde{\beta}^0(K_{\omega_4}) = 0$ .

By Corollary 2.3, we have

$$\beta^1(S) = \tilde{\beta}^0(K_{\omega_1}) + \tilde{\beta}^0(K_{\omega_2}) + \tilde{\beta}^0(K_{\omega_3}) + \tilde{\beta}^0(K_{\omega_4}) = 0 + 0 + 1 + 0 = 1,$$

which coincides with the well-known result of rational homology groups of the Klein bottle. □

### 2.3 The 3–polytopes $nP$

In the following, we assume that  $P$  is the regular right-angled dodecahedron in  $\mathbb{H}^3$  with twelve 2–dimensional facets. We use  $nP$  to denote the stacking of  $n$  copies of  $P$ , ie the polytope made of  $n$  dodecahedra in a row; see Figures 6, 7 and 12. The simplicial complex  $nK$  is the dual of the boundary of  $nP$ . For each polytope  $nP$  with  $n \geq 2$ , there are  $n + 3$  layers of facets of  $nP$ : the first and the last layers are pentagons, the second and the  $(n+2)^{\text{nd}}$  layers consist of five pentagons, and each layer from the third to the  $(n+1)^{\text{st}}$  is made of five hexagons. There is no hexagonal layer in  $1P$ , and the polytope  $nP$  has  $5n + 7$  facets in total. All the polytopes  $nP$ , with  $n \in \mathbb{Z}_+$ , are right-angled hyperbolic polytopes. In addition, we call the  $i$ –layer of a colored 3–polytope  $nP$  a brick, where  $2 \leq i \leq n + 1$  and  $n \geq 2$ . The symbols  $nP$  and  $nK$  are used throughout the paper with this meaning, unless stated otherwise.



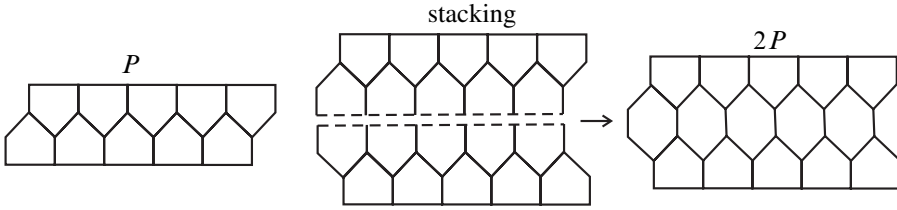


Figure 6: Build up the polytope  $2P$  by stacking.

**Definition 2.5** Given a polytope  $L$  with  $m$  facets, we define  $X(L) = (a_{ij})_{m \times m}$  to be the adjacency matrix of  $L$ , where

$$a_{ij} = \begin{cases} 1 & \text{if } F_i \cap F_j \text{ for } F_i, F_j \in \mathcal{F}(L) \text{ is an } (n-2)\text{-face of } L \text{ or } i = j, \\ 0 & \text{otherwise.} \end{cases}$$

**Definition 2.6** A simple polytope  $L$  is called a flag polytope if every collection of pairwise intersecting facets has a nonempty intersection.

For a flag polytope, all of the information about the intersection of its facets is included in the adjacency matrix. As can be easily checked, the polytope  $nP$  is a flag polytope for every  $n$ . In order to obtain more unified adjacency matrices  $X(nP)$ ,  $n \in \mathbb{Z}_+$ , we order the facets of the polytope  $nP$  in the following manner. The first and the last layer are labeled as 1 and  $5n + 7$ , respectively, while the facets in between are labeled layer by layer. For even layers, we start from the middle and order the rest by left-right double siding, whereas for odd layers, we adopt a right-left double siding. We illustrate the labeling manner on the polytope  $5P$  in Figure 7, where the double sidings of even and odd layers are displayed by the arrow-lines on the second and third layers, respectively.

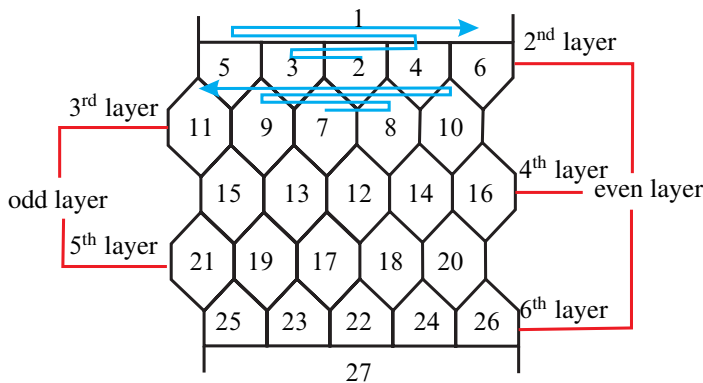


Figure 7: Facet ordering of the polytope  $5P$ .

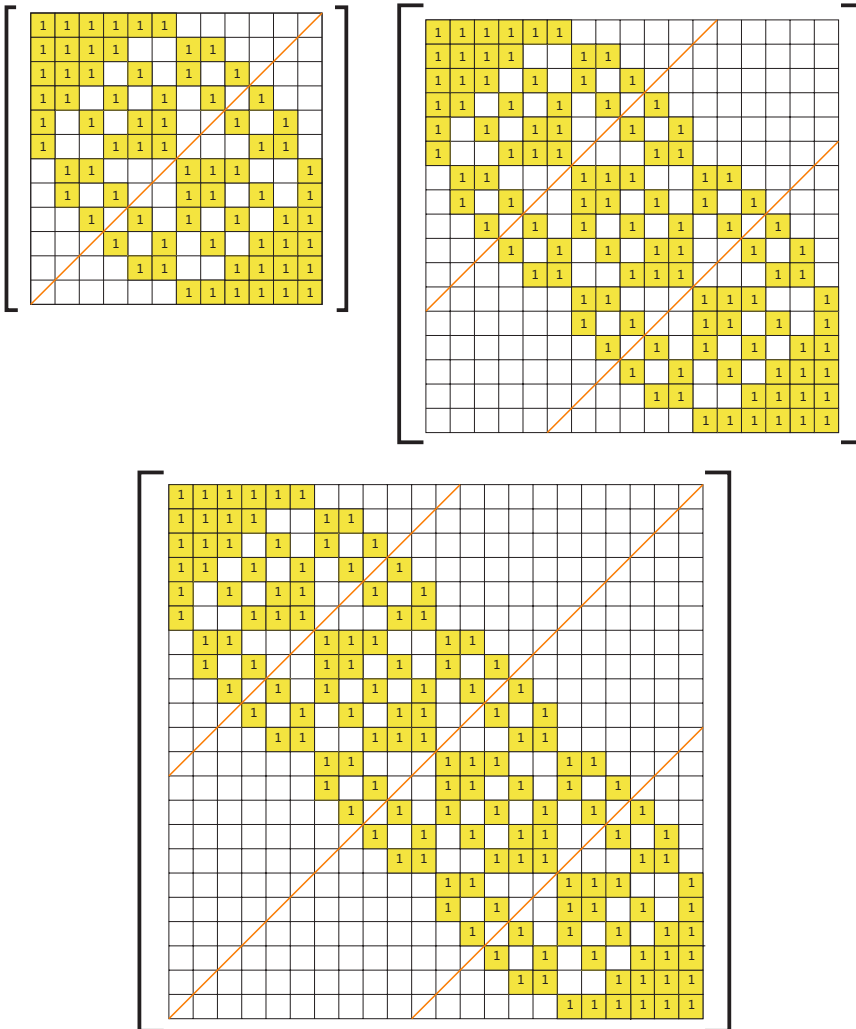


Figure 8: The adjacency matrices of the polytopes  $P$ ,  $2P$  and  $3P$  are given at top left, top right and bottom, respectively.

Using this ordering, we obtain more unified increasing patterns of the adjacency matrices. We display some of them in Figure 8 (the omitted entries are zeros).

### 2.4 Orientability of real toric manifolds

H Nakayama and Y Nishimura discussed the orientability of small covers in [15]. Below we quote their main theorem.

**Theorem 2.7** (Nakayama–Nishimura [15]) *For a simple  $n$ -dimensional polytope  $L$ , and for a basis  $\{e_1, \dots, e_n\}$  of  $\mathbb{Z}_2^n$ , a homomorphism  $\epsilon: \mathbb{Z}_2^n \rightarrow \mathbb{Z}_2 = \{0, 1\}$  is defined by  $\epsilon(e_i) = 1$  for each  $i = 1, \dots, n$ . A small cover  $M(L, \delta)$  is orientable if and only if there exists a basis  $\{e_1, \dots, e_n\}$  of  $\mathbb{Z}_2^n$  such that the image of  $\epsilon\delta$  is  $\{1\}$ .*

The techniques used in proving Theorem 2.7 are actually suitable for all real toric manifolds, not just for small covers. Corollary 2.3 with rational coefficients implies this conclusion as well. The  $n^{\text{th}}$  Betti number of a real toric manifold  $M(L, \delta)$  over the  $n$ -polytope  $L$  is 1 if and only if there is an element in the row space of the characteristic matrix of  $\delta$  with all entries equal to 1.

**Corollary 2.8** (Nakayama–Nishimura [15] and Choi–Park [3]) *For a simple  $n$ -dimensional polytope  $L$ , the real toric manifold  $M(L, \delta)$  is orientable if and only if there is a basis such that the sum of every column of the characteristic matrix  $\Lambda$  of  $\delta$  is 1 mod 2.*

In particular, the four vectors  $(1, 0, 0)$ ,  $(0, 1, 0)$ ,  $(0, 0, 1)$  and  $(1, 1, 1)$ , which are the binary forms of 1, 2, 4 and 7, are the only four elements in  $\mathbb{Z}_2^3$  whose entry sums are 1 mod 2. These four vectors are called *orientable colors*. The three colors left are  $(1, 1, 0)$ ,  $(1, 0, 1)$  and  $(0, 1, 1)$ , which are the binary forms of 3, 5 and 6. An *orientable basis* in  $\mathbb{Z}_2^3$  is defined to be a basis in  $\mathbb{Z}_2^3$  that consists of three linearly independent orientable colors. In particular, the standard basis in  $\mathbb{Z}_2^3$ , ie  $(1, 0, 0)$ ,  $(0, 1, 0)$ ,  $(0, 0, 1)$ , is an orientable basis. If the small cover  $M(L, \lambda)$  is orientable, then there exists an orientable basis such that all the colors of  $\lambda$  are orientable. Note that, for an orientable color, the number of entries with value 1 is always odd. In other words, when changing from one orientable basis to another orientable one, we actually add or remove an even number of 1s from the previous characteristic matrix to form the new one. Hence the parity of the number of 1s in each column is preserved under different orientable bases. Therefore, we have the following corollary.

**Corollary 2.9** *Given a 3-polytope  $nP$  with facets ordered as required in Section 2.3, we fix the colors on first three facets to be  $(1, 0, 0)$ ,  $(0, 1, 0)$  and  $(0, 0, 1)$ . Suppose  $(1, 2, 4, a_1, \dots, a_m)$  is a characteristic vector of  $nP$ . Then the corresponding small cover is nonorientable if there is some  $a_i \in \{3, 5, 6\}$ .*

Starting from a  $\mathbb{Z}_2^3$ -coloring  $\lambda$  on the polytope  $nP$ , we can obtain  $2^m - 1$   $\mathbb{Z}_2^4$ -colorings on  $nP$  by adding a nonzero fourth row to the  $3 \times m$  characteristic matrix  $\Lambda$  of  $\lambda$

as shown:

$$\begin{pmatrix} 1 & 0 & 0 & \cdots & \cdots \\ 0 & 1 & 0 & \cdots & \cdots \\ 0 & 0 & 1 & \cdots & \cdots \\ * & * & * & \cdots & * \end{pmatrix},$$

where  $m = 5n + 7$  and  $* \in \{0, 1\}$ . Those characteristic functions are called the *extensions* of  $\lambda$ , and they naturally satisfy the nonsingularity condition.

**Definition 2.10** A  $\mathbb{Z}_2^3$ -coloring  $\lambda$  on the polytope  $nP$  is *admissible* if there is a  $\mathbb{Z}_2^4$ -coloring extension of  $\lambda$ , denoted by  $\delta$ , such that  $M(nP, \lambda)$  is nonorientable and  $M(nP, \delta)$  is the orientable double cover of  $M(nP, \lambda)$ .

Along with some basic facts about the fundamental group of a double cover we have the following proposition. It is valid for any polytope and we are now interested in the case of polytope  $nP$ .

**Proposition 2.11** A  $\mathbb{Z}_2^3$ -coloring  $\lambda$  over a simple 3-dimensional polytope  $nP$  is *admissible* if  $M(nP, \lambda)$  is nonorientable.

**Proof** Because  $M(nP, \lambda)$  is nonorientable, at least one column of its characteristic matrix  $\Lambda$  has an even sum. Therefore, we can add a nonzero fourth row to the characteristic matrix  $\Lambda$  to obtain a  $\mathbb{Z}_2^4$ -coloring extension of  $\lambda$ , denoted by  $\delta$ , satisfying that the sum of all its columns are odd. By Corollary 2.8,  $M(nP, \delta)$  is orientable.

Let  $W(nP)$  be the Coxeter group of  $nP$  and  $\theta: \mathcal{F}(L) = \{F_1, F_2, \dots, F_m\} \rightarrow \mathbb{Z}_2^m$  be the map that sends each  $F_i$  to  $e_i$ . Now we have the diagram

$$\begin{array}{ccccc} W(nP) & \xrightarrow{l} & \mathbb{Z}_2^m & \xrightarrow{\hat{\delta}} & \mathbb{Z}_2^4 \\ & & & \searrow \hat{\lambda} & \downarrow p \\ & & & & \mathbb{Z}_2^3 \end{array}$$

where  $l$  is the abelianization,  $p$  is the natural projection of  $\mathbb{Z}_2^4$  to  $\mathbb{Z}_2^3$  that keeps only the first three coordinates, and  $\hat{\lambda}$  and  $\hat{\delta}$  are the maps induced by the characteristic functions  $\lambda$  and  $\delta$ , ie  $\lambda = \hat{\lambda} \circ \theta$  and  $\delta = \hat{\delta} \circ \theta$ . It is easy to check that the triangular circuit commutes, namely,  $p \circ \hat{\delta} = \hat{\lambda}$ .

By [6, Corollary 4.5],  $\pi_1(M(nP, \lambda)) = \ker(\hat{\lambda} \circ l) = \ker(p \circ \hat{\delta} \circ l)$  and  $\pi_1(M(nP, \delta)) = \ker(\hat{\delta} \circ l)$ . Thus  $M(nP, \delta)$  is an orientable double cover of  $M(nP, \lambda)$ . □

The  $\mathbb{Z}_2^4$ -coloring  $\delta$  on the polytope  $nP$  in Proposition 2.11 is called an *admissible extension* of  $\lambda$  or a *natural  $\mathbb{Z}_2^4$ -coloring associated to  $\lambda$*  (also referred to as the *natural  $\mathbb{Z}_2^4$ -extension* of  $\lambda$  for short). We use the symbols  $\lambda$  and  $\delta$  with this meaning in the rest of the paper, unless stated otherwise. Moreover, by Corollary 2.3, the Betti numbers of the orientable manifold recovered by the natural  $\mathbb{Z}_2^4$ -extension  $\delta$  can be easily computed, as we are going to show in Example 2.12.

**Example 2.12** Let us calculate the Betti numbers of some orientable real toric manifold  $M(P, \delta)$ .

We show in Figure 9, left, a plane figure of the dodecahedron  $P$  whose facets are ordered in the “double siding” manner introduced in Section 2.3. In Figure 9, right, is the dual simplicial complex  $K = (\partial P)^*$  with its 12 vertices labeled correspondingly.

Color the polytope  $P$  with the characteristic vector  $v = (1, 2, 4, 5, 3, 7, 7, 3, 5, 4, 2, 1)$  and denote the corresponding characteristic function by  $\lambda$ . Then we have a  $\mathbb{Z}_2^3$ -coloring characteristic matrix

$$\Lambda = \begin{pmatrix} 0 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 1 \end{pmatrix}_{3 \times 12}.$$

By Corollary 2.9,  $\lambda$  is nonorientable. The characteristic matrix  $\Delta$  of its admissible extension  $\delta$  is

$$\Delta = \begin{pmatrix} & & & & & \Lambda & & & & & & \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \end{pmatrix}_{4 \times 12}.$$

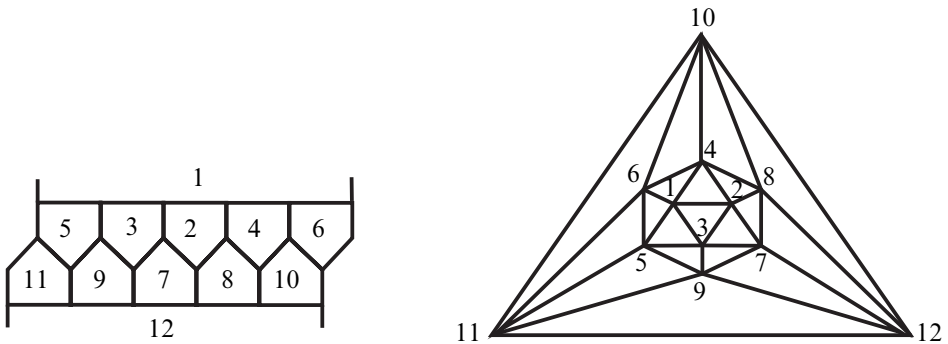


Figure 9: The facet-ordered polytope  $P$ , left, and its dual simplicial complex  $K = (\partial P)^*$ , right.

The row space of  $\Delta$  is given by

$$\text{row } \Delta = \langle (0, 0, 1, 1, 0, 1, 1, 0, 1, 1, 0, 0), (0, 1, 0, 0, 1, 1, 1, 1, 0, 0, 1, 0), \\ (1, 0, 0, 1, 1, 1, 1, 1, 1, 0, 0, 1), (0, 0, 0, 1, 1, 0, 0, 1, 1, 0, 0, 0) \rangle.$$

For each  $\omega_i \in \text{row } \Delta$ , we calculate its reduced 0<sup>th</sup> Betti number in Tables 1–2.

From Tables 1–2 and Corollary 2.3, we have

$$\beta^1(M(P, \delta); \mathbb{Q}) = \sum_{i=1}^{16} \tilde{\beta}^0(K_{\omega_i}; \mathbb{Q}) = \beta^2(M(P, \delta); \mathbb{Q}) \\ = \sum_{i=1}^{16} \tilde{\beta}^1(K_{\omega_i}; \mathbb{Q}) = 7. \quad \square$$

For an orientable 3–manifold  $M(nP, \delta)$ , by Poincaré duality we have  $\beta^0(M(nP, \delta)) = \beta^3(M(nP, \delta)) = 1$  and  $\beta^1(M(nP, \delta)) = \beta^2(M(nP, \delta))$ . So  $\beta^1$  is the only thing we need in order to determine the free part of  $H^*(M(nP, \delta))$ . By Corollary 2.3,  $\beta^1(M(nP, \delta))$  is equal to the sum of the reduced zeroth Betti numbers of the 16 full subcomplexes  $k_{\omega_i}$  of the simplicial complex  $nK = (\partial(nP))^*$ . Each subcomplex  $k_{\omega_i}$  corresponds to a nonzero vector in the row space  $\text{row } \Delta$ .

### 3 The key lemma

The purpose of this section is to prove Lemma 3.1, which is the key element in proving Theorem 1.4. We want to find a special family of admissible  $\mathbb{Z}_2^3$ –colorings over the polytope  $nP$ . According to the correspondence discussed in Section 2, we construct a family of orientable 3–manifolds  $M(nP, \delta)$ .

**Lemma 3.1** *For every positive even integer  $n$ , there is a nonorientable  $\mathbb{Z}_2^3$ –coloring  $\lambda$  over the polytope  $nP$  such that  $\beta^1(M(nP, \delta)) = n + 1$ , where  $\delta$  is the natural associated  $\mathbb{Z}_2^4$ –coloring extension of  $\lambda$ .*

**Proof** We first prove the special case in which  $n = 2$ . We use the notation  $a_1 = 1$ ,  $S_1 = (24247)$  and  $S_2S_1 = (35716\ 24247)$ . By  $[a_1S_1S_2S_1a_1]$ , we mean the colored polytope  $2P$  shown in Figure 10. The corresponding characteristic vector  $C$  is

$$(1, 2, 4, 4, 2, 7, 7, 1, 5, 6, 3, 2, 4, 4, 2, 7, 1).$$

It can be checked with little effort that the nonsingularity condition holds at every vertex. We call  $S_i$ ,  $1 \leq i \leq 2$ , a *brick* and  $a_i$ , which represents the first or the last

$i$	$\omega_i$	$K_{\omega_i}$	$\tilde{\beta}^0(K_{\omega_i})$	$\beta^1(K_{\omega_i})$	$\beta^2(K_{\omega_i})$
1	(0, 0, 1, 1, 0, 1, 1, 0, 1, 1, 0, 0)		1		
2	(0, 1, 0, 0, 1, 1, 1, 1, 0, 0, 1, 0)		1		
3	(1, 0, 0, 1, 1, 1, 1, 1, 1, 0, 0, 1)		0	1	
4	(0, 0, 0, 1, 1, 0, 0, 1, 1, 0, 0, 0)		1		
5	(0, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 0)		0	1	
6	(1, 0, 1, 0, 1, 0, 0, 1, 0, 1, 0, 1)		1		
7	(0, 0, 1, 0, 1, 1, 1, 1, 0, 1, 0, 0)		0	1	
8	(1, 1, 0, 1, 0, 0, 0, 0, 1, 0, 1, 1)		1		

Table 1: The values of  $\tilde{\beta}^0(K_{\omega_i})$  for  $i = 1, \dots, 8$ .

colored facet, an *affix*. They are used for building the coloring. The symbols  $S_i$  and  $a_i$  are used with this meaning in the rest of the paper unless stated otherwise.

$i$	$\omega_i$	$K_{\omega_i}$	$\tilde{\beta}^0(K_{\omega_i})$	$\beta^1(K_{\omega_i})$	$\beta^2(K_{\omega_i})$
9	$(0, 1, 0, 1, 0, 1, 1, 0, 1, 0, 1, 0)$		0	1	
10	$(1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 1)$		1		
11	$(1, 1, 1, 0, 0, 1, 1, 0, 0, 1, 1, 1)$		0	1	
12	$(1, 1, 0, 0, 1, 0, 0, 1, 0, 0, 1, 1)$		0	1	
13	$(1, 0, 1, 1, 0, 0, 0, 0, 1, 1, 0, 1)$		0	1	
14	$(0, 1, 1, 0, 0, 0, 0, 0, 0, 1, 1, 0)$		1		
15	$(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)$	$\emptyset$	no contribution to $\beta^1(M(P, \delta))$		
16	$(1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1)$	$\cong \mathbb{S}^2$	0	0	1

Table 2: The values of  $\tilde{\beta}^0(K_{\omega_i})$  for  $i = 9, \dots, 16$ .

Let us denote by  $\lambda$  the characteristic function of  $C$ . Corollary 2.9 and Proposition 2.11 imply that  $\lambda$  is admissible, and we denote by  $\delta$  its natural  $\mathbb{Z}_2^4$ -extension. It follows that  $M(2P, \lambda)$  is nonorientable, and  $M(2P, \delta)$  is the orientable double cover of  $M(2P, \lambda)$ . The characteristic matrix  $\Delta$  of the coloring  $\delta$  is

$$(3-1) \quad \begin{pmatrix} 0 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$



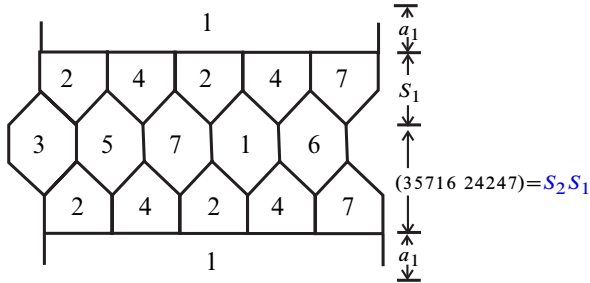


Figure 10: Colored polytope  $2P$ .

Then, the row space row  $\Delta$  is given by

$$(3-2) \quad \begin{pmatrix} 0 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix} .$$

By Corollary 2.3, we can calculate  $\beta^1(M(2P, \delta))$  through its 15 nonempty full sub-complexes  $K_\omega$ . Since  $\tilde{\beta}^0 = \tilde{\beta}_0$ , the reduced zeroth Betti number of each  $K_\omega$  is equal to the number of connected components of  $K_\omega$  minus one.

For every  $i^{\text{th}}$  row  $\omega_i(\Delta) = (w_{i1}, \dots, w_{ij}, \dots, w_{im})$  of the row space row  $\Delta$ , where  $m = 5n + 7$  is the number of facets of  $nP$  and  $1 \leq i \leq 2^4 - 1$ , we define

$$\omega_i^*(\Delta) := \{j \mid 1 \leq j \leq m \text{ and } \omega_{ij} = 1, \text{ where } \omega_{ij} \in \text{row } \Delta\}.$$

Then define  $X(nP, \omega_i(\Delta))$  to be the submatrix of  $X(nP)$  obtained by selecting the  $q^{\text{th}}$  rows and  $q^{\text{th}}$  columns as  $q$  varies in  $\omega_i^*(\Delta)$ .

For example, pick the first row  $\omega_1(\Delta) = (0, 0, 1, 1, 0, 1, 1, 0, 1, 1, 0, 0, 1, 1, 0, 1, 0)$  of the row space row  $\Delta$  shown in matrix (3-2); then  $\omega_1^*(\Delta) = (3, 4, 6, 7, 9, 10, 13, 14, 16)$ .

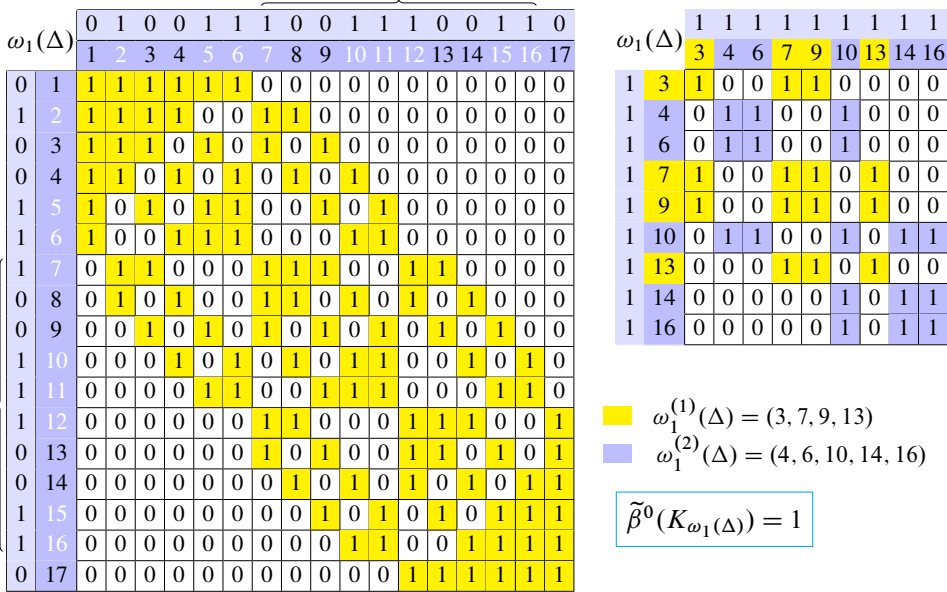


Figure 11: The computation of  $\tilde{\beta}^0(K_{\omega_1(\Delta)})$ . Left:  $X(2P)$ . Right:  $X(2P, \omega_1(\Delta))$ .

Let us consider the submatrix  $X(2P, \omega_1(\Delta))$  which is obtained from the adjacency matrix  $X(2P)$  by selecting the rows and columns set by  $\omega_1^*(\Delta)$ . By examining this matrix, it is obvious that there are two connected components. Use the notation  $\omega_j^{(i)}(\Delta)$  to denote the vertex set of the  $i^{\text{th}}$  connected component of the full subcomplex  $K_{\omega_j(\Delta)}$ . Then, we have  $\omega_1^{(1)}(\Delta) = (3, 7, 9, 13)$  and  $\omega_1^{(2)}(\Delta) = (4, 6, 10, 14, 16)$ ; therefore,  $\tilde{\beta}^0(K_{\omega_1(\Delta)}) = 1$ . The procedure is illustrated in Figure 11.

Likewise, we can calculate all of the  $\tilde{\beta}^0(K_{\omega_i(\Delta)})$ ,  $1 \leq i \leq 15$ , and the computation for  $i = 2, 3, \dots, 7$  is illustrated in (A) and (B) of Figures 16–21 in the online supplement. Finally, we obtain  $\beta^1(M(2P, \delta))=3$ , as shown in the second line in Table 3. This completes the proof of Lemma 3.1 for the case  $n = 2$ .

From the results above, it follows that the first Betti numbers increase by a constant factor if the reduced 0<sup>th</sup> Betti numbers  $\tilde{\beta}^0$  of the full subcomplexes corresponding to  $\omega_i(\Delta)$  increase by a constant factor for  $1 \leq i \leq 15$ . Since the reduced Betti number  $\tilde{\beta}^0(K_{\omega_i(\Delta)})$  is obtained through the matrix  $X(2P, \omega_i(\Delta))$ , we only need to guarantee that matrices  $X(nP, \omega_i)$  for  $n = 2, 4, 6, \dots$  change with a certain pattern for all  $1 \leq i \leq 15$ . Notice that such a submatrix is completely determined by the adjacency matrix and the coloring of the polytope.

$i$	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Betti number
$\tilde{\beta}^0(K_{\omega_i}(\Delta))$	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	$\beta^1(M(2P, \delta)) = 3$
$\tilde{\beta}^0(K_{\omega_i}(\Delta^1))$	1	1	0	1	0	1	1	0	0	0	0	0	0	0	0	$\beta^1(M(4P, \delta^1)) = 5$
$\tilde{\beta}^0(K_{\omega_i}(\Delta^2))$	1	1	0	2	0	2	1	0	0	0	0	0	0	0	0	$\beta^1(M(6P, \delta^2)) = 7$
$\tilde{\beta}^0(K_{\omega_i}(\Delta^3))$	1	1	0	3	0	3	1	0	0	0	0	0	0	0	0	$\beta^1(M(8P, \delta^3)) = 9$
$\tilde{\beta}^0(K_{\omega_i}(\Delta^4))$	1	1	0	4	0	4	1	0	0	0	0	0	0	0	0	$\beta^1(M(10P, \delta^4)) = 11$
$\tilde{\beta}^0(K_{\omega_i}(\Delta^5))$	1	1	0	5	0	5	1	0	0	0	0	0	0	0	0	$\beta^1(M(12P, \delta^5)) = 13$

Table 3: The computation of the first Betti number.

As for the adjacency matrices, they do change in a uniform manner when using the facet ordering described in Section 2.3; see also Figure 22 in the online supplement for the facet ordering and adjacency matrix of the polytopes  $2P$ ,  $4P$  and  $6P$ .

As for the coloring, we duplicate the last two bricks of the colored  $2P$  a total of  $\frac{1}{2}n - 1$  times to construct the desired coloring on  $nP$ , where  $n$  is a positive even integer equal to or greater than 2. It can be easily proved that the nonsingularity condition holds at every vertex. The colorings constructed this way on polytopes  $4P$  and  $6P$  are shown in Figure 12, lower left and lower right, respectively. The colorings are denoted by

$$[a_1 S_1 \underline{S_2 S_1} \underline{S_2 S_1} a_1] \quad \text{and} \quad [a_1 S_1 \underline{S_2 S_1} \underline{S_2 S_1} \underline{S_2 S_1} a_1].$$

Their characteristic functions are written  $\lambda^1$  and  $\lambda^2$ , respectively, where the superscripts denote how many times the last two bricks ( $S_2 S_2$ ) of the coloring  $[a_1 S_1 \underline{S_2 S_1} a_1]$  of  $\lambda$  are repeated. The repeated parts are highlighted in blue and underlined. The nonorientability of these  $\mathbb{Z}_2^3$ -colorings is guaranteed by Corollary 2.9. Moreover, we can obtain their natural  $\mathbb{Z}_2^4$ -extensions  $\delta^1$  and  $\delta^2$ . By Proposition 2.11, the colorings  $\delta^1$  and  $\delta^2$  are admissible. That is,  $M(4P, \delta^1)$  and  $M(6P, \delta^2)$  are the orientable double covers of the nonorientable manifolds  $M(4P, \lambda^1)$  and  $M(6P, \lambda^2)$ , respectively. We denote the characteristic matrices of  $\delta^1$  and  $\delta^2$  by  $\Delta^1$  and  $\Delta^2$ . The three matrices row  $\Delta$ , row  $\Delta^1$  and row  $\Delta^2$  are shown in Figure 23 of the online supplement. Since the coloring on  $nP$  is obtained by duplicating the last two bricks of the coloring  $[a_1 S_1 \underline{S_2 S_1} a_1]$  on  $2P$  a total of  $\frac{1}{2}n - 1$  times, the row space row  $\Delta^i$  can be obtained from row space row  $\Delta$  by duplicating its columns, from the 11<sup>th</sup> to the second columns (counting from right to left),  $\frac{1}{2}n - 1$  times.

By the method outlined before, we also calculate  $\beta^1(M((2 + 2i)P, \delta^i))$  for  $i = 1, 2, \dots, 5$ , as shown in Table 3. We illustrate the calculation of  $\tilde{\beta}^0(K_{\omega_1}(\Delta^1))$  and  $\tilde{\beta}^0(K_{\omega_1}(\Delta^2))$  in Figures 13 and 14, respectively. See also panels (C)–(D) and (E)–(F)

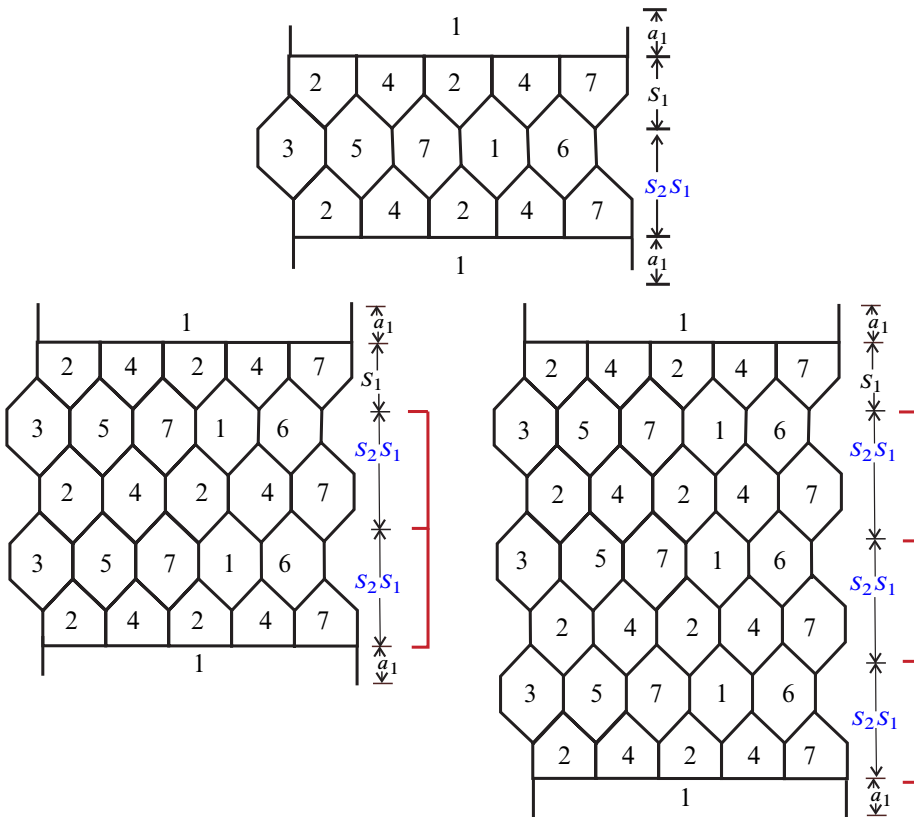


Figure 12: Top: The colored polytope  $2P$ . Bottom left: The colored polytope  $4P$ . Bottom right: The colored polytope  $6P$ . Duplicate the last two bricks of the coloring  $[a_1 S_1 \underline{S_2 S_1} a_1]$  on  $2P$  a total of  $\frac{1}{2}n - 1$  times to construct the desired coloring on  $nP$ .

in Figures 16–21 of the online supplement for the computation of  $\tilde{\beta}^0(K_{\omega_i}(\Delta^1))$  and  $\tilde{\beta}^0(K_{\omega_i}(\Delta^2))$  for  $i = 2, 3, \dots, 7$ . The corresponding results are highlighted in blue in Table 3.

From Figure 11 and Table 3 we can see that the matrices  $X(nP, \omega_i)$  for  $n = 2, 4, 6, \dots$  follow certain patterns for all  $1 \leq i \leq 15$ . In order to guarantee that the sequence  $\{\tilde{\beta}^0(K_{\omega_i}(\Delta^t))\}$  with  $t \in \mathbb{Z}_+$  is an arithmetic progression, we just need to guarantee that the first three items satisfy the relation of an arithmetic progression. For example, since  $\tilde{\beta}^0(K_{\omega_4}(\Delta)) = 0$ ,  $\tilde{\beta}^0(K_{\omega_4}(\Delta^1)) = 1$ ,  $\tilde{\beta}^0(K_{\omega_4}(\Delta^2)) = 2$  and the full subcomplex  $K_{\omega_4}(\Delta^t)$  changes regularly as the colorings are obtained by duplicating  $t$  times the last two bricks of the colored  $2P$  of  $[a_1 S_1 \underline{S_2 S_1} a_1]$ , it follows that  $\{\tilde{\beta}^0(K_{\omega_i}(\Delta^t))\}$  with



$\omega_1(\Delta^2)$		$\begin{matrix} 0 & 0 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 13 & 14 & 15 & 16 & 17 & 18 & 19 & 20 & 21 & 22 & 23 & 24 & 25 & 26 & 27 & 28 & 29 & 30 & 31 & 32 & 33 & 34 & 35 & 36 & 37 \end{matrix}$																																								
0	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

$\omega_1(\Delta^2)$		$\begin{matrix} 1 & 1 \end{matrix}$																																													
1	3	4	6	7	9	10	13	14	16	17	19	20	23	24	26	27	29	30	33	34	36	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

$\omega_1^{(1)}(\Delta^2) = \{3, 7, 9, 13, 17, 19, 23, 27, 29, 33\}$

$\omega_1^{(2)}(\Delta^2) = \{4, 6, 10, 14, 16, 20, 24, 26, 30, 34, 36\}$

$\tilde{\beta}^0(K_{\omega_1(\Delta^2)}) = 1$

Figure 14: The computation of  $\tilde{\beta}^0(K_{\omega_1(\Delta^2)})$ . Top:  $X(6P)$ . Bottom:  $X(6P, \omega_1(\Delta^2))$ .

$t \in \mathbb{Z}_+$  is an arithmetic progression. Namely,  $\tilde{\beta}^0(K_{\omega_4}(\Delta^3)) = 3$ ,  $\tilde{\beta}^0(K_{\omega_4}(\Delta^4)) = 4$ ,  $\tilde{\beta}^0(K_{\omega_4}(\Delta^5)) = 5, \dots$ . As a consequence, if we want to prove that the whole Betti number sequence  $\beta^1(M(nP, \delta^{\frac{1}{2}n}))$ , where  $n$  is an even positive integer, is an arithmetic progression, we only need to verify that  $\beta^1(M(4P, \delta^1)) - \beta^1(M(2P, \delta)) = \beta^1(M(6P, \delta^2)) - \beta^1(M(4P, \delta^1))$ . Summarizing all these findings, we have the following proposition:

**Proposition 3.2** *Let  $\delta$  be a  $\mathbb{Z}_2^4$ -coloring over the polytope  $nP$ . For an arbitrary even number  $s \geq n$ , if*

$$\begin{aligned} \beta^1(M((n+2)P, \delta^{(1)})) - \beta^1(M(nP, \delta)) \\ = \beta^1(M((n+4)P, \delta^{(2)})) - \beta^1(M((n+2)P, \delta^{(1)})), \end{aligned}$$

we have

$$\begin{aligned} \beta^1(M(sP, \delta^{\frac{1}{2}(s-n)})) \\ = \beta^1(M(nP, \delta)) + \frac{1}{2}(s-n)(\beta^1(M(n+1)P, \delta^1) - \beta^1(M(nP, \delta))), \end{aligned}$$

where  $\delta^{(t)}$  represents a  $\mathbb{Z}_2^4$ -coloring over the polytope  $(n+2t)P$ . The coloring vector of  $\delta^{(t)}$  is obtained by duplicating the last two bricks of  $\delta$  exactly  $t$  times.

By Proposition 3.2 and using the facts that  $\beta^1(M(2P, \delta)) = 3$ ,  $\beta^1(M(4P, \delta^1)) = 5$  and  $\beta^1(M(6P, \delta^2)) = 7$ , we can produce Table 4.

This concludes the proof of Lemma 3.1. □

## 4 Proof of Theorem 1.2 for $n$ even

In this section, we prove Theorem 1.2 when  $n$  is even. It is similar to the proof of Lemma 3.1.

**Lemma 4.1** *For any even positive number  $n$ , there is a nonorientable  $\mathbb{Z}_2^3$ -coloring  $\lambda$  over the polytope  $nP$ , such that, for its natural associated  $\mathbb{Z}_2^4$ -coloring  $\delta$ , we have  $\beta^1(M(nP, \delta)) = 5n - 3$ .*

**Proof** Let  $S_1 = (65372)$ ,  $S_2S_3 = (72424\ 65372)$  and  $a_1 = 1$ . By the same idea of Lemma 3.1, we first construct a suitable nonorientable  $\mathbb{Z}_2^3$ -coloring  $\lambda$  over the polytope  $2P$  as follows:

$$(1, 3, 5, 7, 6, 2, 4, 2, 2, 4, 7, 3, 5, 7, 6, 2, 1).$$

	$n = 2$	$n = 4$	$n = 6$	$\dots$	$n = 2 + 2t, t \in \mathbb{N}$
1	1	1	1	$\dots$	1
2	1	1	1	$\dots$	1
3	0	0	0	$\dots$	0
4	0	1	2	$\dots$	$t$
5	0	0	0	$\dots$	0
6	0	1	2	$\dots$	$t$
7	1	1	1	$\dots$	1
8	0	0	0	$\dots$	0
9	0	0	0	$\dots$	0
10	0	0	0	$\dots$	0
11	0	0	0	$\dots$	0
12	0	0	0	$\dots$	0
13	0	0	0	$\dots$	0
14	0	0	0	$\dots$	0
15	0	0	0	$\dots$	0
total $\beta^1$	3	5	7	$\dots$	$3 + 2t = n + 1$

Table 4: The values of  $\beta^1$  in Lemma 3.1.

This colored polytope  $2P$  is denoted by  $[a_1 S_1 \underline{S_2 S_1} a_1]$ . It follows from Corollary 2.9 and Proposition 2.11 that  $\lambda$  is nonorientable and admissible. Denote by  $\delta$  the natural  $\mathbb{Z}_2^4$ -extension of  $\lambda$ . The 3-manifold  $M(2P, \delta)$  is the orientable double cover of the nonorientable 3-manifold  $M(2P, \lambda)$ . By Corollary 2.3, we have  $\beta^1(M(2P, \delta)) = 7$ .

We repeat the last two bricks  $t$  times to construct a coloring over the polytope  $(2 + 2t)P$ , and denote its characteristic function by  $\lambda^t$ . In turn, the colored polytope  $(2 + 2t)P$  is denoted by

$$[a_1 S_1 \underbrace{\underline{S_2 S_1} \cdots \underline{S_2 S_1}}_{t \text{ pairs}} a_1].$$

It can be easily checked that the nonsingularity condition holds at every vertex. Likewise, by Corollary 2.9 and Proposition 2.11, we can obtain an admissible extension  $\delta^t$  of the nonorientable coloring  $\lambda^t$ . Moreover,  $M((2 + 2t)P, \delta^t)$  is the orientable double cover of the nonorientable manifold  $M((2 + 2t)P, \lambda^t)$ . The Betti numbers of  $(M(2P, \delta))$ ,  $(M(4P, \delta^1))$  and  $(M(6P, \delta^2))$  are shown in the second, third and fourth columns of Table 5. By Proposition 3.2 and using the facts that  $\beta^1(M(2P, \delta)) = 7$ ,  $\beta^1(M(4P, \delta^1)) = 17$  and  $\beta^1(M(6P, \delta^2)) = 27$ , we can deduce the last column of Table 5.



	$n = 2$	$n = 4$	$n = 6$	$\dots$	$n = 2 + 2t, t \in \mathbb{N}$
1	0	0	0	$\dots$	0
2	0	0	0	$\dots$	0
3	2	4	6	$\dots$	$2t + 2$
4	1	2	3	$\dots$	$t + 1$
5	0	0	0	$\dots$	0
6	1	3	5	$\dots$	$2t + 1$
7	0	0	0	$\dots$	0
8	1	2	3	$\dots$	$t + 1$
9	0	0	0	$\dots$	0
10	0	1	2	$\dots$	$t$
11	0	0	0	$\dots$	0
12	1	3	5	$\dots$	$2t + 1$
13	1	2	3	$\dots$	$t + 1$
14	0	0	0	$\dots$	0
15	0	0	0	$\dots$	0
total $\beta^1$	7	17	27	$\dots$	$10t + 7 = 5n - 3$

Table 5: The values of  $\beta^1$  for Lemma 4.1.

In other words, we may always find a nonorientable  $\mathbb{Z}_2^3$ -coloring  $\lambda$  such that its natural  $\mathbb{Z}_2^4$ -extension  $\delta$  has  $\beta^1(M(nP, \delta)) = 5n - 3$ . □

**Lemma 4.2** *For any even positive integer  $n$  and any odd integer  $k \in [5n - 1, 5n + 3]$ , there is a nonorientable  $\mathbb{Z}_2^3$ -coloring  $\lambda$  over the polytope  $nP$  such that, for its natural associated  $\mathbb{Z}_2^4$ -coloring  $\delta$ , we have  $\beta^1(M(nP, \delta)) = k$ .*

**Proof** We start at  $n = 2$  and construct suitable characteristic functions of the desired manifolds, whose first Betti numbers increase by  $10t$  when the last pair of their coloring bricks are repeated  $t$  times. First, in Table 6 we prepare an affix and some bricks for constructing the coloring vectors needed.

Let  $\lambda_1^0, \lambda_1^1$  and  $\lambda_1^2$  be the three nonorientable  $\mathbb{Z}_2^3$ -coloring characteristic functions of the coloring vectors

$$[a_1 S_1 \underline{S_2 S_1} a_1], \quad [a_1 S_1 \underline{S_2 S_1 S_2 S_1} a_1], \quad [a_1 S_1 \underline{S_2 S_1 S_2 S_1 S_2 S_1} a_1]$$

over the polytopes  $2P, 4P$  and  $6P$ , respectively. Their characteristic vectors are

- $(1, 2, 4, 4, 3, 6, \underline{5, 1, 6, 3, 2, 2, 4, 4, 3, 6}, 1),$
- $(1, 2, 4, 4, 3, 6, \underline{5, 1, 6, 3, 2, 2, 4, 4, 3, 6}, \underline{5, 1, 6, 3, 2, 2, 4, 4, 3, 6}, 1),$
- $(1, 2, 4, 4, 3, 6, \underline{5, 1, 6, 3, 2, 2, 4, 4, 3, 6}, \underline{5, 1, 6, 3, 2, 2, 4, 4, 3, 6}, \underline{5, 1, 6, 3, 2, 2, 4, 4, 3, 6}, 1).$

to construct	affixes	brick	pair of bricks being repeated
$\lambda_1^t$	$a_1 = 1$	$S_1 = 34246$	$S_2 S_1 = (26513\ 34246)$
$\lambda_2^t, \lambda_3^t$	$a_1 = 1$ $a_2 = 3$ $a_3 = 7$	$S_1 = (24246)$	$S_2 S_3 = (73153\ 14245)$

Table 6: The affixes and bricks for constructing  $\lambda_1^t, \lambda_2^t$  and  $\lambda_3^t$  of Lemma 4.2.

It can be easily checked that the nonsingularity condition holds at every vertex. The natural associated  $\mathbb{Z}_2^4$ -extensions are denoted by  $\delta_1^0, \delta_1^1$  and  $\delta_1^2$ . By Corollary 2.3, we can calculate the first Betti numbers of those manifolds, namely  $\beta^1(M(2P, \delta_1^0)) = 13$ ,  $\beta^1(M(4P, \delta_1^1)) = 23$  and  $\beta^1(M(6P, \delta_1^2)) = 33$ . Thus, according to Proposition 3.2,

$$(4-1) \quad \beta^1(M((2 + 2t)P, \delta_1^t)) = 13 + 10t, \quad \text{where } t \in \mathbb{Z}_+.$$

Similarly, we describe the affixes and bricks for constructing  $\lambda_2^t$  and  $\lambda_3^t$  of Lemma 4.2 in Table 6.

Let us denote by  $\lambda_2^0, \lambda_2^1$  and  $\lambda_2^2$  the three nonorientable  $\mathbb{Z}_2^3$ -coloring characteristic functions of the following colored polytopes  $2P, 4P$  and  $6P$ :

$$[a_1 S_1 \underline{S_2 S_3} a_2], \quad [a_1 S_1 \underline{S_2 S_3 S_2 S_3} a_2], \quad [a_1 S_1 \underline{S_2 S_3 S_2 S_3 S_2 S_3} a_2],$$

and let  $\lambda_3^0, \lambda_3^1$  and  $\lambda_3^2$  be the  $\mathbb{Z}_2^3$ -coloring characteristic functions of the following colored polytopes  $2P, 4P$  and  $6P$ :

$$[a_1 S_1 \underline{S_2 S_3} a_3], \quad [a_1 S_1 \underline{S_2 S_3 S_2 S_3} a_3], \quad [a_1 S_1 \underline{S_2 S_3 S_2 S_3 S_2 S_3} a_3].$$

Their natural associated  $\mathbb{Z}_2^4$ -extensions are denoted as  $\delta_2^0, \delta_2^1, \delta_2^2$  and  $\delta_3^0, \delta_3^1, \delta_3^2$ . The first Betti numbers of these manifolds, namely  $\beta^1(M(2P, \delta_2^0)), \beta^1(M(4P, \delta_2^1)), \beta^1(M(6P, \delta_2^2))$  and  $\beta^1(M(2P, \delta_3^0)), \beta^1(M(4P, \delta_3^1)), \beta^1(M(6P, \delta_3^2))$ , are explicitly calculated to be 15, 25, 35 and 17, 27, 37, respectively.

Thus we have, for each  $t \in \mathbb{Z}_+$ ,

$$(4-2) \quad \beta^1(M((2 + 2t)P, \delta_2^t)) = 15 + 10t,$$

$$(4-3) \quad \beta^1(M((2 + 2t)P, \delta_3^t)) = 17 + 10t.$$

Putting together the results in (4-1), (4-2) and (4-3), we have the proof of Lemma 4.2.  $\square$

**Lemma 4.3** *For any even positive integer  $n$  and any odd integer  $k \in [1, n - 1]$ , there is a nonorientable  $\mathbb{Z}_2^3$ -coloring  $\lambda$  over the polytope  $nP$  such that, for its natural associated  $\mathbb{Z}_2^4$ -coloring  $\delta$ , we have  $\beta^1(M(nP, \delta)) = k$ .*

affixes	bricks	compatible pairs of bricks being repeated
$a_1 = 1$ $a_2 = 3$	$S_1 = (24247)$ $S_2 = (54241)$ $S_3 = (67172)$	$A_1 = (6717254241)$ $A_2 = (7317254241)$

Table 7: The affixes, bricks and compatible pairs for  $\lambda^{(t_1, t_2)}$  of Lemma 4.3.

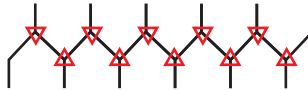


Figure 15: Compatible pair.

**Proof** We consider some affixes and bricks as described in Table 7. For the sake of brevity, we use the symbol  $A_i$  to denote a *compatible pair* of bricks, where “compatible” means the nonsingular condition is satisfied at all ten intersecting vertices of the two bricks as shown in Figure 15.

At first, we construct a nonorientable  $\mathbb{Z}_2^3$ -coloring  $\lambda$  over the polytope  $2P$ , where the colored polytope is  $[a_1 S_1 S_3 S_2 a_2]$ . The nonorientability is guaranteed by Corollary 2.9. The natural  $\mathbb{Z}_2^4$ -extension of  $\lambda$  is denoted by  $\delta$ . Let  $\lambda^{(t_1, t_2)}$  be the  $\mathbb{Z}_2^3$ -coloring characteristic function of the colored polytope  $2(t_1 + t_2 + 1)P$ ,

$$[a S_1 S_3 S_2 \underbrace{A_1, \dots, A_1}_{t_1} \underbrace{A_2, \dots, A_2}_{t_2} a_2].$$

It can be easily checked that the nonsingularity condition holds at every vertex. Moreover,  $\delta^{(t_1, t_2)}$  is the natural  $\mathbb{Z}_2^4$ -extension of  $\lambda^{(t_1, t_2)}$ , which is also defined on the polytope  $2(t_1 + t_2 + 1)P$ . In particular,  $\lambda^{(0,0)} = \lambda$ . The colored  $2Ps$  corresponding to  $\lambda^{(1,0)}$  and  $\lambda^{(0,1)}$  are  $[a_1 S_1 S_3 S_2 A_1 a_2]$  and  $[a_1 S_1 S_3 S_2 A_2 a_2]$ , respectively. In this case, the nonsingularity condition holds at every vertex. The calculated Betti numbers are given in Table 8.

By Proposition 3.2 and

$$\beta^1(M(2P, \delta^{(0,0)})) = 1, \quad \beta^1(M(4P, \delta^{(1,0)})) = 1, \quad \beta^1(M(6P, \delta^{(2,0)})) = 1,$$

$\beta^1(M(2P, \delta^{(0,0)})) = 1$	$\beta^1(M(4P, \delta^{(1,0)})) = 1$	$\beta^1(M(6P, \delta^{(2,0)})) = 1$	...
	$\beta^1(M(4P, \delta^{(0,1)})) = 3$	$\beta^1(M(6P, \delta^{(1,1)})) = 3$	...
		$\beta^1(M(6P, \delta^{(0,2)})) = 5$	...

Table 8: The values of  $\beta^1(M((2(t_1 + t_2 + 2))P, \delta^{(t_1, t_2)}))$  in Lemma 4.3.

we have

$$(4-4) \quad \beta^1(M(2(t_1 + t_2 + 1)P, \delta^{(t_1, t_2)})) = \beta^1(M(2(t_1 + t_2 + 2)P, \delta^{(t_1+1, t_2)})).$$

Likewise, from

$$\beta^1(M(2P, \delta^{(0,0)})) = 1, \quad \beta^1(M(4P, \delta^{(0,1)})) = 3, \quad \beta^1(M(6P, \delta^{(0,2)})) = 5,$$

we have

$$(4-5) \quad \beta^1(M(2(t_1 + t_2 + 1)P, \delta^{(t_1, t_2)})) + 2 = \beta^1(M(2(t_1 + t_2 + 2)P, \delta^{(t_1, t_2+1)})).$$

By (4-4) and (4-5), we obtain

$$(4-6) \quad \beta^1(M(nP, \delta^{(t, \frac{1}{2}n-1-t)})) = n - 2t - 1,$$

where  $n$  is even and  $0 \leq t \leq \frac{1}{2}n - 1$ , which completes the proof of Lemma 4.3.  $\square$

**Lemma 4.4** For any even positive integer  $n$  and any odd integer  $k \in [n + 3, 5n - 5]$ , there is a nonorientable  $\mathbb{Z}_2^3$ -coloring  $\lambda$  over the polytope  $nP$  such that, for its natural associated  $\mathbb{Z}_2^4$ -coloring  $\delta$ , we have  $\beta^1(M(nP, \delta)) = k$ .

**Proof** The considered affixes and bricks are described in Table 9.

First, we construct three nonorientable  $\mathbb{Z}_2^3$ -coloring characteristic functions  $\tilde{\lambda}^0, \tilde{\lambda}^1$  and  $\tilde{\lambda}^2$  of polytopes  $2P, 4P$  and  $6P$ , respectively as below:

$$\begin{aligned} & [a_1 S_1 A_3 a_1], \\ & [a_1 S_1 A_3 A_3 a_1], \\ & [a_1 S_1 A_3 A_3 A_3 a_1]. \end{aligned}$$

Their characteristic vectors are

$$\begin{aligned} & (1, 2, 4, 4, 2, 7, \underline{3}, 7, 5, 2, 6, 2, 4, 4, 2, 7, 1), \\ & (1, 2, 4, 4, 2, 7, \underline{3}, 7, 5, 2, 6, 2, 4, 4, 2, 7, \underline{3}, 7, 5, 2, 6, 2, 4, 4, 2, 7, 1), \\ & (1, 2, 4, 4, 2, 7, \underline{3}, \underline{3}, 7, 5, 2, 6, 2, 4, 4, 2, 7, \underline{3}, 7, 5, 2, 6, 2, 4, 4, 2, 7, \underline{3}, 7, 5, 2, 6, 2, 4, 4, 2, 7, 1). \end{aligned}$$

affixes	brick	compatible pairs of bricks being repeated
$a_1 = 1, a_2 = 4$	$S_1 = (24247)$	$A_1 = (42472 \ 71635)$ $A_2 = (42472 \ 37265)$ $A_3 = (65372 \ 24247)$ $A_4 = (65372 \ 71635)$

Table 9: The affixes, brick and compatible pairs for constructing  $\lambda_i^{(t_1, t_2)}$  of Lemma 4.4.

Also in this case, the nonsingularity condition holds at every vertex. Their natural associated  $\mathbb{Z}_2^4$ -colorings are denoted by  $\tilde{\delta}^0, \tilde{\delta}^1$  and  $\tilde{\delta}^2$ . By Corollary 2.3, we obtain that the first Betti numbers of the corresponding manifolds are 5, 15 and 25, respectively. Thus, we have

$$(4-7) \quad \beta^1(M((2 + 2t)P, \tilde{\delta}^t)) = 5 + 10t$$

for each  $t \in \mathbb{Z}_{\geq 0}$ , where  $t$  is the number of times the last two bricks of  $\tilde{\delta}^0$  are repeated.

Next, we use  $\lambda_i^{(t_1, t_2)}$  to represent the  $\mathbb{Z}_2^3$ -coloring characteristic function of coloring vector

$$[aS_1 \underbrace{A_3, \dots, A_3}_{t_1} A_4 \underbrace{A_1, \dots, A_1}_{t_2} A_i a_j]$$

over the polytope  $2(t_1 + t_2 + 2)P$ . Here  $a_j$  is the affix element and  $j = 2, 1, 1, 2$  when  $i = 1, 2, 3, 4$ , respectively. In particular, the coloring vector of  $\lambda_i^{(0,0)}$  is  $[aS_1 A_4 A_i a_j]$ . The nonsingularity condition holds at every vertex. Moreover,  $\delta_i^{(t_1, t_2)}$  is the natural associated  $\mathbb{Z}_2^4$ -extension of  $\lambda_i^{(t_1, t_2)}$ .

From

$$\begin{aligned} \beta^1(M(4P, \delta_i^{(0,0)})) &= 5 + 2i, \\ \beta^1(M(6P, \delta_i^{(0,1)})) &= 7 + 2i, \\ \beta^1(M(8P, \delta_i^{(0,2)})) &= 9 + 2i \end{aligned}$$

for  $i = 1, 2, 3, 4$ , we have

$$(4-8) \quad \beta^1(M(2(t_1 + t_2 + 2)P, \delta_i^{(t_1, t_2)})) + 2 = \beta^1(M(2(t_1 + t_2 + 3)P, \delta_i^{(t_1, t_2+1)}))$$

for  $i = 1, 2, 3, 4$ . From

$$\begin{aligned} \beta^1(M(4P, \delta_i^{(0,0)})) &= 5 + 2i, \\ \beta^1(M(6P, \delta_i^{(1,0)})) &= 15 + 2i, \\ \beta^1(M(8P, \delta_i^{(2,0)})) &= 25 + 2i \end{aligned}$$

for  $i = 1, 2, 3, 4$ , we have

$$(4-9) \quad \beta^1(M(2(t_1 + t_2 + 2)P, \delta_i^{(t_1, t_2)})) + 10 = \beta^1(M(2(t_1 + t_2 + 3)P, \delta_i^{(t_1+1, t_2)}))$$

By (4-8) and (4-9) it follows that

$$(4-10) \quad \beta^1(M(nP, \delta_i^{(t, \frac{1}{2}n-2-t)})) = n + 8t + 2i + 3$$

for  $n$  even and  $0 \leq t \leq \frac{1}{2}n - 2$ .

2P		4P		6P		8P		10P		...
$\delta$	$\beta^1$	$\delta$	$\beta^1$	$\delta$	$\beta^1$	$\delta$	$\beta^1$	$\delta$	$\beta^1$	...
$\tilde{\delta}^0$	5	$\delta_1^{(0,0)}$	7	$\delta_1^{(0,1)}$	9	$\delta_1^{(0,2)}$	11	$\delta_1^{(0,3)}$	13	...
		$\delta_2^{(0,0)}$	9	$\delta_2^{(0,1)}$	11	$\delta_2^{(0,2)}$	13	$\delta_2^{(0,3)}$	15	...
		$\delta_3^{(0,0)}$	11	$\delta_3^{(0,1)}$	13	$\delta_3^{(0,2)}$	15	$\delta_3^{(0,3)}$	17	...
		$\delta_4^{(0,0)}$	13	$\delta_4^{(0,1)}$	15	$\delta_4^{(0,2)}$	17	$\delta_4^{(0,3)}$	19	...
	$\tilde{\delta}^1$	15	$\delta_1^{(1,0)}$	17	$\delta_1^{(1,1)}$	19	$\delta_1^{(1,2)}$	21	...	
			$\delta_2^{(1,0)}$	19	$\delta_2^{(1,1)}$	21	$\delta_2^{(1,2)}$	23	...	
			$\delta_3^{(1,0)}$	21	$\delta_3^{(1,1)}$	23	$\delta_3^{(1,2)}$	25	...	
			$\delta_4^{(1,0)}$	23	$\delta_4^{(1,1)}$	25	$\delta_4^{(1,2)}$	27	...	
			$\tilde{\delta}^2$	25	$\delta_1^{(2,0)}$	27	$\delta_1^{(2,1)}$	29	...	
					$\delta_2^{(2,0)}$	29	$\delta_2^{(2,1)}$	31	...	
					$\delta_3^{(2,0)}$	31	$\delta_3^{(2,1)}$	33	...	
					$\delta_4^{(2,0)}$	33	$\delta_4^{(2,1)}$	35	...	
					$\tilde{\delta}^3$	35	$\delta_1^{(3,0)}$	37	...	
							$\delta_2^{(3,0)}$	39	...	
							$\delta_3^{(3,0)}$	41	...	
							$\delta_4^{(3,0)}$	43	...	
					$\tilde{\delta}^4$	45	...	...		

Table 10: The values of  $\beta^1 = \beta^1(M(nP, \delta))$  for Lemma 4.4.

By (4-7) and (4-10), we finish the proof of Lemma 4.4. All of the Betti numbers of Lemma 4.4 are listed in Table 10. □

Now, using Lemmas 3.1 and 4.1–4.4, we complete the proof of Theorem 1.2 for  $n$  even.

### 5 Proof of Theorem 1.2 for $n$ odd

In this section, we analogously prove Theorem 1.2 for odd  $n$ .

**Lemma 5.1** *For any odd positive integer  $n$ , there is a nonorientable  $\mathbb{Z}_2^3$ -coloring  $\lambda$  over the polytope  $nP$  such that, for its natural associated  $\mathbb{Z}_2^4$ -coloring  $\delta$ , we have  $\beta^1(M(nP, \delta)) = n$ .*

	$n = 3$	$n = 5$	$n = 7$	$\dots$	$n = 3 + 2t, t \in \mathbb{N}$
1	1	1	1	$\dots$	1
2	1	1	1	$\dots$	1
3	0	0	0	$\dots$	0
4	0	1	2	$\dots$	$t$
5	0	0	0	$\dots$	0
6	0	1	2	$\dots$	$t$
7	1	1	1	$\dots$	1
8	0	0	0	$\dots$	0
9	0	0	0	$\dots$	0
10	0	0	0	$\dots$	0
11	0	0	0	$\dots$	0
12	0	0	0	$\dots$	0
13	0	0	0	$\dots$	0
14	0	0	0	$\dots$	0
15	0	0	0	$\dots$	0
total $\beta^1$	3	5	7	$\dots$	$3 + 2t = n$

Table 11: The values of  $\beta^1(M(nP, \delta^t))$  for  $n = 3 + 2t$  in Lemma 5.1.

**Proof** We first prove the special case in which  $n = 3$ . Consider bricks  $S_1 = (24247)$  and  $S_2 = (35716)$ , and affixes  $a_1 = 1$  and  $a_2 = 4$ . We construct a nonorientable  $\mathbb{Z}_2^3$ -coloring  $\lambda$  over the polytope  $3P$  whose coloring and characteristic vector are

$$[a_1 S_1 S_2 S_1 S_2 a_2] \quad \text{and} \quad (1, 2, 4, 4, 2, 7, 7, 1, 5, 6, 3, 2, 4, 4, 2, 7, 7, 1, 5, 6, 3, 4),$$

respectively.

By Corollary 2.3,  $\beta^1(M(3P, \delta)) = 3$ , where  $\delta$  is the natural  $\mathbb{Z}_2^4$ -extension of  $\lambda$ . We repeat the last two bricks  $t$  times to construct a coloring over the polytope  $(3 + 2t)P$ , and denote its characteristic function by  $\lambda^t$ . It can be easily checked that the nonsingularity condition holds at every vertex. By Corollary 2.9 and Proposition 2.11, we obtain the admissible extension  $\delta^t$  of the nonorientable  $\lambda^t$ . That is,  $M((3 + 2t)P, \delta^t)$  is the orientable double cover of the nonorientable manifold  $M((3 + 2t)P, \lambda^t)$ . The progressions of corresponding Betti numbers are shown in Table 11.

This concludes the proof of Lemma 5.1. □

**Lemma 5.2** *For any odd positive integer  $n$  and any odd integer  $k \in [5n - 9, 5n + 3]$ , there is a nonorientable  $\mathbb{Z}_2^3$ -coloring  $\lambda$  over the polytope  $nP$  such that, for its natural associated  $\mathbb{Z}_2^4$ -coloring  $\delta$ , we have  $\beta^1(M(nP, \delta)) = k$ .*

affix	compatible pairs of bricks being repeated
$a_1 = 1$	$A_1 = (53726\ 71635)$ $A_2 = (24724\ 37265)$ $A_3 = (53726\ 74242)$

Table 12: An affix and compatible pairs of bricks for Lemma 5.2, I.

**Proof** We start at  $n = 3$  and construct six suitable characteristic vectors whose corresponding manifolds' Betti numbers would increase by  $10t$  when repeating the last pair of coloring bricks  $t$  times. An affix and some useful compatible pairs are described in Table 12.

For every  $i = 0, 1, 2$ , let  $\lambda_i^0, \lambda_i^1$  and  $\lambda_i^2$  be the three  $\mathbb{Z}_2^3$ -coloring characteristic functions of the three colorings over the polytopes  $3P, 5P$  and  $7P$  as shown in Table 13. Here  $t$  represents how many times the last compatible pair of  $\lambda_i^0$  is repeated. It can be checked with little effort that the nonsingularity condition holds at every vertex.

Let  $\delta_i^t$  be the natural  $\mathbb{Z}_2^4$ -extensions of  $\lambda_i^t$  for  $i = 0, 1, 2$ . By Corollary 2.3, we may calculate the first Betti numbers of the manifolds corresponding to the coloring vectors in Table 13, namely

$$\beta^1(M(3P, \delta_0^0)) = 7, \quad \beta^1(M(5P, \delta_0^1)) = 17, \quad \beta^1(M(7P, \delta_0^2)) = 27,$$

$$\beta^1(M(3P, \delta_1^0)) = 9, \quad \beta^1(M(5P, \delta_1^1)) = 19, \quad \beta^1(M(7P, \delta_1^2)) = 29$$

and

$$\beta^1(M(3P, \delta_2^0)) = 11, \quad \beta^1(M(5P, \delta_2^1)) = 21, \quad \beta^1(M(7P, \delta_2^2)) = 31.$$

Therefore, according to Proposition 3.2, for each  $t \in \mathbb{N}$ ,

(5-1)  $\beta^1(M((3 + 2t)P, \delta_1^t)) = 7 + 10t,$

(5-2)  $\beta^1(M((3 + 2t)P, \delta_2^t)) = 9 + 10t,$

(5-3)  $\beta^1(M((3 + 2t)P, \delta_3^t)) = 11 + 10t.$

$i$	$t = 0$	1	2
0	$[a_1 A_1 A_2 a_1]$	$[a_1 A_1 A_2 A_2 a_1]$	$[a_1 A_1 A_2 A_2 A_2 a_1]$
1	$[a_1 A_1 A_3 a_1]$	$[a_1 A_1 A_3 A_3 a_1]$	$[a_1 A_1 A_3 A_3 A_3 a_1]$
2	$[a_1 A_1 A_1 a_1]$	$[a_1 A_1 A_1 A_1 a_1]$	$[a_1 A_1 A_1 A_1 A_1 a_1]$

Table 13: The coloring vectors of  $\lambda_i^t$  in Lemma 5.2.



affixes	compatible pairs of bricks being repeated
$a_1 = 1, a_2 = 3$	$A_0 = (34246\ 26513)$ $A_1 = (31245\ 26416)$ $A_2 = (31245\ 16416)$ $A_3 = (31245\ 46452)$

Table 14: The affixes and compatible pairs for Lemma 5.2, II.

Similarly, we prepare the affixes and compatible pairs for constructing the desired characteristic function  $\tilde{\lambda}_i^t$  in Table 14.

For every  $i = 0, 1, 2$ , let  $\tilde{\lambda}_i^0, \tilde{\lambda}_i^1$  and  $\tilde{\lambda}_i^2$  be the three  $\mathbb{Z}_2^3$ -coloring characteristic functions of the three colorings over the polytopes  $3P, 5P$  and  $7P$  as shown in Table 15. Here  $t$  represents how many times the last compatible pair of  $\tilde{\lambda}_i^0$  is repeated. It can be easily checked that the nonsingularity condition holds at every vertex.

Let  $\tilde{\delta}_i^t$  be the natural  $\mathbb{Z}_2^4$ -extensions of  $\tilde{\lambda}_i^t$ , for  $i = 1, 2, 3$ . By Corollary 2.3, we calculate the first Betti numbers of the manifolds corresponding to the coloring vectors in Table 15, namely

$$\beta^1(M(3P, \tilde{\delta}_0^0)) = 13, \quad \beta^1(M(5P, \tilde{\delta}_0^1)) = 23, \quad \beta^1(M(7P, \tilde{\delta}_0^2)) = 33,$$

$$\beta^1(M(3P, \tilde{\delta}_1^0)) = 15, \quad \beta^1(M(5P, \tilde{\delta}_1^1)) = 25, \quad \beta^1(M(7P, \tilde{\delta}_1^2)) = 35,$$

and

$$\beta^1(M(3P, \tilde{\delta}_2^0)) = 17, \quad \beta^1(M(5P, \tilde{\delta}_2^1)) = 27, \quad \beta^1(M(7P, \tilde{\delta}_2^2)) = 37.$$

Thus, according to Proposition 3.2, for each  $t \in \mathbb{N}$ ,

$$(5-4) \quad \beta^1(M((3 + 2t)P, \tilde{\delta}_1^t)) = 13 + 10t,$$

$$(5-5) \quad \beta^1(M((3 + 2t)P, \tilde{\delta}_2^t)) = 15 + 10t,$$

$$(5-6) \quad \beta^1(M((3 + 2t)P, \tilde{\delta}_3^t)) = 17 + 10t.$$

Putting together the results in (5-1)–(5-6), we have the proof of Lemma 5.2. □

$i$	$t = 0$	1	2
0	$[a_1 A_0 \mathbf{A_1} a_2]$	$[a_1 A_0 \mathbf{A_1 A_1} a_2]$	$[a_1 A_0 \mathbf{A_1 A_1 A_1} a_2]$
1	$[a_1 A_0 \mathbf{A_2} a_2]$	$[a_1 A_0 \mathbf{A_2 A_2} a_2]$	$[a_1 A_0 \mathbf{A_2 A_2 A_2} a_2]$
2	$[a_1 A_0 \mathbf{A_3} a_2]$	$[a_1 A_0 \mathbf{A_3 A_3} a_2]$	$[a_1 A_0 \mathbf{A_3 A_3 A_3} a_2]$

Table 15: The coloring vectors of  $\tilde{\lambda}_i^t$  in Lemma 5.2.

affixes	compatible pairs of bricks being repeated
$a_1 = 1, a_2 = 4$	$A_1 = (24247\ 17532)$ $A_2 = (53176\ 17532)$ $A_3 = (53147\ 17532)$

Table 16: The affixes and compatible pairs for  $\lambda^{(t_1, t_2)}$  of Lemma 5.3.

**Lemma 5.3** *For any odd positive integer  $n$  and any odd integer  $k \in [1, n - 1]$ , there is a nonorientable  $\mathbb{Z}_2^3$ -coloring  $\lambda$  over the polytope  $nP$  such that, for its natural associated  $\mathbb{Z}_2^4$ -coloring  $\delta$ , we have  $\beta^1(M(nP, \delta)) = k$ .*

**Proof** We prepare some affixes and compatible pairs as described in Table 16.

At first, we construct a nonorientable  $\mathbb{Z}_2^3$ -coloring characteristic function  $\lambda$ , whose coloring vector is  $[a_1 A_1 A_2 a_2]$ , on the polytope  $3P$ , and denote its natural  $\mathbb{Z}_2^4$ -extension by  $\delta$ . Let  $\lambda^{(t_1, t_2)}$  be the  $\mathbb{Z}_2^3$ -coloring characteristic function of

$$[a_1 A_1 A_2 \underbrace{A_2, \dots, A_2}_{t_1} \underbrace{A_3, \dots, A_3}_{t_2} a_2]$$

over the polytope  $(2(t_1 + t_2) + 3)P$ . We use  $\delta^{(t_1, t_2)}$  to denote the natural associated  $\mathbb{Z}_2^4$ -extension of  $\lambda^{(t_1, t_2)}$ . In particular,  $\lambda^{(0,0)} = \lambda$ . It can be easily checked that the nonsingularity condition holds at every vertex. The results of the calculations of the Betti numbers are reported in Table 17.

According to Proposition 3.2, the Betti number sequence would be an arithmetic progression if the first three numbers satisfy the relation of arithmetic progression.

From

$$\beta^1(M(3P, \delta^{(0,0)})) = 1, \quad \beta^1(M(5P, \delta^{(1,0)})) = 1, \quad \beta^1(M(7P, \delta^{(2,0)})) = 1,$$

we have

$$(5-7) \quad \beta^1(M((2(t_1 + t_2) + 3)P, \delta^{(t_1, t_2)})) = \beta^1(M((2(t_1 + t_2) + 5)P, \delta^{(t_1+1, t_2)})).$$

$\beta^1(M(3P, \delta^{(0,0)})) = 1$	$\beta^1(M(5P, \delta^{(1,0)})) = 1$	$\beta^1(M(7P, \delta^{(2,0)})) = 1$	...
	$\beta^1(M(5P, \delta^{(0,1)})) = 3$	$\beta^1(M(7P, \delta^{(1,1)})) = 3$	...
		$\beta^1(M(7P, \delta^{(0,2)})) = 5$	...

Table 17: The values of  $\beta^1(M((2(t_1 + t_2) + 3)P, \delta^{(t_1, t_2)}))$  in Lemma 5.3.

From

$$\beta^1(M(3P, \delta^{(0,0)})) = 1, \quad \beta^1(M(5P, \delta^{(0,1)})) = 3, \quad \beta^1(M(7P, \delta^{(0,2)})) = 5,$$

we have

$$(5-8) \quad \beta^1(M((2(t_1+t_2)+3)P, \delta^{(t_1,t_2)})) + 2 = \beta^1(M((2(t_1+t_2)+5)P, \delta^{(t_1,t_2+1)})).$$

By (5-7) and (5-8), we obtain

$$(5-9) \quad \beta^1(M(nP, \delta^{(t, \frac{1}{2}(n-3)-t)})) = n - 2 - 2t,$$

for each  $n$  odd with  $n \in \mathbb{Z}_{\geq 3}$  and  $0 \leq t \leq \frac{1}{2}(n - 3)$ .

This concludes the proof of Lemma 5.3. □

**Lemma 5.4** *For any odd positive integer  $n$  and any odd integer  $k \in [n + 1, 5n - 9]$ , there is a nonorientable  $\mathbb{Z}_2^3$ -coloring  $\lambda$  over the polytope  $nP$  such that, for the natural associated  $\mathbb{Z}_2^4$ -coloring  $\delta$ , we have  $\beta^1(M(nP, \delta)) = k$ .*

**Proof** The affixes and compatible pairs of bricks considered are described in Table 18.

At first, we construct a nonorientable  $\mathbb{Z}_2^3$ -coloring  $\lambda$  over the polytope  $3P$  whose coloring vector is  $[a_1 A_4 A_1 a_2]$ . We denote by  $\delta$  the natural associated  $\mathbb{Z}_2^4$ -extension. By calculation, we have

$$(5-10) \quad \beta^1(M(3P, \delta)) = 5.$$

We denote by  $\lambda_i^{t-1}$ , where  $t \in \mathbb{Z}_{\geq 1}$  and  $i = 1, 2, 3, 4$ , the nonorientable  $\mathbb{Z}_2^3$ -coloring characteristic function  $\lambda$  on the polytope  $(2t + 3)P$  corresponding to coloring vector

$$[a_1 A_4 \underbrace{A_1, \dots, A_1}_t A_i a_j],$$

where  $a_j$  is an affix element and  $j$  is given by 2, 1, 1, 2 for  $i = 1, 2, 3, 4$ , respectively.

affixes	compatible pairs of bricks being repeated
$a_1 = 1, a_2 = 4$	$A_1 = (42472 \ 57163)$ $A_2 = (42472 \ 53726)$ $A_3 = (65372 \ 72424)$ $A_4 = (65372 \ 57163)$

Table 18: The affixes and compatible pairs for Lemma 5.4.

3P		5P		7P		9P		11P		...
$\delta$	$\beta^1$	$\delta$	$\beta^1$	$\delta$	$\beta^1$	$\delta$	$\beta^1$	$\delta$	$\beta^1$	...
$\tilde{\delta}$	5	$\delta_1^0$	7	$\delta_1^1$	9	$\delta_1^2$	11	$\delta_1^3$	13	...
$\tilde{\delta}^0$	5	$\delta_2^0$	9	$\delta_2^1$	11	$\delta_2^2$	13	$\delta_2^3$	15	...
		$\delta_3^0$	11	$\delta_3^1$	13	$\delta_3^2$	15	$\delta_3^3$	17	...
		$\delta_4^0$	13	$\delta_4^1$	15	$\delta_4^2$	17	$\delta_4^3$	19	...
				$\delta_1^{(0,0)}$	17	$\delta_1^{(0,1)}$	19	$\delta_1^{(0,2)}$	21	...
		$\tilde{\delta}^1$	15	$\delta_2^{(0,0)}$	19	$\delta_2^{(0,1)}$	21	$\delta_2^{(0,2)}$	23	...
				$\delta_3^{(0,0)}$	21	$\delta_3^{(0,1)}$	23	$\delta_3^{(0,2)}$	25	...
				$\delta_4^{(0,0)}$	23	$\delta_4^{(0,1)}$	25	$\delta_4^{(0,2)}$	27	...
						$\delta_1^{(1,0)}$	27	$\delta_1^{(1,1)}$	29	...
				$\tilde{\delta}^2$	25	$\delta_2^{(1,0)}$	29	$\delta_2^{(1,1)}$	31	...
						$\delta_3^{(1,0)}$	31	$\delta_3^{(1,1)}$	33	...
						$\delta_4^{(1,0)}$	33	$\delta_4^{(1,1)}$	35	...
								$\delta_1^{(2,0)}$	37	...
						$\tilde{\delta}^3$	35	$\delta_2^{(2,0)}$	39	...
								$\delta_3^{(2,0)}$	41	...
								$\delta_4^{(2,0)}$	43	...
								$\tilde{\delta}^4$	45	...

Table 19: The values of  $\beta^1(M(nP, \delta))$ ,  $n = 3, 5, 7, 9, 11, \dots$ , for Lemma 5.4.

In particular,  $\lambda_1^t$  is obtained by inserting  $(t + 1)$  copies of  $A_1$  into the coloring vector of  $\lambda$ . We denote by  $\delta_i^{t-1}$  the natural  $\mathbb{Z}_2^4$ -extension of  $\lambda_i^{t-1}$ . From

$$\beta^1(M(5P, \delta_i^0)) = 5 + 2i, \quad \beta^1(M(7P, \delta_i^1)) = 7 + 2i, \quad \beta^1(M(9P, \delta_i^2)) = 9 + 2i,$$

we have

$$(5-11) \quad \beta^1(M((2t + 3)P, \delta_i^{t-1})) + 2 = \beta^1(M((2t + 5)P, \delta_i^t))$$

for  $i = 1, 2, 3, 4$ .

Next, we construct three nonorientable  $\mathbb{Z}_2^3$ -colorings  $\tilde{\lambda}^0, \tilde{\lambda}^1$  and  $\tilde{\lambda}^2$  on the polytopes  $3P, 5P, 7P$ , whose coloring vectors are, respectively,

$$[a_1 A_1 A_3 a_1], \quad [a_1 A_1 A_3 A_3 a_1], \quad [a_1 A_1 A_3 A_3 A_3 a_1].$$

The natural  $\mathbb{Z}_2^4$ -extensions are denoted by  $\tilde{\delta}^0, \tilde{\delta}^1$  and  $\tilde{\delta}^2$ . By calculation, we have

$$\beta^1(M(3P, \tilde{\delta}^0)) = 5, \quad \beta^1(M(5P, \tilde{\delta}^1)) = 15, \quad \beta^1(M(7P, \tilde{\delta}^2)) = 25.$$

For  $t \in \mathbb{Z}_{\geq 1}$ , we denote by  $\tilde{\lambda}^{t-1}$  the  $\mathbb{Z}_2^3$ -coloring characteristic function of

$$[a_1 A_1 \underbrace{A_3, \dots, A_3}_t a_1]$$

over the polytope  $(2t + 1)P$  and write its natural  $\mathbb{Z}_2^4$ -extension as  $\tilde{\delta}^{t-1}$ . Then we have, for each  $t \in \mathbb{Z}_{\geq 1}$ ,

$$(5-12) \quad \beta^1(M((2t + 1)P, \tilde{\delta}^{t-1})) = 10t - 5.$$

Let  $\lambda_i^{(t_1-1, t_2)}$  denote the  $\mathbb{Z}_2^3$ -coloring characteristic function of the coloring vector

$$[a_1 A_1 \underbrace{A_3, \dots, A_3}_{t_1} A_4 \underbrace{A_1, \dots, A_1}_{t_2} A_i a_j]$$

over the polytope  $(2(t_1 + t_2) + 5)P$ , where  $a_j$  is an affix element and  $j$  is given by 2, 1, 1, 2 for  $i = 1, 2, 3, 4$ , respectively. In particular, the coloring vector of  $\lambda_i^{(0,0)}$  is  $[aA_1A_3A_4A_ia_j]$ . Also  $\delta_i^{(t_1-1, t_2)}$  is the natural  $\mathbb{Z}_2^4$ -extension of  $\lambda_i^{(t_1-1, t_2)}$  over the polytope  $(2(t_1 + t_2) + 5)P$ .

From

$$\begin{aligned} \beta^1(M(7P, \delta_i^{(0,0)})) &= 5 + 2i, \\ \beta^1(M(9P, \delta_i^{(0,1)})) &= 7 + 2i, \\ \beta^1(M(11P, \delta_i^{(0,2)})) &= 9 + 2i \end{aligned}$$

for  $i = 1, 2, 3, 4$ , we have

$$(5-13) \quad \begin{aligned} \beta^1(M((2(t_1 + t_2) + 5)P, \delta^{(t_1-1, t_2)})) + 2 \\ = \beta^1(M((2(t_1 + t_2) + 7)P, \delta^{(t_1-1, t_2+1)})) \end{aligned}$$

for each  $t \in \mathbb{Z}_{\geq 1}$ .

From

$$\begin{aligned} \beta^1(M(7P, \delta_i^{(0,0)})) &= 5 + 2i, \\ \beta^1(M(9P, \delta_i^{(1,0)})) &= 15 + 2i, \\ \beta^1(M(11P, \delta_i^{(2,0)})) &= 25 + 2i \end{aligned}$$

for  $i = 1, 2, 3, 4$ , we have

$$(5-14) \quad \beta^1(M((2(t_1 + t_2) + 5)P, \delta_i^{(t_1-1, t_2)})) + 10 = \beta^1(M((2(t_1 + t_2) + 7)P, \delta_i^{(t_1, t_2)}))$$

for each  $t \in \mathbb{Z}_{\geq 1}$ .

	$\lambda$	$\beta^1(M(P, \delta))$
1	(1, 2, 4, 4, 2, 7, 1, 7, 7, 5, 6, 4)	1
2	(1, 2, 4, 4, 2, 7, 7, 3, 1, 5, 4, 2)	3
3	(1, 2, 4, 4, 2, 7, 3, 5, 5, 6, 3, 1)	5
4	(1, 2, 4, 5, 2, 6, 3, 6, 5, 4, 3, 1)	7

Table 20: The  $\mathbb{Z}_2^3$ -colorings and  $\beta^1$  of their natural  $\mathbb{Z}_2^4$ -extensions of Lemma 5.5.

By (5-13) and (5-14), we have

$$(5-15) \quad \beta^1(M(nP, \delta_i^{(t, \frac{1}{2}(n-1)-3-t)}))) = n + 2i + 8t$$

for  $n \in \mathbb{Z}_{\geq 7}^{\text{odd}}$  and  $0 \leq t \leq \frac{1}{2}(n-7)$ .

Putting together the results in (5-10)–(5-12) and (5-15), we complete the proof of Lemma 5.4. All the Betti numbers of Lemma 5.4 are listed in Table 19.  $\square$

**Lemma 5.5** *For any odd integer  $k \in [1, 7]$ , there is a nonorientable  $\mathbb{Z}_2^3$ -coloring over the dodecahedron  $P$  such that, for its natural associated  $\mathbb{Z}_2^4$ -coloring  $\delta$ , we have  $\beta^1(M(P, \delta)) = k$ .*

**Proof** We report the required characteristic functions in Table 20 to conclude this lemma.  $\square$

Now, using Lemmas 5.1–5.5, we complete the proof of Theorem 1.2 for an odd  $n$ . Thus, together with Section 4, we finish the proof of Theorem 1.2.

## References

- [1] **V M Buchstaber, T E Panov**, *Toric topology*, Math. Surv. Monogr. 204, Amer. Math. Soc., Providence, RI (2015) MR Zbl
- [2] **L Cai**, *On products in a real moment-angle manifold*, J. Math. Soc. Japan 69 (2017) 503–528 MR Zbl
- [3] **S Choi, H Park**, *On the cohomology and their torsion of real toric objects*, Forum Math. 29 (2017) 543–553 MR Zbl
- [4] **M Chu, A Kolpakov**, *A hyperbolic counterpart to Rokhlin’s cobordism theorem*, Int. Math. Res. Not. 2022 (2022) 2460–2483 MR Zbl
- [5] **J F Davis, F Fang**, *An almost flat manifold with a cyclic or quaternionic holonomy group bounds*, J. Differential Geom. 103 (2016) 289–296 MR Zbl

- [6] **M W Davis, T Januszkiewicz**, *Convex polytopes, Coxeter orbifolds and torus actions*, Duke Math. J. 62 (1991) 417–451 MR Zbl
- [7] **F T Farrell, S Zdravkovska**, *Do almost flat manifolds bound?*, Michigan Math. J. 30 (1983) 199–208 MR Zbl
- [8] **M Hochster**, *Cohen–Macaulay rings, combinatorics, and simplicial complexes*, from “Ring theory, II” (B R McDonald, R A Morris, editors), Lect. Notes Pure Appl. Math. 26, Dekker, New York (1977) 171–223 MR Zbl
- [9] **A Kolpakov, B Martelli, S Tschantz**, *Some hyperbolic three-manifolds that bound geometrically*, Proc. Amer. Math. Soc. 143 (2015) 4103–4111 MR Zbl Correction in 144 (2016) 3647–3648
- [10] **A Kolpakov, A W Reid, L Slavich**, *Embedding arithmetic hyperbolic manifolds*, Math. Res. Lett. 25 (2018) 1305–1328 MR Zbl
- [11] **DD Long, A W Reid**, *On the geometric boundaries of hyperbolic 4-manifolds*, Geom. Topol. 4 (2000) 171–178 MR Zbl
- [12] **B Martelli**, *Hyperbolic three-manifolds that embed geodesically*, preprint (2015) arXiv 1510.06325
- [13] **R Meyerhoff, W D Neumann**, *An asymptotic formula for the eta invariants of hyperbolic 3-manifolds*, Comment. Math. Helv. 67 (1992) 28–46 MR Zbl
- [14] **C Millichap**, *Factorial growth rates for the number of hyperbolic 3-manifolds of a given volume*, Proc. Amer. Math. Soc. 143 (2015) 2201–2214 MR Zbl
- [15] **H Nakayama, Y Nishimura**, *The orientability of small covers and coloring simple polytopes*, Osaka J. Math. 42 (2005) 243–256 MR Zbl
- [16] **J G Ratcliffe, S T Tschantz**, *Gravitational instantons of constant curvature*, Classical Quantum Gravity 15 (1998) 2613–2627 MR Zbl
- [17] **J G Ratcliffe, S T Tschantz**, *On the growth of the number of hyperbolic gravitational instantons with respect to volume*, Classical Quantum Gravity 17 (2000) 2999–3007 MR Zbl
- [18] **N Saveliev**, *Lectures on the topology of 3-manifolds: an introduction to the Casson invariant*, 2nd revised edition, de Gruyter, Berlin (2012) MR Zbl
- [19] **L Slavich**, *A geometrically bounding hyperbolic link complement*, Algebr. Geom. Topol. 15 (2015) 1175–1197 MR Zbl
- [20] **L Slavich**, *The complement of the figure-eight knot geometrically bounds*, Proc. Amer. Math. Soc. 145 (2017) 1275–1285 MR Zbl
- [21] **A I Suciu**, *The rational homology of real toric manifolds*, Oberwolfach Rep. 9 (2012) 2972–2976 Part of the conference report “Cohomology rings and fundamental groups of hyperplane arrangements, wonderful compactifications, and real toric varieties”

- [22] **A I Suci**, **A Trevisan**, *Real toric varieties and abelian covers of generalized Davis–Januszkiewicz spaces*, unpublished manuscript (2012)
- [23] **W Thurston**, *Geometry and topology of three-manifolds*, lecture notes, Princeton Univ. (1980) Available at <http://library.msri.org/books/gt3m/>
- [24] **A Y Vesnin**, *Three-dimensional hyperbolic manifolds of Löbell type*, *Sibirsk. Mat. Zh.* 28 (1987) 50–53 MR Zbl In Russian; translated in *Siberian Math. J.* 28 (1987) 731–734

*School of Mathematical Sciences, Fudan University  
Shanghai, China*

*Department of Pure Mathematics, Xi’an Jiaotong-Liverpool University  
Suzhou, China*

majiming@fudan.edu.cn, fangting.zheng@xjtlu.edu.cn

Received: 9 March 2020      Revised: 16 November 2020



# Constrained knots in lens spaces

FAN YE

We study a special family of  $(1, 1)$  knots called constrained knots, which includes 2–bridge knots in the 3–sphere  $S^3$  and simple knots in lens spaces. Constrained knots are parametrized by five integers and characterized by the distribution of  $\text{spin}^c$  structures in the corresponding  $(1, 1)$  diagrams. The knot Floer homology  $\widehat{HFK}$  of a constrained knot is thin. We obtain a complete classification of constrained knots based on the calculation of  $\widehat{HFK}$  and presentations of knot groups. We provide many examples of constrained knots constructed from surgeries on links in  $S^3$ , which are related to 2–bridge knots and 1–bridge braids. We also show many examples of constrained knots whose knot complements are orientable hyperbolic 1–cusped manifolds with simple ideal triangulations.

57K10, 57K14, 57K18, 57K31, 57K32

1. Introduction	1097
2. Preliminaries	1107
3. Parametrization and characterization	1112
4. Knot Floer homology	1119
5. Knots in the same homology class	1128
6. Classification	1134
7. Magic links	1144
8. 1–Bridge braid knots	1154
9. SnapPy manifolds	1159
References	1164

## 1 Introduction

The main object studied in this paper is a special family of knots in lens spaces called constrained knots. Every knot in a closed 3–manifold can be represented by a doubly pointed Heegaard diagram  $(\Sigma, \alpha, \beta, z, w)$  (see Ozsváth and Szabó [29, Section 2]),

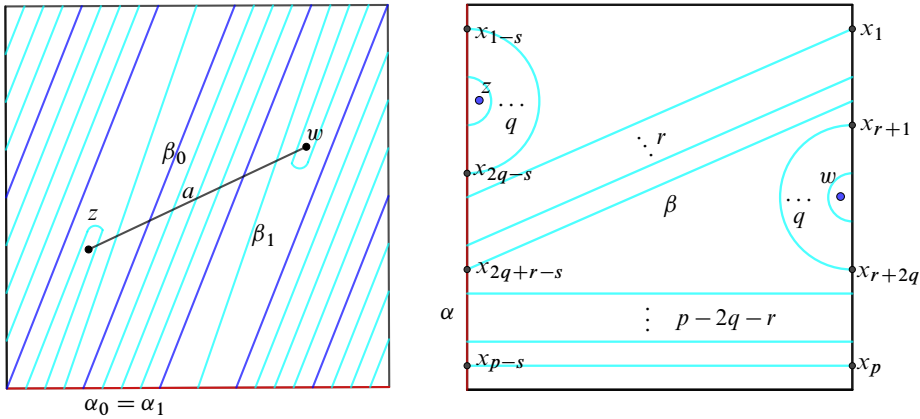


Figure 1: Left: a constrained knot in  $L(5, 2)$ . Right: a  $(1, 1)$  diagram.

where  $\Sigma$  is a closed surface,  $\alpha = \{\alpha_1, \dots, \alpha_g\}$  and  $\beta = \{\beta_1, \dots, \beta_g\}$  are two collections of  $g = g(\Sigma)$  simple closed curves on  $\Sigma$ , and  $z$  and  $w$  are two basepoints on  $\Sigma - (\alpha \cup \beta)$ . Conversely, any doubly pointed Heegaard diagram defines a knot. Explicitly, the knot is the union of an arc  $a$  connecting  $z$  to  $w$  on  $\Sigma - \alpha$ , pushed slightly into the  $\alpha$ -handlebody, and an arc  $b$  connecting  $w$  to  $z$  on  $\Sigma - \beta$ , pushed slightly into the  $\beta$ -handlebody.

Let  $T^2$  be the torus obtained by the quotient map  $\mathbb{R}^2 \rightarrow T^2$  that identifies  $(x, y)$  with  $(x + m, y + n)$  for  $m, n \in \mathbb{Z}$ . Suppose  $p$  and  $q$  are integers satisfying  $p > 0$  and  $\gcd(p, q) = 1$ . Let  $\alpha_0$  and  $\beta_0$  be two simple closed curves on  $T^2$  obtained from two straight lines in  $\mathbb{R}^2$  of slopes  $0$  and  $p/q$ . Then  $(T^2, \alpha_0, \beta_0)$  is called the *standard diagram* of a lens space  $L(p, q)$ . Let  $\alpha_1 = \alpha_0$  and let  $\beta_1$  be a simple closed curve on  $T^2$  that is disjoint from  $\beta_0$  and where  $[\beta_1] = [\beta_0] \in H_1(T^2; \mathbb{Z})$ . Then  $(T^2, \alpha_1, \beta_1)$  is also a Heegaard diagram of  $L(p, q)$ . Let  $z$  and  $w$  be two basepoints in  $T^2 - \alpha_0 \cup \beta_0 \cup \beta_1$ .

The knot defined by the doubly pointed diagram  $(T^2, \alpha_1, \beta_1, z, w)$  is called a *constrained knot* and the diagram is called the *standard diagram* of the constrained knot. We will show that constrained knots are parametrized by five integers, which will be denoted by  $C(p, q, l, u, v)$ . For technical reasons, the knot  $C(p, q, l, u, v)$  is in  $L(p, q')$ , where  $qq' \equiv 1 \pmod{p}$ . An example is shown in Figure 1, left, where  $(T^2, \alpha_0, \beta_0)$  is the standard diagram of  $L(5, 2)$  and  $(T^2, \alpha_1, \beta_1, z, w)$  defines  $C(5, 3, 2, 3, 1)$ .

Roughly speaking, knots defined by doubly pointed Heegaard diagrams with  $g(\Sigma) = 1$  are called  $(1, 1)$  knots and the corresponding diagrams are called  $(1, 1)$  diagrams; for precise definitions see Section 2.1. These  $(1, 1)$  knots are parametrized by four integers (see Goda, Matsuda and Morifuji [14] and Rasmussen [37]), which will be denoted by

$W(p, q, r, s)$ ; see Figure 1, right. After rotation, standard diagrams of constrained knots are special cases of  $(1, 1)$  diagrams. Moreover, the following proposition characterizes constrained knots by the distribution of  $\text{spin}^c$  structures on the ambient 3-manifold in the corresponding  $(1, 1)$  diagrams; for the definition of  $\text{spin}^c$  structures see Ozsváth and Szabó [29] and Rasmussen [38].

**Proposition 1.1** *Let  $K = W(p, q, r, s)$  be a  $(1, 1)$  knot in  $Y = L(a, b)$  with  $a > 1$ , and suppose  $(T^2, \alpha, \beta, z, w)$  is the corresponding  $(1, 1)$  diagram of  $K$ . Let  $\{x_i\}$  be intersection points in  $\alpha \cap \beta$ , ordered by an orientation of  $\alpha$ . Let  $\mathfrak{s}_i = \mathfrak{s}_z(x_i) \in \text{Spin}^c(Y)$  be the  $\text{spin}^c$  structures on  $Y$  corresponding to  $x_i$ . The knot  $K$  is a constrained knot if and only if:*

(i) *For  $k = |\text{Spin}^c(Y)| (= a)$ , there are integers  $p_1, \dots, p_k$  such that*

$$0 < p_1 < p_2 < \dots < p_k \leq p.$$

(ii)  *$\mathfrak{s}_i = \mathfrak{s}_j$  if and only if either  $i, j \in (0, p_1] \cup (p_k, p]$ , or  $i, j \in (p_l, p_{l+1}]$  for some  $l \in \{1, \dots, p-1\}$ .*

A single knot can be represented by  $(1, 1)$  knots  $W(p_1, q_1, r_1, s_1)$  and  $W(p_2, q_2, r_2, s_2)$  with different parameters. For example, both  $W(5, 2, 1, 3)$  and  $W(5, 2, 1, 0)$  represent the figure-8 knot in  $S^3$ . There is no explicit classification of  $(1, 1)$  knots by  $W(p, q, r, s)$  to the author’s knowledge. However, it is possible to classify constrained knots by the parametrization  $C(p, q, l, u, v)$ . In particular, the case  $C(1, 0, 1, u, v)$  consists of 2-bridge knots in  $S^3$  (see Proposition 3.5) and the case  $C(p, q, l, 1, 0)$  consists of simple knots in lens spaces (see Proposition 3.7). Schubert [41] and Rasmussen [38] classify 2-bridge knots and simple knots, respectively. The case  $C(p, q, 1, u, v)$  consists of connected sums of a core knot in a lens space and a 2-bridge knot (see Theorem 7.14). For other constrained knots, the classification is given by:

**Theorem 1.2** *Suppose that  $(p_1, q_1, l_1, u_1, v_1)$  and  $(p_2, q_2, l_2, u_2, v_2)$  are two different collections of integers satisfying, for  $i = 1, 2$ ,*

$$p_i > 1, \quad q_i \in [1, p_i - 1], \quad l_i \in [2, p_i], \quad u_i > 2v_i > 0, \quad u_i \text{ is odd,}$$

$$\gcd(p_i, q_i) = \gcd(u_i, v_i) = 1.$$

*Then constrained knots  $C(p_i, q_i, l_i, u_i, v_i)$  represent the same knot if and only if*

$$p_1 = p_2 = p, \quad q_1 q_2 \equiv 1 \pmod{p}, \quad l_1, l_2 \in \{2, p\} \quad \text{and} \quad (l_1, u_1, v_1) = (l_2, u_2, v_2).$$

The hat version of knot Floer homology  $\widehat{HFK}(Y, K)$  (see Ozsváth and Szabó [29] and Rasmussen [36]) is a powerful invariant for a knot  $K$  in a closed 3-manifold  $Y$ . It decomposes as the direct sum

$$(1) \quad \widehat{HFK}(Y, K) = \bigoplus_{\mathfrak{s} \in \text{Spin}^c(Y)} \widehat{HFK}(Y, K, \mathfrak{s})$$

with respect to  $\text{spin}^c$  structures on  $Y$ . Moreover, the homology  $\widehat{HFK}(Y, K, \mathfrak{s})$  inherits two  $\mathbb{Z}$ -gradings, the Alexander grading and the Maslov grading, from the underlying chain complex  $\widehat{CFK}(Y, K, \mathfrak{s})$ .

**Definition 1.3** A knot  $K \subset Y$  is called an  $\mathfrak{s}$ -thin knot if the difference of the Maslov grading and the Alexander grading on  $\widehat{HFK}(Y, K, \mathfrak{s})$  is constant for homogeneous elements. It is called a thin knot if it is an  $\mathfrak{s}'$ -thin knot for any  $\mathfrak{s}' \in \text{Spin}^c(Y)$ .

Thin knots defined as above generalize  $\delta$ -thin knots in  $S^3$  (see Rasmussen [37]) and Floer homological thin knots; see Manolescu and Ozsváth [23]. Examples of thin knots include all quasialternating knots [23], in particular all 2-bridge knots.

Suppose  $K$  is a thin knot in  $S^3$  and  $\mathfrak{s}_0$  is the unique  $\text{spin}^c$  structure on  $S^3$ . Then the minus version of the knot Floer chain complex  $CFK^-(S^3, K) = CFK^-(S^3, K, \mathfrak{s}_0)$  is determined by the Alexander polynomial  $\Delta_K(t)$  and the signature  $\sigma(K)$  up to chain homotopy; see Petkova [34]. For a compact 3-manifold  $M$  with torus boundary, there exists a set of immersed curves  $\widehat{HF}(M)$  on  $\partial M$  – pt, called the curve invariant (see Hanselman, Rasmussen and Watson [16; 17]) of  $M$ , which encodes the information of Heegaard Floer theory in a diagrammatic way. Based on [17, Section 4; 34, Section 3], it is easy to draw  $\widehat{HF}(E(K))$  of the knot complement  $E(K) = S^3 - \text{int } N(K)$  for a thin knot  $K \subset S^3$ . Roughly speaking, it consists of figure-8 curves and a distinguished curve.

For a  $(1, 1)$  knot  $K \subset Y$  there is a combinatorial method to calculate the chain complex  $CFK^-(Y, K)$ ; see Goda, Matsuda and Morifuji [14]. It applies well to 2-bridge knots and also constrained knots. From the standard diagram of a constrained knot  $K \subset Y$ , if we focus on intersection points corresponding to the same  $\text{spin}^c$  structure  $\mathfrak{s} \in \text{Spin}^c(Y)$ , we can obtain an explicit relation between  $CFK^-(Y, K, \mathfrak{s})$  and  $CFK^-(S^3, K', \mathfrak{s}_0)$ , where  $K'$  is some 2-bridge knot. In particular, for  $K = C(p, q, l, u, v) \subset Y$  and  $\mathfrak{s} \in \text{Spin}^c(Y)$ , the group  $\widehat{HFK}(Y, K, \mathfrak{s})$  is determined by Alexander polynomials of 2-bridge knots  $K_1 = \mathfrak{b}(u, v)$  and  $K_2 = \mathfrak{b}(u - 2v, v)$ . Hence:

**Proposition 1.4** *Constrained knots are thin.*

Results about thin complexes in [34, Section 3] apply directly to  $CFK^-(Y, K, \mathfrak{s})$  for a constrained knot  $K \subset Y$ . Then we can draw the part of the curve invariant corresponding to each  $\text{spin}^c$  structure following the approach in [17, Section 4]. Similar to the case of a 2-bridge knot, the curve invariant part consists of figure-8 curves and a distinguished curve; see Figure 7.

To connect the distinguished curves of different parts, we should study the following grading, which relates elements in  $\widehat{HFK}(Y, K, \mathfrak{s})$  for different  $\text{spin}^c$  structures. The total homology  $\widehat{HFK}(Y, K)$  inherits a relative  $H_1(E(K); \mathbb{Z})$ -grading (see Rasmussen [38, Section 3.3]):

$$(2) \quad \widehat{HFK}(Y, K) = \bigoplus_{h \in H_1(E(K); \mathbb{Z})} \widehat{HFK}(Y, K, h).$$

This grading generalizes the Alexander grading on  $\widehat{HFK}(S^3, K, \mathfrak{s}_0)$  and corresponds to  $\text{spin}^c$  structures on  $E(K)$  with some boundary conditions; see Juhász [20, Section 4]. Under the map  $H_1(E(K); \mathbb{Z}) \rightarrow H_1(Y; \mathbb{Z})$ , this grading reduces to the grading in (1).

Similar to the Alexander grading on  $\widehat{HFK}(S^3, K, \mathfrak{s}_0)$ , summands of  $\widehat{HFK}(Y, K)$  in the opposite  $H_1(E(K); \mathbb{Z})$ -gradings are isomorphic (see Section 3 of Ozsváth and Szabó [29]), up to a global grading shift in  $H_1(E(K); \mathbb{Z})$ . This symmetry is called the *global symmetry*. If it is not mentioned, this  $H_1(E(K); \mathbb{Z})$ -grading is also called the *Alexander grading*, and denoted by  $\text{gr}(x) \in H_1(E(K); \mathbb{Z})$  for a homogeneous element  $x \in \widehat{HFK}(Y, K)$ . To fix the ambiguity of the global grading shift, a specific grading shift will be used so that under the global symmetry, the absolute value of the Alexander grading is left invariant. The Alexander grading in this specific grading shift is called the *absolute Alexander grading*. To be clear, when considering the Alexander grading on  $\widehat{HFK}(Y, K, \mathfrak{s})$  mentioned before, the  $\text{spin}^c$  structure  $\mathfrak{s}$  will be specified.

Following [38, Section 3.3], for a constrained knot, the Alexander grading can be calculated from the standard diagram. The Alexander grading on  $\widehat{HFK}(Y, K)$  indicates an explicit way to connect different parts of the curve invariant. Then it is not hard to draw the whole curve invariant of a constrained knot. As an application, much information about the Heegaard Floer theory of a constrained knot can be obtained from the curve invariant of the knot complement.

For a constrained knot  $K \subset Y$  and the corresponding 2-bridge knots  $K_1$  and  $K_2$  mentioned before, the symmetry on  $\widehat{HFK}(S^3, K_i, \mathfrak{s}_0)$  for  $i = 1, 2$  induces a symmetry

on  $\widehat{HFK}(Y, K, \mathfrak{s})$ , which is called the *local symmetry*. For  $\mathfrak{s} \in \text{Spin}^c(Y)$ , the average  $A(K, \mathfrak{s})$  of any two homogeneous elements  $x, y \in \widehat{HFK}(Y, K, \mathfrak{s})$  that are symmetric under the local symmetry is called the *middle grading* of  $\mathfrak{s}$ :

$$A(K, \mathfrak{s}) = \frac{1}{2}(\text{gr}(x) + \text{gr}(y)) \in H_1(E(K); \mathbb{Z}).$$

**Theorem 1.5** For  $i = 1, 2$  let  $K_i = C(p_i, q_i, l_i, u_i, v_i)$  be constrained knots in the same lens space  $Y$  with  $[K_1] = [K_2] \in H_1(Y; \mathbb{Z})$ . Consider the absolute Alexander grading on  $\widehat{HFK}(Y, K_i)$ . Then there are isomorphisms

$$H_1(E(K_1); \mathbb{Z}) \cong H_1(E(K_2); \mathbb{Z}) \cong H_1,$$

so that  $A(K_1, \mathfrak{s}) = A(K_2, \mathfrak{s}) \in H_1$  for any  $\mathfrak{s} \in \text{Spin}^c(Y)$ .

**Theorem 1.6** Suppose  $K$  is a knot in  $Y = L(p, q)$ . Let  $K'$  be a simple knot in the same manifold  $Y$  with  $[K'] = [K] \in H_1(Y; \mathbb{Z})$ . Consider the absolute Alexander gradings on  $\widehat{HFK}(Y, K)$  and  $\widehat{HFK}(Y, K')$ . We know  $\widehat{HFK}(Y, K') \cong \mathbb{Z}^p$ . If  $\widehat{HFK}(Y, K) \cong \mathbb{Z}^p$ , then there are isomorphisms  $H_1(E(K); \mathbb{Z}) \cong H_1(E(K'); \mathbb{Z}) \cong H_1$ , so that there is a one-to-one correspondence between generators of  $\widehat{HFK}(Y, K)$  and  $\widehat{HFK}(Y, K')$  with the same absolute Alexander grading in  $H_1$ .

Theorem 1.6 provides a clue for the following conjecture, which is related to Berge's conjecture [2] claiming that any knot in  $S^3$  admitting lens space surgeries falls into Berge's list.

**Conjecture 1.7** (Baker, Grigsby and Hedden [1] and Hedden [19]) Suppose  $K$  is a knot in  $Y = L(p, q)$ . If  $\widehat{HFK}(Y, K) \cong \mathbb{Z}^p$ , then  $K$  is a simple knot, ie a  $(1, 1)$  knot  $W(p', q', r', s')$  with  $q' = 0$ .

Though constrained knots are defined by doubly pointed Heegaard diagrams, there are many other ways to construct constrained knots, at least for some special families of parameters. In the following, we introduce two approaches based on Dehn surgeries on links in  $S^3$ .

The first approach is inspired by the relation between knot Floer homologies of constrained knots and 2–bridge knots. A *magic link* is a 3–component link as shown in Figure 2, left, where  $K_0$  is a 2–bridge knot, and  $K_1$  and  $K_2$  are unknots. Dehn surgeries on  $K_1$  and  $K_2$  induce a lens space, in which  $K_0$  becomes a knot  $K'_0$ .

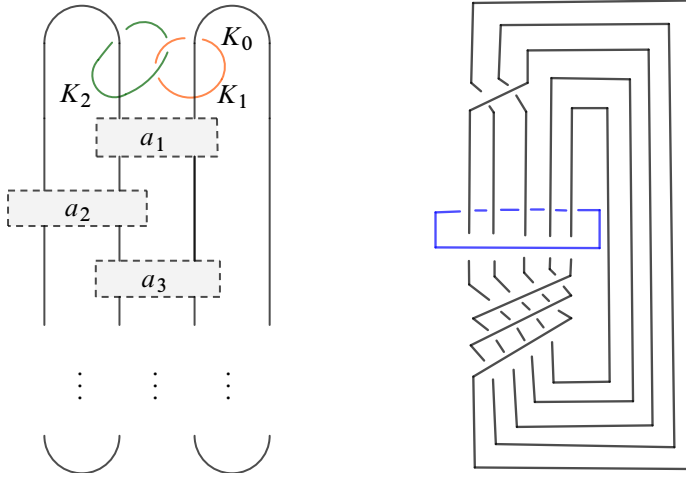


Figure 2: Left: magic link. Right: 1-bridge braid.

**Theorem 1.8** Suppose that integers  $p$  and  $q$  satisfy  $p > q > 0$  and  $\gcd(p, q) = 1$ . Suppose integers  $n_1, n_2$  and  $l$  satisfy

$$n_1 \in \left[0, \frac{p}{q}\right), \quad n_2 \in \left[0, \frac{p}{p-q}\right),$$

$$l \in \{n_1q + 1, p - n_1q + 1, n_2(p - q) + 1, p - n_2(p - q) + 1\}.$$

Let  $L = K_0 \cup K_1 \cup K_2$  be a magic link with  $K_0 = \mathfrak{b}(u, v)$ . Then the knot  $C(p, q, l, u, v)$  is equivalent to the knot  $K'_0$  obtained by performing some Dehn surgeries on  $K_1$  and  $K_2$ .

The second approach arises from 1-bridge braids. Suppose the solid torus  $H = S^1 \times D^2$  is embedded in  $\mathbb{R}^3 \subset S^3$  in a standard way, and suppose  $K_1$  is the core of  $S^3 - H$ . Let  $K_0 \subset H$  be a 1-bridge braid; see Gabai [12; 13]. Then  $L = K_0 \cup K_1$  is a 2-component link in  $S^3$ ; an example is given in Figure 2, right. Dehn filling along a simple closed curve on  $\partial H$  is equivalent to Dehn surgery on  $K_1$ . The resulting manifold is a lens space and  $K_0$  becomes a knot  $K'_0$  in the lens space. A knot  $K'_0$  constructed from this approach is called a 1-bridge braid knot.

**Theorem 1.9** The knots  $C(p, q, l, u, \pm 1)$  are equivalent to 1-bridge braid knots, where  $C(p, q, l, u, -1)$  means  $C(p, q, l, u, u - 1)$ .

Other than Dehn surgeries, constrained knots can also be constructed by Dehn filling the boundary of (orientable hyperbolic) 1-cusped manifolds. Many 1-cusped manifolds

are knot complements of constrained knots. *SnapPy* by Culler, Dunfield and Weeks [7] provides a list of 59 068 1-cusped manifolds admitting ideal triangulations with at most nine tetrahedra. Using the code in Ye [45], we show 21 922 of them are complements of constrained knots. Table 1 shows examples of 1-cusped manifolds that are complements of constrained knots. The names of manifolds in the table are from *SnapPy*. The slopes in the table are considered in the basis from *SnapPy* and the integers indicate the parametrization of the constrained knot that is equivalent to the core of the filling solid torus. For example, Dehn filling along the curve of slope  $1/0$  on the boundary of  $m003$  gives  $C(10, 3, 3, 1, 0)$ . If different parametrizations correspond to the same knot (see Theorem 1.2), we only show one collection of parameters. The complete list can be found in [45].

**Proposition 1.10** *Curve invariants  $\widehat{HF}(M)$  of knot complements  $M$  of all 1-cusped manifolds that have ideal triangulations with at most 5 tetrahedra can be drawn explicitly, except the manifolds in Table 2.*

**Proof** There are 286 orientable 1-cusped manifolds that have ideal triangulations with at most five ideal tetrahedra. Of these, 232 manifolds are complements of constrained knots, whose curve invariants can be calculated by the method in Section 4. Other than examples from constrained knots, 37 manifolds are Floer simple (by the list in Dunfield [10]), whose curve invariants can be calculated by the approach in Hanselman, Rasmussen and Watson [17, Section 1]. Other manifolds are listed in Table 2 ((1, 1) parameters are from Dunfield's code [45]). The chain complex  $CFK^-(Y, K)$  of a (1, 1) knot can be calculated by the method in Goda, Matsuda and Morifuji [14]. Then the curve invariant can be calculated by [17, Section 4]. Note that chain complexes of  $8_{20}$ ,  $9_{42}$  and  $9_{46}$  in the table were calculated in Ozsváth and Szabó [32].  $\square$

It is known that a 2-bridge knot  $b(u, v)$  is a torus knot if  $v = 1$  or  $v = u - 1$ . The latter case is written as  $v = -1$ . If  $v \neq \pm 1$ , the 2-bridge knot  $b(u, v)$  is hyperbolic, that is, the interior of the knot complement admits a hyperbolic metric of finite volume. We may generalize the results about 2-bridge knots to constrained knots. Note that the knot complement of a torus knot is a Seifert fibered space.

**Theorem 1.11** *If  $C(p, q, l, u, v)$  has Seifert fibered complement, then  $v = \pm 1$ .*

Since  $C(p, q, 1, u, v)$  is a connected sum of two knots, there is an essential torus in the knot complement, and hence  $C(p, q, 1, u, v)$  is not hyperbolic. Using the code in [45]



name	slope + $(p, q, l, u, v)$
m003	$(1, 0) + (10, 3, 3, 1, 0), (-1, 1) + (5, 4, 5, 3, 1), (0, 1) + (5, 4, 5, 3, 1)$
m004	$(1, 0) + (1, 0, 1, 5, 2)$
m006	$(0, 1) + (15, 4, 2, 1, 0), (1, 0) + (5, 3, 4, 3, 1)$
m007	$(1, 0) + (3, 1, 2, 3, 1)$
m009	$(1, 0) + (2, 1, 2, 5, 2)$
m010	$(1, 0) + (6, 5, 6, 3, 1)$
m011	$(1, 0) + (13, 3, 3, 1, 0), (0, 1) + (9, 4, 9, 3, 1)$
m015	$(1, 0) + (1, 0, 1, 7, 2)$
m016	$(0, 1) + (18, 5, 3, 1, 0), (-1, 1) + (19, 7, 2, 1, 0)$
m017	$(0, 1) + (14, 3, 5, 1, 0), (-1, 1) + (21, 8, 21, 1, 0), (1, 0) + (7, 5, 6, 3, 1)$
m019	$(0, 1) + (17, 5, 4, 1, 0), (1, 1) + (11, 7, 11, 3, 1), (1, 0) + (6, 5, 5, 3, 1)$
m022	$(1, 0) + (7, 6, 7, 3, 1)$
m023	$(1, 0) + (3, 1, 3, 5, 2)$
m026	$(0, 1) + (19, 4, 2, 1, 0), (1, 0) + (8, 3, 7, 3, 1)$
m027	$(1, 0) + (16, 3, 3, 1, 0), (0, 1) + (13, 4, 13, 3, 1)$
m029	$(1, 0) + (5, 2, 3, 3, 1)$
m030	$(1, 0) + (7, 4, 5, 3, 1)$
m032	$(1, 0) + (1, 0, 1, 9, 2)$
m033	$(0, 1) + (18, 5, 5, 1, 0), (1, 0) + (9, 7, 8, 3, 1)$
m034	$(1, 0) + (4, 1, 3, 3, 1)$
m035	$(1, 0) + (4, 1, 2, 3, 1)$
m036	$(-1, 1) + (21, 8, 2, 1, 0), (1, 0) + (3, 2, 3, 5, 1)$
m037	$(1, 1) + (24, 7, 2, 1, 0), (1, 0) + (8, 5, 6, 3, 1)$
m038	$(1, 0) + (3, 2, 3, 5, 2)$
m039	$(1, 0) + (4, 1, 4, 5, 2)$
m040	$(1, 0) + (8, 7, 8, 3, 1)$
m043	$(0, 1) + (25, 7, 24, 1, 0), (-1, 1) + (25, 9, 2, 1, 0)$
m044	$(0, 1) + (24, 7, 23, 1, 0), (-1, 1) + (17, 10, 17, 3, 1), (1, 0) + (7, 6, 5, 3, 1)$
m045	$(1, 0) + (2, 1, 2, 7, 2)$
m046	$(-1, 1) + (30, 11, 30, 1, 0), (1, 0) + (10, 7, 8, 3, 1)$
m047	$(0, 1) + (23, 4, 2, 1, 0), (1, 0) + (11, 3, 10, 3, 1)$
m049	$(1, 0) + (19, 3, 3, 1, 0), (0, 1) + (17, 13, 17, 3, 1)$
m052	$(0, 1) + (26, 7, 3, 1, 0), (1, 0) + (7, 5, 4, 3, 1)$
m053	$(1, 0) + (1, 0, 1, 11, 2)$
m054	$(0, 1) + (22, 5, 7, 1, 0), (1, 0) + (11, 8, 9, 3, 1)$
m055	$(1, 0) + (23, 7, 5, 1, 0), (0, 1) + (14, 11, 13, 3, 1)$

Table 1: 1-cusped manifolds and constrained knots.

name	comments	name	comments
<i>m136</i>	no lens space filling	<i>m305</i>	no lens space filling
<i>m137</i>	$W(8, 2, 3, 1) \subset S^1 \times S^2$	<i>m306</i>	no lens space filling
<i>m199</i>	$9_{42} = W(9, 2, 2, 3) \subset S^3$	<i>m345</i>	$W(10, 3, 1, 5) \subset L(2, 1)$
<i>m201</i>	$10_{132} = W(11, 2, 1, 3) \subset S^3$	<i>m370</i>	$(1, 0)$ filling gives $L(8, 3)$
<i>m206</i>	$(1, 0)$ filling gives $L(5, 2)$	<i>m372</i>	$9_{46} = \text{Pretzel}(-3, 3, 3) \subset S^3$
<i>m222</i>	$8_{20} = W(9, 3, 0, 2) \subset S^3$	<i>m389</i>	$10_{139} = W(11, 3, 1, 4) \subset S^3$
<i>m224</i>	$11_{190} = W(13, 2, 1, 8) \subset S^3$	<i>m390</i>	$(1, 0)$ filling gives $L(7, 2)$
<i>m235</i>	no lens space filling	<i>m410</i>	no lens space filling
<i>m304</i>	$W(12, 3, 0, 5) \subset L(2, 1)$		

Table 2: Exceptions of 1-cusped manifolds.

and the `verify_hyperbolicity()` function in *SnapPy*, we verified that  $C(p, q, l, u, v)$  is hyperbolic for  $p \leq 10$ ,  $l > 1$ ,  $u < 20$  and  $v \neq \pm 1$ .

**Conjecture 1.12**  $C(p, q, l, u, v)$  with  $l > 1$  and  $v \neq \pm 1$  is hyperbolic.

**Organization** The remainder of this paper is organized as follows. In Section 2, we collect some conventions and definitions in 3-dimensional topology, and facts about  $(1, 1)$  knots, simple knots and 2-bridge knots. In Section 3, we describe the parametrization of constrained knots and prove Proposition 1.1. Many propositions about constrained knots are also given in Section 3. In Section 4, an algorithm for the knot Floer homology of a constrained knot is obtained, which induces Proposition 1.4 and the necessary part of Theorem 1.2. In Section 5, we study knots in the same homology class and prove Theorems 1.5 and 1.6 by Turaev torsions of 3-manifolds. In Section 6, we finish the proof of Theorem 1.2 by constructing isomorphisms between fundamental groups of knot complements and applying the fact that knots are determined by their fundamental groups. The last three sections discuss magic links, 1-bridge braid knots and *SnapPy* manifolds, respectively.

**Acknowledgements** The author would like to thank his supervisor Jacob Rasmussen for introducing him to this project and guiding him on his research. The author is grateful to Nathan M Dunfield for sharing code about  $(1, 1)$  knots, and to Sirui Lu and Muge Chen for helping him calculate many examples by computer. The author is grateful to Zekun Chen, Zhenkun Li, Donghao Wang, Zipei Nie and Wenzhao Chen for helpful conversations, and anonymous referees for the helpful comments. The author

would also like to thank his parents and relatives for their support. Calculations are based on *Mathematica*, *SageMath* [40] and *SnapPy* [7].

## 2 Preliminaries

We begin with basic conventions. For  $r \in \mathbb{R}$ , let  $\lceil r \rceil$  and  $\lfloor r \rfloor$  denote the minimum integer and the maximum integer satisfying  $\lceil r \rceil \geq r$  and  $\lfloor r \rfloor \leq r$ , respectively. For a group  $H$ , let  $\text{Tors } H$  denote the set of torsion elements in  $H$ .

If it is not mentioned, all manifolds are smooth, connected and oriented, and orientations of knots are omitted. The fundamental group of a manifold  $M$  is denoted by  $\pi_1(M)$ , where the basepoint is omitted. For a submanifold  $A$  in a manifold  $Y$ , let  $N(A)$  denote the regular neighborhood of  $A$  in  $Y$  and let  $\text{int } N(A)$  denote its interior. Suppose  $Y$  is a closed 3-manifold and  $K$  is a knot in  $Y$ . Let  $E(K) = Y - \text{int } N(K)$  denote the knot complement of  $K$ .

For a simple closed curve  $\alpha$  on a surface  $\Sigma$ , let  $[\alpha]$  denote its homology class in  $H_1(\Sigma; \mathbb{Z})$ . If it is clear, we do not distinguish  $\alpha$  and  $[\alpha]$ . The algebraic intersection number of two curves  $\alpha$  and  $\beta$  on a surface  $\Sigma$  is denoted by  $[\alpha] \cdot [\beta]$  or  $\alpha \cdot \beta$ , while the number of intersection points of  $\alpha$  and  $\beta$  is denoted by  $|\alpha \cap \beta|$ .

A basis  $(m, l)$  of  $H_1(T^2; \mathbb{Z})$  always satisfies  $m \cdot l = -1$ . Suppose  $K$  is a knot in a closed 3-manifold  $Y$ . A basis of  $\partial E(K)$  means a basis of  $H_1(\partial E(K); \mathbb{Z}) \cong H_1(T^2; \mathbb{Z})$ . In practice, there are two standard choices of the basis of  $\partial E(K)$ :

- (i) Let  $m$  and  $l$  be simple closed curves on  $\partial E(K)$  such that Dehn filling along  $m$  gives  $Y$ ,  $m \cdot l = -1$ , and the orientation of  $m$  is determined from the orientation of  $K$  by the “right-hand rule”. The curves  $m$  and  $l$  are called the *meridian* and the *longitude* of the knot  $K$ , respectively. The basis  $(m, l)$  is called the *regular basis* of  $\partial E(K)$ .
- (ii) Let  $m^*$  and  $l^*$  be simple closed curves on  $\partial E(K)$  such that  $l^*$  represents the generator of  $\text{Ker}(H_1(E(K); \mathbb{Q}) \rightarrow H_1(Y; \mathbb{Q}))$  and  $m^* \cdot l^* = -1$ . They are called the *homological meridian* and the *homological longitude* of the knot  $K$ , respectively. The basis  $(m^*, l^*)$  is called the *homological basis* of  $\partial E(K)$ .

The choices of  $l$  and  $m^*$  are not unique. The longitude  $l$  is isotopic to  $K$ , while  $m^*$  does not have any geometric meaning. Sometimes (eg for knots in  $S^3$ ) these two choices of the basis are equivalent. If it is not mentioned, we choose the regular basis  $(m, l)$  as

the basis of  $\partial E(K)$ . The slope  $p/q$  of a Dehn surgery indicates that the meridian of the filling solid torus is glued to the curve corresponding to  $pm + ql$ .

Suppose  $M$  is an oriented manifold. Let  $-M$  denote the same manifold with the reverse orientation, called the *mirror manifold* of  $M$ . Suppose  $K$  is an (oriented) knot in a 3-manifold  $M$ . Then it is specified by the knot complement  $E(K)$  and the (oriented) meridian  $m$  of the knot. The *mirror image* of  $K$  is the knot in  $-M$  specified by  $(-M, -m)$ .

When mentioning that  $Y = L(p, q)$  is a lens space, we always suppose that  $p$  and  $q$  are integers satisfying  $\gcd(p, q) = 1$  and  $(p, q) \neq (0, 1)$ . In particular, the manifold  $S^1 \times S^2$  is not considered as a lens space. The lens space is oriented as follows. Let  $(T^2, \alpha_0, \beta_0)$  be the standard diagram of a lens space. Then the orientation on the  $\alpha_0$ -handlebody is induced from the standard embedding of a solid torus in  $\mathbb{R}^3$ . With this convention, the lens space  $L(p, q)$  is obtained from the  $p/q$  Dehn surgery on the unknot in  $S^3$ .

We recall some definitions about knots in closed 3-manifolds. Suppose  $K$  is a knot in a lens space  $Y$ .

The knot  $K$  is called a *trivial knot* or an *unknot* if it bounds a disk embedded in  $Y$ . It is called a *core knot* if  $E(K)$  is homeomorphic to a solid torus. It is called a *split knot* if  $Y$  contains a sphere which decomposes  $Y$  into a punctured lens space and a ball containing  $K$  in its interior. It is called a *composite knot* if  $Y$  contains a 2-sphere  $S$  which intersects  $K$  transversely in two points and  $S \cap E(K)$  is  $\partial$ -incompressible in  $E(K)$ . It is called a *prime knot* if it is not a composite knot.

The torus  $T^2 \subset Y$  in the standard diagram  $(T^2, \alpha_0, \beta_0)$  is called the *Heegaard torus* of  $Y$ . The knot  $K$  is called a  $(p, q)$  *torus knot in  $Y$*  if  $K$  can be isotoped to lie on the Heegaard torus as an essential curve with slope  $p/q$  in the standard diagram of  $Y$ . The unknot is considered as a torus knot. Complements of torus knots in lens spaces are Seifert fibered spaces.

The knot  $K$  is called a *satellite knot* if  $E(K)$  has an essential torus. For  $q > 1$ , the space  $C_{p,q}$  is obtained by removing a regular fiber from a solid torus with a  $(p, q)$  fibering, which is called a *cable space of type  $(p, q)$* . The knot  $K$  is called a  $(p, q)$  *cable knot* on  $K_0$  if  $K_0$  is knot in  $Y$  such that  $E(K) = E(K_0) \cup C_{p,q}$ . In this case, the knot  $K$  lies as an essential curve on  $\partial N(K_0)$ , and  $K$  is neither a longitude nor a meridian of  $K_0$ . It is well-known that composite knots are satellite knots. A cable knot on  $K_0$  with  $E(K_0)$  having an incompressible boundary is also a satellite knot.

### 2.1 (1, 1) knots

In this subsection, we review some facts about (1, 1) knots. Proofs are omitted.

A knot  $K$  in a closed 3–manifold  $Y$  has *tunnel number one* if there is a properly embedded arc  $\gamma$  in  $E(K)$  such that  $E(K) - N(\gamma)$  is a genus two handlebody. Equivalently, the knot complement  $E(K)$  admits a genus two Heegaard splitting. The arc  $\gamma$  is called an *unknotting tunnel* for  $K$ . A properly embedded arc  $\gamma$  in a handlebody  $H$  is called a *trivial arc* if there is an embedded disk  $D \subset H$  such that  $\partial D = \gamma \cup (D \cap \partial H)$ . The disk  $D$  is called the *canceling disk* of  $\gamma$ . A knot  $K$  in a 3–manifold  $Y$  admits a (1, 1) *decomposition* if there is a genus one Heegaard splitting  $Y = H_1 \cup_{T^2} H_2$  such that  $K \cap H_i$  is a properly embedded trivial arc  $k_i$  in  $H_i$  for  $i = 1, 2$ . In this case,  $Y$  is either a lens space (including  $S^3$ ), or  $S^1 \times S^2$ . A knot  $K$  that admits a (1, 1) decomposition is called a (1, 1) *knot*. We do not consider (1, 1) knots in  $S^1 \times S^2$ . Note that any (1, 1) knot has tunnel number one.

**Proposition 2.1** [44, Proposition 3.2] *If a nontrivial knot in a lens space has tunnel number one, then the complement is irreducible. Consequently, the complement is a Haken manifold.*

Doubly pointed Heegaard diagrams parametrize their corresponding (1, 1) knots. The orientation of the knot is unimportant in this paper so we may swap the two basepoints.

**Proposition 2.2** [14; 37] *For  $p, q, r, s \in \mathbb{N}$  satisfying  $2q + r \leq p$  and  $s < p$ , a (1, 1) decomposition of a knot determines and is determined by a doubly pointed Heegaard diagram. After isotopy, such a diagram looks like  $(T^2, \alpha, \beta, z, w)$  in Figure 1, right, where  $p$  is the total number of intersection points,  $q$  is the number of strands around each basepoint,  $r$  is the number of strands in the middle band, and the  $i^{\text{th}}$  point on the right-hand side is identified with the  $(i + s)^{\text{th}}$  point on the left-hand side.*

Let  $W(p, q, r, s) = W(p, q, r, s)_+$  denote the (1, 1) knot defined by Figure 1, right, and let  $W(p, q, r, s)_-$  denote the knot defined by the diagram that is vertically symmetric to Figure 1, right. These doubly pointed Heegaard diagrams are called (1, 1) *diagrams*. In the diagrams, strands around basepoints are called *rainbows* and strands in the bands are called *stripes*. The roles of the curves  $\alpha$  and  $\beta$  here are different from in [37]. For the same parameters, the knot  $W(p, q, r, s)$  is the mirror image of  $K(p, q, r, s)$  in [37].

**Proposition 2.3** *There are relations among (1, 1) knots:*

- (i)  $W(p, q, r, s)_+$  is the mirror image of  $W(p, q, r, p - s)_-$ .

(ii)  $W(p, q, r, s)_+$  is equivalent to  $W(p, q, p - 2q - r, s - 2q)_-$ .

Thus, we know that  $W(p, q, r, s)_+$  is the mirror image of  $W(p, q, p - 2q - r, p - s + 2q)_+$ .

**Proof** The first relation is from the vertical symmetry. The second relation is from redrawing the diagram so that the lower band becomes the middle band and the middle band becomes the lower band.  $\square$

**Definition 2.4** For a closed 3-manifold  $Y$ , consider the hat version of Heegaard Floer homology  $\widehat{HF}(Y)$  defined in [31]. A closed 3-manifold  $Y$  is called an  $L$ -space if  $\widehat{HF}(Y, \mathfrak{s}) \cong \mathbb{Z}$  for any  $\mathfrak{s} \in \text{Spin}^c(Y)$ . A knot  $K$  in an  $L$ -space  $Y$  is called an  $L$ -space knot if a nontrivial Dehn surgery on  $K$  gives an  $L$ -space.

**Theorem 2.5** [15, Theorem 1.2] *A  $(1, 1)$  knot is an  $L$ -space knot if and only if, in the corresponding  $(1, 1)$  diagram with any orientation of  $\beta$ , all of rainbows around a fixed basepoint are oriented in the same way.*

**Definition 2.6** [38, Section 2.1] Let  $(T^2, \alpha_0, \beta_0)$  be the standard Heegaard diagram of  $L(p, q)$  and let  $P_i$  for  $i \in \mathbb{Z}/p\mathbb{Z}$  be components of  $T^2 - \alpha_0 \cup \beta_0$ , ordered from left to right. Let  $z \in P_1$  and  $w \in P_{k+1}$  be two points. The knot defined by  $(T^2, \alpha_0, \beta_0, z, w)$  is called a *simple knot*, and is denoted by  $S(p, q, k)$  (or by  $K(p, q, k)$  in [38]). The orientation of the knot is induced by the orientation of the arc connecting  $z$  to  $w$ .

**Proposition 2.7** [38, Lemma 2.5] *There are relations among simple knots  $S(p, q, k)$ :*

- (i)  $S(p, q, -k)$  is the orientation-reverse of  $S(p, q, k)$ .
- (ii)  $S(p, -q, -k)$  is the mirror image of  $S(p, q, k)$ .
- (iii)  $S(p, q, k) \cong S(p, q', kq')$ , where  $qq' \equiv 1 \pmod{p}$ .

Note that a simple knot is homotopic to an immersed curve on  $T^2$ . The homology class  $[S(p, q, k)]$  in  $H_1(L(p, q); \mathbb{Z})$  is  $k[b]$ , where  $b$  is the core curve of the  $\beta_0$ -handlebody. The simple knots  $S(p, q, k_1)$  and  $S(p, q, k_2)$  represent the same homology class if and only if  $k_1 \equiv k_2 \pmod{p}$ . Thus, there is no relation other than those in Proposition 2.7.

## 2.2 2-Bridge knots

In this subsection, we review some facts about 2-bridge links from [6; 25; 35].

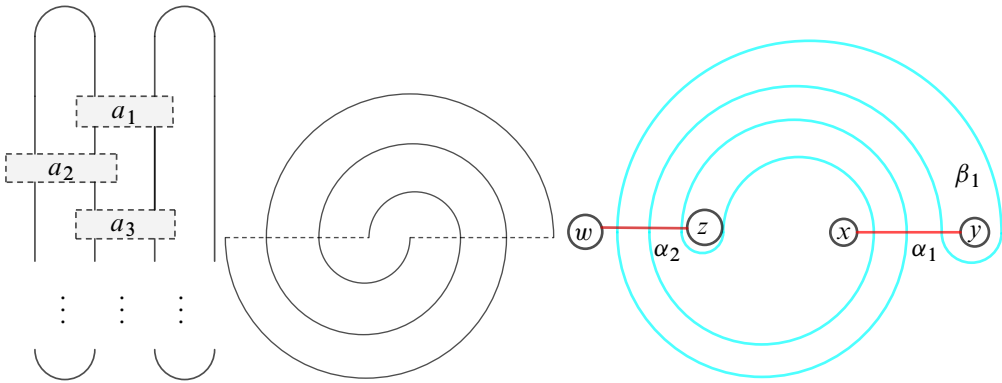


Figure 3: Left: 2-bridge. Center:  $b(3, 1)$ . Right: diagram of  $E(b(3, 1))$ .

**Definition 2.8** Suppose  $h$  is the height function given by the  $z$ -coordinate in  $\mathbb{R}^3 \subset S^3$ . A knot or a link in  $S^3$  is called a *2-bridge knot* or a *2-bridge link* if it can be isotoped in a presentation so that  $h$  has two maxima and two minima on it. Such a presentation is called the *standard presentation* of the knot.

A 2-bridge link has two components. Each component is equivalent to the unknot. Suppose integers  $a$  and  $b$  satisfy  $\gcd(a, b) = 1$  and  $a > 1$ . For every oriented lens space  $L(a, b)$ , there is a unique 2-bridge knot or link whose branched double cover space is diffeomorphic to  $L(a, b)$ . Let  $b(a, b)$  denote the knot or link related to  $L(a, b)$ . It is a knot if  $a$  is odd, and a link if  $a$  is even. Thus, the classification of 2-bridge knots or links depends on the classification of lens spaces [5]. For  $i = 1, 2$ , two 2-bridge knots or links  $b(a_i, b_i)$  are equivalent if and only if  $a_1 = a_2 = a$  and  $b_1 \equiv b_2^{\pm 1} \pmod{a}$ .

Suppose  $a/b$  is represented as the continued fraction

$$[0; a_1, -a_2, \dots, (-1)^{m+1}a_m] = 0 + \frac{1}{a_1 - \frac{1}{a_2 - \frac{1}{a_3 - \dots}}}$$

Moreover, suppose  $m$  is odd. The standard presentation of a 2-bridge knot or link  $b(a, b)$  looks like Figure 3, left, where the  $|a_i|$  for  $i \in [1, m]$  represent numbers of half-twists in the boxes, and signs of the  $a_i$  represent signs of half-twists. Different choices of continued fractions give the same knot or link. For any 2-bridge knot or link, the numbers  $(-1)^{i+1}a_i$  can be all positive, which implies any 2-bridge knot or link is alternating.

The knot or link  $b(a, b)$  admits another canonical presentation known as the *Schubert normal form*. It induces a Heegaard diagram of  $E(b(a, b))$  and a doubly pointed

Heegaard diagram of  $\mathfrak{b}(a, b)$ . Figure 3, center, gives an example of the Schubert normal form of  $\mathfrak{b}(3, 1)$  and Figure 3, right, is the corresponding Heegaard diagram of the knot complement. The corresponding doubly pointed Heegaard diagram is obtained by replacing  $\alpha_2$  by two basepoints,  $z$  and  $w$ . Two horizontal strands in the Schubert normal form are arcs near two maxima in the standard presentation. Thus, two 1–handles attached to points  $w$  and  $z$ , and  $x$  and  $y$  in Figure 3, right, are neighborhoods of these arcs.

**Proposition 2.9** [35] *Suppose  $K = \mathfrak{b}(a, b)$  with  $b$  odd and  $|b| < a$ . The symmetrized Alexander polynomial  $\Delta_K(t)$  and the signature  $\sigma(K)$  satisfy*

$$\Delta_K(t) = t^{-\frac{1}{2}\sigma(K)} \sum_{i=0}^{a-1} (-1)^i t^{\sum_{j=0}^i (-1)^{\lfloor ib/a \rfloor}} \quad \text{and} \quad \sigma(K) = \sum_{i=1}^{a-1} (-1)^{\lfloor ib/a \rfloor}.$$

**Proposition 2.10** [8; 18] *Let  $K$  be a  $(1, 1)$  knot in a lens space. Then  $K$  is a split knot if and only if  $K$  is the unknot. The knot  $K$  is a composite knot if and only if it is a connected sum of a 2–bridge knot and a core knot of a lens space.*

### 3 Parametrization and characterization

For a constrained knot  $K$ , there is a standard diagram  $(T^2, \alpha_1, \beta_1, z, w)$  of  $K$ , defined in the introduction. Based on standard diagrams, we describe the parametrization of constrained knots. For integers  $p, q$  and  $q'$  satisfying

$$\gcd(p, q) = \gcd(p, q') = 1 \quad \text{and} \quad qq' \equiv 1 \pmod{p}$$

we know that  $L(p, q)$  is diffeomorphic to  $L(p, q')$  [5]. Suppose  $(T^2, \alpha_0, \beta_0)$  is the standard diagram of  $L(p, q')$ , ie the curve  $\beta_0$  is obtained from a straight line of slope  $p/q'$  in  $\mathbb{R}^2$ , and suppose that the diagram  $(T^2, \alpha_1, \beta_1, z, w)$  is induced by  $(T^2, \alpha_0, \beta_0)$  as in the introduction. The curves  $\alpha_0$  and  $\beta_0$  divide  $T^2$  into  $p$  regions, which are parallelograms in Figure 1, left; see also Figure 4, left. A new diagram  $C$  is obtained by gluing top edges and bottom edges of parallelograms. We can shape  $C$  into a square. An example is shown in Figure 4, where  $p = 5$ ,  $q = 3$  and  $q' = 2$ .

For  $i \in \mathbb{Z}/p\mathbb{Z}$ , let  $D_i$  denote rectangles in  $C$ , ordered from the bottom edge to the top edge. Since  $qq' \equiv 1 \pmod{p}$  and we start with the standard diagram of  $L(p, q')$ , we know that the right edge of  $D_j$  is glued to the left edge of  $D_{j+q}$ . The bottom edge  $e_b$  of  $D_1$  is glued to the top edge  $e_t$  of  $D_p$ . By definition of a constrained knot, the curve



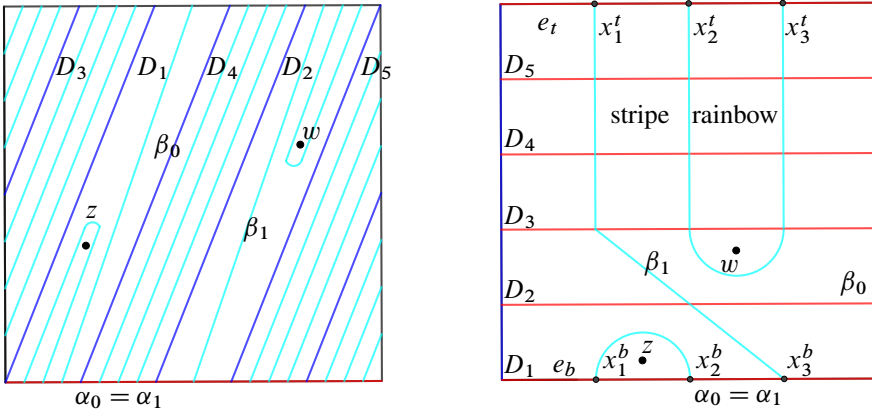


Figure 4: Heegaard diagrams of  $C(5, 3, 2, 3, 1)$ .

$\alpha_1$  is the same as  $\alpha_0$  and the curve  $\beta_1$  is disjoint from  $\beta_0$ . Thus, in this new diagram  $C$ , the curve  $\alpha_1$  is the union of  $p$  horizontal lines and  $\beta_1$  is the union of strands which are disjoint from vertical edges of  $D_i$  for  $i \in \mathbb{Z}/p\mathbb{Z}$ .

Similar to the definitions for  $(1, 1)$  knots, strands in the standard diagram of a constrained knot are called *rainbows* and *stripes*. Boundary points of a rainbow and a stripe are called *rainbow points* and *stripe points*, respectively. A rainbow must bound a basepoint, otherwise it can be removed by isotopy. Numbers of rainbows on  $e_b$  and  $e_t$  are the same since the numbers of rainbow points are the same. Without loss of generality, suppose  $z$  is in all rainbows on  $e_b$  and  $w$  is in all rainbows on  $e_t$ . Let  $x_i^b$  and  $x_i^t$  for  $i \in [1, u]$  be boundary points on the bottom edge and the top edge, respectively, ordered from left to right in Figure 4, right.

**Lemma 3.1** *The number  $u$  of boundary points on  $e_b$  or  $e_t$  is odd. When  $u = 1$ , there is no rainbow and only one stripe. When  $u > 1$ , there exists an integer  $v \in (0, \frac{1}{2}u)$  such that one of the following cases happens:*

- (i) *The set  $\{x_i^b \mid i \leq 2v\} \cup \{x_i^t \mid i > u - 2v\}$  contains all rainbow points.*
- (ii) *The set  $\{x_i^t \mid i \leq 2v\} \cup \{x_i^b \mid i > u - 2v\}$  contains all rainbow points.*

**Proof** The algebraic intersection number of  $\beta_1$  and  $e_b$  is odd. Hence  $u$  is also odd. If  $u = 1$ , then the argument is clear.

Suppose  $u > 1$ ; we show the last argument in three steps. Firstly, if both  $x_i^b$  and  $x_j^b$  are boundary points of the same rainbow  $R$ , then  $x_k^b$  for  $i < k < j$  are all rainbow

points, otherwise the stripe corresponding to the stripe point  $x_k^b$  would intersect  $R$ . Thus, rainbow points on  $e_b$  are consecutive. The same assertion holds for  $x_i^t$ .

Secondly, one of  $x_1^b$  and  $x_1^t$  must be a rainbow point. Indeed, if this were not true then both  $x_1^b$  and  $x_1^t$  would be stripe points. They cannot be boundary points of the same stripe, otherwise  $\beta_1$  would not be connected. They cannot be boundary points of different stripes, otherwise two corresponding stripes would intersect each other. Thus, the assumption is false. Similarly, one of  $x_u^b$  and  $x_u^t$  must be a rainbow point.

Finally, if  $x_1^b$  is a rainbow point then  $x_u^b$  cannot be a rainbow point, otherwise all points would be rainbow points. As discussed above, the point  $x_u^t$  is a rainbow point. Since the number of rainbow points on  $e_t$  is even, there exists an integer  $v$  satisfying case (i). If  $x_1^t$  is a rainbow point, similar argument implies there exists  $v$  satisfying case (ii).  $\square$

When  $u = 1$ , after isotoping  $\beta_1$ , suppose the unique stripe is a vertical line in  $C - \{z, w\}$ . By moving  $z$  through the left edge or the right edge if necessary, suppose basepoints  $z$  and  $w$  are in different sides of the stripe. If  $z$  is on the left of the stripe, set  $v = 0$ . If  $z$  is on the right of the stripe, set  $v = 1$ .

Then suppose  $u > 1$ . When in case (i) of Lemma 3.1, rainbows on  $e_b$  connect  $x_i^b$  to  $x_{2v+1-i}^b$  for  $i \in [1, v]$ , rainbows on  $e_t$  connect  $x_{u+1-i}^t$  to  $x_{u-2v+i}^t$  for  $i \in [1, v]$ , and stripes connect  $x_j^b$  to  $x_{u+1-j}^t$  for  $j \in [2v + 1, u]$ . When in case (ii) of Lemma 3.1, the setting is obtained by replacing  $i$  and  $j$  by  $u + 1 - i$  and  $u + 1 - j$ , respectively. Without loss of generality, suppose  $z$  is in  $D_1$ , and  $w$  is in  $D_l$ . Note that now basepoints cannot be moved through vertical edges of  $C$ . Otherwise the rainbows would intersect the vertical edges, which contradicts the definition of the constrained knot. Then we parametrize constrained knots in  $L(p, q')$  by the tuple  $(l, u, v)$  for case (i) and  $(l, u, u - v)$  for case (ii). Since  $\beta_1$  is connected, we have  $\gcd(u, v) = 1$ . In summary, the following theorem holds:

**Theorem 3.2** *Constrained knots are parametrized by five integers  $(p, q, l, u, v)$ , where  $p > 0, q \in [1, p - 1], l \in [1, p], u > 0, v \in [0, u - 1], \gcd(p, q) = \gcd(u, v) = 1$  and  $u$  is odd. Moreover,  $v \in [1, u - 1]$  when  $u > 1$  and  $v \in \{0, 1\}$  when  $u = 1$ .*

Note that the parameter  $v$  in Theorem 3.2 is different from the integer  $v$  in case (ii) of Lemma 3.1. Intuitively, for  $v \in [1, u - 1]$  in the parametrization  $(p, q, l, u, v)$  with  $u > 1$ , the number  $\min\{v, u - v\}$  is the number of rainbows around a basepoint.

For parameters  $(p, q, l, u, v)$ , let  $C(p, q, l, u, v)$  denote the corresponding constrained knot. When considering the orientation, let  $C(p, q, l, u, v)^+$  denote the knot induced by

$(T, \alpha_1, \beta_1, z, w)$  and let  $C(p, q, l, u, v)^-$  denote the knot induced by  $(T, \alpha_1, \beta_1, w, z)$ . For  $q \notin [1, p - 1]$  and  $l \notin [1, p]$ , consider the integers  $q$  and  $l$  modulo  $p$ . If  $u > 1$  and  $v \notin [1, u - 1]$ , consider the integer  $v$  modulo  $u$ . For  $p < 0$ , let  $C(p, q, l, u, v)$  denote  $C(-p, -q, l, u, v)$ .

**Remark 3.3** The knot  $C(p, q, l, u, v)$  is in  $L(p, q')$ , where  $qq' \equiv 1 \pmod{p}$ . Even though  $L(p, q)$  is diffeomorphic to  $L(p, q')$ , constrained knots  $C(p, q, l, u, v)$  and  $C(p, q', l, u, v)$  are not necessarily equivalent. For example, Theorem 1.2 implies that constrained knots  $C(5, 2, 3, 3, 1)$  and  $C(5, 3, 3, 3, 1)$  are not equivalent.

We now provide some basic propositions of constrained knots. Also, we indicate the relationship of constrained knots with other families of knots mentioned in Section 2.

**Proposition 3.4**  $C(p, -q, l, u, -v)$  is the mirror image of  $C(p, q, l, u, v)$  for  $u > 1$ .  $C(p, -q, l, 1, 1)$  is the mirror image of  $C(p, q, l, 1, 0)$ .

**Proof** This follows from the vertical reflection of the standard diagram. □

Hence we only consider  $C(p, q, l, u, v)$  with  $0 \leq 2v < u$  in the rest of the paper.

**Proposition 3.5**  $C(1, 0, 1, u, v) \cong \mathfrak{b}(u, v)$ .

**Proof** By cutting along  $\alpha_1$  and a small circle around  $x$  in Figure 3, right, the doubly pointed diagram of a 2-bridge knot can be shaped into a square. This proposition is clear by comparing this diagram with the new diagram  $C$  related to  $C(1, 0, 1, u, v)$ . □

**Proposition 3.6** For any fixed orientations of  $\alpha_1$  and  $\beta_1$  in the standard diagram of a constrained knot, intersection points  $x_i^b$  have alternating signs and adjacent strands of  $\beta_1$  in the new diagram  $C$  have opposite orientations.

**Proof** From a similar observation in the proof of Proposition 3.5, for  $C(p, q, l, u, v)$ , the curve  $\beta_1$  in the new diagram  $C$  is same as the curve  $\beta$  in the doubly pointed Heegaard diagram of  $\mathfrak{b}(u, v)$ . Thus, it suffices to consider the 2-bridge knot  $\mathfrak{b}(u, v)$ . The Schubert normal form of  $\mathfrak{b}(u, v)$  is the union of two dotted horizontal arcs behind the plane and two winding arcs on the plane. Suppose  $\gamma$  is one of the winding arcs. Then  $\beta_1 = \partial N(\gamma)$  cuts the plane into two regions, the inside region  $\text{int } N(\gamma)$  and the outside region  $\mathbb{R}^2 - N(\gamma)$ . Points  $x$  and  $y$  in Figure 3, right, are in different regions

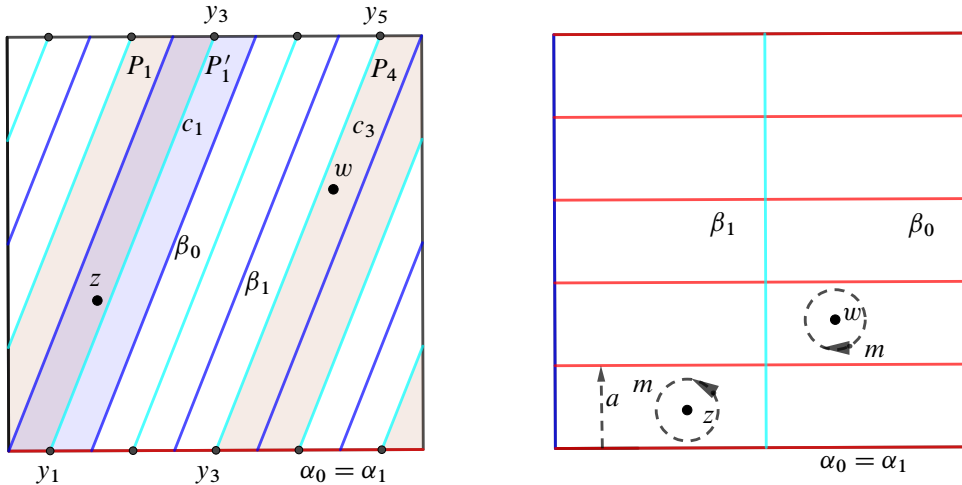


Figure 5:  $S(5, 2, 3) \cong C(5, 3, 2, 1, 0)^+$ , where regions  $P_1, P_4$  and  $P'_1$  are indicated by shadow.

and points  $x_i^b$  are on the arc connecting  $x$  to  $y$ . Since regions on different sides of  $\beta_1$  must be different, the arc connecting  $x$  to  $y$  is cut by  $x_i^b$  into pieces that lie in the inside region and the outside region alternately. For each piece of the arc, the endpoints are boundary points of a connected arc in  $\beta_1$ . Thus, signs of  $x_i^b$  are alternating. The orientations on strands of  $\beta_1$  are induced by signs of  $x_i^b$ . Hence adjacent strands of  $\beta_1$  have opposite orientations. □

**Proposition 3.7** For  $p, q, q' \in \mathbb{Z}$  satisfying  $qq' \equiv 1 \pmod{p}$ , there are relations

- (i)  $S(p, q', k) \cong C(p, q, l, 1, 0)^+$ , where  $k - 1 \equiv (l - 1)q' \pmod{p}$ ,
- (ii)  $S(p, q', k) \cong C(p, q, l, 1, 1)^+$ , where  $k + 1 \equiv (l - 1)q' \pmod{p}$ .

**Proof** Consider curves  $\alpha = \alpha_0 = \alpha_1, \beta_0$  and  $\beta_1$  in the definition of a constrained knot. When  $u = 1$ , the curve  $\beta_1$  is parallel to  $\beta_0$ . Consider the new diagram  $C$  and regions  $D_i$  for  $i \in \mathbb{Z}/p\mathbb{Z}$  as in Figure 4, right. Suppose components of  $T^2 - \alpha \cup \beta_1$  are  $P_i$  and components of  $T^2 - \alpha \cup \beta_0$  are  $P'_i$ , ordered from left to right as in Figure 5 so that  $z \in P_1 \cap P'_1$ . Suppose  $y_i$  are intersection points of  $\alpha$  and  $\beta_1$  on the bottom edge of  $P'_i$ . The strand  $c_i = \beta_1 \cap P'_i$  connects  $y_i$  to  $y_{i+q'}$ , so the strand  $\beta_1 \cap D_l$  in the new diagram  $C$  is  $c_{1+(l-1)q'}$ . When  $v = 0$ , the other basepoint  $w$  is in  $P_{(l-1)q'+2}$ , so  $k \equiv (l - 1)q' + 1 \pmod{p}$ . When  $v = 1$ , the other basepoint  $w$  is in  $P_{(l-1)q'}$ , so  $k \equiv (l - 1)q' - 1 \pmod{p}$ . □

**Corollary 3.8** For  $p, q, q' \in \mathbb{Z}$  satisfying  $qq' \equiv 1 \pmod{p}$ , there are relations

- (i)  $C(p, q, l, 1, 0) \cong C(p, q, l + 2q, 1, 1)$ ,
- (ii)  $C(p, q, l, 1, 0)^+ \cong C(p, q, -2q + 2 - l, 1, 0)^-$ ,
- (iii)  $C(p, q, l, 1, 0)$  is the mirror image of  $C(p, -q, l + 2q, 1, 0) \cong C(p, -q, l, 1, 1)$ ,
- (iv)  $C(p, q, l, 1, 0) \cong C(p, q', q'l - 2q' + 2, 1, 0)$ ,
- (v)  $S(p, q, k) \cong S(p, q', kq') \cong C(p, q, k - q + 1, 1, 0)^+$ .

**Proof** These relations follow from Propositions 2.7 and 3.7. □

**Corollary 3.9** The knot  $C(p, q, -q + 1, 1, 0)$  is an unknot in a lens space. The knot  $C(p, q, l, 1, 0)$  for  $l = 1, -2q + 1, -q + 2$  or  $-q$  is a core knot of a lens space.

**Proof** The unknot case is obtained by substituting  $k = 0$  in case (v) of Corollary 3.8. Note that  $S(p, q, 0)$  is the unknot: the knot is isotopic to a circle bounding a disk on  $T^2$ . The core knot cases are obtained by substituting  $k = \pm 1, \pm q$  in case (v) of Corollary 3.8. Note that  $S(p, q, q)$  is isotopic to a simple closed curve on  $T^2$  that intersects  $\alpha$  once, which also is isotopic to the core curve of the  $\alpha$ -handlebody. By Proposition 3.7, simple knots  $S(p, q, \pm q)$  and  $S(p, q, \pm 1)$  are also core knots. □

**Proposition 3.10** For  $K = C(p, q, l, 1, 0)$ , we have a presentation of the homology

$$H_1(E(K); \mathbb{Z}) \cong \langle [a], [m] \rangle / (p[a] + k[m]) \cong \mathbb{Z} \oplus \mathbb{Z} / \gcd(p, k)\mathbb{Z},$$

where  $m$  is the circle in Figure 5,  $a$  is the core curve of  $\alpha_0$ -handle and  $k \in (0, p]$  satisfies  $k - 1 \equiv (l - 1)q^{-1} \pmod{p}$ .

**Proof** This follows from Proposition 3.7 and results in [38, Section 3.3]. □

**Proposition 3.11** Suppose  $C(p, q, l, u, v)$  is a constrained knot in  $L(p, q')$  with  $0 \leq 2v < u$ . Let  $q_i \in [0, p)$  be integers satisfying  $q_i \equiv iq' \pmod{p}$  and let  $k \in [1, p]$  be the integer satisfying  $k - 1 \equiv (l - 1)q' \pmod{p}$ . Moreover, let

$$n_1 = \#\{i \in [0, l - 1] \mid q_i \in [0, k - 1]\} \quad \text{and} \quad n_2 = \#\{i \in [0, l - 1] \mid q_i \in [1, q' - 1]\}.$$

Then  $C(p, q, l, u, v) \cong W(pu - 2v(l - 1), v, uk - 2vn_1, uq' - 2vn_2)$ .

**Proof** The parameters  $(p - l + 1)u + (l - 1)(u - 2v) = pu - 2v(l - 1)$  and  $v$  are from counting the numbers of intersection points and rainbows in the standard

diagram of a constrained knot, respectively. Suppose that  $P'_i$  are components of  $T^2 - \alpha_0 \cup \beta_0$  in the standard diagram of  $L(p, q')$ , ordered from left to right so that  $z \in P'_1$ . Similar to the proof of Proposition 3.7, we know  $w \in P'_k$ . Then the parameter  $(k - n_1)u + n_1(u - 2v) = uk - 2vn_1$  counts the number of stripes between rainbows and the parameter  $(q' - n_2)u + n_2(u - 2v) = uq' - 2vn_2$  counts the twisting number.  $\square$

**Proof of Theorem 1.11** For a knot  $K$  in a lens space with Seifert fibered complement, any Dehn surgery other than the one along homological longitude gives a Seifert fibered space. By discussion in [39, Section 5], all oriented Seifert fibered spaces over  $\mathbb{R}P^2$  are L-spaces and the classification of L-spaces over  $S^2$  is given by [39, Theorem 5.1]. Moreover, no higher genus Seifert fibered spaces are L-spaces. The classification in [39, Theorem 5.1] indicates there are at least two Dehn fillings on the knot complement that are L-spaces, ie  $K$  is always an L-space knot. By Proposition 3.11, we can transform standard diagrams of constrained knots into  $(1, 1)$  diagrams. By Proposition 3.6 and Theorem 2.5, a constrained knot is an L-space knot if and only if  $(u, v) = (1, 0)$  or  $(1, 1)$ , or  $u > 1$  and  $v = \pm 1$ .  $\square$

**Proof of Proposition 1.1** The necessary part of the proposition follows directly from the definition of constrained knots: the intersection points of  $\alpha_1$  and  $\beta_1$  between two consecutive intersection points of  $\alpha_0$  and  $\beta_0$  correspond the same  $\text{spin}^c$  structure on the lens space, where  $\alpha_1, \beta_1, \alpha_0$  and  $\beta_0$  are curves in the standard diagrams of the constrained knot and the lens space.

We prove the sufficient part of this proposition. For simplicity, intervals are considered in  $\mathbb{Z}/p\mathbb{Z}$ . In particular, let  $(p_k, p_1]$  denote  $(0, p_1] \cup (p_k, p]$ . Consider intersection points  $x_i$  for  $i \in [1, p]$  as shown in Figure 1, right.

Firstly,  $\text{spin}^c$  structures  $\mathfrak{s}_i$  are equal for all  $i \in [r + 1, r + 2q]$ . Indeed, for  $i \in [1, q]$ , the points  $x_{r+i}$  and  $x_{r+2q+1-i}$  are boundary points of a rainbow, that is there is a holomorphic disk connecting  $x_{r+i}$  to  $x_{r+2q+1-i}$ . Thus  $\mathfrak{s}_{r+i} = \mathfrak{s}_{r+2q+1-i}$ . If  $q = 1$ , this assertion is trivial. If  $q > 1$  and the assertion did not hold, then there must be an integer  $q_0$  and two  $\text{spin}^c$  structures  $\mathfrak{s}_A$  and  $\mathfrak{s}_B$  such that  $\mathfrak{s}_i = \mathfrak{s}_A$  for all  $i \in [r + q_0, r + 2q + 1 - q_0]$  and  $\mathfrak{s}_j = \mathfrak{s}_B$  for all  $j \notin [r + q_0, r + 2q + 1 - q_0]$ , which implies  $a = 2$ . Since  $\text{spin}^c$  structures of two boundary points of a stripe are different, for all  $i \in [2q + 1 - s, p - s]$ ,  $\text{spin}^c$  structures  $\mathfrak{s}_i$  are different from  $\mathfrak{s}_B$ . Thus  $\mathfrak{s}_i = \mathfrak{s}_A$  for all  $i \in [2q + 1 - s, p - s]$ . For  $i \in [1, q]$ , points  $x_{i-s}$  and  $x_{2q+1-i-s}$  are boundary points of a rainbow, so  $\mathfrak{s}_{i-s} = \mathfrak{s}_{2q+1-i-s}$ . Since there are  $2q_0$  points corresponding

to  $\mathfrak{s}_A$ , integers  $q_0$  should satisfy the inequality  $2q_0 > p - 2q$ . For  $i \in [q + q_0 - \frac{1}{2}p, q]$ , points  $x_{i-s}$  and  $x_{2q+1-i-s}$  correspond to  $\mathfrak{s}_B$ . In particular, points  $x_{r+1}$  and  $x_{r+2q}$  are identified with  $x_{2q+1-i_0-s}$  and  $x_{i_0-s}$  for  $i_0 = q + q_0 - \frac{1}{2}p$ , respectively. Let  $R_1$  be the rainbow with boundary points  $x_{r+1}$  and  $x_{r+2q}$ , and let  $R_2$  be the rainbow with boundary points  $x_{2q+1-i_0-s}$  and  $x_{i_0-s}$ . The union of  $R_1$  and  $R_2$  becomes a component of  $\beta$ , which contradicts the assumption that  $\beta$  only has one component.

We can similarly show that the  $\text{spin}^c$  structures  $\mathfrak{s}_i$  are equal for all  $i \in [1 - s, 2q - s]$ . From this discussion, for any  $i \in [1, k]$ , we have

$$p_i \neq r + 1, r + 2, \dots, r + 2q - 1, 1 - s, 2 - s, \dots, 2q - 1 - s.$$

Suppose  $y_i$  for  $i \in [1, k]$  are points on  $\alpha$  between  $x_{p_i}$  and  $x_{p_i+1}$ . If  $p_i \neq r, r + 2q$  or  $p$ , then  $p_i$  and  $p_i + 1$  must be boundary points of two successive stripes. Suppose  $x_j$  and  $x_{j+1}$  are the other boundary points of these stripes, respectively. There must be a point  $y_j$  between  $x_j$  and  $x_{j+1}$  because  $\mathfrak{s}_j - \mathfrak{s}_{j+1} = \mathfrak{s}_{p_i} - \mathfrak{s}_{p_i+1} \neq 0$ . Let  $b_i$  be a strand connecting  $y_i$  to  $y_j$  which is disjoint from  $\beta$ .

Suppose  $p_i = p$ . If  $r \neq 0$  and  $p - 2q - r \neq 0$  there are stripes connecting  $\mathfrak{s}_p$  to  $\mathfrak{s}_{p-s}$  and connecting  $\mathfrak{s}_1$  to  $\mathfrak{s}_{2q+1-s}$ , respectively. Thus  $\mathfrak{s}_{p-s} - \mathfrak{s}_{2q+1-s} = \mathfrak{s}_p - \mathfrak{s}_1 \neq 0$ . There is a point  $y_j$  either between  $x_{p-s}$  and  $x_{1-s}$  or between  $x_{2q-s}$  and  $x_{2q+1-s}$  for some  $j$ . Only one case will happen because the number of intersection points corresponding to any fixed  $\text{spin}^c$  structure is odd. Let  $b_i$  be a strand connecting  $y_i$  to  $y_j$  which is disjoint from  $\beta$ . If either  $r = 0$  or  $p - 2q - r = 0$ , by choosing different stripes,  $\mathfrak{s}_{p-s} - \mathfrak{s}_{2q+1-s} \neq 0$  still holds. The point  $y_j$  and the strand  $b_i$  can also be found. By a similar argument, this is also true for  $p_i = r$  and  $r + 2q$ .

Let  $\beta_0$  be the union of  $b_i$ . Without considering basepoints,  $\beta_0$  is isotopic to  $\beta$ . Thus, it has only one component. Finally, the curves  $\beta_0$ ,  $\alpha$  and  $\beta$  can be identified with  $\beta_0$ ,  $\alpha_1$  and  $\beta_1$  in the definition of a constrained knot. Thus, we conclude that the  $(1,1)$  knot is a constrained knot. □

## 4 Knot Floer homology

Heegaard Floer homology is an invariant for closed 3-manifolds discovered by Ozsváth and Szabó [30; 31]. It has been generalized to knot Floer homology [29; 36], sutured Floer homology [20], bordered Floer homology [22] and immersed curves for manifolds

with torus boundary [16; 17]. See [38, Section 3] for a brief review of knot Floer homology for rationally null-homologous knots. See also [33].

In this section suppose  $K = C(p, q, l, u, v)$  is a constrained knot in  $Y = L(p, q')$ , where  $qq' \equiv 1 \pmod{p}$ . Write  $H_1 = H_1(E(K); \mathbb{Z})$  and  $\widehat{HFK}(K) = \widehat{HFK}(Y, K)$  for short. For any homogeneous element  $x \in \widehat{HFK}(K)$ , let  $\text{gr}(x) \in H_1$  be the Alexander grading of  $x$  mentioned in the introduction. Note that the Alexander grading is well-defined up to a global grading shift [11], ie up to multiplication by an element in  $H_1$ . However, the difference  $\text{gr}(x) - \text{gr}(y)$  for two homogeneous elements  $x$  and  $y$  is always well-defined. This difference can be calculated explicitly by the doubly pointed Heegaard diagram of the knot by the approach in [38, Section 3.3].

Consider the group ring  $\mathbb{Z}[H_1]$ . Two elements  $f_1$  and  $f_2$  in  $\mathbb{Z}[H_1]$  are *equivalent*, denoted by  $f_1 \sim f_2$ , if there exists an element  $g \in \pm H_1$  such that  $f_1 = gf_2$ . For any element  $h \in H_1$ , there is a grading summand  $\widehat{HFK}(K, h)$  of  $\widehat{HFK}(K)$  as in (2). There is also a relative  $\mathbb{Z}/2$  grading on  $\widehat{HFK}(K)$  induced by signs of the intersection numbers in the Heegaard diagram (see [11, Section 2.4]) and related to the modulo 2 Maslov grading on  $\widehat{HFK}(K, \mathfrak{s})$ . This grading respects the Alexander grading and induces a  $\mathbb{Z}/2$  grading on  $\widehat{HFK}(K, h)$ . Then the Euler characteristic  $\chi(\widehat{HFK}(K, h))$  is well-defined up to sign. We can consider the (graded) Euler characteristic of  $\widehat{HFK}(K)$ :

$$\begin{aligned} \chi(\widehat{HFK}(K)) &= \sum_{h \in H_1} \chi(\widehat{HFK}(K, h)) \cdot h \\ &= \sum_{h \in H_1} (\text{rk } \widehat{HFK}_{\text{even}}(K, h) - \text{rk } \widehat{HFK}_{\text{odd}}(K, h)) \cdot h. \end{aligned}$$

From the above discussion, we know  $\chi(\widehat{HFK}(K))$  is an element in  $\mathbb{Z}[H_1]$  up to equivalence. For  $\mathfrak{s} \in \text{Spin}^c(Y)$ , we consider  $\widehat{HFK}(K, \mathfrak{s})$  as a subgroup of  $\widehat{HFK}(K)$  so that it also has an  $H_1$ -grading and  $\chi(\widehat{HFK}(K, \mathfrak{s}))$  is also an element in  $\mathbb{Z}[H_1]$  up to equivalence.

For a constrained knot  $K$ , we will show  $\widehat{HFK}(K)$  totally depends on  $\chi(\widehat{HFK}(K))$ . Explicitly this means that, for any  $h \in H_1$ , the dimension of  $\widehat{HFK}(K, h)$  is the same as the absolute value  $|\chi(\widehat{HFK}(K, h))|$ .

As shown in Figures 4 and 6, suppose  $e^j$  is the top edge of  $D_j$  and  $x_i^j$  is the intersection point of  $e^j$  and  $\beta_1$  for  $j \in \mathbb{Z}/p\mathbb{Z}$  and  $i \in [1, u(j)]$ . Let  $x_{\text{middle}}^j = x_{(u(j)+1)/2}^j$  be middle points. It is clear that  $\mathfrak{s}_z(x_{i_1}^{j_1}) = \mathfrak{s}_z(x_{i_2}^{j_2})$  if and only if  $j_1 = j_2$ . For any integer  $j \in [1, p]$ , define  $\mathfrak{s}_j = \mathfrak{s}_z(x_{\text{middle}}^j) \in \text{Spin}^c(Y)$ .



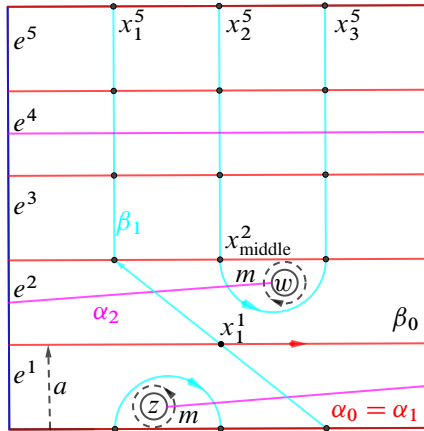


Figure 6: Heegaard diagram of  $E(C(5, 3, 2, 3, 1))$ .

**Lemma 4.1** For  $K = C(p, q, l, u, v)$  with  $u > 2v > 0$ , suppose  $k \in (0, p]$  is the integer satisfying  $k - 1 \equiv (l - 1)q^{-1} \pmod{p}$ . Define

$$k' = \begin{cases} k - 2 & \text{if } v \text{ is odd,} \\ k & \text{if } v \text{ is even.} \end{cases}$$

Suppose  $d = \gcd(p, k')$ . Then there is a presentation of the homology  $H_1$ :

$$H_1 = H_1(E(K); \mathbb{Z}) \cong \langle [a], [m] \rangle / (p[a] + k'[m]) \cong \mathbb{Z} \oplus \mathbb{Z} / d\mathbb{Z},$$

where  $m$  is the circle in Figure 6 and  $a$  is the core curve of  $\alpha_0$ -handle.

**Proof** Suppose  $\beta_1$  is oriented so that the orientation of the middle stripe is from bottom to top. Let  $[\beta_1(p, q, l, u, v)]$  denote the homology class of  $\beta_1$  corresponding to  $C(p, q, l, u, v)$ . By Proposition 3.6, orientations of rainbows around a basepoint are alternating. Note that moving all rainbows of  $\beta_1$  across basepoints gives the diagram of the simple knot  $C(p, q, l, 1, 0)$ . Then

$$\begin{aligned} [\beta_1(p, q, l, u, v)] + 2[m] &= [\beta_1(p, q, l, 1, 0)] & \text{if } v \text{ is odd,} \\ [\beta_1(p, q, l, u, v)] &= [\beta_1(p, q, l, 1, 0)] & \text{if } v \text{ is even.} \end{aligned}$$

Then this proposition follows from Proposition 3.10. Note that  $[a]$  and  $[m]$  correspond to core curves of  $\alpha_1$  and  $\alpha_2$ , and the relation in the presentation of  $H_1$  corresponds to algebraic intersection numbers  $\alpha_1 \cdot \beta$  and  $\alpha_2 \cdot \beta$ ; see Section 6 for the approach to obtain a presentation of  $\pi_1(E(K))$  and note that  $H_1$  is the abelianization of  $\pi_1(E(K))$ .  $\square$

**Lemma 4.2** (Proposition 1.4) *For  $K = C(p, q, l, u, v)$  with  $u > 2v \geq 0$ , suppose  $H_1$  is presented as in Lemma 4.1. For any integer  $j \in [1, p]$ , let  $\mathfrak{s}_j = \mathfrak{s}_z(x_{\text{middle}}^j)$  for intersection points  $x_{\text{middle}}^j$  in Figure 6. Then for any  $j$ , the group  $\widehat{HFK}(K, \mathfrak{s}_j)$  is determined by its Euler characteristic.*

Moreover, suppose integers  $u'$  and  $v'$  satisfy  $u' = u - 2v$  and  $v' \equiv v \pmod{u'}$ . Let  $\Delta_1(t)$  and  $\Delta_2(t)$  be Alexander polynomials of  $\mathfrak{b}(u, v)$  and  $\mathfrak{b}(u', v')$ , respectively. Then

$$\chi(\widehat{HFK}(K, \mathfrak{s}_j)) \sim \begin{cases} \Delta_1([m]) & \text{if } j \in [l, p], \\ \Delta_2([m]) & \text{if } j \in [1, l - 1]. \end{cases}$$

**Proof** For  $j \in [1, p]$ , consider the edge  $e^j$  and the intersection numbers  $x_i^j$  of  $e^j$  and  $\beta_1$  in the diagram  $C$ . Suppose  $(e^j)'$  is the curve obtained by identifying two endpoints of  $e^j$ . For  $j \in [l, p]$ , the diagram  $(T^2, (e^j)', \beta_1, z, w)$  is the same as the diagram of  $K_1 = \mathfrak{b}(u, v)$ . For  $j \in [1, l - 1]$ , by case (iii) of Lemma 7.8, the diagram  $(T^2, (e^j)', \beta_1, z, w)$  is isotopic to the diagram of  $K_2 = \mathfrak{b}(u', v')$ . For the readers' convenience, we sketch the proof.

The fact that  $u' = u - 2v$  follows directly from the number of intersection points of  $(e^j)'$  and  $\beta_1$ , which is the same as the number of stripes. Then we consider  $v'$ . Let  $D = N(x_{\text{middle}}^p)$  be a neighborhood of  $x_{\text{middle}}^p$  such that  $D$  contains all rainbows. Consider the isotopy obtained by rotating  $D$  counterclockwise. If  $v > u'$ , after rotation the resulting diagram has  $v - u'$  rainbows. The formula for  $v'$  follows by induction.

Since 2-bridge knots are alternating they are thin [28] in the sense of Definition 1.3. By comparing the number of generators of  $\widehat{CFK}(K_i)$  for  $i = 1, 2$  from  $(T^2, (e^j)', \beta_1, z, w)$  and the dimension of  $\widehat{HFK}(K_i)$  from the Alexander polynomial (see Proposition 2.9), we know there is no differential on  $\widehat{CFK}(K_i)$ . This fact can also be shown by a direct calculation following the method in [14]. Thus, the constrained knot  $K$  is also thin and there is no differential on  $\widehat{CFK}(K, \mathfrak{s}_j)$ . In particular, the group  $\widehat{CFK}(K, \mathfrak{s}_j)$  is determined by its Euler characteristic.

As discussed at the start of this section, the characteristic  $\chi(\widehat{HFK}(K, \mathfrak{s}_j))$  is an element in  $H_1$  up to equivalence. Similar to the proof of [35, Lemma 3.4], for  $j \in [l, p]$ ,

$$\text{gr}(x_{i+1}^j) - \text{gr}(x_i^j) = [m]^{(-1)^{lv/u}}.$$

For  $j \in [1, l - 1]$ , just replace  $u$  and  $v$  by  $u'$  and  $v'$  in the above formula, respectively. Comparing the formula for the Alexander polynomial in Proposition 2.9, we conclude the formula for  $\chi(\widehat{HFK}(K, \mathfrak{s}_j))$ . □

**Lemma 4.3** Consider integers  $k$  and  $k'$ , and the presentation of  $H_1$  as in Lemma 4.1. For  $j \neq 0, l - 1$ ,

$$\text{gr}(x_{\text{middle}}^{j+1}) - \text{gr}(x_{\text{middle}}^j) = \begin{cases} [a] + [m] & \text{if } jq^{-1} \equiv 1, \dots, k - 2 \pmod{p}, \\ [a] & \text{otherwise.} \end{cases}$$

For  $l \neq 1$  and  $j = 0, l - 1$ ,

$$\text{gr}(x_{\text{middle}}^{j+1}) - \text{gr}(x_{\text{middle}}^j) = \begin{cases} [a] + [m] & \text{if } v \text{ is even,} \\ [a] & \text{if } v \text{ is odd.} \end{cases}$$

For  $l = 1$ ,

$$\text{gr}(x_{\text{middle}}^{j+1}) - \text{gr}(x_{\text{middle}}^j) = \begin{cases} [a] + [m] & \text{if } v \text{ is even,} \\ [a] - [m] & \text{if } v \text{ is odd.} \end{cases}$$

**Proof** For simple knots, the proof is based on Fox calculus; see [38, Proposition 6.1]. For a general constrained knot and  $j \neq 0, l - 1$ , the proof in [38] still works because orientations of strands are alternating. The differences of gradings for  $j = 0$  and  $j = l - 1$  are the same because  $z$  and  $w$  are symmetric by rotation. The proof follows from

$$\sum_{j=0}^{p-1} \text{gr}(x_{\text{middle}}^{j+1}) - \text{gr}(x_{\text{middle}}^j) = 0 \in H_1 \quad \text{and} \quad p[a] + k'[m] = 0 \in H_1. \quad \square$$

**Corollary 4.4** Suppose  $K = C(p, q, l, u, v)$  is a constrained knot in  $Y = L(p, q')$ , where  $qq' \equiv 1 \pmod{p}$ . For any integer  $j \in [1, p]$ , let  $\mathfrak{s}_j = \mathfrak{s}_z(x_{\text{middle}}^j) \in \text{Spin}^c(Y)$  for intersection points  $x_{\text{middle}}^j$  in Figure 6. Then  $\mathfrak{s}_{j+1} - \mathfrak{s}_j$  only depends on  $p$  and  $q$ .

**Proof** By the map

$$H_1(E(K); \mathbb{Z}) / ([m]) \rightarrow H_1(Y; \mathbb{Z}),$$

the grading difference  $\text{gr}(x_{\text{middle}}^{j+1}) - \text{gr}(x_{\text{middle}}^j)$  is mapped to  $\mathfrak{s}_{j+1} - \mathfrak{s}_j$ , which only depends on the image of  $[a]$ . □

**Lemma 4.5** Consider  $\mathfrak{b}(u, v)$  and  $\mathfrak{b}(u', v')$  as in Lemma 4.2. Then

$$\sigma(\mathfrak{b}(u', v')) = \begin{cases} \sigma(\mathfrak{b}(u, v)) & \text{if } v \text{ is even,} \\ \sigma(\mathfrak{b}(u, v)) + 2 & \text{if } v \text{ is odd.} \end{cases}$$

**Proof** Consider standard presentations of 2-bridge knots in Section 2.2. It is easy to see  $\mathfrak{b}(u, v)$  and  $\mathfrak{b}(u', v')$  form two knots in the skein relation. By the skein relation formula of signatures of knots, we can conclude this lemma. Moreover, we provide another proof based on the Alexander grading.

By the algorithm of the Alexander grading, we have

$$\text{gr}(x_{u'}^1) - \text{gr}(x_u^0) = [a] + [m].$$

From the rotation symmetry and the formula of the signature in Proposition 2.9,

$$\begin{aligned} \text{gr}(x_u^0) - \text{gr}(x_{\text{middle}}^0) &= \text{gr}(x_{\text{middle}}^0) - \text{gr}(x_1^0) = \frac{1}{2}\sigma(b(u, v))[m], \\ \text{gr}(x_{u'}^1) - \text{gr}(x_{\text{middle}}^1) &= \text{gr}(x_{\text{middle}}^1) - \text{gr}(x_1^1) = \frac{1}{2}\sigma(b(u', v'))[m]. \end{aligned}$$

Then this lemma follows from these equations and Lemma 4.3. □

**Theorem 4.6** *For a constrained knot  $K = C(p, q, l, u, v)$ , consider the Alexander polynomials  $\Delta_1(t)$  and  $\Delta_2(t)$  in Lemma 4.2. Then  $\widehat{HFK}(K)$  is determined by its Euler characteristic, which is calculated by*

$$(3) \quad \chi(\widehat{HFK}(K)) = \Delta_1([m]) \sum_{j=l}^p \text{gr}(x_{\text{middle}}^j) + \Delta_2([m]) \sum_{j=1}^{l-1} \text{gr}(x_{\text{middle}}^j).$$

**Proof** By the result of Lemma 4.2, we only need to consider the (relative) signs of intersection points corresponding to different  $\text{spin}^c$  structures. By Proposition 3.6, signs of intersection points  $x_i^j$  for fixed  $j$  are alternating. Since  $u$  and  $u' = u - 2v$  are odd, signs of  $x_1^j$  and  $x_{u(j)}^j$  are the same, where  $u(j)$  is either  $u$  or  $u'$  by Lemma 4.2. From the diagram, signs of  $x_{u(j)}^j$  for  $j \in [0, l]$  are the same and signs of  $x_1^k$  for  $k \in [l, p]$  are the same. Thus, we obtain (3). □

All terms in (3) can be calculated by Lemmas 4.3 and 4.5. Thus, we obtain an algorithm for  $\widehat{HFK}(K)$  for a constrained knot  $K$ .

Let signs of  $x_1^j$  be positive. The Alexander grading can be fixed by the global symmetry, ie we consider the absolute Alexander grading. Note that the global symmetry corresponds to switching the roles of  $z$  and  $w$ , which is equivalent to a rotation of the standard diagram of a constrained knot. Then we have

$$\text{gr}(x_{\text{middle}}^j) = -\text{gr}(x_{\text{middle}}^{2l-j}) \text{ for any } j.$$

In this assumption we may use square roots of elements in  $H_1$  to achieve the symmetry, and the Euler characteristic  $\chi(\widehat{HFK}(K))$  is a well-defined element in  $(\frac{1}{2}\mathbb{Z})[H_1]$  for this case. The group  $\widehat{HFK}(K)$  with the Alexander grading fixed as above is called the *canonical representative*.

**Proof of the necessary part of Theorem 1.2** For  $i = 1, 2$ , if  $K_i = C(p_i, q_i, l_i, u_i, v_i)$  are equivalent, then  $p_1 = p_2 = p$  and  $q_1 \equiv q_2^{\pm 1} \pmod{p}$  by the classification of lens spaces [5]. Suppose  $Y$  is the lens space containing  $K_1$  and  $K_2$ . For  $i = 1, 2$ , consider  $(u'_i, v'_i)$  as in Lemma 4.2. By comparing knot Floer homologies, we know  $l_1 = l_2$  and

$$u_1 = |\Delta_{\mathfrak{b}}(u_1, v_1)(-1)| = |\Delta_{\mathfrak{b}}(u_2, v_2)(-1)| = u_2,$$

$$u_1 - 2v_1 = |\Delta_{\mathfrak{b}}(u'_1, v'_1)(-1)| = |\Delta_{\mathfrak{b}}(u'_2, v'_2)(-1)| = u_2 - 2v_2.$$

Thus, we have  $(l_1, u_1, v_1) = (l_2, u_2, v_2) = (l, u, v)$ . Moreover, the sets of  $\text{spin}^c$  structures corresponding to  $\mathfrak{b}(u, v)$  for two constrained knots should be the same. By Corollary 4.4, it suffices to consider simple knots. Let  $\mathfrak{s}_j^i$  be  $\text{spin}^c$  structures related to diagrams of  $K_i$  for  $i = 1, 2$ . Traveling along  $\alpha_1$  of  $K_1$ , middle points are in the order

$$x_{\text{middle}}^0, x_{\text{middle}}^{q_1}, \dots, x_{\text{middle}}^{(p-1)q_1}.$$

Thus, we have

$$\mathfrak{s}_{q_1+j}^1 - \mathfrak{s}_j^1 = \mathfrak{s}_{j+1}^2 - \mathfrak{s}_j^2 \in H^2(Y; \mathbb{Z}).$$

Then the following sets are the same:

$$\{\mathfrak{s}_j^1 - \mathfrak{s}_0^1 + \mathfrak{s}_j^1 - \mathfrak{s}_1^1 \mid j \in [l, p]\}, \quad \{\mathfrak{s}_j^2 - \mathfrak{s}_0^2 + \mathfrak{s}_j^2 - \mathfrak{s}_1^2 \mid j \in [l, p]\}.$$

Equivalently, numbers in  $\{0, q_1, \dots, (p-l)q_1\}$  should be consecutive congruence classes modulo  $p$ . By the following proposition, this can only happen when  $l \in \{2, p\}$ .  $\square$

**Proposition 4.7** *Suppose that integers  $p, q$  and  $k$  satisfy  $1 < q < p-1$ ,  $\text{gcd}(p, q) = 1$  and  $0 \leq k < p-1$ . Then there exists an integer  $x$  such that the sets  $\{x, x+1, \dots, x+k\}$  and  $\{0, q, \dots, kq\}$  can be identified modulo  $p$  if and only if  $k = 0, p-2$ .*

**Proof** If  $k = 0, p-2$ , this proposition is trivial. Suppose  $k \neq 0, p-2$ . Assume elements in sets are in  $\mathbb{Z}/p\mathbb{Z}$  in this proof. Define

$$T = \{0, 1, \dots, p-1\}, \quad S^q = \{0, q, \dots, kq\} \quad \text{and} \quad S_x = \{x, x+1, \dots, x+k\}.$$

Suppose  $S^q = S_x$  for some  $x$  and  $n = \lfloor p/q \rfloor \geq 2$ . If  $k \leq n$ , then the set  $S^q$  cannot be identified with  $S_x$ . Thus  $k \geq n+1$  and  $\{0, q, \dots, nq\} \subset S^q = S_x$ . Suppose  $T - S_x = \{y, y+1, \dots, y+p-k-2\}$ , where  $y = x+k+1$ . Since  $(T - S_x) \cap S^q$  is empty by assumption, the set  $T - S_x$  must be either a subset of  $\{iq+1, iq+2, \dots, (i+1)q-1\}$  for some integer  $i \in [0, n-1]$  or a subset of  $\{nq+1, nq+2, \dots, p-1\}$ . If  $q = 2$ , then  $k = 0$ , which contradicts the assumption. Suppose  $q > 2$ . Since  $k \neq 0, p-2$ , we know  $y, y+1 \in T - S_x$ .

If the first case happens with  $i = 0$ , then we know  $\{q + 1, q + 2, \dots, 2q - 1\} \subset S_x = S_q$  because  $n \geq 2$ . Since  $y + q, y + 1 + q \in \{q + 1, q + 2, \dots, 2q - 1\}$ , there exist different integers  $k_0, k_1 \in [1, k]$  such that

$$y + q \equiv k_0 q \quad \text{and} \quad y + 1 + q \equiv k_1 q \pmod{p}.$$

If  $k_0 > k_1$ , then  $k_0 - 1 \in [1, k - 1]$  and  $y = (k_0 - 1)q \in S^q$ . If  $k_0 < k_1$ , then  $k_1 - 1 \in [1, k - 1]$  and  $y + 1 = (k_1 - 1)q \in S^q$ . Both contradict the assumption.

If the first case happens with  $i > 0$  or the second case happens, then there exist different integers  $k_0, k_1 \in [1, k]$  such that

$$y - q \equiv k_0 q \quad \text{and} \quad y + 1 - q \equiv k_1 q \pmod{p}.$$

If  $k_0 > k_1$ , then  $k_1 + 1 \in [2, k]$  and  $y + 1 = (k_1 + 1)q \in S^q$ . If  $k_0 < k_1$ , then  $k_0 + 1 \in [2, k]$  and  $y = (k_0 + 1)q \in S^q$ . Both contradict the assumption.

In summary, for  $p > 2q$ , there is a contradiction if  $k \neq 0, p - 2$ . If  $p < 2q$  and  $S^q = S_x$ , then we consider

$$S^{p-q} = \{-x, -x - 1, \dots, -x - k\} = S_{-x-k}.$$

Note that  $p > 2(p - q)$ . From a similar discussion, there is also a contradiction.  $\square$

In the rest of this section, we indicate how to draw the curve invariant [16; 17] of the knot complement of a constrained knot. Readers who are not familiar with the curve invariant can safely skip the following discussion since there is no further result in this paper relying on it.

Suppose that  $K = C(p, q, l, u, v)$  is a constrained knot in  $Y = L(p, q')$ , where  $qq' \equiv 1 \pmod{p}$ . Let  $M = E(K)$ . From the standard diagram of the constrained knot, we know  $[K] = k'[b] \in H_1(Y; \mathbb{Z})$ , where  $b$  is the core curve of  $\beta_0$ -handle and  $k'$  is the integer in Lemma 4.1. Since  $K$  is thin, the curve invariant  $\widehat{HF}(M)$  can be drawn as follows.

The curve invariant can be decomposed with respect to  $\text{Spin}^c(M)$ , which is affine over  $H^2(M; \mathbb{Z})$ . By Poincaré duality and the long exact sequence for  $(M, \partial M)$ , we know  $|H^2(M; \mathbb{Z})| = |H_1(M, \partial M; \mathbb{Z})| = |H_1(M; \mathbb{Z}) / \text{Im}(H_1(\partial M; \mathbb{Z}))| = |\text{Tors } H_1(M; \mathbb{Z})|$ .

For simplicity, suppose  $H_1(M; \mathbb{Z}) \cong \mathbb{Z}$ . Then  $|\text{Spin}^c(M)| = 1$  and  $\gcd(p, k') = 1$ .

The curve invariant can be lifted to the universal cover  $\mathbb{R}^2$  of  $\partial M$ . Suppose the basis is  $([l^*], -[m^*])$ , where the homological meridian  $m^*$  (see Section 2) is chosen so that

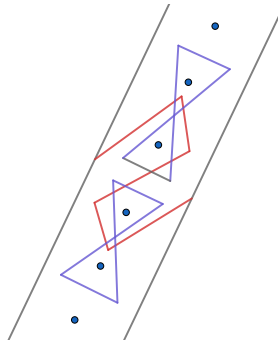


Figure 7: Part of the curve invariant of  $C(p, q, l, 11, 3)$ .

$[m] = p[m^*] - k_0[l^*]$  for some  $k_0 \in [0, p)$ . Consider parallel lines with slope  $p/k_0$  away from the basepoint on  $M$ . They cut  $\mathbb{R}^2$  into bands. Suppose that lifts of the basepoint are integer points and lie on a line with slope  $p/k_0$  in each band. Since  $\widehat{HF}(Y, \mathfrak{s}) \cong \mathbb{Z}$  for any  $\mathfrak{s} \in \text{Spin}^c(Y)$ , the curve invariant intersects each line once.

Based on the proof of Lemma 4.2, the chain complex  $\widehat{CFK}(K, \mathfrak{s})$  for any  $\mathfrak{s} \in \text{Spin}^c(Y)$  is similar to the chain complex related to a 2-bridge knot. Moreover, from the relation of the standard diagram of  $K$  and the Heegaard diagram of a 2-bridge knot, the minus version of the knot Floer chain complex  $CFK^-(K, \mathfrak{s})$  is also related to  $CFK^-$  of a 2-bridge knot. From the results in [34, Section 3] about thin complexes and the results in [17, Section 4] about how to draw the curve invariant from  $CFK^-$ , the part of the curve invariant of  $K$  in a band is the union of some purple figure-8 curves and a distinguished red arc as shown in Figure 7, which totally depends on the Alexander polynomial and the signature of the related 2-bridge knot.

**Lemma 4.8** *Suppose  $H_1(M; \mathbb{Z}) \cong \mathbb{Z}$  and consider  $k_0$  and  $k'$  as above. Suppose  $a$  and  $b$  are core curves of  $\alpha_0$  and  $\beta_0$  handles corresponding to the standard diagram of  $Y = L(p, q')$ . Then  $k_0q(k')^2 \equiv -1 \pmod{p}$ . Hence  $k_0$  is determined by  $k'$ .*

**Proof** The homology  $H_1(M; \mathbb{Z})$  is generated by  $[m^*]$ . Let  $\tilde{m}^*$  denote the image of  $[m^*]$  in  $H_1(Y; \mathbb{Z})$ . By Lemma 4.1,  $[a] = -k'\tilde{m}^*$ . The relation  $[b] = q[a]$  implies  $[K] = -q(k')^2\tilde{m}^*$ . Then a lift of  $[K]$  in  $H_1(T^2; \mathbb{Z})$  equals  $-q(k')^2[m^*] + k_1[l^*]$  for some  $k_1$ . Since  $l$  is isotopic to  $K$ , we have  $[K] = [l] \in H_1(T^2; \mathbb{Z})$ . Then since  $[m] = p[m^*] - k_0[l^*]$  and  $[m] \cdot [l] = [m^*] \cdot [l^*] = -1$ , we have

$$[m] \cdot [l] = (p[m^*] - k_0[l^*]) \cdot (-q(k')^2[m^*] + k_1[l^*]) = (pk_1 - k_0q(k')^2)[m^*] \cdot [l^*].$$

Hence we conclude the congruence result for  $k_0$ . □

For  $i \in \mathbb{Z}/p\mathbb{Z}$ , suppose  $B_i$  are bands in  $\mathbb{R}^2$  mentioned above, ordered from left to right. Suppose  $\mathfrak{s}_i \in \text{Spin}^c(Y)$  are  $\text{spin}^c$  structures corresponding to  $B_i$ . Since the slope of parallel lines is  $p/k_0$ , the difference  $\mathfrak{s}_{i+1} - \mathfrak{s}_i$  is  $k'_0 \tilde{m}^*$  for the integer  $k'_0$  satisfying  $k_0 k'_0 \equiv -1 \pmod{p}$ . By the above lemma, we have  $k'_0 \equiv q(k')^2 \pmod{p}$ . By definition of  $k'$  in Lemma 4.1, we have

$$-qk' \equiv \begin{cases} -q-l+1 & \text{if } v \text{ is even modulo } p, \\ q-l+1 & \text{if } v \text{ is odd modulo } p. \end{cases}$$

Since  $[a] = -k' \tilde{m}^*$ , bands  $B_{-iqk'}$  for  $i \in [1, l-1]$  correspond to  $\mathfrak{b}(u', v')$  and  $B_{-iqk'}$  for  $i \in [l, p]$  correspond to  $\mathfrak{b}(u, v)$  in  $\widehat{HF}(M)$ . Finally, the Alexander grading indicates the relative height of the curves in bands and there is a unique way to connect curves in different bands.

## 5 Knots in the same homology class

For fixed  $(p, q, u, v)$  and each  $h \in H_1(L(p, q'); \mathbb{Z}) \cong \mathbb{Z}/p\mathbb{Z}$ , where  $qq' \equiv 1 \pmod{p}$ , there is a parameter  $l \in [1, p]$  such that  $C(p, q, l, u, v)$  is a representative of  $h$ , ie  $[C(p, q, l, u, v)] = h$ . In other words, for any knot  $K$  in  $L(p, q')$  there are infinitely many constrained knots  $K'$  satisfying  $[K'] = [K] \in H_1(L(p, q'); \mathbb{Z})$ .

In this section we focus on knots representing the same homology class in a lens space. The main results are Theorems 1.5 and 1.6. Since we will not use the parameters of a constrained knot, we denote a lens space by  $L(p, q)$  rather than  $L(p, q')$  as in other sections. Many results in this section are related to the Turaev torsion  $\tau(M)$  of a 3-manifold  $M$  with torus boundary [42], which can be calculated by any presentation of  $\pi_1(M)$ . For simplicity, write  $\tau(K) = \tau(E(K))$ . The following proposition enables us to compare elements in homology groups of different knot complements:

**Proposition 5.1** [5] *Let  $K$  be a knot in a 3-manifold  $Y$ . The isomorphism class of the homology  $H_1(E(K); \mathbb{Z})$  only depends on the homology class  $[K] \in H_1(Y; \mathbb{Z})$ .*

Suppose  $Y = L(p, q)$  and  $K$  is a knot in  $Y$ . By Proposition 3.10, Lemma 4.1 and Proposition 5.1, there exists a positive integer  $d$  satisfying  $H_1(E(K); \mathbb{Z}) \cong \mathbb{Z} \oplus \mathbb{Z}/d\mathbb{Z}$ . Let  $m$  be the meridian of  $K$  in the sense of Section 2. Suppose  $t$  and  $r$  are generators of  $\mathbb{Z} \oplus \mathbb{Z}/d\mathbb{Z}$  such that

$$H_1(E(K); \mathbb{Z}) \cong \langle t, r \rangle / (dr).$$

Then there exist  $p_0, a \in \mathbb{Z}$  such that the above isomorphism sends  $[m]$  to  $p_0 t + ar$ .



**Lemma 5.2** *The integer  $p$  is divisible by  $d$ , and  $p_0 = \pm p/d$ . Moreover, the greatest common divisor of  $p_0, d$  and  $a$  is 1.*

**Proof** By the isomorphism  $H_1(E(K); \mathbb{Z})/([m]) \cong H_1(Y; \mathbb{Z})$ , the order  $p$  of  $H_1(Y; \mathbb{Z})$  is the same as

$$\left| \det \left( \begin{bmatrix} p_0 & a \\ 0 & d \end{bmatrix} \right) \right| = |dp_0|.$$

If the greatest common divisor of  $p_0, d$  and  $a$  is not 1, then the Smith normal form of this matrix cannot be

$$\begin{bmatrix} 1 & 0 \\ 0 & p \end{bmatrix}$$

because elementary transformations in the algorithm of the Smith normal form do not decrease the common divisor of all entries. □

**Lemma 5.3** *Let  $K_1$  and  $K_2$  be knots in  $Y = L(p, q)$  representing the same homology class  $h \in H_1(Y; \mathbb{Z})$ . Let  $m_1$  and  $m_2$  be meridians of  $K_1$  and  $K_2$  in the sense of Section 2. For  $i = 1, 2$ , there are isomorphisms  $j_i: H_1(E(K_i); \mathbb{Z}) \rightarrow \mathbb{Z} \oplus \mathbb{Z}/d\mathbb{Z}$  such that  $j_1([m_1]) = j_2([m_2])$ .*

**Proof** For  $i = 1, 2$ , by the discussion after Proposition 5.1, there exists an isomorphism  $j'_i: H_1(E(K_i); \mathbb{Z}) \rightarrow \mathbb{Z} \oplus \mathbb{Z}/d\mathbb{Z}$  such that

$$j'_1([m_1]) = p_0t + ar \quad \text{and} \quad j'_2([m_2]) = p'_0t + br.$$

Then it suffices to find an automorphism  $f$  of  $\mathbb{Z} \oplus \mathbb{Z}/d\mathbb{Z}$  such that

$$f(p_0t + ar) = p'_0t + br.$$

By Lemma 5.2, the integers  $p_0$  and  $p'_0$  are in  $\{p/d, -p/d\}$ . Let  $f_0$  be the automorphism of  $\mathbb{Z} \oplus \mathbb{Z}/d\mathbb{Z}$  sending  $(t, r)$  to  $(-t, r)$ . If  $p_0 = -p/d$ , the map  $j'_1$  can be replaced by  $f_0 \circ j'_1$ . The same assertion holds for  $p'_0$ . Without loss of generality, suppose  $p_0 = p'_0 = p/d$ . Let  $g = \gcd(p_0, d)$ ,  $p_0 = gp_1$  and  $d = gd_0$ . Then  $\gcd(p_1, d_0) = 1$ , and there exist integers  $x_0$  and  $k_0$  satisfying  $x_0p_1 + k_0d_0 = 1$ . By Lemma 5.2,  $\gcd(g, a) = \gcd(g, b) = 1$ . There exist integers  $a_0$  and  $k_1$  satisfying  $a_0a + k_1g = b$  and  $\gcd(a_0, g) = 1$ . Suppose  $x = (k_1 - k_2a)x_0$  and  $y = k_2g + a_0$  for some integer  $k_2$ . Then

$$\begin{aligned} xp_0 + ya &\equiv (k_1 - k_2a)x_0gp_1 + (k_2g + a_0)a \\ &\equiv (k_1 - k_2a)(1 - k_0d_0)g + (k_2g + a_0)a \equiv k_1g + a_0a \equiv b \pmod{gd_0}. \end{aligned}$$

The map

$$f: \mathbb{Z} \oplus \mathbb{Z}/d\mathbb{Z} \rightarrow \mathbb{Z} \oplus \mathbb{Z}/d\mathbb{Z} \quad \text{given by } t \mapsto t + xr \text{ and } r \mapsto yr$$

is an isomorphism if and only if  $\gcd(y, d) = 1$ . Since  $f(t + ar) = t + (xp_0 + ya)r$ , this lemma follows from the next proposition.  $\square$

**Proposition 5.4** *Suppose integers  $a_0$  and  $g$  satisfying  $\gcd(a_0, g) = 1$ . For any integer  $d$  there exists an integer  $k_2$  satisfying  $\gcd(y, d) = 1$ , where  $y = k_2g + a_0$ .*

**Proof** If  $q$  is a prime number satisfying  $p \mid \gcd(g, d)$ , then  $a_0$  is not divisible by  $q$  and neither is  $y$  because  $\gcd(a_0, g) = 1$ . Then  $\gcd(y, d) = \gcd(y, d/q)$ . Without loss of generality, suppose  $\gcd(g, d) = 1$ . By the Chinese remainder theorem, the following congruence equations have a solution  $y$ :

$$y \equiv a_0 \pmod{g}, \quad y \equiv 1 \pmod{d}.$$

Then  $\gcd(y, d) = 1$ . We know that  $k_2 = (y - a_0)/g$  satisfies the proposition.  $\square$

From now on, let us fix isomorphisms  $j_1$  and  $j_2$  as in Lemma 5.3. Then the homology classes of meridians and their images under  $j_i$  for  $i = 1, 2$  can be identified, ie  $[m_1]$  and  $[m_2]$  are regarded as the same element  $[m]$  in  $\mathbb{Z} \oplus \mathbb{Z}/d\mathbb{Z}$ . The following is the key lemma in this section, and is based on results in [42].

**Lemma 5.5** *Let  $K_1$  and  $K_2$  be two knots in  $Y = L(p, q)$  representing the same homology class. Let  $j_i$  be the isomorphisms  $H_1(E(K_i); \mathbb{Z}) \cong \mathbb{Z} \oplus \mathbb{Z}/d\mathbb{Z} = H_1$  as in Lemma 5.3. Then  $\tau(K_1) - \tau(K_2)$  can be regarded as an element in  $\mathbb{Z}[H_1]/\pm H_1$ . Moreover, we have*

$$\tau(K_1) - \tau(K_2) = (1 - [m])g \quad \text{for some } g \in \mathbb{Z}[H_1]/\pm H_1.$$

**Proof** Note that  $\tau(K_i)$  is not a priori an element in  $\mathbb{Z}[H_1(E(K_i); \mathbb{Z})]/\pm H_1(E(K_i); \mathbb{Z})$  (see [42, Corollary II.4.3]). However, the difference  $\tau(K_1) - \tau(K_2)$  is a well-defined element in  $\mathbb{Z}[H_1]/\pm H_1$  under the isomorphisms of group rings induced by  $j_1$  and  $j_2$ . To resolve the ambiguity of  $\pm H_1$ , we can choose an Euler structure and a homology orientation on  $E(K_i)$  (see [42, Section I.1]). For any compact 3-manifold with torus boundary, Euler structures are in one-to-one correspondence with  $\text{spin}^c$  structures related to the Alexander grading. For any closed 3-manifold, Euler structures are in one-to-one correspondence with  $\text{spin}^c$  structures on the manifold. We omit the choice

of the homology orientation that determines the sign of  $\tau(K_i)$ , and only consider the choice of the Euler structure for simplicity. For an Euler structure  $e$  on  $M$ , the Turaev torsion  $\tau(M)$  has a representative  $\tau(M, e)$ .

For  $i = 1, 2$ , let  $e_i$  be Euler structures on  $E(K_i)$  inducing the same Euler structure  $e^Y$  on  $Y$ . Adapting notation from [42, Section II.4.5], suppose the integer  $K(e_i)$  satisfies

$$c(e_i) = \frac{e_i}{e_i^{-1}} \in t^{K(e_i)} \text{Tors } H_1.$$

We can also consider  $c(e_i)$  as the Chern class of the  $\text{spin}^c$  structure on  $E(K_i)$  corresponding to  $e_i$ . Note that  $t$  is the generator of the free part of  $H_1$ .

From the correspondence between Euler structures and  $\text{spin}^c$  structures, it is possible to choose  $e_i$  so that  $K(e_1) = K(e_2)$ . In the proof of [42, Lemma II.4.5.1(i)], we have

$$\tau(E(K_i), e_i) \in \frac{-\Sigma_{H_1}}{t-1} + \mathbb{Z}[H_1],$$

where  $\Sigma_H = \Sigma_{h \in \text{Tors } H_1} h$ , so  $\tau(K_1, e_1) - \tau(K_2, e_2) \in \mathbb{Z}[H_1]$ . Also, in [42, Section II.4], for a 3-manifold  $M$  with  $b_1(M) = 1$ , the polynomial part  $[\tau](M, e) \in (\frac{1}{2}\mathbb{Z})[H_1]$  of  $\tau(M, e)$  is defined by

$$(4) \quad [\tau](M, e) = \left( \tau(M, e) + \frac{\Sigma_{H_1}(M)}{t-1} \right) \times \begin{cases} t^{\frac{1}{2}(K(e)+1)} & \text{if } K(e) \text{ is odd,} \\ t^{\frac{1}{2}K(e)} \left( \frac{1}{2}(t+1) \right) & \text{if } K(e) \text{ is even.} \end{cases}$$

By [42, Remark II.4.5.2], for any Euler structure  $e$  on  $M$ , the polynomial part  $[\tau](M, e)$  is in the kernel of the map  $\text{aug}: \mathbb{Z}[H_1] \rightarrow \mathbb{Z}$  that sends elements in  $H_1$  to  $1 \in \mathbb{Z}$ . Thus,

$$\text{aug}(\tau(K_1, e_1) - \tau(K_2, e_2)) = \text{aug}([\tau](K_1, e_1) - [\tau](K_2, e_2)) = 0.$$

By the  $m = 1$  case in [42, Theorem X.4.1], since the map  $\kappa: \mathbb{Q}[H_1] \rightarrow \mathbb{Q}[H_1]$  that sends  $x$  to  $x - \text{aug}(x)\Sigma_{H_1}/|H_1|$  is trivial, we have

$$\text{pr}(\tau(K_1, e_1) - \tau(K_2, e_2)) = -([K_1] - 1)\tau(Y, e^Y) + ([K_2] - 1)\tau(Y, e^Y) = 0,$$

where  $\text{pr}$  is the map in the following proposition. Also from the following proposition, there is an element  $g \in \mathbb{Z}[H_1]$  such that

$$\tau(K_1, e_1) - \tau(K_2, e_2) = (1 - [m])g.$$

Since  $\tau(K_1, e_1) - \tau(K_2, e_2)$  reduces to  $\tau(K_1) - \tau(K_2)$  in  $\mathbb{Z}[H_1]/\pm H_1$ , we obtain the equation for elements in  $\mathbb{Z}[H_1]/\pm H_1$ . □

**Proposition 5.6** Let  $\text{pr}: \mathbb{Z}[\mathbb{Z} \oplus \mathbb{Z}/d\mathbb{Z}] \rightarrow \mathbb{Z}[\mathbb{Z}/p\mathbb{Z}]$  be the map between group rings induced by the composition of maps

$$\mathbb{Z} \oplus \mathbb{Z}/d\mathbb{Z} \xrightarrow{\cong} H_1(E(K_i); \mathbb{Z}) \rightarrow H_1(E(K_i); \mathbb{Z})/([m_i]) \xrightarrow{\cong} H_1(Y; \mathbb{Z}) \xrightarrow{\cong} \mathbb{Z}/p\mathbb{Z}.$$

Then the kernel of  $\text{pr}$  is the ideal generated by  $1 - [m]$ .

**Proof** Suppose  $\mathbb{Z}/p\mathbb{Z} = \{s_1, \dots, s_p\}$  and suppose  $H = \sum_{i=1}^k a_i h_i$  is an element in the kernel of  $\text{pr}$ , where  $a_i \in \mathbb{Z}$  and  $h_i \in \mathbb{Z} \oplus \mathbb{Z}/d\mathbb{Z}$ . Let  $\Theta_H(s_j)$  be the set consisting of all elements  $h_i$  satisfying  $\text{pr}(h_i) = s_j$  in the summation defining  $H$ . Then  $\sum_{h_i \in \Theta_H(s_j)} a_i h_i$  is also in the kernel of  $\text{pr}$  for any  $j$ . Without loss of generality, suppose  $\text{pr}(h_i) = s_1$  for any  $h_i$  in the summation of  $H$ . By definition of the map  $\text{pr}$ , for any  $i$ , we have  $h_i = [m]^{\alpha(i)} h_1$  for some integer  $\alpha(i)$ . Then

$$H = \sum_{j=0}^{k'} b_j [m]^j h_1$$

for some integer  $k'$ . Since  $H$  is in the kernel of  $\text{pr}$ , we have

$$\sum_{j=0}^{k'} b_j = 0.$$

Thus, the polynomial

$$\sum_{j=0}^{k'} b_j x^j$$

has a root  $x = 1$ . In other words,  $\sum_{j=0}^{k'} b_j x^j = (1 - x)g(x)$  for some polynomial  $g(x)$ . Then we have  $H = (1 - [m])g([m])h_1$  and conclude the proposition.

There is another quick proof from the referee. The functor that takes a group to its group ring is left-adjoint to the functor that takes a commutative ring to its group of units. The quotient  $\mathbb{Z}/p\mathbb{Z}$  is the colimit of the diagram  $\mathbb{Z} \rightrightarrows \mathbb{Z} \oplus \mathbb{Z}/d\mathbb{Z}$ , where one map is  $1 \mapsto [m]$  and the other is the zero map. Then the proposition follows from the fact that left-adjoints preserve colimits. □

**Lemma 5.7** [39, Proposition 2.1] Suppose  $K$  is a knot in  $Y = L(p, q)$  and let  $H_1 = H_1(E(K); \mathbb{Z})$ . Then

$$\chi(\widehat{HFK}(Y, K)) = (1 - [m])\tau(K) \in \mathbb{Z}[H_1]/\pm H_1.$$

**Theorem 5.8** Let  $K_1$  and  $K_2$  be two knots representing the same homology class in  $Y = L(p, q)$ . Suppose  $H_1(E(K_i); \mathbb{Z}) \cong \mathbb{Z} \oplus \mathbb{Z}/d\mathbb{Z} = H_1$  as in Lemma 5.3. After shifting Alexander gradings on  $\widehat{HF\bar{K}}(Y, K_i)$  for  $i = 1, 2$ , the difference of their Euler characteristics satisfies the following condition: for any  $\mathfrak{s} \in \text{Spin}^c(Y)$ , there exists a Laurent polynomial  $f(x) \in \mathbb{Z}[x, x^{-1}]$  and an element  $\tilde{s} \in H_1$  such that

$$\chi(\widehat{HF\bar{K}}(Y, K_1, \mathfrak{s})) - \chi(\widehat{HF\bar{K}}(Y, K_2, \mathfrak{s})) = ([m] - 1)^2 f([m])\tilde{s}.$$

**Proof** Note that  $\chi(\widehat{HF\bar{K}}(Y, K_i))$  is an element in  $\mathbb{Z}[H_1]$  up to equivalence. Fixing the Alexander grading on  $\widehat{HF\bar{K}}(Y, K_i)$  is equivalent to choosing a representative of  $\chi(\widehat{HF\bar{K}}(Y, K_i))$  in  $\mathbb{Z}[H_1]$ . By Lemma 5.5 and Lemma 5.7, after shifting Alexander gradings, there exists some  $g \in \mathbb{Z}[H_1]/\pm H_1$  such that

$$\chi(\widehat{HF\bar{K}}(Y, K_1)) - \chi(\widehat{HF\bar{K}}(Y, K_2)) = (1 - [m])(\tau(K_1) - \tau(K_2)) = ([m] - 1)^2 g.$$

Choose a lift  $\tilde{g}$  of  $g$  in  $\mathbb{Z}[H_1]$ . It can be written as the sum  $\tilde{g} = \sum_{j=1}^p g_j$ , where  $g_j$  contains terms that are in the preimage of  $s_j \in H_1(Y; \mathbb{Z})$  under the map

$$H_1 \rightarrow H_1(Y; \mathbb{Z}) = \{s_1, \dots, s_p\}.$$

For any  $j$ , there exists a Laurent polynomial  $f_j(x)$  and an element  $\tilde{s}_j \in H_1$  such that  $g_j = f_j([m])\tilde{s}_j$ . Thus, the above equation can be decomposed into  $\text{spin}^c$  structures, which induces the theorem. □

**Remark 5.9** For constrained knots  $K_1$  and  $K_2$ , the group  $\widehat{HF\bar{K}}(Y, K_i)$  can be chosen as the canonical representative in Section 4, meaning we consider the absolute Alexander grading mentioned in the introduction.

**Proof of Theorem 1.5** We choose the isomorphisms  $H_1(E(K_i); \mathbb{Z}) \cong H_1$  considered in Lemma 5.3. By Lemma 4.2, for a constrained knot  $K_i \subset Y$  and a  $\text{spin}^c$  structure  $\mathfrak{s}$  on  $Y$ , there is a symmetrized Alexander polynomial  $\Delta_i(t)$  of a 2-bridge knot, so that

$$\chi(\widehat{HF\bar{K}}(Y, K_i, \mathfrak{s})) \sim \Delta_i([m]).$$

Since the Alexander grading reduces to the grading induced by  $\text{spin}^c$  structures under the map  $H_1(E(K_i); \mathbb{Z}) \rightarrow H_1(Y; \mathbb{Z})$ , we know Alexander gradings of nontrivial summands of  $\widehat{HF\bar{K}}(Y, K_i, \mathfrak{s})$  correspond to the  $\text{spin}^c$  structure  $\mathfrak{s}$ . By definition of the equivalence on  $\mathbb{Z}[H_1]$ , there exists an element  $\tilde{s} \in H_1$  in the preimage of  $\mathfrak{s}$  such that

$$(5) \quad \chi(\widehat{HF\bar{K}}(Y, K_i, \mathfrak{s})) = \pm \Delta_i([m])[m]^{\gamma_i} \tilde{s},$$

where  $\gamma_i$  is an integer. Write  $f_i(x) = \pm \Delta_i(x)x^{\gamma_i}$  for simplicity. Since  $\Delta_i(t)$  is symmetrized, the middle grading is the grading of  $[m]^{\gamma_i} \tilde{s}$ . Note that the multiplication in  $\mathbb{Z}[H_1]$  corresponds to the addition in  $H_1$ . Then we have

$$A(K_1, \mathfrak{s}) - A(K_2, \mathfrak{s}) = (\gamma_1[m] + \tilde{s}) - (\gamma_2[m] + \tilde{s}) = (\gamma_1 - \gamma_2)[m] \in H_1.$$

By Theorem 5.8, there is a Laurent polynomial  $f(x) \in \mathbb{Z}[x, x^{-1}]$  such that

$$f_1(x) - f_2(x) = (x - 1)^2 f(x).$$

Hence for a large integer  $N$ , there is a polynomial  $f_0(x)$  such that

$$x^N (f_1(x) - f_2(x)) = (x - 1)^2 f_0(x).$$

Substituting  $x = 1$  gives  $f_1(1) = f_2(1)$ , that is signs in (5) are the same for  $i = 1, 2$ . Consider derivatives at  $x = 1$ :

$$\begin{aligned} 0 &= \frac{d(x^N (f_1(x) - f_2(x)))}{dx} = N(f_1(1) - f_2(1)) + \frac{df_1}{dx}(1) - \frac{df_2}{dx}(1) \\ &= \pm \left( \frac{d\Delta_1(x)}{dx}(1) - \frac{d\Delta_2(x)}{dx}(1) + \gamma_1 \Delta_1(1) - \gamma_2 \Delta_2(1) \right) = \gamma_1 - \gamma_2, \end{aligned}$$

where the last equation follows from  $\Delta_i(t) = \Delta_i(t^{-1})$  and  $\Delta_i(1) = 1$ . Thus, we have  $A(K_1, \mathfrak{s}) = A(K_2, \mathfrak{s})$ .  $\square$

**Proof of Theorem 1.6** This follows from the proof of Theorem 1.5 with  $\Delta_i(t) = 1$ .  $\square$

## 6 Classification

The main result in this section is the proof of the sufficient part of Theorem 1.2. The following lemma enables us to prove it by considering knot groups, ie fundamental groups of knot complements.

**Lemma 6.1** [43] *Let  $M_1$  and  $M_2$  be Haken manifolds with torus boundaries. If there is an isomorphism  $\psi: \pi_1(M_1) \rightarrow \pi_1(M_2)$  that induces an isomorphism*

$$\psi|_{\pi_1(\partial M_1)}: \pi_1(\partial M_1) \rightarrow \pi_1(\partial M_2),$$

*then there exists a diffeomorphism  $\psi_0: (M_1, \partial M_1) \rightarrow (M_2, \partial M_2)$  inducing  $\psi$ .*

*In addition, if  $M_1$  and  $M_2$  are two knot complements and  $\psi$  sends the meridian of one knot to the meridian of the other knot, then two knots are equivalent.*

A constrained knot is defined by a doubly pointed Heegaard diagram, from which it is easy to obtain a Heegaard diagram of the knot complement similar to the case in Figure 3, right. The Heegaard diagram is related to the handlebody decomposition of the corresponding 3–manifold, and then also related to the cell complex of the corresponding 3–manifold. Thus, it is possible to obtain a presentation of the fundamental group from the Heegaard diagram. We show this presentation explicitly:

Suppose  $K = C(p, q, l, u, v)$  is a constrained knot with  $u > 2v \geq 0$ . Suppose  $(T^2, \alpha_1, \beta_1, z, w)$  is the standard diagram of  $K$ . Let  $\Sigma$  be the surface of genus two obtained by attaching a 1–handle at basepoints  $z$  and  $w$ . Suppose  $\alpha_2$  is the curve on  $\Sigma$  that is a union of an arc connecting  $z$  to  $w$  in  $T^2 - \alpha_1$  and an arc on the attached handle; see Figure 6. Suppose  $\beta = \beta_1$ . Then  $(\Sigma, \{\alpha_1, \alpha_2\}, \beta)$  is a Heegaard diagram of  $E(K)$ .

Let the innermost rainbow  $R_0$  around  $w$  be oriented from the right boundary point  $x_r$  to the left boundary point  $x_l$ . This induces an orientation of  $\beta$ . Let  $\alpha_1$  and  $\alpha_2$  be oriented from the left vertical edge to the right vertical edge in the new diagram  $C$  of the constrained knot.

Suppose  $s$  and  $t$  correspond to cores of  $\alpha_1$ –handle and  $\alpha_2$ –handle, respectively. In the above orientation, we can obtain a presentation  $\pi_1(E(K)) \cong \langle s, t \mid \omega = 1 \rangle$ , where the word  $\omega$  is given in the following way:

- (i) Starting at  $x_l$  and traveling along  $\beta$ , suppose intersection points of  $\beta \cap (\alpha_1 \cup \alpha_2)$  are ordered as  $x_1, x_2, \dots, x_m$ .
- (ii) If  $x_i$  is an intersection point of  $\alpha_1$  and  $\beta$  it corresponds to a word  $s^{\pm 1}$ , where the sign depends on the contribution of  $x_i$  in the algebraic intersection number  $\alpha_1 \cap \beta$ .
- (iii) If  $x_i$  is an intersection point of  $\alpha_2$  and  $\beta$  it corresponds to a word  $t^{\pm 1}$ , where the sign depends on the contribution of  $x_i$  in the algebraic intersection number  $\alpha_2 \cap \beta$ .
- (iv) The word  $\omega$  is obtained from  $x_1 x_2 \cdots x_m$  by replacing  $x_i$  by corresponding words in  $\{s, s^{-1}, t, t^{-1}\}$ .

The word  $\omega(p, q, l, u, v) = \omega(C(p, q, l, u, v))$  in the above setting is called the *standard relation* of a constrained knot  $C(p, q, l, u, v)$ . We begin by understanding the standard relation of a 2–bridge knot. For fixed integers  $(u, v)$ , let  $\epsilon_i = (-1)^{\lfloor iv/u \rfloor}$ .

**Lemma 6.2** For the constrained knot  $C(1, 0, 1, u, v) \cong \mathfrak{b}(u, v)$ , the standard relation  $\omega$  is  $s^{\epsilon_1} t^{\epsilon_2} s^{\epsilon_3} \dots s^{\epsilon_{2u-1}} t^{\epsilon_{2u}}$ .

**Proof** This is from the relation between the Schubert normal form and the Heegaard diagram of the 2-bridge knot. Note that the formula of the Alexander polynomial in Proposition 2.9 follows from this presentation and Fox calculus [42, Chapter II].  $\square$

For fixed integers  $(p, q, l)$  with  $q \in [1, p - 1]$ ,  $l \in [1, p]$  and  $\gcd(p, q) = 1$ , suppose the integer  $k \in (0, p]$  satisfies  $k - 1 \equiv (l - 1)q \pmod{p}$  and the integer  $q_i \in [0, p)$  satisfies  $q_i \equiv iq \pmod{p}$ . Define

$$\theta_i = \theta_i(p, q, l) = \begin{cases} 1 & \text{if } q_i \in [0, k), \\ 0 & \text{if } q_i \in [k, p), \end{cases}$$

$$s_*(p, q, l) = s t^{\theta_l} s t^{\theta_{l+1}} s \dots s t^{\theta_{p-1}} s,$$

$$t_*(p, q, l) = t^{\theta_0} s t^{\theta_1} s \dots s t^{\theta_{l-1}}.$$

In particular, we have  $\theta_0 = 1$  and  $\theta_{l-1} = 1$ . Note that the integer  $q$  in  $s_*(p, q, l)$  or  $t_*(p, q, l)$  does not correspond to the parameter  $q$  in  $C(p, q, l, u, v)$ . Indeed, the constrained knot  $C(p, q, l, u, v)$  corresponds to  $s_*(p, q', l)$  and  $t_*(p, q', l)$ , where  $qq' \equiv 1 \pmod{p}$ . We can see this fact from the following proposition.

**Proposition 6.3** For  $K = C(p, q, l, u, v)$ , suppose that the integer  $q' \in [0, p)$  satisfies  $qq' \equiv 1 \pmod{p}$ . Suppose  $s_* = s_*(p, q', l)$  and  $t_* = t_*(p, q', l)$ . Define

$$t_{\#}^{\epsilon_i} = \begin{cases} t^{\epsilon_i} & \text{if } \epsilon_{i-1} = -\epsilon_{i+1}, \\ t_*^{\epsilon_i} & \text{if } \epsilon_{i-1} = \epsilon_{i+1}. \end{cases}$$

Then the standard relation of  $K$  is  $\omega(p, q, l, u, v) = s_*^{\epsilon_1} t_{\#}^{\epsilon_2} s_*^{\epsilon_3} \dots s_*^{\epsilon_{2u-1}} t_{\#}^{\epsilon_{2u}}$ .

**Proof** The standard diagram of  $C(p, q, l, u, v)$  generalizes the standard diagram of  $C(1, 0, 1, u, v)$ . Then  $\omega(p, q, l, u, v)$  can be obtained from  $\omega(1, 0, 1, u, v)$  by replacing  $s$  and  $t$  by some words. We figure out the replacement as follows.

Suppose that the integer  $k \in (0, p]$  satisfies  $(k - 1)q \equiv l - 1 \pmod{p}$ , which coincides with the definition of  $k$  for  $(p, q', l)$  before this proposition. Note that we define  $q'_i$  by  $qq'_i \equiv i \pmod{p}$  since we consider  $(p, q', l)$  rather than  $(p, q, l)$ .

Consider the new diagram  $C$  of  $K$  mentioned in Section 3; see Figure 4. There are regions  $D_j$  for  $j \in \mathbb{Z}/p\mathbb{Z}$ , where the right edge of  $D_j$  is glued to the left edge of  $D_{j+q}$ . Consider the part of  $\alpha_2$  on  $T^2$  that connects  $z$  to  $w$ . It goes across regions in the order

$$D_1, D_{q+1}, D_{2q+1}, \dots, D_l.$$



By definition of  $k$ , there are  $k$  regions in the above sequence. By definition of  $q'_i$ , any region  $D_j$  in the above sequence lies at the  $(q'_{j-1} + 1)^{\text{th}}$  position, so  $q'_j < k$ . For example,  $q'_0 = 0$  implies that  $D_1$  lies at the first position and  $q'_{l-1} = k - 1$  implies that  $D_l$  lies at the  $k^{\text{th}}$  position. Then  $\theta_j = 1$  if and only if  $\alpha_2 \cap D_{j+1}$  is nonempty.

Then the word  $s_*(p, q', l)$  corresponds to intersection points of  $\beta \cap (\alpha_1 \cup \alpha_2)$  on an arc component of  $\beta \cap (\bigcup_{j=l+1}^{p-1} D_j)$ . The word  $t_*(p, q', l)$  corresponds to intersection points of  $\beta \cap (\alpha_1 \cup \alpha_2)$  on an arc component of  $\beta \cap (\bigcup_{j=1}^l D_j)$  that is also a subarc of a stripe.

Thus, we can replace  $s$  by  $s_* = s_*(p, q', l)$ . When  $\epsilon_{i-1} = -\epsilon_{i+1}$ , the corresponding intersection point related to  $t^{\epsilon_i}$  is on the rainbow, so we just replace  $t^{\epsilon_i}$  by  $t^{\epsilon_i}$  itself. When  $\epsilon_{i-1} = \epsilon_{i+1}$ , the corresponding intersection point related to  $t^{\epsilon_i}$  is on the stripe, so we replace  $t^{\epsilon_i}$  by  $t_*^{\epsilon_i} = t_*(p, q', l)$ . This is how  $t_{\#}^{\epsilon_i}$  is defined. □

Suppose  $K_1 = C(p, q, l, u, v)$  and  $K_2 = C(p, q', l, u, v)$ , where  $qq' \equiv 1 \pmod p$  and  $l \in \{2, p\}$ . Proposition 6.3 provides presentations of  $\pi_1(E(K_1))$  and  $\pi_1(E(K_2))$ . We will construct an explicit isomorphism  $\pi_1(E(K_1)) \cong \pi_1(E(K_2))$  based on the standard relations. First of all, let us introduce some notation:

Given words  $w_1$  and  $w_2$  made by  $s$  and  $t$ , let  $h_{w_1, w_2} = h(w_1, w_2)$  be a map on words such that for any word  $\omega$  made by  $s$  and  $t$ , the word  $h_{w_1, w_2}(\omega)$  is obtained from  $\omega$  by replacing  $s$  and  $t$  by  $w_1$  and  $w_2$ , respectively. For any integer  $n$ , define maps

$$f_1^n = h(s, s^n t), \quad f_2^n = h(t^n s, t), \quad g_1^n = h(s, t s^n), \quad g_2^n = h(s^n t, t)$$

and

$$h_0 = h(t, s), \quad h_1 = h(t, s^{-1}), \quad h_2 = h_1 \circ h_1 = h(s^{-1}, t^{-1}).$$

The map  $f_1^n$  induces an isomorphism  $\langle s, t \mid \omega \rangle \cong \langle s, t \mid f_1^n(\omega) \rangle$  by mapping  $t$  to  $s^n t$  and  $s$  to  $s$ , which is still denoted by  $f_1^n$ . A similar argument applies to  $f_2^n$ . For  $m$  odd, let  $f_m^n = f_1^n$ . For  $m$  even, let  $f_m^n = f_2^n$ . Given integers  $p, q > 0$ , suppose

$$\frac{q}{p} = [a_0; a_1, a_2, \dots, a_m] = a_0 + \frac{1}{a_1 + \frac{1}{a_2 + \frac{1}{a_3 + \dots}}}$$

is the unique continued fraction of  $q/p$  with  $a_i > 0$  and  $a_m > 1$ . Define

$$f^{q/p} = f_m^{-a_m+1} \circ f_{m-1}^{-a_{m-1}} \circ \dots \circ f_1^{-a_1} \circ f_0^{-a_0} \quad \text{and} \quad F^{q/p} = f_1^1 \circ f_2^{-1} \circ f^{q/p}.$$

The maps  $g_m^n, g^{q/p}$  and  $G^{q/p}$  are defined similarly based on  $g_1^n$  and  $g_2^n$ .

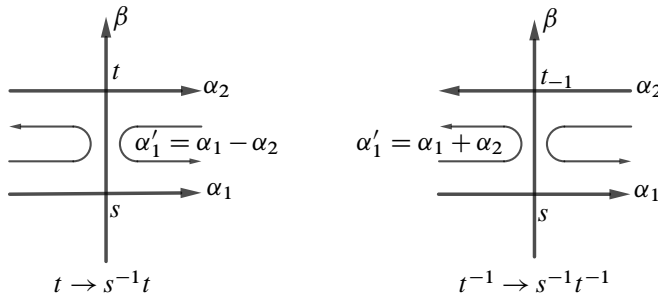


Figure 8: Examples of handle slides.

**Remark 6.4** The isomorphisms  $f_m^n$  and  $g_m^n$  can be achieved by handle slides of  $\alpha$  curves in the Heegaard diagram of the knot complement. Indeed, if there are two consecutive intersection points  $x_i$  and  $x_{i+1}$  in the definition of the standard relation that correspond to  $s$  and  $t$ , respectively, then the arc of  $\beta$  between  $x_i$  and  $x_{i+1}$  can be used for the handle slide. If  $\alpha_1$  is slid over  $\alpha_2$ , then the relation  $\omega$  becomes  $f_1^{-1}(\omega)$ . If  $\alpha_2$  is slid over  $\alpha_1$ , then the relation  $\omega$  becomes  $g_2^{-1}(\omega)$ . Moreover, when  $(x_i, x_{i+1}) \rightarrow (s, t), (s, t^{-1}), (s^{-1}, t), (s^{-1}, t^{-1}), (t, s), (t, s^{-1}), (t^{-1}, s), (t^{-1}, s^{-1})$ , where  $\rightarrow$  implies the replacement considered in the definition of the standard relation, then the corresponding maps are

$$\begin{aligned} & (f_1^{-1}, g_2^{-1}), \quad (g_1^1, g_2^1), \quad (f_1^1, f_2^1), \quad (g_1^{-1}, f_2^{-1}), \\ & (f_2^{-1}, g_1^{-1}), \quad (g_2^1, g_1^1), \quad (f_2^1, f_1^1), \quad (g_2^{-1}, f_1^{-1}), \end{aligned}$$

respectively. Two examples are shown in Figure 8.

The proof of the following lemma follows directly from definitions of maps.

**Lemma 6.5** *There are relations between maps*

- (i)  $h_0 \circ h_0 = h_2 \circ h_2 = \text{id}$ ,
- (ii)  $f_1^n \circ h_1 = h_1 \circ f_2^{-n}$  and  $f_2^n \circ h_1 = h_1 \circ g_1^{-n}$ ,
- (iii)  $g_1^n \circ h_1 = h_1 \circ g_2^{-n}$  and  $g_2^n \circ h_1 = h_1 \circ f_1^{-n}$ .

In the following lemmas, integers  $p, q$  and  $q'$  satisfy

$$p > 0, \quad q, q' \in [1, p - 1], \quad \text{gcd}(p, q) = 1 \quad \text{and} \quad qq' \equiv 1 \pmod{p}.$$

**Lemma 6.6** *The following equations hold:*

$$(6) \quad f^{q/p}(s_*(p, q, 2)ts) = ts \quad \text{and} \quad f^{q/p}(s_*(p, q, 2)st) = st,$$

$$(7) \quad g^{q/p}(tss_*(p, q, 2)) = ts \quad \text{and} \quad g^{q/p}(stss_*(p, q, 2)) = st.$$

**Proof** If  $l = 2$ , by definition  $s_*(p, q, 2) = st^{\theta_2}st^{\theta_3}s \dots st^{\theta_{p-1}}s$ , where  $\theta_i = \theta_i(p, q, 2)$ . Suppose that the integer  $k$  satisfies  $k - 1 \equiv (l - 1)q \pmod{p}$ . We know that  $k = q + 1$ . Suppose

$$\frac{q}{p} = [0; a_1, a_2, \dots, a_m]$$

with  $a_i > 0$  and  $a_m > 1$ . We prove (6) by induction on  $m$ .

If  $m = 1$ , then  $q = 1$  and  $p = a_1$ . Thus  $s_*(p, q, 2) = s^{a_1-1}$  and  $f^{q/p} = f_1^{-(a_1-1)}$  by definition. It can be checked directly that (6) holds.

Suppose (6) holds for  $m = m_0 - 1$ . Consider integers  $q_i$  satisfying  $q_i \equiv iq \pmod{p}$ . Since  $\gcd(p, q) = 1$ , if  $q_i \equiv iq \pmod{p}$ , then  $i = 1$ . So  $q_i \neq q$  for  $i \in [2, p - 1]$ . Since  $k = q + 1$ , the condition  $q_i \in [0, k)$  is the same as  $q_i \in [0, q)$  for  $i \in [2, p - 1]$ . Thus  $\theta_i(p, q, 2) = 1$  if and only if

$$\left\lfloor \frac{iq}{p} \right\rfloor - \left\lfloor \frac{(i-1)q}{p} \right\rfloor = 1.$$

In other words, we have

$$\theta_i(p, q, 2) = \left\lfloor \frac{iq}{p} \right\rfloor - \left\lfloor \frac{(i-1)q}{p} \right\rfloor \text{ for } i \in [2, p - 1].$$

If  $\theta_i(p, q, 2) = 1$ , there is some integer  $j \in [1, q - 1]$  such that

$$i = \left\lfloor \frac{jP}{q} \right\rfloor + 1 = ja_1 + \left\lfloor \frac{j_r}{q} \right\rfloor + 1,$$

where

$$\frac{r}{q} = [0; a_2, a_3, \dots, a_{m_0}].$$

Let  $j_1 = j$  and  $j_2 = j - 1$  for  $j \in [2, q - 1]$ . Then we have

$$\left( j_1 a_1 + \left\lfloor \frac{j_1 r}{q} \right\rfloor + 1 \right) - \left( j_2 a_1 + \left\lfloor \frac{j_2 r}{q} \right\rfloor + 1 \right) = a_1 + \left\lfloor \frac{j_1 r}{q} \right\rfloor - \left\lfloor \frac{j_2 r}{q} \right\rfloor = a_1 + \theta_j(q, r, 2).$$

Thus

$$\begin{aligned} s_*(p, q, 2)ts &= s^{a_1}t s^{a_1} s^{\theta_2(q,r,2)} t s^{a_1} s^{\theta_3(q,r,2)} t \dots s^{a_1} s^{\theta_{q-2}(q,r,2)} t s^{a_1} s^{\theta_{q-1}(q,r,2)} t s^{a_1} t s \\ &= (s^{a_1}t) s^{\theta_2(q,r,2)} (s^{a_1}t) s^{\theta_3(q,r,2)} (s^{a_1}t) \dots s^{\theta_{q-2}(q,r,2)} (s^{a_1}t) s^{\theta_{q-1}(q,r,2)} \\ &\quad \cdot (s^{a_1}t) (s^{a_1}t) s \\ &= h_{s^{a_1}t, s}(s_*(q, r, 2)st) = f_1^{a_1} \circ h_0(s_*(q, r, 2)st), \end{aligned}$$

where the second equality follows from the fact that  $\theta_i(p, q, 2) = 0$  if  $i < a_1$ . Similarly,

$$\begin{aligned} s_*(p, q, 2)st &= s^{a_1}t s^{a_1} s^{\theta_2(q,r,2)}t s^{a_1} s^{\theta_3(q,r,2)}t \dots s^{a_1} s^{\theta_{q-2}(q,r,2)}t s^{a_1} s^{\theta_{q-1}(q,r,2)}t s^{a_1}st \\ &= (s^{a_1}t) s^{\theta_2(q,r,2)}(s^{a_1}t) s^{\theta_3(q,r,2)}(s^{a_1}t) \dots s^{\theta_{q-2}(q,r,2)}(s^{a_1}t) s^{\theta_{q-1}(q,r,2)} \\ &\qquad \qquad \qquad \cdot (s^{a_1}t) s^{a_1}st \\ &= h_{s^{a_1}t, s}(s_*(q, r, 2)ts) = f_1^{a_1} \circ h_0(s_*(q, r, 2)ts). \end{aligned}$$

By the inductive assumption, we have

$$f^{r/q}((s_*(p, q, 2)ts) = ts \quad \text{and} \quad f^{r/q}((s_*(p, q, 2)st) = st.$$

Since  $f^{q/p} = h_0 \circ f^{r/q} \circ h_0 \circ f_1^{-a_1}$  and  $h_0 \circ h_0 = \text{id}$ , we have

$$\begin{aligned} f^{q/p}((s_*(p, q, 2)ts) &= h_0 \circ f^{r/q}(s_*(q, r, 2)st) = h_0(st) = ts, \\ f^{q/p}((s_*(p, q, 2)st) &= h_0 \circ f^{r/q}(s_*(q, r, 2)ts) = h_0(ts) = st. \end{aligned}$$

By a similar method, it can be proven that

$$tss_*(p, q, 2) = g_1^{a_1} \circ h_0(sts_*(q, r, 2)) \quad \text{and} \quad sts_*(p, q, 2) = g_1^{a_1} \circ h_0(tss_*(q, r, 2)).$$

Then by induction, (7) holds. □

**Lemma 6.7** *The following equations hold:*

$$\begin{aligned} F^{q/p}(t) &= f_1^1 \circ f_2^{-1} \circ f^{q/p}(t) = h_0(s_*(p, q', 2)ts), \\ G^{q/p}(t) &= g_1^1 \circ g_2^{-1} \circ g^{q/p}(t) = h_0(sts_*(p, q', 2)). \end{aligned}$$

**Proof** The proofs of the two equations are similar. We only show the proof of the first equation. By the proof of Lemma 6.6, we know

$$\theta_i(p, q, 2) = \left\lfloor \frac{iq}{p} \right\rfloor - \left\lfloor \frac{(i-1)q}{p} \right\rfloor \quad \text{for } i \in [2, p-1].$$

Thus  $\theta_i(p, q, 2) = 0$  if and only if

$$\left\lfloor \frac{i(q-p)}{p} \right\rfloor - \left\lfloor \frac{(i-1)(q-p)}{p} \right\rfloor = \left\lfloor \frac{iq}{p} \right\rfloor - \left\lfloor \frac{(i-1)q}{p} \right\rfloor - 1 = -1.$$

This is equivalent to

$$\left\lfloor \frac{i(p-q)}{p} \right\rfloor - \left\lfloor \frac{(i-1)(p-q)}{p} \right\rfloor = 1,$$

ie  $\theta_i(p, p-q, 2) = 1$ . Then

$$\begin{aligned} f_1^{-1} \circ h_0(s_*(p, q, 2)ts) &= s^{-1}t s^{-\theta_2(p,p-q,2)}t \dots t s^{-\theta_{p-1}(p,p-q,2)}t t \\ &= s^{-1}h_1(s_*(p, p-q)st)s = t^{-1}h_1(sts_*(p, p-q))t. \end{aligned}$$

Suppose  $q/p = [0; a_1, a_2, \dots, a_m]$  with  $a_i > 0$  and  $a_m > 1$ . We have

$$f_2^{-1} \circ f^{q/p} = \begin{cases} f_2^{-1} \circ f_1^{-a_m+1} \circ f_2^{-a_{m-1}} \circ \dots \circ f_2^{-a_2} \circ f_1^{-a_1} & \text{if } m \text{ is odd,} \\ f_2^{-a_m} \circ f_1^{-a_{m-1}} \circ \dots \circ f_2^{-a_2} \circ f_1^{-a_1} & \text{if } m \text{ is even.} \end{cases}$$

By the extended Euclidean algorithm,

$$\frac{p-q'}{p} = \begin{cases} [0; 1, a_m-1, a_{m-1}, \dots, a_2, a_1] & \text{if } m \text{ is odd,} \\ [0; a_m, a_{m-1}, \dots, a_2, a_1] & \text{if } m \text{ is even.} \end{cases}$$

It can be proven by induction on  $n$  that for  $b/a = [0; b_1, b_2, \dots, b_{2n-1}, b_{2n}]$ ,

$$(8) \quad f_2^{-b_1} \circ f_1^{-b_2} \circ \dots \circ f_2^{-b_{2n-1}} \circ f_1^{-b_{2n}}(t) = h_1(ts_*(a, b)s).$$

Indeed, if  $n = 1$ , then  $f_2^{-b_2} \circ f_1^{-b_1}(t) = (t^{-b_2}s)^{-b_1}t = (s^{-1}t^{b_2})^{b_1}t$ . Equation (8) is clear.

Suppose (8) holds for  $n = n_0 - 1$ . Let

$$\frac{b'}{a'} = [0; b_2, \dots, b_{2n_0-1}, b_{2n_0}] \quad \text{and} \quad \frac{b''}{a''} = [0; b_3, \dots, b_{2n_0-1}, b_{2n_0}].$$

By the proof of Lemma 6.6,

$$\begin{aligned} t f_1^{b_1}(s_*(a'', b'', 2)st)t^{-1} &= t h_0(s_*(a', b', 2)ts)t^{-1} \\ &= s^{-1} h_0(t s s_*(a', b', 2))s = s^{-1} g_1^{b_1}(s t s_*(a'', b'', 2))s. \end{aligned}$$

Thus

$$\begin{aligned} f_2^{-b_1} \circ f_1^{-b_2} \circ h_1(ts_*(a, b, 2)stt^{-1}) &= f_2^{-b_1} \circ h_1 \circ f_2^{b_2}(t f_2^{-b_2} \circ f_1^{-b_1}(s_*(a'', b'', 2)st)t^{-1}) \\ &= f_2^{-b_1} \circ h_1(t f_1^{-b_1}(s_*(a'', b'', 2)st)t^{-1}) \\ &= h_1 \circ g_1^{b_1} \circ (s^{-1} g_1^{b_1}(s t s_*(a'', b'', 2))s) \\ &= h_1(s^{-1}(s t s_*(a'', b'', 2))s) = h_1(ts_*(a'', b'', 2)s). \quad \square \end{aligned}$$

**Remark 6.8** By Remark 6.4, the map  $f^{q/p}$  can be regarded as a sequence of handle slides. Consider the matrix of algebraic intersection points

$$\begin{bmatrix} [\alpha_1] \cdot p[a] & [\alpha_2] \cdot p[a] \\ [\alpha_1] \cdot [m] & [\alpha_2] \cdot [m] \end{bmatrix},$$

where  $a$  and  $m$  are curves in Figure 6. The maps  $f_1^n$  and  $f_2^n$  induce column transformations of this matrix, which are still denoted by  $f_1^n$  and  $f_2^n$ . Then

$$f^{q/p} \left( \begin{bmatrix} p & q \\ 0 & 1 \end{bmatrix} \right) = \begin{bmatrix} 1 & 1 \\ q' - p & q' \end{bmatrix} \quad \text{and} \quad F^{q/p} \left( \begin{bmatrix} p & q \\ 0 & 1 \end{bmatrix} \right) = \begin{bmatrix} 1 & 0 \\ q' & p \end{bmatrix}.$$

Indeed, the definitions of  $f^{q/p}$  and  $F^{q/p}$  come from the extend Euclidean algorithm for calculating  $\gcd(p, q)$  (see the proof of Lemma 6.7).

**Proposition 6.9** *Up to circular permutation,*

$$h_0 \circ F^{q/p}(\omega(p, q', 2, u, v)) = \begin{cases} h_2(\omega(p, q, 2, u, v)) & \text{if } v \text{ is odd,} \\ \omega(p, q, 2, u, v) & \text{if } v \text{ is even.} \end{cases}$$

**Proof** Suppose  $a = s_*(p, q, 2)$  and  $b = s_*(p, q', 2)$ . Then

$$t_* = t_*(p, q, 2) = t_*(p, q', 2) = tst$$

and

$$\omega(p, q', 2, u, v) = a^{\epsilon_1} t_{\#}^{\epsilon_2} a^{\epsilon_3} \dots t_{\#}^{\epsilon_{2u}}, \quad \omega(p, q, 2, u, v) = b^{\epsilon_1} t_{\#}^{\epsilon_2} b^{\epsilon_3} \dots t_{\#}^{\epsilon_{2u}}.$$

The word  $a^{\epsilon_{i-1}} t_{\#}^{\epsilon_i} a^{\epsilon_{i+1}}$  is one of

- (i)  $atsta = (ats)ta$  and  $a^{-1}(tst)^{-1}a^{-1} = a^{-1}t^{-1}(ats)^{-1}$ ,
- (ii)  $ata^{-1} = (ats)t(ast)^{-1}$  and  $at^{-1}a^{-1} = (ast)t^{-1}(ats)^{-1}$ ,
- (iii)  $a^{-1}ta$  and  $a^{-1}t^{-1}a$ .

Thus  $\omega(p, q', 2, u, v) = a_{\#}^{\epsilon_1} t^{\epsilon_2} a_{\#}^{\epsilon_3} \dots t^{\epsilon_{2u}}$ , where

$$a_{\#}^{\epsilon_i} = \begin{cases} (ats)^{\epsilon_i} & \text{if } \epsilon_i = \epsilon_{\epsilon_i+i}, \\ (ast)^{\epsilon_i} & \text{if } \epsilon_i = -\epsilon_{\epsilon_i+i}. \end{cases}$$

By Lemma 6.6 and Lemma 6.7,

$$F^{q/p}(ats) = s = h_0(t), \quad F^{q/p}(ast) = t^{-1}st = h_0(s^{-1}ts) \quad \text{and} \quad F^{q/p}(t) = h_0(bts).$$

Thus  $h_0 \circ F^{q/p}(\omega(p, q', 2, u, v)) = c_{\#}^{\epsilon_1} (bts)^{\epsilon_2} c_{\#}^{\epsilon_3} \dots (bts)^{\epsilon_{2u}}$ , where

$$c_{\#}^{\epsilon_i} = \begin{cases} t^{\epsilon_i} & \text{if } \epsilon_i = \epsilon_{\epsilon_i+i}, \\ (s^{-1}ts)^{\epsilon_i} & \text{if } \epsilon_i = -\epsilon_{\epsilon_i+i}. \end{cases}$$

The word  $(bts)^{\epsilon_{i-1}} c_{\#}^{\epsilon_i} (bts)^{\epsilon_{i+1}}$  is one of

- (i)  $(bts)t(bts) = b(tst)bt$  and  $(bts)^{-1}(t)^{-1}(bts)^{-1} = (bts)^{-1}(tst)^{-1}b^{-1}$ ,
- (ii)  $(bts)(s^{-1}ts)(bts)^{-1} = btb^{-1}$  and  $(bts)(s^{-1}ts)^{-1}(bts)^{-1} = bt^{-1}b^{-1}$ ,
- (iii)  $(bts)^{-1}t(bts)$  and  $(bts)^{-1}t^{-1}(bts)$ .

Thus

$$h_0 \circ F^{q/p}(\omega(p, q', 2, u, v)) = t_{\#}^{\epsilon_1} b^{\epsilon_2} t_{\#}^{\epsilon_3} \dots b^{\epsilon_{2u}} = b_{\#}^{\epsilon_{u+1}} t^{\epsilon_{u+2}} b_{\#}^{\epsilon_{u+3}} \dots b^{\epsilon_{3u}},$$

where the last equality holds up to circular permutation. The proposition follows from the fact that  $\epsilon_{u+i} = (-1)^v \epsilon_i$ . □

**Proposition 6.10** Up to circular permutation,

$$h_0 \circ G^{(p-q)/p}(\omega(p, q', p, u, v)) = \begin{cases} h_2(\omega(p, q, p, u, v)) & \text{if } v \text{ is odd,} \\ \omega(p, q, p, u, v) & \text{if } v \text{ is even.} \end{cases}$$

**Proof** The essential idea of the proof is the same as that of Proposition 6.9. Now

$$s_*(p, q, p) = s \quad \text{and} \quad t_*(p, q, p) = ts_*(p, p - q, 2)t.$$

Suppose  $a = s_*(p, p - q, 2)$  and  $b = s_*(p, p - q', 2)$ . By analyzing cases of  $s^{\epsilon_i - 1} t_{\#}^{\epsilon_i} s^{\epsilon_i}$  we get  $\omega(p, q', p, u, v) = a_{\#}^{\epsilon_1} t^{\epsilon_2} a_{\#}^{\epsilon_3} \dots t^{\epsilon_{2u}}$ , where

$$a_{\#}^{\epsilon_i} = \begin{cases} (sta)^{\epsilon_i} & \text{if } \epsilon_i = \epsilon_{\epsilon_i + i}, \\ (sat)^{\epsilon_i} & \text{if } \epsilon_i = -\epsilon_{\epsilon_i + i}. \end{cases}$$

Note that  $\epsilon_i \in \{\pm 1\}$  and  $\epsilon_{\epsilon_i + i} = \epsilon_{\pm i}$  in the definition of  $a_{\#}^{\epsilon_i}$ . By Lemma 6.6 and Lemma 6.7, we have  $G^{(p-q)/p}(t) = stb$ . Thus

$$h_0 \circ G^{(p-q)/p}(\omega(p, q', p, u, v)) = c_{\#}^{\epsilon_1} (stb)^{\epsilon_2} c_{\#}^{\epsilon_3} \dots (stb)^{\epsilon_{2u}},$$

where

$$c_{\#}^{\epsilon_i} = \begin{cases} t^{\epsilon_i} & \text{if } \epsilon_i = \epsilon_{\epsilon_i + i}, \\ ((stb)^{-1} (sts^{-1}) (stb))^{\epsilon_i} & \text{if } \epsilon_i = -\epsilon_{\epsilon_i + i}. \end{cases}$$

By analyzing cases of  $(stb)^{\epsilon_i - 1} c_{\#}^{\epsilon_i} (stb)^{\epsilon_i - 1} (stb)^{\epsilon_i + 1}$  we get

$$h_0 \circ G^{(p-q)/p}(\omega(p, q', p, u, v)) = t_{\#}^{\epsilon_1} b^{\epsilon_2} t_{\#}^{\epsilon_3} \dots b^{\epsilon_{2u}}.$$

Then this proposition follows from a similar argument as in Proposition 6.9. □

**Proof of the sufficient part of Theorem 1.2** For  $i = 1, 2$ , let  $K_i = C(p_i, q_i, l_i, u_i, v_i)$ ,  $M_i = E(K_i)$  and suppose  $(\mu_i, \lambda_i)$  is the regular basis of  $\partial M_i$ . Suppose

$$(p_1, u_1, v_1) = (p_2, u_2, v_2) = (p, u, v), q_1 q_2 \equiv 1 \pmod{p} \quad \text{and} \quad l_1 = l_2 \in \{2, p\}.$$

By knot Floer homology, constrained knots  $K_i$  are not unknots in lens spaces. By Proposition 2.1, we know that the  $M_i$  are Haken manifolds.

Let  $q' = q_1$  and  $q = q_2$  in Propositions 6.9 and 6.10. Let  $\psi$  be the map from  $\pi_1(M_1)$  to  $\pi_1(M_2)$  induced by

$$\begin{aligned} h_0 \circ F^{q/p} & \quad \text{if } l_1 = l_2 = 2, \\ h_0 \circ G^{(p-q)/p} & \quad \text{if } l_1 = l_2 = p. \end{aligned}$$

By Propositions 6.9 and 6.10, the map  $\psi$  is an isomorphism. The meridians  $\mu_i$  and longitudes  $\lambda_i$  can be isotoped to lie on Heegaard diagrams of  $M_i$  so that  $\mu_1 = m$  and  $\mu_2 = pa$ , where  $a$  and  $m$  are curves in Figure 6. Moreover, suppose that meridians and longitudes are disjoint from  $\beta_1$ . By Remarks 6.4 and 6.8, the map

$\psi$  can be achieved by handle slides of  $\alpha$  curves. After handle slides, the meridian and the longitude are still disjoint from  $\beta_1$ , which implies  $\psi$  induces an isomorphism  $\psi|_{\pi_1(\partial M_1)}: \pi_1(\partial M_1) \rightarrow \pi_1(\partial M_2)$ .

Moreover, for the case  $l_1 = l_2 = 2$ , note that  $t$  corresponds to  $\mu_1 \cap (\alpha_1 \cup \alpha_2)$  and  $s_*(p, q', 2)ts$  corresponds to  $\mu_2 \cap (\alpha_1 \cup \alpha_2)$  in the presentations of the fundamental groups. By Lemma 6.7,

$$\psi|_{\pi_1(\partial M_1)}(\mu_1) = \psi(t) = s_*(p, q', 2)ts = \mu_2.$$

Thus, by Lemma 6.1, we know  $K_1$  is equivalent to  $K_2$ .

For the case  $l_1 = l_2 = p$ , based on Lemma 6.7, the proof is similar. □

## 7 Magic links

A constrained knot is defined by a doubly pointed Heegaard diagram  $(T^2, \alpha_1, \beta_1, z, w)$ , where  $\beta_1$  looks similar to the  $\beta$  curve in the diagram of a 2-bridge knot (see Lemma 4.2 and Proposition 3.5). In this section we provide Dehn surgery descriptions for some families of constrained knots, which is inspired by the relation between constrained knots and 2-bridge knots. The main objects in this section are magic links.

**Definition 7.1** Suppose integers  $u$  and  $v$  satisfy  $0 \leq v < u$  and  $\gcd(u, v) = 1$ , and  $u$  is odd. Especially,  $(u, v) = (1, 0)$  is allowed. A *magic link*  $\mathfrak{L}(u, v) = K_0 \cup K_1 \cup K_2$  is a 3-component link linked as shown in Figure 2, left, where  $K_0$  is the 2-bridge knot  $\mathfrak{b}(u, v)$  in the standard presentation, and  $K_1$  and  $K_2$  are unknots. For  $-u < v < 0$ , let  $\mathfrak{L}(u, v)$  be the mirror link of  $\mathfrak{L}(u, -v)$ . Let  $\mathfrak{L}(1, 1)$  be the mirror link of  $\mathfrak{L}(1, 0)$ .

**Remark 7.2** The name of magic links is from the fact that the link complement  $S^3 - \mathfrak{L}(3, 1)$  is diffeomorphic to the magic manifold studied in [24].

For  $i = 1, 2$ , suppose integers  $p_i$  and  $q_i$  satisfy  $p_i > 0$  and  $\gcd(p_i, q_i) = 1$ . Let  $M(u, v, p_1/q_1, p_2/q_2)$  and  $K_0(u, v, p_1/q_1, p_2/q_2)$  denote the manifold and the resulting knot  $K'_0$  obtained by  $p_i/q_i$  Dehn surgery on  $K_i$ .

**Proposition 7.3** *The manifolds  $M(u, v, p_1/q_1, p_2/q_2)$  and  $M(u, v, p_2/q_2, p_1/q_1)$  are diffeomorphic. Moreover, the knots  $K'_0$  in these manifolds are equivalent.*

**Proof** The components  $K_1$  and  $K_2$  in the magic link switch their positions under the rotation around a vertical line, while  $K_0$  remains unchanged. □



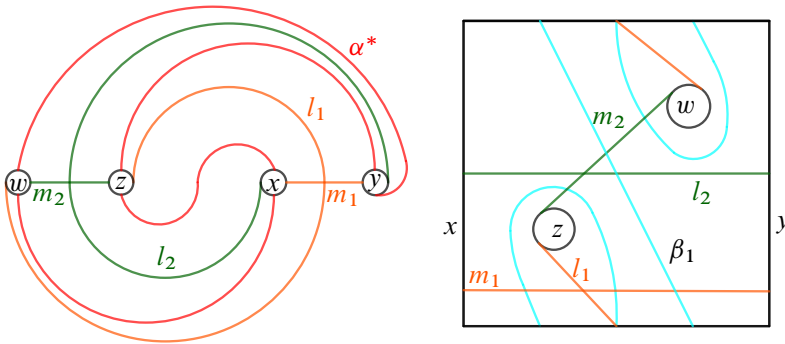


Figure 9: Heegaard diagrams of  $E(\mathcal{L}(3, 1))$ , where  $\beta_1$  is omitted in the left figure and  $\alpha^*$  is omitted in the right figure.

**Remark 7.4** Manifolds  $M(u, v, p_1/q_1, p_2/q_2)$  and  $M(u, v, p_2/q_2, p_1/q_1)$  will not be distinguished in the rest of the paper. Neither will the corresponding knots  $K_0$  of these manifolds.

**Proposition 7.5** For integers  $u$  and  $v$  satisfying  $0 < v < u$  and  $\gcd(u, v) = 1$ , and where  $u$  is odd, the link  $\mathcal{L}(u, u - v)$  is the mirror link of  $\mathcal{L}(u, v)$ . Thus

$$\mathcal{L}(u, u - v) \cong \mathcal{L}(u, -v)$$

and  $K_0(u, v, p_1/q_1, p_2/q_2)$  is the mirror image of  $K_0(u, u - v, p_1/(-q_1), p_2/(-q_2))$ .

**Proof** Suppose  $\mathfrak{b}(u, v)$  is in the standard presentation for

$$\frac{v}{u} = [0; a_1, a_2, \dots, a_m].$$

Since  $(u - v)/u = 1 - v/u$ , by adding one positive half-twist on the two left strands, the standard presentation for  $[0; -a_1, -a_2, \dots, -a_m]$  becomes a standard presentation of  $\mathfrak{b}(u, u - v)$ . After isotoping the link outside twists related to  $a_i$ , the link  $\mathcal{L}(u, u - v)$  becomes the mirror link of  $\mathcal{L}(u, v)$ . □

**Lemma 7.6** In Figure 9,  $(\Sigma_2, \alpha^*, \beta_1)$  are Heegaard diagrams of  $E(\mathcal{L}(3, 1))$ . For  $i = 1, 2$ , the meridian  $m_i$  and the longitude  $l_i$  of  $K_i$  can be isotoped to lie on  $\Sigma_2$  as in the diagrams. For general integers  $u$  and  $v$  satisfying  $0 < v < u$  and  $\gcd(u, v) = 1$ , and where  $u$  and  $v$  are odd, the similar assertion holds when  $\beta_1$  is replaced by  $\beta$  in the doubly pointed Heegaard diagram of  $\mathfrak{b}(u, v)$ .

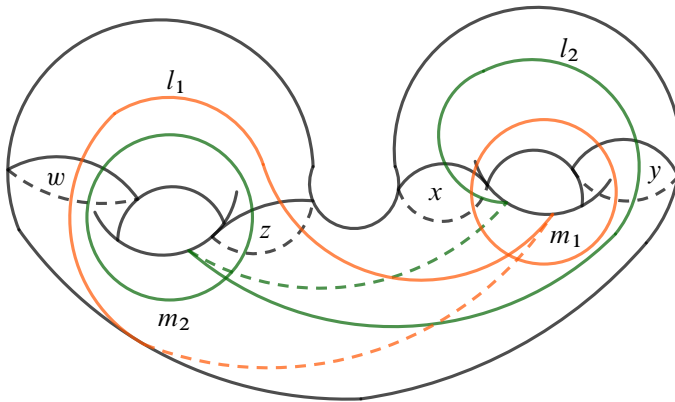


Figure 10: Meridians and longitudes on the Heegaard surface.

**Proof** Consider  $(u, v) = (3, 1)$ . The curve  $\alpha^*$  is separating and  $\beta_1$  is nonseparating. Therefore, the manifold obtained from  $\Sigma_2 \times [-1, 1]$  by attaching 2–handles along  $\alpha^* \times \{-1\}$  and  $\beta_1 \times \{1\}$  has three boundary components, each of which is a torus. Moreover, if two more 2–handles are attached along  $m_1 \times \{-1\}$  and  $m_2 \times \{-1\}$ , the resulting manifold is  $E(\mathfrak{b}(3, 1))$ . The longitude  $l_0$  of  $\mathfrak{b}(3, 1)$  can be isotoped to lie on  $\Sigma_2$  as shown in the Schubert normal form (see Figure 3, center). Note that the geometric intersection number of  $m_i$  and  $l_i$  is one.

On the other hand, components of the link corresponding to the Heegaard diagrams in Figure 9 can be obtained by pushing  $l_i$  slightly into the handlebody corresponding to  $\alpha = \{\alpha^*, m_1, m_2\}$  and pushing  $l_0$  slightly into the handlebody corresponding to  $\beta = \{\beta_1, m_0\}$ , where  $m_0$  is the meridian of  $\mathfrak{b}(3, 1)$  on  $\Sigma_2$ . This can be seen explicitly if we redraw the Heegaard surface as in Figure 10. After isotoping unknot components, it is easy to see the link from these diagrams is equivalent to  $\mathfrak{L}(3, 1)$ . For general  $(u, v)$ , the proof applies without change.  $\square$

For integers  $u$  and  $v$  satisfying  $-u < v < 0$ , and where  $u$  and  $v$  odd, the corresponding diagram is obtained by reflecting the diagram of  $\mathfrak{L}(u, -v)$  along a vertical line. Since  $\mathfrak{L}(u, u - v) \cong \mathfrak{L}(u, -v)$ , Heegaard diagrams for all  $v \in (-u, u)$  with  $\gcd(u, v) = 1$  and  $(u, v) = (1, 0), (1, 1)$  are obtained from this approach. Such a diagram is called a *standard diagram* of  $E(\mathfrak{L}(u, v))$ .

A resolution of an intersection point of a meridian and a longitude on the Heegaard surface is called a *positive resolution* or a *negative resolution* when the meridian turns left or right, respectively, to the longitude in any direction; see Figure 11.

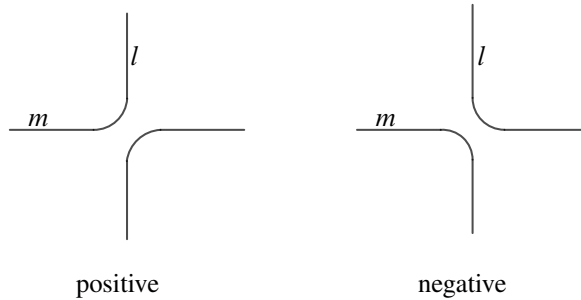


Figure 11: Positive and negative resolutions.

**Corollary 7.7** For  $i = 1, 2$  suppose integers  $p_i$  and  $q_i$  satisfy  $\gcd(p_i, q_i) = 1$  and  $p_i > 0$ . The Heegaard diagram  $(\Sigma_2, \{\alpha_1, \alpha_2\}, \beta_1)$  of  $E(K_0(u, v, p_1/q_1, p_2/q_2))$  is obtained in the following way:  $\alpha_i$  is obtained by resolving intersection points of  $|p_i|$  copies of  $m_i$  and  $|q_i|$  copies of  $l_i$  positively or negatively if  $q_i$  is positive or negative, respectively. Especially when  $(p_i, q_i) = (1, 0)$ , the corresponding  $\alpha_i$  is  $m_i$ .

**Proof** This follows from the definition of Dehn surgery. Note that  $\alpha_i$  is the meridian of the filling solid torus for  $i = 1, 2$ . □

Consider cyclic covers of the diagram of a 2–bridge knot  $b(u, v)$  as shown in Figure 12. For  $i \in \mathbb{Z}$ , let  $a_i = a_i(u, v)$  be a red strand connecting the left edge to the right edge and passing through  $|i|$  copies of the fundamental domains, where the sign of  $i$  determines the direction of the strand; see Figure 12 for examples of strands. Let  $A_i = A_i(u, v)$  be the set consisting of strands that can be isotoped into the neighborhood of  $a_i(u, v)$  in the complement of basepoints. Some intersection points of  $a_i(u, v)$  and  $\beta_1$  can be removed by isotopy. Intersection points that cannot be removed are shown in Figure 12. Identifying endpoints of  $a_i$ , a diagram of a 2–bridge knot  $b(U(u, v, i), V(u, v, i))$  can be obtained for some integers  $U(u, v, i)$  and  $V(u, v, i)$ .

Let  $a_* = a_*(u, v)$  and  $a_\# = a_\#(u, v)$  be the strands in Figure 12. For  $i = *, \#$ , the set  $A_i(u, v)$  and the functions  $U(u, v, i)$  and  $V(u, v, i)$  are defined similarly. For  $i \in \mathbb{Z}$  or  $i = *, \#$ , consider  $V(u, v, i) \in \mathbb{Z}/U\mathbb{Z} - \{0\}$  for  $U = U(u, v, i) > 1$ . When  $U(u, v, i) = 1$ , consider  $V(u, v, i) \in \{0, 1\}$ . In the latter case, we use the following conventions:

$$n \equiv \begin{cases} 1 & \text{if } n \text{ is odd modulo } 1, \\ 0 & \text{if } n \text{ is even modulo } 1, \end{cases} \quad \text{and} \quad \pm n \equiv \mp m \pmod{1} \text{ for } n \text{ odd and } m \text{ even.}$$

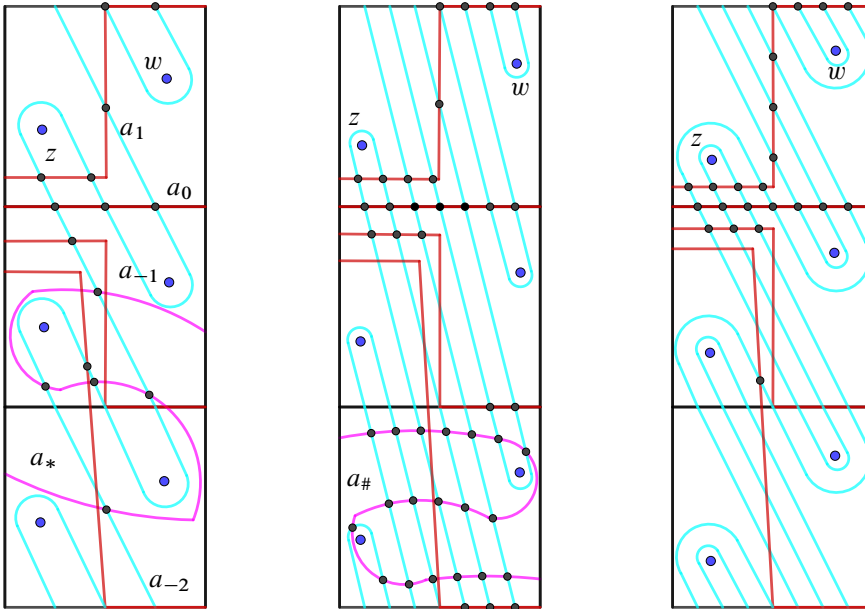


Figure 12: Cyclic covers of Heegaard diagrams corresponding to  $(u, v) = (3, 1), (7, 1), (7, 2)$ .

**Lemma 7.8** Suppose  $u$  and  $v$ , for  $u$  odd, are integers satisfying  $(u, v) = (1, 0)$  or  $0 < 2v < u$ , and  $\gcd(u, v) = 1$ . For  $i \in \{1, 0, -1, -2, *, \#\}$ , the functions  $U(u, v, i)$  and  $V(u, v, i)$  can be expressed explicitly:

- (i)  $U(u, v, 1) = u + 2v$  and  $V(u, v, 1) = v$ .
- (ii)  $U(u, v, 0) = u$  and  $V(u, v, 0) = v$ .
- (iii)  $U(u, v, -1) = u - 2v$  and  $V(u, v, -1) \equiv v \pmod{u - 2v}$ .
- (iv)  $U(u, v, -2) = |u - 4v|$  and  $V(u, v, -2) \equiv v \operatorname{sign}(u - 4v) \pmod{|u - 4v|}$  for  $u > 3$ , and  $U(3, 1, -2) = 1$  and  $V(3, 1, -2) = 1$ .
- (v)  $U(u, v, *) = 3u - 4v$  and  $V(u, v, *) = u - v$ .
- (vi)  $U(u, v, \#) = 3u - 2v$  and  $V(u, v, \#) = 2u - v$ .

**Proof** For fixed  $(u, v)$ , let  $R_i$  and  $S_i$  be numbers of rainbows and stripes in the diagram of  $\mathfrak{b}(U(u, v, i), V(u, v, i))$ . Case (ii) is trivial, where  $R_0 = v$  and  $S_0 = u - 2v$ . Suppose  $V'$  satisfies

$$0 < V' < U(u, v, i) \quad \text{and} \quad V' \equiv V(u, v, i) \pmod{U(u, v, i)}.$$

Define

$$\epsilon_i = \begin{cases} -1 & \text{if } 2V' < U(u, v, i), \\ 1 & \text{if } 2V' > U(u, v, i). \end{cases}$$

Then  $(U(u, v, i), V(u, v, i))$  can be recovered from  $(R_i, S_i, \epsilon_i)$  by

$$(9) \quad U(u, v, i) = 2R_i + S_i \quad \text{and} \quad V(u, v, i) = \epsilon_i R_i.$$

Suppose that all isotopies on the surface move basepoints in the following discussion.

For cases (i) and (vi), let  $x_1$  be the center of the fundamental domain and let  $D_1 = N(x_1)$  be the neighborhood containing two basepoints  $z$  and  $w$ . Straightening strands isotopes the diagram by rotating  $D_1$  clockwise and counterclockwise by  $\pi$  for cases (i) and (vi), respectively. Equivalently, the new  $\beta$  is obtained by pushing rainbows on the top edge to the bottom right and bottom left, respectively. Rainbows and stripes satisfy the following equations and we obtain the results by formulae in (9):

$$\begin{aligned} R_1 &= R_0, & S_1 &= 2R_0 + S_0, & \epsilon_1 &= 1, \\ R_\# &= R_0 + S_0, & S_\# &= 2R_0 + S_0, & \epsilon_\# &= -1. \end{aligned}$$

For case (v), let  $x_2$  be the middle intersection point on the top edge and let  $D_2 = N(x_2)$  be the neighborhood containing all rainbows. Straightening the strand isotopes the diagram by rotating  $D_2$  clockwise by  $\pi$ . Then we have

$$R_* = R_0 + S_0, \quad S_* = S_0, \quad \epsilon_* = 1.$$

For case (iii), the number  $U(u, v, -1)$  is the same as  $S_0$ . Straightening the strand isotopes the diagram by rotating  $D_2$  counterclockwise, which induces the formula of  $V(u, v, -1)$ . This isotopy can also be regarded as pulling back rainbows once.

For case (iv), if  $(u, v) = (3, 1)$ , then the formula is obtained directly from Figure 12. If  $u > 3$ , then there are three subcases where  $S_0 > 2R_0$ ,  $2R_0 > S_0 > R_0$  and  $R_0 > S_0$ , equivalently  $u > 4v$ ,  $4v > u > 3v$  and  $3v > u > 2v$ , respectively. Note that  $u$  is odd, so  $u \neq 4v$ .

Suppose  $S_0 > 2R_0$  (eg  $(u, v) = (7, 1), (13, 3)$ ). In this subcase  $V(u, v, -1) = v$ . Straightening the strand isotopes the diagram by pulling back rainbows twice. Then  $(U(u, v, -2), V(u, v, -2))$  is obtained by applying case (iii) twice, ie

$$U(u, v, -2) = u - 4v, \quad V(u, v, -2) \equiv v \pmod{u - 4v}.$$

Suppose  $2R_0 > S_0 > R_0$  (eg  $(u, v) = (7, 2), (15, 4)$ ). Straightening the strand isotopes the diagram by rotating  $D_2$  counterclockwise by  $\pi$ . After isotopy, the number of

intersection points of  $a_{-2}$  and  $\beta$  is  $U(u, v, -2) = 2R_0 - S_0 = 4v - u$ . The number of rainbows is  $R_{-2} = S_0 - R_0$  and  $\epsilon_{-2} = -1$ . Hence

$$V(u, v, -2) = U(u, v, -2) - (S_0 - R_0) = 7v - 2u.$$

Suppose  $R_0 > S_0$  (eg  $(u, v) = (7, 3)$ ). Straightening the strand isotopes the diagram by rotating  $D_2$  counterclockwise by  $\pi$ . In this subcase, this isotopy is obtained by reversing the isotopy in case (v). Then

$$R_{-2} = R_0 - S_0, \quad S_{-2} = S_0, \quad \epsilon_{-2} = 1, \quad U(u, v, -2) = 4v - u, \quad V(u, v, -2) = 3v - u.$$

The formula for case (iv) then follows from summarizing the above subcases.  $\square$

**Remark 7.9** Indeed, for any  $i \in \mathbb{Z}$ , functions  $U(u, v, i)$  and  $V(u, v, i)$  might be expressed explicitly. For example, we have  $U(u, v, i) = u + 2iv$  and  $V(u, v, i) = v$  for  $i > 0$ . However, for  $i < 0$ , functions are more complicated so we omit the discussion.

The following lemma is a basic result from the Dehn surgery on the Hopf link.

**Lemma 7.10** *The manifold  $M(u, v, p_1/q_1, p_2/q_2)$  is diffeomorphic to the lens space*

$$L(p_1 p_2 - q_1 q_2, p_1 p'_2 - q_1 q'_2) \quad \text{where } p_2 q'_2 - q_2 p'_2 = -1.$$

**Theorem 7.11** *Suppose integers  $u_0$  and  $v_0$  satisfy  $(u_0, v_0) = (1, 0)$  or  $0 < 2v_0 < u_0$ , and  $\gcd(u_0, v_0) = 1$ , where  $u$  is odd. Suppose  $U_i = U(u_0, v_0, i)$  and  $V_i = V(u_0, v_0, i)$ . The knot  $K_0 = K_0(u_0, v_0, p_1/q_1, p_2/q_2)$  is equivalent to  $C(p, q, l, u, v)$  for  $(l, u, v)$  in Table 3 and some  $(p, q)$ . In cases (i)–(iv),  $(p, q) = (p_1 p_2 - q_1 q_2, q_1)$ . In cases (v)–(viii),  $(p, q) = (p_1 p_2 - q_1 q_2, q_1 p_2)$ . In cases (ix) and (x),  $p = p_1 p_2 - q_1 q_2$  and  $q \in \{\pm q_0^{\pm 1}\}$ , where  $q_0 = p_1 p'_2 - q_1 q'_2$  is calculated in Lemma 7.10.*

**Proof** First, we make some comments on the parameters  $(p, q)$ . Lemma 7.10 provides a way to specify the ambient lens space of  $K_0$ . Explicitly, the lens space is  $L(p, q)$ , where  $p = p_1 p_2 - q_1 q_2$  and  $q \in \{\pm q_0^{\pm 1}\}$ .

In cases (i)–(iv),  $p_2 = 1$ . Hence we can choose  $q'_2 = -1$  and  $p'_2 = 0$  in Lemma 7.10. Then we can set  $q_0 = p_1 p'_2 - q_1 q'_2 = q_1$ . In cases (v)–(viii),  $|q_1| = 1$ . By Proposition 7.3, we can switch the roles of  $(p_1, q_1)$  and  $(p_2, q_2)$  in Lemma 7.10. So we can pick  $p'_1 = q_1$  and  $q'_1 = 0$  so that  $p_1 p'_1 - q_1 q'_1 = -1$ . Then we can set  $q_0 = p_2 p'_1 - q_2 q'_1 = q_1 p_2$ .

From Remark 3.3 we know that  $C(p, q, l, u, v)$  may be different from  $C(p, q^{-1}, l, u, v)$ . Hence to define a constrained knot, we need to fix the choice of  $q$  in the set  $\{\pm q_0^{\pm 1}\}$ .

case	conditions	$(l - 1, u, v)$
(i)	$p_2 = 1, q_1q_2 < 0$	$(-q_1q_2, U_0, V_0)$
(ii)	$p_2 = 1, q_2 > 1, q_1 > p_1 > 0, U_{-1} \geq U_{-2}$	$(p_1, U_{-1}, V_{-1})$
(ii')	$p_2 = 1, q_2 > 1, q_1 > p_1 > 0, U_{-1} < U_{-2}$	$(q_1q_2 - 2p_1, U_{-2}, V_{-2})$
(iii)	$p_2 = 1, q_2 < -1, -q_1 > p_1 > 0$	$(q_1q_2 - 2p_1, U_*, V_*)$
(iv)	$(p_2, q_2) = (1, 0)$	$(0, U_0, V_0)$
(v)	$p_1 > 1,  q_1  = 1, q_1q_2 < 0$	$(-q_1q_2, U_0, V_0)$
(vi)	$p_1 > 1, q_1 = 1, p_2 > q_2 > 0$	$(p_1p_2 - 2q_2, U_1, V_1)$
(vii)	$p_1 > 1, q_1 = -1, p_2 > -q_2 > 0$	$(-p_2, p_1p_2 + 2q_2, U_\#, V_\#)$
(viii)	$(p_1, q_1) = (0, 1)$	$(0, U_{-1}, V_{-1})$
(ix)	$(p_2, q_2) = (1, 1), q_1 > 0, (p_1, q_1) \neq (1, 1)$	$(\pm q_1, U_n, V_n)$ for $n \in \mathbb{Z}$
(x)	$(p_2, q_2) = (1, -1), q_1 < 0, (p_1, q_1) \neq (1, -1)$	$l - 1 = \pm q_1$

Table 3: Cases where Dehn surgeries on magic links induce constrained knots.

For cases (i)–(viii), the later proof shows  $q = q_0$ . However, for cases (ix)–(x) it is hard to provide a general formula for the choice of  $q$  since the proof is not constructive.

We prove the theorem case by case:

For case (i), we consider two subcases:

- (a)  $p_2 = 1$  and  $q_2 > 0, q_1 < 0$ ,
- (b)  $p_2 = 1$  and  $q_2 < 0, q_1 > 0$ .

The proofs of these two subcases are similar so we only prove case (a).

In case (a),  $|q_2| = q_2$  and  $|q_1| = -q_1$ . Consider curves  $m_1, l_1, m_2$  and  $l_2$  in Figure 10 and the Heegaard diagram  $(\Sigma_2, \{\alpha_1, \alpha_2\}, \beta_1)$  of  $E(K_0)$  in Corollary 7.7. For example, if  $q_2 = 3$ , then  $\alpha_2$  is obtained by resolving intersection points of  $m_2$  and three copies of  $l_2$  positively. Let  $l'_1$  be the curve obtained by sliding  $l_1$  over  $\alpha_2$  along an arc  $a$  around  $z$ ; see Figure 13, top left. Let  $\alpha'_1$  be obtained by taking  $|p_1|$  copies of  $m_1$  and  $|q_1|$  copies of  $l'_1$  and resolving negatively. Then  $(\Sigma_2, \{\alpha'_1, \alpha_2\}, \beta_1)$  is also a Heegaard diagram of  $E(K_0)$  since  $l'_1$  is isotopic to  $l_1$  in the link complement. Consider the genus 1 surface  $\Sigma_1$  obtained from  $\Sigma_2$  by removing the 1–handle attaching to  $z$  and  $w$ . Then  $(\Sigma_1, \alpha'_1, \beta_1, z, w)$  is a doubly pointed Heegaard diagram of  $K_0$ . We can compare this diagram with the standard diagram of a constrained knot.

By construction, there are  $q_2$  strands in  $l'_1$  connecting the left edge to the right edge, where  $(q_2 - 1)$  strands do not intersect the top edge and one strand intersects the top edge.

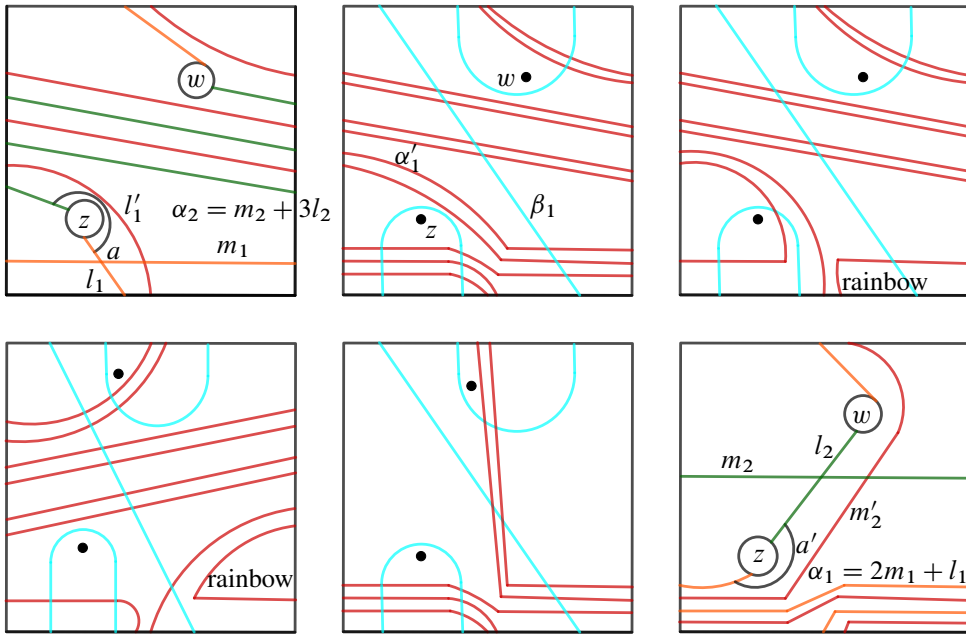


Figure 13: Examples of  $K_0$ .

Since  $m_1$  is a strand connecting the left edge to the right edge, there are  $(p_1 - q_1q_2)$  strands in  $\alpha'_1$  connecting the left edge to the right edge. These strands can be divided into two parts, in which the strands are isotopic to the strands  $a_0$  and  $a_{-1}$  defined before Lemma 7.8, respectively. By counting the number of strands, we have

$$|A_0(u_0, v_0)| = p_1 \quad \text{and} \quad |A_{-1}(u_0, v_0)| = -q_1q_2.$$

Hence  $(\Sigma_1, \alpha'_1, \beta_1, z, w)$  is the same as the standard diagram (see Figure 6) of

$$C(p_1 - q_1q_2, q_1, -q_1q_2 + 1, u_0, v_0).$$

Thus, the two knots are equivalent. For example, Figure 13, top middle, corresponds to  $C(9, -2, 7, 3, 1) = C(9, 7, 7, 3, 1)$ , where

$$(p_1, q_1, p_2, q_2, u, v) = (3, -2, 1, 3, 3, 1).$$

Cases (ii)–(iv) are proven by a similar strategy. Indeed, we can compare  $\alpha'_1$  in the doubly pointed diagram of  $K_0$  with the standard diagram of a constrained knot to obtain the parameters. In particular, the type and the number of strands in  $\alpha'_1$  are important, so we only state the main difference about the curve  $\alpha'_1$ .



For cases (ii) and (ii'), let  $\alpha'_1$  be the curve as defined in case (i). It is the union of strands with endpoints on the left edge and the right edge. By the assumption  $q_1 > p_1$ , we may have rainbows in  $\alpha'_1$ , ie strands whose endpoints are on the same edge. Since the rainbows on the right edge do not bound a basepoint, we can isotopy  $\alpha'_1$  to remove them. After removing  $p_1$  rainbows on the right edge, there are  $(q_1q_2 - 2p_1)$  strands and  $p_1$  strands isotopic to  $a_{-1}$  and  $a_{-2}$ , respectively:

$$|A_{-1}| = q_1q_2 - 2p_1 \quad \text{and} \quad |A_{-2}| = p_1.$$

The choice of case (ii) and case (ii') depends on if  $U_{-1} \geq U_{-2}$  or  $U_{-1} < U_{-2}$ , respectively. This is because the parameter  $u$  is the greater number in  $\{U_{-1}, U_{-2}\}$ .

For case (iii), the pair of sets  $(A_{-1}, A_{-2})$  in the above proof is replaced by  $(A_{-1}, A_*)$ . Counting the number of strands, we have

$$|A_{-1}| = q_1q_2 - 2p_1 \quad \text{and} \quad |A_*| = p_i.$$

By Lemma 7.8, the number  $U_*$  is always greater than  $U_{-1}$ .

For case (iv), all strands are isotopic to  $a_0$ .

Examples can be found in Figure 13. In all examples,  $(u, v) = (3, 1)$ . In the top right subfigure, the diagram of  $K_0$  is in case (ii) with  $(p_1, q_1, p_2, q_2) = (1, 2, 1, 3)$ , which corresponds to  $C(-5, 2, 2, 1, 0) = C(5, 3, 2, 1, 0)$ . In the bottom left subfigure, the diagram of  $K_0$  is in case (iii) with  $(p_1, q_1, p_2, q_2) = (1, -2, 1, -3)$ , which corresponds to  $C(-5, -2, 5, 5, 2) = C(5, 2, 5, 5, 2)$ . In the bottom middle subfigure, the diagram of  $K_0$  is in case (iii) with  $(p_1, q_1, p_2, q_2) = (3, -2, 1, 0)$ , which corresponds to  $C(3, -2, 1, 3, 1) = C(3, 1, 1, 3, 1)$ .

For proofs of cases (v)–(viii), we consider the curve  $m'_2$  obtained by sliding  $m_2$  over  $\alpha_1$  along an arc  $a'$  around  $z$ ; see Figure 13, bottom right, for  $p_1 = 2$ . Now the resulting diagram of  $E(K_0)$  is  $(\Sigma_2, \{\alpha_1, \alpha'_2\}, \beta_1)$ , where  $\alpha'$  is obtained from  $m'_2$  and  $l_2$  by resolution. The proofs are similar to cases (i)–(iv).

For cases (ix) and (x), the diagrams are more complicated. By Proposition 1.1, we can check by the distribution of the spin<sup>c</sup> structures of intersection points that the knot  $K_0$  is a constrained knot. The parameter  $l$  can be obtained by counting the number of strands. □

The following corollary is obtained by changing parameters in Table 3.

case	$l - 1$
(i) and (v) with $q_2 > 0$ and (iv), (viii) and (ix)	$\pm nq$ where $nq \in [0, p)$
(i) and (v) with $q_2 < 0$ and (iv), (viii) and (x)	$\pm n(p - q)$ where $n(p - q) \in [0, p)$
(ii)	$\lceil p/q \rceil q - p$
(ii') and (vi)	$2p - \lceil p/q \rceil q$
(iii) and (vii)	$2p - \lceil p/(p - q) \rceil (p - q)$

Table 4: Choices of the parameter  $l$ .

**Corollary 7.12** Suppose integers  $p$  and  $q$  satisfy  $p > q > 0$  and  $\gcd(p, q) = 1$ . The choices of  $l$  from Theorem 7.11 are in Table 4. Note that Theorem 1.8 follows from the first two rows in Table 4.

**Remark 7.13** For integers  $u_0$  and  $v_0$  satisfying  $(u_0, v_0) = (1, 1)$  or  $0 < -2v_0 < u_0$ , and  $\gcd(u_0, v_0) = 1$ , where  $u_0$  is odd, the surgery description can be induced similarly to Table 3. We omit the explicit description.

We describe some special examples of Table 3 as follows.

Consider integers  $u_0$  and  $v_0$  satisfying  $(u_0, v_0) = (1, 0)$  in Theorem 7.11. We know that the manifold  $E(\mathcal{L}(1, 0))$  is diffeomorphic to  $S^1 \times F$ , where  $F$  is a disk with two holes. For integers  $p_1, p_2, q_1$  and  $q_2$  satisfying  $p_1 p_2 \neq q_1 q_2$ , the knot  $K_0(1, 0, p_1/q_1, p_2/q_2)$  is a torus knot in a lens space.

Cases (iii) and (vii) in Table 3 cover the cases  $(u, v) = (3, \pm 1)$ . By Corollary 7.12, for  $p, q \in \mathbb{Z}$  with  $p > q > 0$ , the knot  $C(p, \pm q, 2p - \lceil p/q \rceil q + 1, 3, \pm 1)$  is a torus knot.

**Theorem 7.14** The knot  $C(p, q, 1, u, v)$  is the connected sum of the 2–bridge knot  $\mathfrak{b}(u, v)$  and the core knot  $C(p, q, 1, 1, 0)$  of  $L(p, q')$ , where  $qq' \equiv 1 \pmod{p}$ .

**Proof** By case (iv) in Theorem 7.11, the knot  $C(p, q, 1, u, v)$  is identified with  $K_0(u, v, p/q, 1/0)$ , which is obtained by the  $p/q$  surgery on the meridian of  $\mathfrak{b}(u, v)$ . By Corollary 3.9, the knot  $C(p, q, 1, 1, 0)$  is the core knot, which is obtained by the  $p/q$  surgery on one component of the Hopf link. □

## 8 1–Bridge braid knots

In this section we describe another approach to construct constrained knots by Dehn surgeries. Many results are based on [15, Section 3]. The main objects in this section are 1–bridge braids, defined below.

**Definition 8.1** A knot in the solid torus  $S^1 \times D^2$  is called a 1-bridge braid if it is isotopic to a union of two arcs  $\gamma \cup \delta$  such that  $\gamma \subset \partial(S^1 \times D^2)$  is braided, meaning it is transverse to each meridian  $\{pt\} \times \partial D^2$ , and  $\delta$  is a bridge, meaning it is properly embedded in some meridional disk  $\{pt\} \times D^2$ .

We denote 1-bridge braids by  $B(w, b, t)$  [13], where  $w > 0$  is the winding number,  $b \in [0, w - 2]$  is the bridge width, and  $t \in [1, w - 1]$  is the twist number. When  $b = 0$ , the 1-bridge braid can be isotoped to lie on  $\partial(S^1 \times D^2)$ . Let  $B(w, w - 1, t)$  denote  $B(w, 0, t + 1)$ .

As mentioned in [15, Section 3], after isotopy, the arc  $\gamma$  can be lifted to a straight line (a geodesic) in the universal cover  $\mathbb{R}^2$  of  $\partial(S^1 \times D^2)$ , which is still denoted by  $\gamma$ . Suppose that  $\gamma$  connects  $(0, 0)$  to  $(t', w)$ , where  $t' \in \mathbb{Q} \cap [t, t + 1)$ . Let  $B(w, s(\gamma))$  denote this 1-bridge braid, where  $s(\gamma) = t'/w$  is called the inverse slope of  $\gamma$ . Suppose  $s = n/d$  with  $\gcd(n, d) = 1$ . Suppose that the integer  $n_i \in [0, d)$  satisfies  $n_i \equiv ni \pmod{d}$ . Then  $b$  is given by the formula

$$b = \#\{i \in [1, w - 1] \mid n_i < n_w\}.$$

**Definition 8.2** Suppose integers  $p$  and  $q$  satisfy  $0 < q < p$  and  $\gcd(p, q) = 1$ . Denote the knot in  $L(p, q)$  obtained by Dehn filling  $(S^1 \times D^2, B(w, s(\gamma)))$  along the curve on  $\partial(S^1 \times D^2)$  with slope  $p/q$  by  $B(w, s(\gamma), p, q)$ . This is called a 1-bridge braid knot.

**Proposition 8.3** For a 1-bridge braid  $B(w, s(\gamma))$ , suppose  $s$  represents the core of the solid torus, and suppose  $t$  represents the meridian of the braid. For  $j \in [1, w - 1]$ , define

$$\theta_j = \begin{cases} 1 & \text{if } n_j < n_w, \\ 0 & \text{if } n_j > n_w. \end{cases}$$

Then the 2-variable Alexander polynomial of  $B(w, s(\gamma))$  is

$$\Delta(s, t) = \sum_{i=0}^{w-1} s^i t^{\sum_{j=1}^i \theta_j}.$$

**Proof** Suppose  $H_2 = S^1 \times D^2 - N(\delta)$ , which is diffeomorphic to a genus two handlebody. Let  $D$  be the canceling disk of  $\delta$ . There are two meridian disks  $\{pt\} \times \partial D^2$  and  $D$  of  $H_2$ . Suppose their boundaries are  $\alpha_1$  and  $\alpha_2$ , respectively, then suppose that  $\beta = \partial N(\gamma)$  and  $\Sigma = \partial H_2$ . Then  $(\Sigma, \{\alpha_1, \alpha_2\}, \beta)$  is a Heegaard diagram of

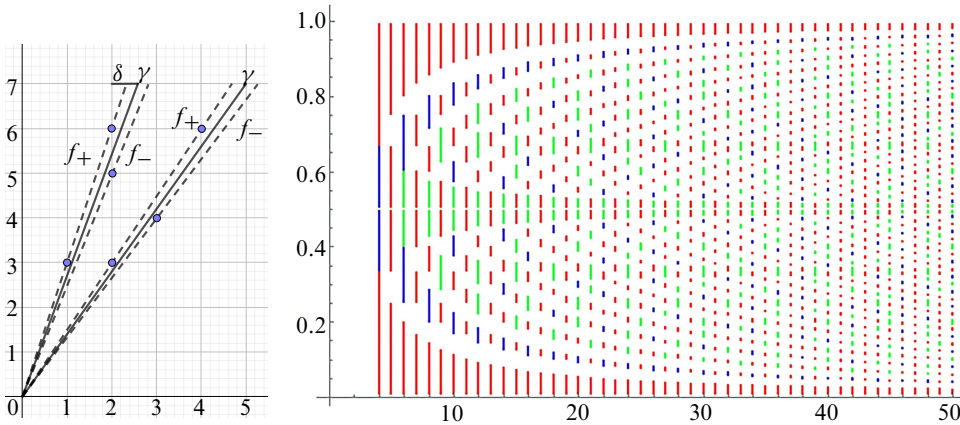


Figure 14: Left: 1-bridge braid in  $\mathbb{R}^2$ . Right: simple intervals.

$S^1 \times D^2 - N(B(w, s(\gamma)))$ . It induces a presentation of the fundamental group by the method in Section 6:

$$\pi_1(S^1 \times D^2 - N(B(w, s(\gamma)))) = \langle s, t \mid \omega t \omega^{-1} t^{-1} = 1 \rangle,$$

where  $\omega = st^{\theta_1} st^{\theta_2} s \dots st^{\theta_{w-1}} s$ . Then we can calculate the Alexander polynomial by Fox calculus [42, Chapter II]. □

Let  $F_n$  be the  $n^{\text{th}}$  Farey sequence, ie the sequence containing all rational numbers  $x/y$  with  $0 \leq x \leq y \leq n$  and  $\gcd(x, y) = 1$ , listed in the increasing order. For example:

$$(10) \quad F_1 = \left(\frac{0}{1}, \frac{1}{1}\right), F_2 = \left(\frac{0}{1}, \frac{1}{2}, \frac{1}{1}\right), F_3 = \left(\frac{0}{1}, \frac{1}{3}, \frac{1}{2}, \frac{2}{3}, \frac{1}{1}\right), F_4 = \left(\frac{0}{1}, \frac{1}{4}, \frac{1}{3}, \frac{1}{2}, \frac{2}{3}, \frac{3}{4}, \frac{1}{1}\right).$$

For a fixed integer  $w$ , suppose  $f_-$  and  $f_+$  are successive terms in  $F_{w-1}$ . For any two 1-bridge braids with inverse slopes  $s_2, s_2 \in (f_-, f_+)$  there is an isotopy between them [15, Section 3]. If  $s(\gamma) \in (f_-, f_+)$ , the interval  $\mathbb{S}(\gamma) = [f_-, f_+]$  is called the *simple interval* of  $\gamma$ . Two examples are shown in Figure 14, left.

For integers  $w$  and  $t$  satisfying  $\gcd(w, t) = 1$ , the 1-bridge braid knot  $B(w, t/w, p, q)$  is the  $(w, t)$  torus knot in  $L(p, q)$  defined in Section 2. Suppose  $f_{\pm} = n_{\pm}/d_{\pm}$ , where  $n_{\pm}$  and  $d_{\pm}$  are integers satisfying  $\gcd(n_{\pm}, d_{\pm}) = 1$ . If  $d_{\pm} \mid w$ , then the 1-bridge braid knot  $B(w, s(\gamma), p, q)$  with  $s(\gamma) \in (f_-, f_+)$  is the  $(1, \mp w/d_{\pm})$  cable knot of the  $(d_{\pm}, nw/d_{\pm})$  torus knot in  $L(p, q)$ , respectively; see [15, Section 3.1]. The braids  $B(w, s(\gamma))$  in the above two cases are called *torus braids* and *cable braids*, respectively. In other cases, the braid  $B(w, s(\gamma))$  is called a *strict braid*.

**Theorem 8.4** *The 1–bridge braid knot  $B(w, s(\gamma), p, q)$  is a simple knot if and only if  $q/p \in \mathbb{S}(\gamma)$ . In this case, it is the simple knot  $S(p, q, wq)$ .*

**Proof** The sufficient part follows from the discussion before [15, Theorem 3.2]. Indeed, the arc  $\gamma$  can be isotoped to have the inverse slope  $q/p$  (if  $q/p = f_{\pm}$ , then let the slope of  $\gamma$  be  $f_{\pm} \mp \epsilon$  for small  $\epsilon > 0$ ). Then the knot is the union of two arcs of slopes 0 and  $q/p$ , respectively. Then it is straightforward to check that the knot is a simple knot. Note that the knot is homologous to  $wq$  of the core of the filling solid torus. Thus, the knot is  $S(p, q, wq)$ .

The necessary part for a strict braid is shown by [15, Theorem 3.2]. When  $B(w, s(\gamma))$  is not strict, the proof of [15, Theorem 3.2] still applies because  $d_{\pm} < w$ .  $\square$

Let us consider special cases of simple knots obtained from Theorem 8.4. Consider examples of Farey sequences in (10). For  $w \leq 3$ , all simple knots are from torus braids. For  $w \leq 4$ , all simple knots are from either torus braids or cable braids. For  $w \geq 4$ , the union of the simple intervals for torus braids and cable braids are shown in Figure 14, right, where red arcs represent torus braids (they are Berge–Gabai knots of type I; see [12; 3; 2]), blue arcs represent  $(1, \pm 2)$  cable braids (they are Berge–Gabai knots of type II), and green arcs represent other cable braids.

**Proof of Theorem 1.9** By Theorem 8.4, simple knots are 1–bridge braid knots. For constrained knots that are not simple knots, we show  $C(p, q, l, u, 1)$  is equivalent to a 1–bridge braid knot. The case  $C(p, q, l, u, -1) = C(p, q, l, u, u - 1)$  is the mirror image of  $C(p, -q, l, u, 1)$  by Proposition 3.4 so is also equivalent to a 1–bridge braid knot.

The proof is inspired by Figure 15, left. Suppose  $(T^2, \alpha_1, \beta_1, z, w)$  is the standard diagram of  $C(p, q, l, u, 1)$ . By definition, the constrained knot is the union of two arcs  $a$  and  $b$  connecting  $z$  to  $w$  in  $T^2 - \alpha_1$  and  $T^2 - \beta_1$ , pushed slightly into the  $\alpha_1$ –handlebody and the  $\beta_1$ –handlebody, respectively. The arc  $a$  can be chosen as a horizontal one, and there are infinitely many choices of isotopy classes of  $b$  on  $T^2$ . Let  $\gamma_i$  denote different choices of  $b$  for  $i \in \mathbb{Z}$ . All choices induce equivalent knots because they are isotopic in the  $\beta_1$ –handlebody.

Since there is only one rainbow for  $\beta_1$ , the arc  $\gamma_i$  does not have any rainbows. For a large integer  $i$ , the arc  $\gamma_i$  can be isotoped to a straight line. Then  $\gamma_i$  is transverse to each meridian disk of the  $\alpha_1$ –handlebody and the union of  $a$  and  $\gamma_i$  is a 1–bridge braid in the  $\alpha_1$ –handlebody. Hence  $C(p, q, l, u, 1)$  is equivalent to a 1–bridge braid knot.  $\square$

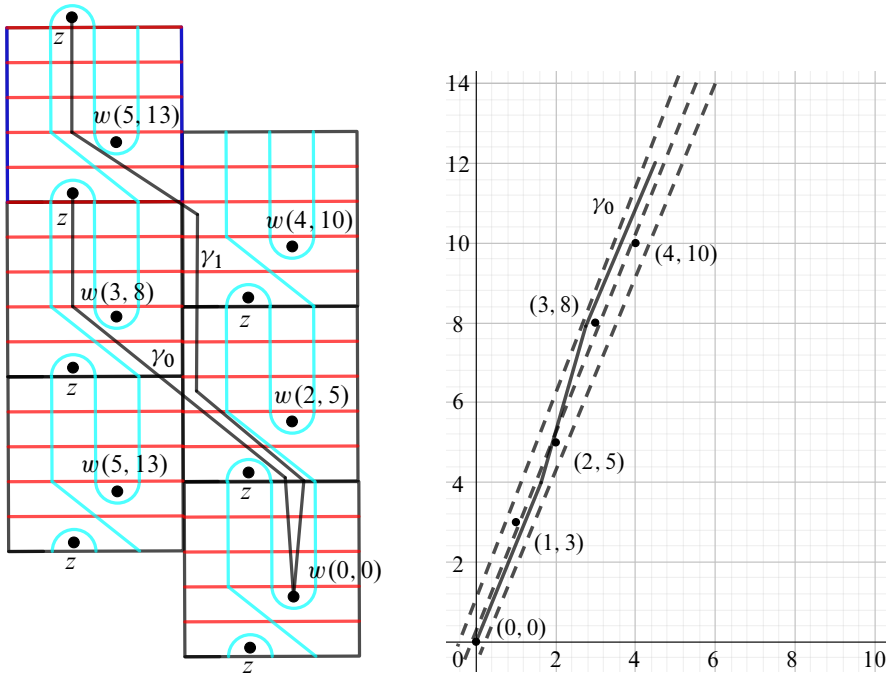


Figure 15: Arcs  $\gamma_i$  for  $C(5, 3, 2, 3, 1)$  (left) and  $\gamma_0$  for  $C(5, 3, 2, 3, 1)$  in  $\mathbb{R}^2$  (right).

It is possible to find the explicit formula of  $B(w(\gamma_i), s)$  in Theorem 1.9 as follows:

Suppose lifts of  $w$  in the universal cover  $\mathbb{R}^2$  of  $T^2$  are lattice points of  $\mathbb{Z}^2$  as in Figure 15, right. Then domains in Figure 15, left, lie in the narrow bands with dotted boundaries in Figure 15, right. From the parametrization of the constrained knot, we know  $C(p, q, l, u, 1)$  is in  $L(p, q')$ , where  $qq' \equiv 1 \pmod{p}$ . Then the slope of the dotted boundaries is  $p/q'$ . Indeed, these boundaries are  $\beta_0$  in the standard diagram  $(T^2, \alpha_0, \beta_0)$  of  $L(p, q')$ .

Suppose

$$\lambda = \frac{qq' - 1}{p} \quad \text{and} \quad \epsilon = \begin{cases} 0 & \text{if } l + q \leq p, \\ 1 & \text{if } l + q > p. \end{cases}$$

Suppose  $\gamma_0$  is the first arc that can be straightened in the lift of  $T^2 - \beta_1$ . Suppose  $D_j$  for  $j \in \mathbb{Z}/p\mathbb{Z}$  are regions in the new diagram  $C$  mentioned in Section 3. The part of  $\gamma_i$  that lies in  $\bigcup_{i=l+1}^p D_i$  and the disk bounded by the unique rainbow of  $\beta_1$  around a basepoint is called the part *in the generalized rainbow*. There are two parts of  $\gamma_i$  in generalized rainbows related to  $z$  and  $w$ .

The parameter  $w(\gamma_i)$  is the same as  $|\gamma_i \cap \alpha_1|$ . Thus

$$w(\gamma_i) = p(u - 3) + 2(p - l + 1) + (q + l - 1 - p\epsilon) + pi = p(u - 1 - \epsilon + i) + q - l + 1,$$

where  $p(u - 3) + 2(p - l + 1)$  is from parts of  $\gamma_i$  in the two generalized rainbows and  $(q + l - 1 - p\epsilon) + pi$  is from the remaining part. Any lift of  $w$  in the left band in Figure 15 has the coordinates

$$(\lambda + nq', q + np) \quad \text{for some } n \in \mathbb{Z}.$$

The closest lift of  $w$  near  $\gamma_i$  other than  $(0, 0)$  has the coordinates

$$(\lambda + n_0q', q + n_0p) \quad \text{where } n_0 = (u - 1)/2 - \epsilon + i.$$

It lies at a lift of the region  $D_l$  that intersects the endpoint of the part of  $\gamma_i$  in the generalized rainbow related to  $z$ . Thus, the inverse slope of  $\gamma_i$  is

$$\frac{\lambda + n_0q'}{q + n_0p} - r$$

for a small rational number  $r$ .

In practice, for given parameters  $(p, q, l, u, 1)$ , it is possible to determine if a constrained knot  $C(p, q, l, u, 1)$  is from a torus braid or a cable braid. For example, consider  $(l, u, v) = (2p - \lceil p/q \rceil q + 1, 3, 1)$  and  $i = 0$ . Then

$$\epsilon = 1, \quad n_0 = 0 \quad \text{and} \quad \omega = \left(1 + \left\lceil \frac{p}{q} \right\rceil\right)q - p.$$

The inverse slope is  $\lambda/q - r$ . Suppose  $x = (1 + \lceil p/q \rceil)\lambda - q'$ . Since

$$\frac{\lambda}{q} = \frac{x + q'}{w + p},$$

the rational number  $x/w$  is in the simple interval  $\mathbb{S}(\gamma_0)$ , ie  $\gamma_0$  is isotopic to the arc with inverse slope  $x/w$ . Thus  $C(p, q, 2p - \lceil p/q \rceil q + 1, 3, 1)$  is a torus knot. This is consistent with the example from the magic link  $\mathfrak{L}(1, 0)$  mentioned in Section 7.

## 9 SnapPy manifolds

A compact orientable manifold  $M$  with torus boundary is called a (*hyperbolic*) 1-cusped manifold if the interior of  $M$  admits a hyperbolic metric of finite volume. All 1-cusped manifolds that have ideal triangulations with at most nine ideal tetrahedra are included in SnapPy [7]. They are called *SnapPy manifolds*. In this section we explain

the strategy used to study the relation between 1-cusped manifolds and constrained knots by computer program. Code and results can be found in [45].

Suppose  $M$  is a 1-cusped manifold and suppose  $\gamma$  is a simple closed curve on  $\partial M$ . The pair  $(M, \gamma)$  is called an *exceptional filling* if Dehn filling along  $\gamma$  gives a nonhyperbolic manifold  $M(\gamma)$ . For such  $(M, \gamma)$ , the core of the filling solid torus induces a knot in  $M(\gamma)$ . The induced knot  $K(M, \gamma)$  is called a *SnapPy knot* if  $M$  is a SnapPy manifold. Dunfield provided a census of exceptional fillings for SnapPy manifolds [9]. In this census, there are 44 487 exceptional fillings  $(M, \gamma)$ , covering 38 056 different SnapPy manifolds, for which  $M(\gamma)$  is a lens space.

Suppose  $M(\gamma) = L(p, q)$  and  $m$  is the meridian of  $K = K(M, \gamma)$ . If  $H_1(M; \mathbb{Z}) \cong \mathbb{Z}$  and it is generated by  $t$ , then  $\tau(K) = \Delta_K(t)/(1-t)$  [42]. The Alexander polynomial only depends on  $M$  and can be found in *SnapPy*. The Euler characteristic  $\chi(\widehat{HFK}(M(\gamma), K))$  can be calculated by Lemma 5.7. Suppose it is  $\sum a_i t^i$ . Since

$$H_1(M; \mathbb{Z})/([m]) \cong H_1(M(\gamma); \mathbb{Z}) \cong \mathbb{Z}/p\mathbb{Z},$$

we know  $[m] = t^p$ . Then  $\chi(\widehat{HFK}(M(\gamma), K))$  can be decomposed into  $p$  polynomials

$$\sum_{i \equiv i_0 \pmod{p}} a_i t^i \text{ for } i_0 \in [0, p).$$

Suppose

$$F_{i_0}(t) = \sum_{i \equiv i_0 \pmod{p}} a_i t^{(i-i_0)/p}$$

and let  $f_i(t)$  be images of  $F_i(t)$  in  $\mathbb{Z}[t]/\pm(t)$ . The exceptional filling  $(M, \gamma)$  has  $n$  *form(s)* if the set  $\{f_i(t) \mid i \in [0, p)\}$  has  $n$  elements.

If  $F_i(t)$  is a monomial for any  $i$ , then  $(M, \gamma)$  has 1 form. By Theorem 1.6, the Euler characteristic must be the same as the simple knot in the same homology class. Such an  $(M, \gamma)$  is called a *simple filling*. It does not necessarily induce a simple knot since Conjecture 1.7 has not been proven yet.

For  $l = 1$ , the constrained knot  $C(p, q, l, u, v)$  is not hyperbolic since it is satellite by Theorem 7.14. If  $F_i(t)$  is symmetric, coefficients of  $F_i(t)$  are alternating for any  $i$ , and  $(M, \gamma)$  has 2 forms, then  $K$  might be a constrained knot  $C(p, q', l, u, v)$ , where  $l > 1$ ,  $u > 1$  and  $q' \equiv \pm q^{\pm 1} \pmod{p}$ . As in the proof of the necessary part of Theorem 1.2, a tuple of *virtual parameters*  $(l, u, v)$  can be calculated by  $F_i(-1)$ . Conversely, given  $(p, q', l, u, v)$ , the characteristic of the corresponding constrained knot is given by Theorem 4.6. If  $\chi(\widehat{HFK}(K))$  is equivalent to  $\chi(\widehat{HFK}(C(p, q', l, u, v)))$  as elements



in  $\mathbb{Z}[H_1(M; \mathbb{Z})]$  for virtual parameters  $(l, u, v)$ , then  $(M, \gamma)$  is called a *constrained filling*. If symmetrized Alexander polynomials of  $K$  and  $C(p, q', l, u, v)$  are the same, then  $(M, \gamma)$  is called a *general constrained filling*. If  $H_1(M; \mathbb{Z}) \cong \mathbb{Z}$ , then  $(M, \gamma)$  is a constrained filling if and only if it is a general constrained filling.

If  $\text{Tors } H_1(M; \mathbb{Z})$  is nontrivial, then the Turaev torsion  $\tau(M)$  can be calculated by a presentation of  $\pi_1(M)$ . *SnapPy* provides a presentation of  $\pi_1(M)$  and the related words of the preferred meridian and the preferred longitude (they are not necessarily the same as the meridian and the longitude mentioned in Section 2). By the filling slope from Dunfield's census, the homology class  $[m] \in H_1(M; \mathbb{Z})$  is obtained. The algorithm described above also works and definitions also apply to this case.

The code in [45] constructs complements of constrained knots in *SnapPy* by functions in the *Twister* package. Then the function *M.identify()* in *SnapPy* tells us if the manifold with a constrained filling is indeed the complement of a constrained knot. Mirror manifolds are not distinguished here.

In Dunfield's census, there are 16 355 simple fillings and 8537 constrained fillings, covering 15 262 and 8508 *SnapPy* manifolds, respectively. All 15 262 and 8421 of 8508 *SnapPy* manifolds are complements of simple knots and constrained knots, respectively. There are 1838 manifolds that are both complements of simple knots and constrained knots with  $u > 1$ . Thus, there are 21 845 *SnapPy* manifolds that are complements of constrained knots in lens spaces. Other than these manifolds, there are 77 *SnapPy* manifolds that are complements of 2-bridge knots, which are special cases of constrained knots.

The choice of the slope in a constrained filling is subtle. For example, suppose  $M = m003$ , and  $\gamma_1 = (-1, 1)$  and  $\gamma_2 = (0, 1)$  in the basis from *SnapPy*. Then both  $M(\gamma_1)$  and  $M(\gamma_2)$  are diffeomorphic to  $L(5, 4)$  and  $M$  is the complement of  $C(5, 4, 5, 3, 1)$ . Indeed, there is an isometry of  $M$  sending  $\gamma_1$  to  $\gamma_2$ . Both  $M(\gamma_1)$  and  $M(\gamma_2)$  induce the same knot,  $C(5, 4, 5, 3, 1)$ . All nine pairs of slopes in Dunfield's census with this subtlety are from isometries, except the case where  $M = m172$ ,  $\gamma_1 = (0, 1)$  and  $\gamma_2 = (1, 1)$ . Manifolds  $M(\gamma_1)$  and  $M(\gamma_2)$  are oppositely oriented copies of the same lens space. The first slope induces  $S(49, 18, 7)$  and the second induces  $S(49, 18, 21)$  (up to mirror image), which are not equivalent. This example is interesting in the study of cosmetic surgery [4]. To summarize: the *SnapPy* knots induced by  $15\,262 + 8421 = 23\,683$  constrained fillings in the above discussion are all constrained knots.

There are 87 SnapPy manifolds with constrained fillings that are not complements of constrained knots. For such a manifold either the constrained knot with corresponding virtual parameters is not hyperbolic, or there is another SnapPy manifold which is the complement of the constrained knot with the same parameters. For example, the manifold  $m390$  has a constrained filling  $(1, 0)$  with virtual parameters  $(7, 4, 7, 5, 2)$ , while  $E(C(7, 4, 7, 5, 2))$  is diffeomorphic to  $s090$ .

If  $\text{Tors } H_1(M; \mathbb{Z})$  is nontrivial, then there are 54 general constrained fillings that are not constrained fillings. For example, manifolds  $M_1 = m400$  and  $M_2 = m141$  satisfy  $|\text{Tors } H_1(M_i; \mathbb{Z})| = 2$  and  $\Delta_{M_i}(t) = t^5 - t^4 + t^2 + t^{-2} - t^{-4} + t^{-5}$  for  $i = 1, 2$ , and  $M_1(1,1) \cong M_2(-1, 1) \cong L(18, 13)$ . Both manifolds have general constrained fillings, and  $M_2 \cong E(C(18, 3, 18, 3, 1))$ . Calculation shows  $(M_1, (1, 1))$  is not a constrained filling, ie the Euler characteristic of the induced knot is different from that of  $C(18, 3, 18, 3, 1)$ .

For the exceptional manifolds in Proposition 1.10, manifolds  $m206$  and  $m370$  have exceptional fillings with 2 forms and have virtual parameters  $(l, u, v) = (5, 5, 2)$  and  $(8, 5, 2)$ , respectively. Unfortunately, both exceptional fillings are not even general constrained fillings. The manifold  $m390$  is discussed above. For other 5-manifolds there is no lens space filling (or even  $S^1 \times S^2$  filling). It is harder to obtain information in Heegaard Floer theory.

In the rest of this section we discuss the ways to obtain the genus and the fiberness of a knot. The genera and fiberness of Snappy knots can also be found in [45].

**Definition 9.1** Suppose  $K$  is a knot in  $Y = L(p, q)$  and suppose

$$H_1(E(K); \mathbb{Z}) \cong \mathbb{Z} \oplus \mathbb{Z}/d\mathbb{Z} \cong \langle t, r \rangle \langle dr \rangle.$$

By the excision theorem, Poincaré duality and the universal coefficient theorem,

$$H_2(Y, K; \mathbb{Z}) \cong H_2(E(K), \partial E(K); \mathbb{Z}) \cong H^1(E(K); \mathbb{Z}) \cong \text{Hom}(H_1(E(K); \mathbb{Z}), \mathbb{Z}) \cong \mathbb{Z}.$$

Suppose  $S$  is a connected, oriented and properly embedded surface representing the generator of  $H_2(E(K), \partial E(K); \mathbb{Z})$ . It is called a *Seifert surface* of  $K$ . Let the *genus*  $g(K)$  and the *Thurston norm*  $x([S])$  be the minimum values of  $g(S)$  and  $-\chi(S)$  among all Seifert surfaces, respectively.

**Definition 9.2** For a homogeneous element  $x$  of  $\widehat{HFK}(Y, K)$ , suppose

$$\text{gr}(x) = at + br \in H_1(E(K); \mathbb{Z}).$$

Let  $\text{gr}_0(x)$  be the number  $a$ . The width of  $\widehat{HFK}(Y, K)$  is the maximum value of  $|\text{gr}_0(x) - \text{gr}_0(y)|$  among all pairs of homogeneous elements  $(x, y)$ . Suppose homogeneous elements  $x_0$  and  $y_0$  satisfy

$$\text{width } \widehat{HFK}(Y, K) = |\text{gr}_0(x_0) - \text{gr}_0(y_0)|.$$

Suppose  $H(x_0)$  is the subgroup of  $\widehat{HFK}(Y, K)$  generated by homogeneous elements  $x$  satisfying  $\text{gr}_0(x) = \text{gr}_0(x_0)$ . The top rank of  $\widehat{HFK}(Y, K)$  is  $\dim_{\mathbb{Q}} H(x_0) \otimes \mathbb{Q}$ .

**Theorem 9.3** [21; 27] Consider  $Y, K$  and  $S$  as in Definition 9.1 such that  $E(K)$  is irreducible. Suppose  $m$  is the meridian of  $K$ . Then the width of  $\widehat{HFK}(Y, K)$  equals  $x([S]) + |[m] \cdot [\partial S]|$ , where  $[m] \cdot [\partial S]$  is the algebraic intersection number on  $\partial E(K)$ .

**Proposition 9.4** Let  $Y, K$  and  $S$  be as in Definition 9.1. Suppose  $E(K)$  is irreducible. Suppose  $(m, l)$  is the regular basis of  $K$ . Let  $n$  be the minimum number of boundary components of a Seifert surface. Then  $|[m] \cdot [\partial S]| = p/d$  and  $n = \text{gcd}(d, p/d)$ . Thus,

$$x([S]) = \text{width}(\widehat{HFK}(Y, K)) - \frac{p}{d} \quad \text{and} \quad g(K) = 1 + \frac{1}{2} \left( x([S]) - \frac{p}{d} \right).$$

**Proof** Suppose  $[K] = k[b]$ , where  $[b]$  is a generator of  $H_1(Y; \mathbb{Z})$ . Since  $d = \text{gcd}(p, k)$ , the order of  $[K]$  in  $H_1(Y; \mathbb{Z})$  is  $p/d$ . By Poincaré duality and the universal coefficient theorem, we have

$$H_2(E(K); \mathbb{Z}) \cong H^1(E(K), \partial E(K); \mathbb{Z}) \cong \text{Hom}(H_1(E(K), \partial E(K)), \mathbb{Z}) = 0.$$

By the long exact sequence from  $(E(K), \partial E(K))$ , the boundary map

$$\partial_*: H_2(E(K), \partial E(K); \mathbb{Z}) \rightarrow H_1(\partial E(K); \mathbb{Z})$$

is injective and the image of  $\partial_*$  is the same as the kernel of the map

$$i_*: H_1(\partial E(K); \mathbb{Z}) \rightarrow H_1(E(K); \mathbb{Z}).$$

Since  $H_1(E(K); \mathbb{Z})/[m] \cong H_1(Y; \mathbb{Z})$ , we have  $[\partial S] = \pm(x[m] + p/d[l])$  for some  $x \in \mathbb{Z}$ . Then  $|[m] \cdot [\partial S]| = p/d$  and  $n = \text{gcd}(x, p/d)$ .

Let  $[m]$  and  $[l]$  also denote their images in  $H_1(E(K); \mathbb{Z})$ . By Lemma 5.2, we have

$$[m] = \pm \left( \frac{p}{d} \right) t + ar \quad \text{and} \quad \text{gcd} \left( \frac{p}{d}, d, a \right) = 1.$$

Suppose  $[l] = yt + zr$  for some  $y, z \in \mathbb{Z}$ . Since  $[\partial S] \in \text{Ker}(i_*)$ , we know  $xa + (p/d)z$  is divisible by  $d$ . Suppose  $n_0 = \text{gcd}(d, p/d)$ . Then  $\text{gcd}(n_0, a) = 1$  and  $n_0 \mid xa + (p/d)z$ . Thus  $n_0 \mid x$  and  $n_0 \mid n$ . Suppose  $l^*$  is the homological longitude. Then  $n[l^*] = [\partial S]$  and the image of  $[l^*]$  in  $H_1(E(K); \mathbb{Z})$  is  $wr$  for some  $w \in \mathbb{Z}$ . Thus  $n \mid d$  and  $n \mid n_0$ . This induces  $n = n_0$ .  $\square$

**Theorem 9.5** [21; 26] *Consider  $Y, K$  and  $S$  as in Definition 9.1 such that  $E(K)$  is irreducible. If the top rank of  $\widehat{HF}K(Y, K)$  is 1, then  $K$  is fibered with the fiber  $S$ .*

**Proof** Suppose  $Y(S)$  is the balanced sutured manifold  $(N, \nu)$ , where  $N = Y - \text{Int}(S \times I)$  and  $\nu = \partial S \times I$ . Lemma 3.9 and the proof of Theorem 1.5 in [21] imply that the rank of  $SFH(Y(S))$  is the same as the top rank of  $\widehat{HF}K(Y, K)$ . Then  $Y(S)$  is a product sutured manifold by [21, Theorem 9.7], which implies  $K$  is fibered with fiber  $S$ .  $\square$

## References

- [1] **K L Baker, J E Grigsby, M Hedden**, *Grid diagrams for lens spaces and combinatorial knot Floer homology*, Int. Math. Res. Not. 2008 (2008) art. id. rnm024 MR Zbl
- [2] **J Berge**, *Some knots with surgeries yielding lens spaces*, unpublished manuscript (1990) arXiv 1802.09722
- [3] **J Berge**, *The knots in  $D^2 \times S^1$  which have nontrivial Dehn surgeries that yield  $D^2 \times S^1$* , Topology Appl. 38 (1991) 1–19 MR Zbl
- [4] **S A Bleiler, C D Hodgson, J R Weeks**, *Cosmetic surgery on knots*, from “Proceedings of the Kirbyfest” (J Hass, M Scharlemann, editors), Geom. Topol. Monogr. 2, Geom. Topol. Publ., (1999) 23–34 MR Zbl
- [5] **E J Brody**, *The topological classification of the lens spaces*, Ann. of Math. 71 (1960) 163–184 MR Zbl
- [6] **G Burde, H Zieschang**, *Knots*, 2nd edition, De Gruyter Studies in Mathematics 5, de Gruyter, Berlin (2003) MR Zbl
- [7] **M Culler, N M Dunfield, J R Weeks**, *SnapPy, a computer program for studying the topology of 3-manifolds* Available at <http://snappy.computop.org>
- [8] **H Doll**, *A generalized bridge number for links in 3-manifolds*, Math. Ann. 294 (1992) 701–717 MR Zbl
- [9] **N M Dunfield**, *A census of exceptional Dehn fillings*, from “Characters in low-dimensional topology” (O Collin, S Friedl, C Gordon, S Tillmann, L Watson, editors), Contemp. Math. 760, Amer. Math. Soc., Providence, RI (2020) 143–155 MR Zbl
- [10] **N M Dunfield**, *Floer homology, group orderability, and taut foliations of hyperbolic 3-manifolds*, Geom. Topol. 24 (2020) 2075–2125 MR Zbl

- [11] **S Friedl, A Juhász, J Rasmussen**, *The decategorification of sutured Floer homology*, J. Topol. 4 (2011) 431–478 MR Zbl
- [12] **D Gabai**, *Surgery on knots in solid tori*, Topology 28 (1989) 1–6 MR Zbl
- [13] **D Gabai**, *1–bridge braids in solid tori*, Topology Appl. 37 (1990) 221–235 MR Zbl
- [14] **H Goda, H Matsuda, T Morifuji**, *Knot Floer homology of  $(1, 1)$ -knots*, Geom. Dedicata 112 (2005) 197–214 MR Zbl
- [15] **J E Greene, S Lewallen, F Vafaee**,  *$(1, 1)$   $L$ -space knots*, Compos. Math. 154 (2018) 918–933 MR Zbl
- [16] **J Hanselman, J Rasmussen, L Watson**, *Bordered Floer homology for manifolds with torus boundary via immersed curves*, preprint (2016) arXiv 1604.03466
- [17] **J Hanselman, J Rasmussen, L Watson**, *Heegaard Floer homology for manifolds with torus boundary: properties and examples*, Proc. Lond. Math. Soc. 125 (2022) 879–967 MR
- [18] **C Hayashi**, *Satellite knots in 1–genus 1–bridge positions*, Osaka J. Math. 36 (1999) 711–729 MR Zbl
- [19] **M Hedden**, *On Floer homology and the Berge conjecture on knots admitting lens space surgeries*, Trans. Amer. Math. Soc. 363 (2011) 949–968 MR Zbl
- [20] **A Juhász**, *Holomorphic discs and sutured manifolds*, Algebr. Geom. Topol. 6 (2006) 1429–1457 MR Zbl
- [21] **A Juhász**, *Floer homology and surface decompositions*, Geom. Topol. 12 (2008) 299–350 MR Zbl
- [22] **R Lipshitz, P Ozsvath, D Thurston**, *Bordered Heegaard Floer homology: invariance and pairing*, preprint (2008) arXiv 0810.0687
- [23] **C Manolescu, P Ozsváth**, *On the Khovanov and knot Floer homologies of quasi-alternating links*, from “Proceedings of Gökova Geometry–Topology Conference 2007”, GGT, Gökova (2008) 60–81 MR Zbl
- [24] **B Martelli, C Petronio**, *Dehn filling of the “magic” 3–manifold*, Comm. Anal. Geom. 14 (2006) 969–1026 MR Zbl
- [25] **K Murasugi**, *Knot theory and its applications*, Birkhäuser, Boston, MA (1996) MR Zbl
- [26] **Y Ni**, *Knot Floer homology detects fibred knots*, Invent. Math. 170 (2007) 577–608 MR Zbl
- [27] **Y Ni**, *Link Floer homology detects the Thurston norm*, Geom. Topol. 13 (2009) 2991–3019 MR Zbl
- [28] **P Ozsváth, Z Szabó**, *Heegaard Floer homology and alternating knots*, Geom. Topol. 7 (2003) 225–254 MR Zbl
- [29] **P Ozsváth, Z Szabó**, *Holomorphic disks and knot invariants*, Adv. Math. 186 (2004) 58–116 MR Zbl

- [30] **P Ozsváth, Z Szabó**, *Holomorphic disks and three-manifold invariants: properties and applications*, Ann. of Math. 159 (2004) 1159–1245 MR Zbl
- [31] **P Ozsváth, Z Szabó**, *Holomorphic disks and topological invariants for closed three-manifolds*, Ann. of Math. 159 (2004) 1027–1158 MR Zbl
- [32] **P Ozsváth, Z Szabó**, *Knot Floer homology, genus bounds, and mutation*, Topology Appl. 141 (2004) 59–85 MR Zbl
- [33] **P S Ozsváth, Z Szabó**, *Knot Floer homology and rational surgeries*, Algebr. Geom. Topol. 11 (2011) 1–68 MR Zbl
- [34] **I Petkova**, *Cables of thin knots and bordered Heegaard Floer homology*, Quantum Topol. 4 (2013) 377–409 MR Zbl
- [35] **J A Rasmussen**, *Floer homology of surgeries on two-bridge knots*, Algebr. Geom. Topol. 2 (2002) 757–789 MR Zbl
- [36] **J A Rasmussen**, *Floer homology and knot complements*, PhD thesis, Harvard University (2003) MR arXiv math/0306378
- [37] **J Rasmussen**, *Knot polynomials and knot homologies*, from “Geometry and topology of manifolds” (HU Boden, I Hambleton, A J Nicas, B D Park, editors), Fields Inst. Commun. 47, Amer. Math. Soc., Providence, RI (2005) 261–280 MR Zbl
- [38] **J Rasmussen**, *Lens space surgeries and L-space homology spheres*, preprint (2007) arXiv 0710.2531
- [39] **J Rasmussen, S D Rasmussen**, *Floer simple manifolds and L-space intervals*, Adv. Math. 322 (2017) 738–805 MR Zbl
- [40] *SageMath, version 9.0* Available at <https://www.sagemath.org>
- [41] **H Schubert**, *Knoten mit zwei Brücken*, Math. Z. 65 (1956) 133–170 MR Zbl
- [42] **V Turaev**, *Torsions of 3-dimensional manifolds*, Progr. Math. 208, Birkhäuser, Basel (2002) MR Zbl
- [43] **F Waldhausen**, *On irreducible 3-manifolds which are sufficiently large*, Ann. of Math. 87 (1968) 56–88 MR Zbl
- [44] **M J Williams**, *On nonsimple knots in lens spaces with tunnel number one*, preprint (2009) arXiv 0908.1765
- [45] **F Ye**, data and code to accompany this paper, Harvard Dataverse (2020) Available at <https://doi.org/10.7910/DVN/GLFLHI>

Department of Pure Mathematics and Mathematical Statistics, University of Cambridge  
Cambridge, United Kingdom

[fy260@cam.ac.uk](mailto:fy260@cam.ac.uk)

Received: 8 July 2020      Revised: 1 July 2021

## Convexity in hierarchically hyperbolic spaces

JACOB RUSSELL  
DAVIDE SPRIANO  
HUNG CONG TRAN

Hierarchically hyperbolic spaces (HHSs) are a large class of spaces that provide a unified framework for studying the mapping class group, right-angled Artin and Coxeter groups, and many 3–manifold groups. We investigate strongly quasiconvex subsets in this class and characterize them in terms of their contracting properties, relative divergence, the coarse median structure, and the hierarchical structure itself. Along the way, we obtain new tools to study HHSs, including two new equivalent definitions of hierarchical quasiconvexity and a version of the bounded geodesic image property for strongly quasiconvex subsets. Utilizing our characterization, we prove that the hyperbolically embedded subgroups of hierarchically hyperbolic groups are precisely those that are almost malnormal and strongly quasiconvex, producing a new result in the case of the mapping class group. We also apply our characterization to study strongly quasiconvex subsets in several specific examples of HHSs. We show that while many commonly studied HHSs have the property that every strongly quasiconvex subset is either hyperbolic or coarsely covers the entire space, right-angled Coxeter groups exhibit a wide variety of strongly quasiconvex subsets.

20F65, 20F67

1. Introduction	1168
2. Coarse geometry	1177
3. Divergence of contracting subsets	1181
4. Hierarchically hyperbolic spaces	1186
5. Constructing hulls with hierarchy paths	1199
6. Characterization of strongly quasiconvex subsets in HHSs	1207
7. Strongly quasiconvex subsets in familiar examples	1227
8. Hyperbolically embedded subgroups of HHGs	1238
Appendix. Subsets with arbitrary reasonable lower relative divergence	1243
References	1245

## 1 Introduction

From Gromov's original work on hyperbolic groups to the resolution of the virtual Haken conjecture, quasiconvex subsets have played a central role in the study of hyperbolic metric spaces and groups; see Agol [2], Gromov [31; 32] and Wise [55]. A subset  $Y$  is *quasiconvex* if every geodesic based on  $Y$  is contained in a fixed neighborhood of  $Y$ . A central feature of quasiconvex subsets of hyperbolic spaces is their quasi-isometry invariance, ie the image of a quasiconvex subset of a hyperbolic space under a quasi-isometry is quasiconvex.

Outside of hyperbolic spaces, quasiconvexity fails to be a quasi-isometry invariant. However, a strengthening of this definition to require “quasiconvexity with respect to quasigeodesics” and not just geodesics is sufficient to ensure quasi-isometry invariance. A subset  $Y$  of a quasigeodesic metric space  $X$  is *strongly quasiconvex* if every quasigeodesic based in  $Y$  is contained in a bounded neighborhood of  $Y$ , where the radius of the neighborhood is determined by the quasigeodesic constants. Strong quasiconvexity provides a “coarsification” of the classical definition of a convex subset that ensures that the image of a strongly quasiconvex subset under a quasi-isometry will be strongly quasiconvex, regardless of the geometry of the ambient space. Strongly quasiconvex subsets are therefore an avenue to study the geometry of any space up to quasi-isometry.

The study of strongly quasiconvex geodesics in nonhyperbolic spaces (often called *Morse geodesics*) has been a vibrant and fruitful area of research over the last decade; for example, Arzhantseva, Cashen, Gruber and Hume [6], Charney and Sultan [19], Druţu, Mozes and Sapir [22] and Ol'shanskii, Osin and Sapir [44]. Recently, considerable interest has arisen in understanding general strongly quasiconvex subsets in nonhyperbolic spaces.

The third author studied strongly quasiconvex subsets and subgroups in [54] and showed that many important properties of quasiconvex subsets in hyperbolic spaces persist for strongly quasiconvex subsets of any geodesic metric space. These results have found applications in understanding the cell stabilizers of groups acting on CAT(0) cube complexes — see Groves and Manning [33] — and the splittings of groups over codimension 1 subgroups — see Petrosyan [46]. Using the name Morse instead of strongly quasiconvex, Genevois studied strongly quasiconvex subsets of CAT(0) cube complexes in [28] and Kim studied strongly quasiconvex subgroups of the mapping class groups in [38]. Strongly quasiconvex subgroups that are also hyperbolic were introduced by Durham and Taylor as *stable* subgroups [25] and have received considerable



study; for a sampling see Abbott, Behrstock and Durham [1], Antolín, Mj, Sisto and Taylor [3], Aougab, Durham and Taylor [4], Behrstock [7], and Koberda, Mangahas and Taylor [39].

In this paper, we are primarily interested in understanding the strongly quasiconvex subsets of hierarchically hyperbolic spaces (HHSs). Introduced by Behrstock, Hagen and Sisto in [9] and refined in [10], examples of hierarchically hyperbolic spaces include hyperbolic spaces, the mapping class group of a surface, Teichmüller space with either the Weil–Petersson or Teichmüller metrics, many cocompactly cubulated groups, and the fundamental groups of 3–manifolds without Nil or Sol components. Important consequences of hierarchical hyperbolicity include a Masur–Minsky style distance formula [10], a quadratic isoperimetric inequality [10], restrictions on morphisms from higher rank lattices (Haettel [34]), a largest acylindrical action on a hyperbolic space [1], rank-rigidity and Tits alternative theorems (Durham, Hagen and Sisto [24]), control over the top-dimensional quasiflats (Behrstock, Hagen and Sisto [11]), and bounds on the asymptotic dimension (Durham, Hagen and Sisto [8]). The definition and much of the theory of hierarchically hyperbolic spaces is inspired by the Masur–Minsky subsurface projection machinery for the mapping class group. Our investigation is therefore a natural extension of the problem purposed by Farb in [27, Problem 2.3.8] to study convexity in the mapping class group.

Heuristically, a hierarchically hyperbolic space consists of a metric space  $X$  with an associated collection of hyperbolic spaces  $\mathfrak{S}$ , such that for each space  $Z$  in  $\mathfrak{S}$ , there is a projection map  $X \rightarrow Z$ . The philosophy of hierarchically hyperbolic spaces is that one can study the coarse geometry of  $X$  by studying the projection of  $X$  to each of the spaces in  $\mathfrak{S}$ . In this paper, we shall consider hierarchically hyperbolic spaces satisfying the *bounded domain dichotomy*; a minor regularity condition requiring every space in  $\mathfrak{S}$  to have either infinite or uniformly bounded diameter. The bounded domain dichotomy simplifies the statements and proofs of our results while being satisfied by all of the examples of hierarchically hyperbolic spaces given above and more broadly by all hierarchically hyperbolic groups.

**Equivalent conditions to being strongly quasiconvex** The main goal of this paper is to provide several equivalent conditions for a subset of a hierarchically hyperbolic space to be strongly quasiconvex. A major theme is that several different notions of convexity that coincide with being quasiconvex in a hyperbolic space, coincide with being strongly quasiconvex in a hierarchically hyperbolic spaces. One such notion of convexity is

that of contracting subsets. A subset  $Y \subseteq X$  of a quasigeodesic space is *contracting* if there exists a coarsely Lipschitz retraction  $r: X \rightarrow Y$  under which large balls far from  $Y$  have images with uniformly bounded diameter. Being contracting generalizes the strong contracting behavior of the closest point projection onto a convex subset of the hyperbolic plane. In general, strongly quasiconvex subsets are not contracting (see Example 3.8); however these two notions of convexity tend to agree in the presence of nonpositive curvature. Indeed, it is a classical fact that a subset of a hyperbolic space is strongly quasiconvex if and only if it is contracting; the same is true for subsets of a CAT(0) cube complex [19; 28]. The first of our equivalent condition is to extend these results to hierarchically hyperbolic spaces.

**Theorem 1.1** (strongly quasiconvex and contracting are equivalent) *Let  $X$  be a hierarchically hyperbolic space with the bounded domain dichotomy. A subset  $Y \subseteq X$  is strongly quasiconvex if and only if  $Y$  is contracting.*

In [6], a different notion of contracting subset is considered, and it is shown that a subset of a geodesic metric space is strongly quasiconvex if and only if the subset is *sublinearly contracting*. Example 3.8 demonstrates that our definition of contracting (Definition 2.10) is strictly stronger than sublinear contracting, but the two notions agree in the setting of hierarchically hyperbolic spaces. Another key difference between our definition of contracting and that in [6] is that we do not require the contracting map  $r: X \rightarrow Y$  to be the closest point projection, but allow for any coarsely Lipschitz retraction that has the contracting property. This has the advantage of turning contracting into a quasi-isometry invariant directly from the definition and is crucial in allowing us to utilize a naturally occurring retraction map in hierarchically hyperbolic spaces that is far more tractable than the closest point projection.

The third notion of convexity considered is *hierarchical quasiconvexity*, which is specific to hierarchically hyperbolic spaces. Introduced in [10] by Behrstock, Hagen and Sisto, hierarchically quasiconvex subsets have played a central role in the study of hierarchically hyperbolic space [8; 10; 11]. Notably, a hierarchical quasiconvex subset of an HHS is itself an HHS. While hierarchically quasiconvex subsets are not always strongly quasiconvex, we classify precisely when the two concepts agree. Strongly quasiconvex subsets are exactly the hierarchically quasiconvex subsets that satisfy the *orthogonal projection dichotomy* (Definition 6.2), which describes how the projections of a strongly quasiconvex subset to each of the associated hyperbolic spaces must look.

**Theorem 1.2** (strongly quasiconvex subsets are hierarchically hyperbolic) *Let  $X$  be a hierarchically hyperbolic space with the bounded domain dichotomy. A subset  $Y \subseteq X$  is strongly quasiconvex if and only if  $Y$  is hierarchically quasiconvex and has the orthogonal projection dichotomy. In particular, if  $Y \subseteq X$  is strongly quasiconvex, then  $Y$  is hierarchically hyperbolic.*

Theorem 1.2 is truly the central result of this paper as it explains how the strongly quasiconvex subsets interact with the projections defining the hierarchically hyperbolic structure of the ambient space. Further, this characterization is complete as the theorem fails whenever any of the hypotheses are weakened; see Remark 6.14.

In [1], Abbott, Behrstock and Durham give several equivalent conditions for quasigeodesics in a hierarchically hyperbolic space to be strongly quasiconvex and for a map from a quasigeodesic space  $Y$  into a hierarchically hyperbolic space to be a stable embedding; see Proposition 2.8. Theorems 1.1 and 1.2 generalize these results to general strongly quasiconvex subsets and do not require the hypothesis of unbounded products utilized by Abbott, Behrstock and Durham. This generalization to all strongly quasiconvex subsets is essential to our applications in Sections 7 and 8.

Part of the proof of Theorem 1.2 involves studying hierarchically quasiconvex hulls in hierarchically hyperbolic spaces. The *hierarchically quasiconvex hull* of a subset  $Y$  is (coarsely) the smallest hierarchically quasiconvex set containing  $Y$ . We show that the hull of any subset of a hierarchically hyperbolic space can be constructed using special quasigeodesics called hierarchy paths (see Theorem 5.2 for the precise statement).

**Theorem 1.3** (constructing hulls with hierarchy paths) *If  $Y$  is a subset of a hierarchically hyperbolic space  $X$ , then the hierarchically quasiconvex hull of  $Y$  can be constructed in a uniformly finite number of steps by iteratively connecting points by hierarchy paths.*

This construction is reminiscent of the construction of quasiconvex hulls in hyperbolic spaces by connecting pairs of points by geodesics and is similar to the join construction of hulls in coarse median spaces presented by Bowditch in [16]. The main purpose of Theorem 1.3 in this article is to establish that hierarchically quasiconvex subsets are exactly the subsets that are “quasiconvex with respect to hierarchy paths”. However, we expect this construction to have further applications in the study of hierarchically hyperbolic spaces. Indeed, Hagen and Petyt have used this construction to build quasi-isometries from some hierarchically hyperbolic groups to cube complexes [35], and

in Section 5.1 we apply Theorem 1.3 to provide a characterization of hierarchical quasiconvexity in terms of the coarse median structure on a hierarchically hyperbolic space. This later result allows us to conclude that, in the setting of hierarchically hyperbolic spaces, the coarse median hull constructed in [16] is coarsely equal to the hierarchically quasiconvex hull; extending [16, Lemma 7.3] from finite to arbitrary subsets.

Charney and Sultan proved that strongly quasiconvex geodesics in a CAT(0) space are characterized by having at least quadratic *lower divergence* [19]. The third author introduced a generalization of lower divergence to all subsets [53] and studied its relationship with strong quasiconvexity [54]. If  $Y$  is a subset of the quasigeodesic space  $X$ , the *lower relative divergence of  $X$  with respect to  $Y$*  (or the divergence of  $Y$  in  $X$ ) is a family of functions that measures how efficiently one can travel in  $X$  while avoiding  $Y$ . Building on the work in [54], we establish the following.

**Theorem 1.4** (contracting subsets have at least quadratic divergence) *Let  $X$  be a quasigeodesic metric space. If  $Y \subseteq X$  is contracting, then the lower relative divergence of  $X$  with respect to  $Y$  is at least quadratic. Further, if  $X$  is a hierarchically hyperbolic space with the bounded domain dichotomy, then the lower relative divergence of  $X$  with respect to  $Y$  is at least quadratic if and only if  $Y$  is strongly quasiconvex (equivalently if and only if  $Y$  is contracting).*

Since the lower relative divergence of  $X$  with respect to  $Y$  agrees with Charney and Sultan's lower divergence when  $Y$  is a geodesic in  $X$ , Theorem 1.4 proves that strongly quasiconvex geodesics (aka Morse geodesics) in hierarchically hyperbolic spaces with the bounded domain dichotomy are also characterized by having at least quadratic lower divergence.

After proving Theorems 1.1 through 1.4, we establish several HHS analogues of the “bounded geodesic image property” of quasiconvex subsets of hyperbolic spaces. One of these analogues is the following.

**Theorem 1.5** *Let  $Y$  be a strongly quasiconvex subset of a hierarchically hyperbolic space  $X$  with the bounded domain dichotomy. There is a contracting map  $g_Y : X \rightarrow Y$  such that for each  $\lambda \geq 1$  there exists a constant  $r_\lambda > 0$  such that, for all  $x, y \in \mathcal{X}$ , if  $d(g_Y(x), g_Y(y)) > r_\lambda$ , then any  $\lambda$ -hierarchy path from  $x$  to  $y$  must intersect the  $r_\lambda$ -neighborhood of  $Y$ .*

**Strongly quasiconvex subsets in specific examples** After characterizing the strongly quasiconvex subsets of hierarchically hyperbolic spaces, we apply our results to study the strongly quasiconvex subsets of some of the most common examples of hierarchically hyperbolic spaces: the mapping class group, Teichmüller space, right-angled Artin and Coxeter groups, and the fundamental groups of graph manifolds.

It has been shown that strongly quasiconvex subgroups of the mapping class group [38], right-angled Artin groups with connected defining graph [28; 54], and certain  $\mathcal{CFS}$  right-angled Coxeter groups (Nguyen and Tran [43]) are either hyperbolic or finite-index. We give sufficient conditions for a hierarchically hyperbolic space to have the property that all its strongly quasiconvex subsets are either hyperbolic or coarsely cover the entire space; see Proposition 7.2. Applying this criteria to specific examples yields a new, unified proof of the work of Kim, Genevois, Nguyen and Tran as well as the following new results for Teichmüller space, graph manifolds, and a class of right-angled Coxeter groups that we call strongly  $\mathcal{CFS}$ .

**Corollary 1.6** *The following HHSs have the property that every strongly quasiconvex subset is either hyperbolic or coarsely covers the entire space:*

- (a) *The Teichmüller space of a finite-type surface with the Teichmüller metric.*
- (b) *The Teichmüller space of a finite-type surface of complexity at least 6 with the Weil–Petersson metric.*
- (c) *The mapping class group of an oriented, connected, finite type surface.*
- (d) *A right-angled Artin group with connected defining graph*
- (e) *A right-angled Coxeter group with strongly  $\mathcal{CFS}$  defining graph.*
- (f) *The fundamental group of a nongeometric graph manifold.*

*In particular, if  $H$  is a strongly quasiconvex subgroup in any of the groups (c)–(f), then  $H$  is either stable or finite-index.*

Stable subgroups of the mapping class group and right-angled Artin groups have been studied extensively and have several interesting equivalent characterizations including convex cocompactness in the mapping class group and purely loxodromic in right-angled Artin groups [25; 39].

We also use HHS theory and Theorem 1.2 to give a new proof of [54, Theorem 1.11] and [28, Proposition 4.9] characterizing when a special subgroup of a right-angled

Coxeter group is strongly quasiconvex. We then utilize this characterization, along with a construction of Behrstock, to demonstrate the large variety of different strongly quasiconvex subsets that can be found in the class of  $\mathcal{CFS}$  right-angled Coxeter groups.

**Theorem 1.7** *Every right-angled Coxeter group is an infinite-index strongly quasiconvex subgroup of some  $\mathcal{CFS}$  right-angled Coxeter group.*

**Hyperbolically embedded subgroups** As a final application of our characterization of strongly quasiconvex subsets, we study the hyperbolically embedded subgroups of hierarchically hyperbolic groups. Hyperbolically embedded subgroups are generalizations of peripheral subgroups in relatively hyperbolic groups (see Dahmani, Guirardel and Osin [20]) and are a key component of studying acylindrically hyperbolic groups, a large class of groups exhibiting hyperbolic-like behavior (see Osin [45]). Work of Dahmani, Guirardel and Osin [20] and Sisto [50] showed that if a finite collection of subgroups  $\{H_i\}$  is hyperbolically embedded in a finitely generated group  $G$ , then  $\{H_i\}$  is an almost malnormal collection and each  $H_i$  is strongly quasiconvex. While the converse of this statement is false in general (see the beginning of Section 8 for a counterexample), the converse does hold in the case of hyperbolic groups — see Bowditch [13, Theorem 7.11] — and cocompactly cubulated groups [28, Theorem 6.31]. We prove the converse in the setting of hierarchically hyperbolic groups.

**Theorem 1.8** (characterization of hyperbolically embedded subgroups) *Let  $G$  be a hierarchically hyperbolic group. A finite collection of subgroups  $\{H_i\}$  is hyperbolically embedded in  $G$  if and only if  $\{H_i\}$  is an almost malnormal collection and each  $H_i$  is strongly quasiconvex.*

By [38, Theorem A], an infinite-index subgroup of the mapping class group of a surface is strongly quasiconvex if and only if it is convex cocompact (this fact can also be deduced from Corollary 1.6). Thus, as a specific case of Theorem 1.8, we have the following new result for the mapping class group.

**Corollary 1.9** *If  $S$  is an oriented, connected, finite-type surface of complexity at least 2 and  $\{H_i\}$  is a finite collection of subgroups of the mapping class group of  $S$  then the following are equivalent:*

- $\{H_i\}$  is hyperbolically embedded.
- $\{H_i\}$  is an almost malnormal collection and each  $H_i$  is strongly quasiconvex.
- $\{H_i\}$  is an almost malnormal collection and each  $H_i$  is convex cocompact.

## 1.1 Open questions

We believe that strongly quasiconvex subgroups are a rich area of study with many interesting open questions both in the setting of hierarchically hyperbolic groups and beyond. In light of Theorem 1.1, it is natural to wonder which results for strongly quasiconvex subgroups of hyperbolic groups can be extended to strongly quasiconvex subgroups of hierarchically hyperbolic groups (or even finitely generated groups). As a starting point, one may aim to extend work of Gromov [31], Arzhantseva [5], and Gitik [30] on combination theorems for strongly quasiconvex subgroups of hyperbolic groups.

**Question 1** *Prove combination theorems for strongly quasiconvex subgroups of hierarchically hyperbolic groups (or even finitely generated groups). In particular, investigate conditions guaranteeing that the subgroup generated by two strongly quasiconvex subgroups,  $Q_1$  and  $Q_2$ , is strongly quasiconvex and isomorphic to  $Q_1 *_{Q_1 \cap Q_2} Q_2$ .*

As strongly quasiconvex subsets are invariant under quasi-isometry, they have the potential to play an important role in the quasi-isometric classification of hierarchically hyperbolic spaces. The following would be an interesting first step in this direction.

**Question 2** *Provide necessary conditions for an HHS to have the property that all its strongly quasiconvex subsets are either hyperbolic or coarsely cover the entire space. Using defining graphs, characterize all right-angled Coxeter groups whose strongly quasiconvex subsets are hyperbolic or coarsely cover the entire group.<sup>1</sup>*

Looking beyond hierarchically hyperbolic spaces, we wonder about the possibilities of understanding strongly quasiconvex subsets in other spaces with a notion of nonpositive curvature. Specifically we ask the following.

**Question 3** *For what other spaces are strongly quasiconvex subsets contracting (in the sense of Definition 2.10)?*

Some of the first spaces one could consider are CAT(0) spaces, coarse median spaces, and the outer automorphism groups of free groups. Sultan [52] shows that strongly quasiconvex geodesics in CAT(0) spaces are always contracting. We conjecture the

---

<sup>1</sup>The case of right-angled Coxeter groups has been resolved by Genevois [29].

same holds for all strongly quasiconvex subsets of a CAT(0) space.<sup>2</sup> A possible starting point for coarse median spaces could be the recently posted paper [16], in which Bowditch constructs hulls for subsets of coarse median spaces and produces a number of results similar to our work in Section 5.

Our proof of Theorem 1.8 rests strongly upon the equivalence between strongly quasiconvex and contracting subsets. One may then presume that any group that is an answer to Question 3 is also an answer for the following question.

**Question 4** *For what other finitely generated groups are almost malnormal, strongly quasiconvex subgroups hyperbolically embedded?*

A long-standing open question in the study of quasiconvex subgroups of hyperbolic group is whether or not finitely generated, almost malnormal subgroups of hyperbolic groups must be quasiconvex. Accordingly, we ask the analogous question for the larger class of hierarchically hyperbolic groups.

**Question 5** *Are finitely generated, almost malnormal subgroups of hierarchically hyperbolic groups strongly quasiconvex?*

## Outline

In Section 2, we begin with the basic definitions and properties of strongly quasiconvex subsets and the related notions of stability and contracting subsets of general quasigeodesic spaces. In Section 3, we define lower relative divergence and study the relationship between contracting subsets, strongly quasiconvex subsets, and lower relative divergence in any quasigeodesic space. We move on to hierarchically hyperbolic spaces in Section 4, where we give the definition of an HHS and detail the relevant tools and constructions we will need from the theory. In Section 5, we explain how to construct hierarchically quasiconvex hulls using hierarchy paths. As applications of this construction, we give a characterization of hierarchically quasiconvex sets in terms of the coarse median structure on the HHS and prove that strongly quasiconvex subsets are also hierarchically quasiconvex. In Section 6, we state and prove our equivalent characterizations of strongly quasiconvex subsets, finishing the proofs of Theorems 1.1, 1.2, and 1.4. The remaining sections are devoted to applications of this characterization. We give a generalization of the bounded geodesic image property

<sup>2</sup>This conjecture has been confirmed by Cashen [18].



for strongly quasiconvex subsets in Section 6.3, study strongly quasiconvex subsets in specific examples in Section 7, and characterize hyperbolically embedded subgroups of HHGs in Section 8.

**Acknowledgements** We gratefully acknowledge Mark F Hagen for suggesting the strategy to attack Theorem 5.2; Mark V Sapir for an example of an almost malnormal strongly quasiconvex subgroup that is not hyperbolically embedded; and Johanna Mangahas for suggesting the relation between hierarchical quasiconvexity and coarse median quasiconvexity in Section 5.1. We thank Brian Bowditch, Tai-Danae Bradley, Heejoung Kim, Chris Hruska, and Dan Beryne for their comments on early versions of this paper. We are also grateful to Kevin Schreve for pointing out an error in the first version of this paper. Russell and Spriano thank the organizers of YGGT 2018 and GAGTA 2018 where some of the work on this paper was completed. They also give special thanks to their respective advisors, Jason Behrstock and Alessandro Sisto, for their ongoing support and their many helpful comments on early drafts of this paper. Finally, we would like to thank the referees for a number of comments that improved this paper.

## 2 Coarse geometry

### 2.1 Quasigeodesic spaces, conventions, and notation

This paper focuses on understanding the geometry of metric spaces up to quasi-isometry. While many of the metric spaces we are interested in applying our results to are geodesic metric spaces, many of the subspaces we will be studying will be quasigeodesic, but not geodesic metric spaces. Thus, we will almost always assume our metric spaces are *quasigeodesic metric spaces*.

**Definition 2.1** A metric space  $X$  is a  $(K, L)$ -*quasigeodesic metric space* if for all  $x, y \in X$  there exists a  $(K, L)$ -quasigeodesic  $\gamma: [a, b] \rightarrow X$  with  $\gamma(a) = x$  and  $\gamma(b) = y$ .

Given a  $(K, L)$ -quasigeodesic metric space  $X$ , we can construct a geodesic metric space quasi-isometric to  $X$  as follows: fix an  $\epsilon$ -separated net  $N \subseteq X$  and connect a pair of points  $x, y \in N$  by an edge of length  $d(x, y)$  if  $d(x, y) < 2\epsilon$ . The resulting metric graph will be quasi-isometric to  $X$ . Since  $\epsilon$  can be chosen to depend only on  $K$  and  $L$ , this graph can be constructed such that the quasi-isometry constants will also depend only on  $K$  and  $L$ . When convenient, we will exploit this fact to reduce proofs to the geodesic case.

A particularly important collection of metric spaces in geometric group theory is the class of  $\delta$ -hyperbolic metric spaces, introduced by Gromov in [31; 32]. While  $\delta$ -hyperbolic spaces are usually required to be geodesic, the following is a direct extension of the definition to the setting of quasigeodesic metric spaces.

**Definition 2.2** A  $(K, L)$ -quasigeodesic metric space is  $\delta$ -hyperbolic if for every  $(K, L)$ -quasigeodesic triangle the  $\delta$ -neighborhood of the union of any two of the edges contains the third.

Gromov's four-point condition can also be used to define a hyperbolic quasigeodesic metric space; however as shown in [21, Example 11.36], this definition fails to be a quasi-isometry invariant if the spaces are not geodesic. In contrast, Definition 2.2 is a quasi-isometry invariant among quasigeodesic spaces. In particular, using the "guessing geodesic" criterion, from [42, Theorem 3.15] or [15, Theorem 3.1], one can show that a quasigeodesic space is hyperbolic in the sense of Definition 2.2 if and only if it is quasi-isometric to a geodesic metric space that is hyperbolic in the usual sense.

When referring to a property defined by a parameter (eg  $\delta$ -hyperbolic), we will often suppress that parameter when its specific value is not needed. To reduce the proliferation of additive and multiplicative constants throughout this paper, we will adopt the following notation.

**Notation 2.3** Let  $A, B, K, L$  be real numbers. We write

$$A \preceq_{K,L} B \quad \text{if} \quad A \leq KB + L.$$

If  $A \preceq_{K,L} B$  and  $B \preceq_{K,L} A$ , we write  $A \asymp_{K,L} B$ .

We say two subsets of a metric space  $K$ -coarsely coincide if their Hausdorff distance is at most  $K$ .

## 2.2 Strong quasiconvexity, contracting, and stability

The primary notion of convexity we will consider is the following notion of strong quasiconvexity.

**Definition 2.4** (strongly quasiconvex subset) A subset  $Y$  of a quasigeodesic metric space  $X$  is *strongly quasiconvex* if there is a function  $Q : [1, \infty) \times [0, \infty) \rightarrow [0, \infty)$  such that for every  $(K, L)$ -quasigeodesic  $\gamma$  with endpoints in  $Y$ , we have  $\gamma \subseteq N_{Q(K,L)}(Y)$ . We call the function  $Q$  the *convexity gauge* for  $Y$ .

It follows directly from the definition that strong quasiconvexity is a quasi-isometry invariant in the following sense.

**Lemma 2.5** *Let  $X$  and  $Z$  be a quasigeodesic metric spaces and  $f: X \rightarrow Z$  be a  $(K, L)$ -quasi-isometry. If  $Y$  is a  $Q$ -strongly quasiconvex subset of  $X$ , then  $f(Y)$  is a  $Q'$ -strongly quasiconvex subset of  $Z$ , with  $Q'$  depending only on  $Q, K$  and  $L$ .*

In the setting of hyperbolic spaces, strong quasiconvexity is equivalent to the weaker condition of quasiconvexity.

**Definition 2.6** A subset  $Y$  of a geodesic metric space  $X$  is *quasiconvex* if there exists  $D \geq 0$  such that for any geodesic  $\gamma$  with endpoints on  $Y$ , we have  $\gamma \subseteq N_D(Y)$ . We call the constant  $D$  the *convexity constant* for  $Y$ .

If  $Y$  is a  $Q$ -strongly quasiconvex subset of the  $(K, L)$ -quasigeodesic space  $X$ , then any two points in  $Y$  can be joined by a  $(K, L)$ -quasigeodesic in  $X$  that lies uniformly close to  $Y$ . Thus  $Y$  equipped with the metric inherited from  $X$  will be a  $(K', L')$ -quasigeodesic metric space where  $K'$  and  $L'$  depend only on  $K, L$ , and  $Q$ . For the rest of the paper, when discussing geometric properties (such as hyperbolicity) of a strongly quasiconvex subset, we shall implicitly do so with respect to the metric inherited from the ambient space. In particular, if  $f: X \rightarrow Z$  is a quasi-isometry between quasigeodesic spaces and  $Y$  is a strongly quasiconvex subset of  $X$ , then  $Y$  is quasi-isometric to  $f(Y)$ .

In [25], Durham and Taylor introduced the following related notion of convexity.

**Definition 2.7** A quasi-isometric embedding  $\Phi$  from a quasigeodesic metric space  $Y$  into a quasigeodesic metric space  $X$  is a *stable embedding* if there is a function  $R: [1, \infty) \times [0, \infty) \rightarrow [0, \infty)$  such that if  $\alpha$  and  $\beta$  are two  $(K, L)$ -quasigeodesics of  $X$  with the same endpoints in  $\Phi(Y)$ , then  $d_{\text{Haus}}(\alpha, \beta) \leq R(K, L)$ .

While the images of stable embeddings maintain many of the features of quasiconvex subsets of hyperbolic spaces, the definition is highly restrictive. In particular, as the next proposition records, stable embeddings must always be onto hyperbolic subsets.

**Proposition 2.8** *Let  $\Phi: Y \rightarrow X$  be a quasi-isometric embedding from a quasigeodesic metric space  $Y$  to a quasigeodesic metric space  $X$ . Then  $\Phi$  is a stable embedding if and only if  $Y$  is hyperbolic and  $\Phi(Y)$  is strongly quasiconvex. In particular, if  $Y$  is a*

strongly quasiconvex subset of  $X$ , then the inclusion  $i: Y \hookrightarrow X$  is a stable embedding if and only if  $Y$  is hyperbolic with respect to the metric inherited from  $X$ .

In [54, Proposition 4.3], the third author proves the above proposition for the case of geodesic spaces. The more general statement above follows immediately from the fact that a quasigeodesic space is always quasi-isometric to a geodesic space plus the fact that strong quasiconvexity, stability, and hyperbolicity are all quasi-isometry invariants.

One class of metric spaces we are particularly interested in are finitely generated groups equipped with a word metric. In this setting we are particularly interested in understanding the strongly quasiconvex and stable subgroups.

**Definition 2.9** Let  $G$  be a finitely generated group equipped with a word metric from some finite generating set. A subgroup  $H < G$  is a *strongly quasiconvex subgroup* of  $G$  if  $H$  is a strongly quasiconvex subset of  $G$  with respect to the word metric on  $G$ . A subgroup  $H < G$  is a *stable subgroup* if  $H$  is a strongly quasiconvex subgroup and  $H$  is a hyperbolic group.

The above definition of stable subgroup is different than the one originally given in [25], but it is equivalent by Proposition 2.8.

If  $H$  is a strongly quasiconvex subgroup of  $G$ , then  $H$  is also finitely generated and undistorted in  $G$ . Further, since strongly quasiconvex is a quasi-isometry invariant, being a strongly quasiconvex or a stable subgroup is independent of the choice of finite generating set for  $G$ .

It is common in the literature to study various “contracting” properties of strongly quasiconvex subsets. We compare strongly quasiconvex subsets with the following notion of a contracting subset.

**Definition 2.10** Let  $X$  be a quasigeodesic metric space and  $Y \subseteq X$ . A map  $g: X \rightarrow Y$  is said to be  $(A, D)$ -contracting for some  $A \in (0, 1]$  and  $D \geq 1$  if

- (1)  $g$  is  $(D, D)$ -coarsely Lipschitz;
- (2) for any  $y \in Y$ ,  $d(y, g(y)) \leq D$ ;
- (3) for all  $x \in X$ , if we set  $R = Ad(x, Y)$ , then  $\text{diam}(g(B_R(x))) \leq D$ .

A subset  $Y$  is said to be  $(A, D)$ -contracting if there is an  $(A, D)$ -contracting map from  $X$  to  $Y$ .

The above definition is motivated by [40, Definition 2.2] and generalizes the usual definition of contracting in hyperbolic and CAT(0) spaces to include maps that are not the closest point projection. This is critical to our study of hierarchically hyperbolic spaces in Section 6 and allows quasi-isometry invariance to be established directly from the definition.

**Lemma 2.11** *Let  $X$  and  $Z$  be quasigeodesic metric spaces and  $f: X \rightarrow Z$  be a  $(K, L)$ -quasi-isometry. If  $Y$  is an  $(A, D)$ -contracting subset of  $X$ , then  $f(Y)$  is an  $(A', D')$ -contracting subset of  $Z$ , where  $A'$  and  $D'$  depend only on  $A, D, K$  and  $L$ .*

In the setting of hyperbolic spaces, strongly quasiconvex subsets are contracting. The contracting map will be the following coarse closest point projection: if  $X$  is a  $\delta$ -hyperbolic metric space and  $Y \subseteq X$  is  $Q$ -strongly quasiconvex, then there exist  $K$  depending on  $\delta$  and  $Q$  and a  $(1, K)$ -coarsely Lipschitz map  $p_Y: X \rightarrow Y$  such that for all  $x \in X$ ,  $d(x, p_Y(x)) \leq d(x, Y) + 1$ . By an abuse of language, we will refer to  $p_Y$  as the *closest point projection* of  $X$  onto  $Y$ . For any  $Q$ -strongly quasiconvex subset  $Y$  of a  $\delta$ -hyperbolic space, the map  $p_Y$  is  $(1, D)$ -contracting where  $D$  depends only on  $Q$  and  $\delta$ .

### 3 Divergence of contracting subsets

In this section we show that contracting subsets are always strongly quasiconvex. Without some negative curvature hypotheses, such as being hierarchically hyperbolic, the converse is not always true as we show in Example 3.8. Both of these statements are proved using lower relative divergence which was originally introduced by the third author in [53]. The lower relative divergence is a family of functions that measures how efficiently one can travel in  $X$  while avoiding a subset  $Y$ ; see Figure 1.

**Definition 3.1** (lower relative divergence) Let  $X$  be a geodesic space and  $Y \subseteq X$ . For  $r > 0$  we adopt the notation

- (1)  $\partial N_r(Y) = \{x \in X \mid d(x, Y) = r\}$ ,
- (2)  $d_r$  is the induced path metric on  $X - N_r(Y)$ .

The *lower relative divergence of  $X$  with respect to  $Y$*  (or the *divergence of  $Y$  in  $X$* ), denoted by  $\text{div}(X, Y)$ , is the set of functions  $\{\sigma_\rho^n\}$  defined as follows: For each  $\rho \in (0, 1]$ , integer  $n \geq 2$  and  $r \in (0, \infty)$ , if there is no pair of  $x_1, x_2 \in \partial N_r(Y)$

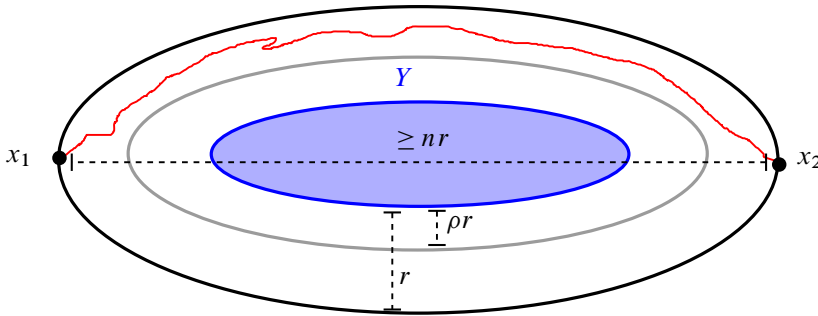


Figure 1: A sketch of a step in the construction of the function  $\sigma_\rho^n$ . The points  $x_1, x_2 \in \partial N_r(Y)$  are at least  $nr$  far apart, so we measure the distance between  $x_1$  and  $x_2$  in the complement of the  $\rho r$ -neighborhood of  $Y$ . We then take the infimum of these distances over all such pairs of points to obtain  $\sigma_\rho^n(r)$ .

such that  $d_r(x_1, x_2) < \infty$  and  $d(x_1, x_2) \geq nr$ , we define  $\sigma_\rho^n(r) = \infty$ . Otherwise, we define  $\sigma_\rho^n(r) = \inf d_{\rho r}(x_1, x_2)$  where the infimum is taken over all  $x_1, x_2 \in \partial N_r(Y)$  such that  $d_r(x_1, x_2) < \infty$  and  $d(x_1, x_2) \geq nr$ .

The lower relative divergence is often characterized by how the asymptotics of the functions  $\{\sigma_\rho^n\}$  compare to linear, polynomial and exponential functions. Such descriptions are described in detail in [53]. We will restrict our attention to the following two properties of  $\text{div}(X, Y)$ .

**Definition 3.2** Let  $X$  be a geodesic metric space and  $Y \subseteq X$ .

The lower relative divergence of  $X$  with respect to  $Y$  is *completely superlinear* if there exists  $n_0 \geq 3$  such that for every  $\rho \in (0, 1]$  and  $C > 0$  the set  $\{r \in [0, \infty) \mid \sigma_\rho^{n_0}(r) \leq Cr\}$  is bounded.

The lower relative divergence of  $X$  with respect to  $Y$  is *at least quadratic* if there exists a positive integer  $M$  such that for every  $\rho \in (0, 1]$  and  $n \geq 2$  there exist  $C > 0$  and  $r_0 > 0$  such that  $\sigma_\rho^{Mn}(r) > Cr^2$  for all  $r > r_0$ .

The properties of being completely superlinear and at least quadratic are preserved under quasi-isometry in the following sense.

**Lemma 3.3** (consequence of [53, Proposition 4.9]) *Let  $f: X \rightarrow Z$  be a quasi-isometry between geodesic spaces. If  $Y \subseteq X$  and  $W \subseteq Z$  with  $d_{\text{Haus}}(f(Y), W) < \infty$ ,*

then  $\text{div}(X, Y)$  is completely superlinear (resp. at least quadratic) if and only if  $\text{div}(Z, W)$  is completely superlinear (resp. at least quadratic).

In [53], the lower relative divergence was defined only for geodesic ambient spaces; however the definition can be extended to include quasigeodesic metric spaces as follows.

**Definition 3.4** (lower relative divergence in quasigeodesic spaces) Let  $X$  be a quasigeodesic space and  $Y \subseteq X$ . Let  $Z$  be a geodesic space and  $f: X \rightarrow Z$  be a quasi-isometry. Then *the lower relative divergence of  $X$  with respect to  $Y$*  (or the *divergence of  $Y$  in  $X$* ), denoted by  $\text{div}(X, Y)$ , is the lower relative divergence of  $Z$  with respect to  $f(Y)$ .

We say  $\text{div}(X, Y)$  is *completely superlinear* (resp. *at least quadratic*) if  $\text{div}(Z, f(Y))$  is completely superlinear (resp. at least quadratic).

While the definition of  $\text{div}(X, Y)$  in a quasigeodesic space depends on a choice of  $Z$  and  $f$ ,  $\text{div}(X, Y)$  being completely superlinear (resp. at least quadratic) is independent of this choice by Lemma 3.3. In fact, while it will not be relevant for the content of this paper,  $\text{div}(X, Y)$  is independent of the choice of  $Z$  and  $f$  in a much stronger sense. In [53] the third author defined an equivalence relation  $\sim$  between the collections of functions used to define the lower relative divergence. If  $f_1: X \rightarrow Z_1$  and  $f_2: X \rightarrow Z_2$  are two quasi-isometries with  $Z_1$  and  $Z_2$  geodesic spaces, then by [53, Proposition 4.9],  $\text{div}(Z_1, f_1(Y)) \sim \text{div}(Z_2, f_2(Y))$ . Thus  $\text{div}(X, Y)$  is well defined up to this notion of equivalence.

The following proposition shows that contracting subsets always have at least quadratic divergence.

**Proposition 3.5** *If  $X$  is a quasigeodesic space and  $Y$  is a contracting subset of  $X$ , then the lower relative divergence of  $X$  with respect to  $Y$  is at least quadratic.*

**Proof** Since every quasigeodesic space is quasi-isometric to a geodesic metric space, Lemma 2.11 allows us to assume  $X$  is geodesic. Assume that  $Y$  is  $(A, D)$ -contracting and let  $g: X \rightarrow Y$  be an  $(A, D)$ -contracting map. We first show that for all  $x \in X$ ,

$$d(x, g(x)) \leq 2Dd(x, Y) + 4D.$$

Let  $y \in Y$  be such that  $d(x, y) \leq d(x, Y) + 1$ . Then from the definition of  $(A, D)$ -contracting,

$$\begin{aligned} d(x, g(x)) &\leq d(x, y) + d(y, g(y)) + d(g(y), g(x)) \\ &\leq d(x, Y) + 1 + D + Dd(x, y) + D \\ &\leq (D + 1)d(x, Y) + 3D + 1 \\ &\leq 2Dd(x, Y) + 4D. \end{aligned}$$

Now, let  $\{\sigma_\rho^n\}$  be the lower relative divergence of  $X$  with respect to  $Y$ . We claim that for each  $n \geq 4D + 2$  and  $\rho \in (0, 1]$ ,

$$\sigma_\rho^n(r) \geq \left(\frac{A\rho}{4D}\right)r^2 \quad \text{for each } r > 8D.$$

Let  $r > 8D$ ,  $n$  be an integer greater than  $4D + 2$ , and  $\rho \in (0, 1]$ . If  $\sigma_\rho^n(r) = \infty$ , then the above inequality is true. Otherwise, let  $x_1, x_2 \in \partial N_r(Y)$  be such that  $d(x_1, x_2) \geq nr$  and  $d_r(x_1, x_2) \leq \infty$ . The distances  $d(x_1, g(x_1))$  and  $d(x_2, g(x_2))$  are bounded above by  $2Dr + 4D$ . Therefore,

$$d(g(x_1), g(x_2)) \geq d(x_1, x_2) - d(x_1, g(x_1)) - d(x_2, g(x_2)) \geq nr - 4Dr - 8D \geq r.$$

Let  $\gamma$  be a rectifiable path in  $N_{\rho r}(Y)$  connecting  $x_1$  and  $x_2$  and  $R = A\rho r/2$ . There exist  $t_0 < t_1 < t_2 < \dots < t_{m-1} < t_m$  such that  $\gamma(t_0) = x_1, \gamma(t_m) = x_2$  and

$$\frac{1}{2}R \leq \ell(\gamma|_{[t_{i-1}, t_i]}) \leq R,$$

where  $\ell(\cdot)$  denotes the length of a path. This implies

$$(1) \quad \ell(\gamma) = \sum_{i=1}^m \ell(\gamma|_{[t_{i-1}, t_i]}) \geq \frac{1}{2}mR.$$

Since  $g$  is an  $(A, D)$ -contracting map and  $d(\gamma(t_{i-1}), \gamma(t_i)) < Ad(\gamma(t_{i-1}), Y)$ , we have  $d(g(\gamma(t_{i-1})), g(\gamma(t_i))) \leq D$  for each  $1 \leq i \leq m$ . Thus

$$(2) \quad d(g(x_1), g(x_2)) \leq \sum_{i=1}^m d(g(\gamma(t_{i-1})), g(\gamma(t_i))) \leq mD.$$

Since  $d(g(x_1), g(x_2)) \geq r$ , inequality (2) implies  $m \geq r/D$ . Combining this with inequality (1), we have

$$\ell(\gamma) \geq \frac{1}{2}mR \geq \left(\frac{A\rho}{4D}\right)r^2.$$

Therefore,

$$\sigma_\rho^n(r) \geq \left(\frac{A\rho}{4D}\right)r^2$$



for  $n \geq 4D + 2$ ,  $\rho \in (0, 1]$ , and  $r > 8D$ . This implies that the lower relative divergence of  $X$  with respect to  $Y$  is at least quadratic.  $\square$

In [53], the third author classified strongly quasiconvex subsets in terms of their lower relative divergence. This result continues to hold in the slightly more general setting of quasigeodesic spaces.

**Theorem 3.6** [54, Theorem 3.1] *Let  $X$  be a quasigeodesic space and  $Y \subseteq X$ . Then  $Y$  is strongly quasiconvex if and only if the lower relative divergence of  $X$  with respect to  $Y$  is completely superlinear.*

**Proof** Since every quasigeodesic metric space is quasi-isometric to a geodesic metric space, the result follows immediately from [54, Theorem 1.5] when  $Y$  is infinite diameter. If  $\text{diam}(A) = r_0 < \infty$ , then for all  $r > r_0$ ,  $\partial N_{r_0}(Y) = \emptyset$  and thus  $\sigma_\rho^n(r) = \infty$ . Hence  $\text{div}(X, Y)$  is completely superlinear and  $Y$  is strongly quasiconvex.  $\square$

Proposition 3.5 and Theorem 3.6 combine to say that if a subset  $Y \subseteq X$  is  $(A, D)$ -contracting, then  $Y$  is strongly quasiconvex. A direct proof of this result was shown by Sultan for the case of quasigeodesics, but the proof extends to any subset without modification [52, Lemma 3.3]. For completeness, we include a proof using the bound on the lower relative divergence of  $Y$  from Proposition 3.5.

**Corollary 3.7** *Let  $X$  be a  $(K, L)$ -quasigeodesic space and  $Y \subseteq X$ . If  $Y$  is  $(A, D)$ -contracting, then  $Y$  is  $Q$ -strongly quasiconvex where  $Q$  is determined by  $A, D, K$  and  $L$ .*

**Proof** Let  $Y$  be a  $(A, D)$ -contracting subset of  $X$ . We first assume that  $X$  is a geodesic metric space. Let  $\{\sigma_\rho^n\}$  be the lower relative divergence of  $X$  with respect to  $Y$ . The proof of Proposition 3.5 shows that for each  $n \geq 4D + 2$  and  $\rho \in (0, 1]$ ,

$$\sigma_\rho^n(r) \geq \left(\frac{A\rho}{4D}\right)r^2 \quad \text{for all } r > 8D.$$

Therefore, by fixing  $n = n_0 = 4D + 3$  and  $\rho = 1$ ,

$$\sigma_1^{n_0}(r) \geq \left(\frac{A}{4D}\right)r^2 \quad \text{for all } r > 8D.$$

If  $\gamma$  is a  $(\lambda, \epsilon)$ -quasigeodesic with endpoints on  $Y$ , let  $m = \inf\{B \in \mathbb{R} \mid \gamma \subseteq N_B(Y)\}$ . The proof of [54, Proposition 3.1] establishes that if  $m$  is larger than a fixed constant

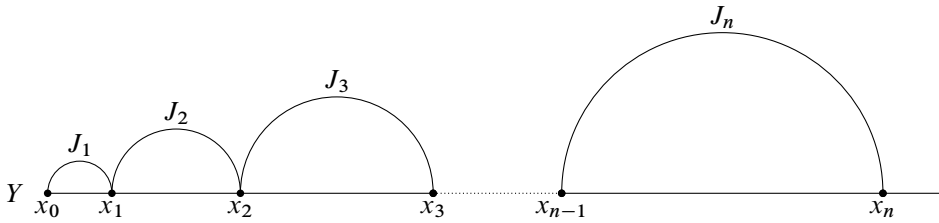


Figure 2: The space  $X$  of Example 3.8.

depending on  $\lambda$  and  $\epsilon$ , then there exist constants  $C_0$  and  $C_1$  depending only on  $\lambda$ ,  $\epsilon$  and  $n_0$ , such that  $\sigma_1^{n_0}(C_0m) \leq C_1m$ . Thus,

$$\left(\frac{A}{4D}\right)(C_0m)^2 \leq \sigma_1^{n_0}(C_0m) \leq C_1m,$$

and hence  $m$  is bounded by some constant depending only on  $\lambda$ ,  $\epsilon$ ,  $A$  and  $D$ . Thus, there exists a function  $Q$  depending only on  $A$  and  $D$  such that  $Y$  is  $Q$ -strongly quasiconvex.

When  $X$  is a  $(K, L)$ -quasigeodesic space, there exist a geodesic metric space  $Z$  and a quasi-isometry  $f : X \rightarrow Z$  with constants determined by  $K$  and  $L$ . The result follows from the geodesic case by Lemmas 2.5 and 2.11. □

We finish this section by adapting [6, Example 3.4] to give a counterexample to the converse of Corollary 3.7.

**Example 3.8** (strongly quasiconvex subsets need not be contracting) Let  $Y$  be a ray with initial point  $x_0$  and let  $(x_n)$  be the sequence of points along  $Y$  such that for each  $n \geq 1$  the distance between  $x_{n-1}$  and  $x_n$  is equal to  $n$ . We connect each pair  $(x_{n-1}, x_n)$  by an additional segment  $J_n$  of length  $n^{3/2}$  as shown in Figure 2. Let  $X$  be the resulting geodesic space.

By Proposition A.2 the lower relative divergence of  $X$  with respect to  $Y$  is completely superlinear, but not at least quadratic — heuristically,  $\text{div}(X, Y)$  behaves like  $r^{3/2}$ . So  $Y$  is strongly quasiconvex, but not contracting by Proposition 3.5 and Theorem 3.6.

## 4 Hierarchically hyperbolic spaces

We now recall the main definitions of hierarchically hyperbolic groups and spaces. The main references, where not specified, are [9; 10]. While we give the entire definition

of an HHS for completeness, we advise the reader that we shall only directly utilize axioms (1), (2), (3), (5), (8), and (10) of Definition 4.1 in the remainder of the paper.

**Definition 4.1** (hierarchically hyperbolic space) Let  $\mathcal{X}$  be a quasigeodesic space. A *hierarchically hyperbolic space (HHS) structure* on  $\mathcal{X}$  consists of constants  $E \geq \kappa_0 > 0$ , an index set  $\mathfrak{S}$ , and a set  $\{CW \mid W \in \mathfrak{S}\}$  of geodesic  $\delta$ -hyperbolic spaces  $(CW, d_W)$ , such that the following conditions are satisfied.

(1) **Projections** For each  $W \in \mathfrak{S}$ , there exists a *projection*  $\pi_W : \mathcal{X} \rightarrow 2^{CW}$  such that for all  $x \in \mathcal{X}$ ,  $\pi_W(x) \neq \emptyset$ , and  $\text{diam}(\pi_W(x)) < E$ . Moreover, there exists a  $K$  such that each  $\pi_W$  is  $(K, K)$ -coarsely Lipschitz and  $\pi_W(\mathcal{X})$  is  $K$ -quasiconvex in  $CW$ .

(2) **Nesting**  $\mathfrak{S}$  is equipped with a partial order  $\sqsubseteq$ , and either  $\mathfrak{S} = \emptyset$  or  $\mathfrak{S}$  contains a unique  $\sqsubseteq$ -maximal element; when  $V \sqsubseteq W$ , we say  $V$  is *nested* in  $W$ . For each  $W \in \mathfrak{S}$ , we denote by  $\mathfrak{S}_W$  the set of  $V \in \mathfrak{S}$  such that  $V \sqsubseteq W$ . Moreover, for all  $V, W \in \mathfrak{S}$  with  $V \not\sqsubseteq W$  there is a specified nonempty subset  $\rho_W^V \subseteq CW$  with  $\text{diam}_{CW}(\rho_W^V) \leq E$ . There is also a *projection*  $\rho_V^W : CW \rightarrow 2^{CV}$ .

(3) **Orthogonality**  $\mathfrak{S}$  has a symmetric and antireflexive relation called *orthogonality*; we write  $V \perp W$  when  $V$  and  $W$  are orthogonal. Whenever  $V \sqsubseteq W$  and  $W \perp U$ , we require that  $V \perp U$ . Additionally, if  $V \perp W$ , then  $V$  and  $W$  are not  $\sqsubseteq$ -comparable.

(4) **Containers** For each  $T \in \mathfrak{S}$  and each  $U \in \mathfrak{S}_T$  for which  $\{V \in \mathfrak{S}_T \mid V \perp U\} \neq \emptyset$ , there exists a  $W \in \mathfrak{S}_T - \{T\}$ , such that whenever  $V \perp U$  and  $V \sqsubseteq T$ , we have  $V \sqsubseteq W$ . We say  $W$  is a *container* for  $U$  in  $\mathfrak{S}_T$ .

(5) **Transversality and consistency** If  $V, W \in \mathfrak{S}$  are not orthogonal and neither is nested in the other, then we say  $V$  and  $W$  are *transverse*, denoted by  $V \pitchfork W$ . If  $V \pitchfork W$ , then there are nonempty sets  $\rho_W^V \subseteq CW$  and  $\rho_V^W \subseteq CV$ , each of diameter at most  $E$ , satisfying

$$\min\{d_W(\pi_W(x), \rho_W^V), d_V(\pi_V(x), \rho_V^W)\} \leq \kappa_0$$

for all  $x \in \mathcal{X}$ .

For  $V, W \in \mathfrak{S}$  satisfying  $V \sqsubseteq W$  and for all  $x \in \mathcal{X}$ ,

$$\min\{d_W(\pi_W(x), \rho_W^V), \text{diam}_{CV}(\pi_V(x) \cup \rho_V^W(\pi_W(x)))\} \leq \kappa_0.$$

Finally, if  $U \sqsubseteq V$ , then  $d_W(\rho_W^U, \rho_W^V) \leq \kappa_0$  whenever  $W \in \mathfrak{S}$  satisfies either  $V \not\sqsubseteq W$  or  $V \pitchfork W$  and  $W \not\sqsubseteq U$ .

(6) **Finite complexity** There exists  $n \geq 0$  such that any set of pairwise  $\sqsubseteq$ -comparable elements has cardinality at most  $n$ .

(7) **Large links** There exists  $\zeta \geq 1$  such that the following holds. Let  $W \in \mathfrak{S}$  and  $x, x' \in \mathcal{X}$ . There exist  $\{U_i\}_{i=1, \dots, m} \subseteq \mathfrak{S}_W - \{W\}$  such that

$$m \leq \zeta d_W(\pi_W(x), \pi_W(x')) + \zeta$$

and for all  $V \in \mathfrak{S}_W - \{W\}$ , either  $V \in \mathfrak{S}_{U_i}$  for some  $i$ , or  $d_V(\pi_V(x), \pi_V(x')) < E$ . Also,  $d_W(\pi_W(x), \rho_W^{U_i}) \leq \zeta d_W(\pi_W(x), \pi_W(x')) + \zeta$  for each  $i$ .

(8) **Bounded geodesic image** For all  $W \in \mathfrak{S}$ , all  $V \in \mathfrak{S}_W - \{W\}$ , and all geodesics  $\gamma$  of  $CW$ , either  $\text{diam}(\rho_V^W(\gamma)) \leq E$  or  $\gamma \cap N_E(\rho_V^W) \neq \emptyset$ .

(9) **Partial realization** There exists a constant  $\alpha$  with the following property. Let  $\{V_j\}$  be a family of pairwise orthogonal elements of  $\mathfrak{S}$ , and let  $p_j \in \pi_{V_j}(\mathcal{X}) \subseteq CV_j$ . Then there exists  $x \in \mathcal{X}$  such that

- $d_{V_j}(x, p_j) \leq \alpha$  for all  $j$ ;
- for each  $j$  and each  $V \in \mathfrak{S}$  with  $V_j \sqsubseteq V$ , we have  $d_V(x, \rho_V^{V_j}) \leq \alpha$ ;
- if  $W \pitchfork V_j$  for some  $j$ , then  $d_W(x, \rho_W^{V_j}) \leq \alpha$ .

(10) **Uniqueness** For each  $\kappa \geq 0$ , there exists  $\theta_u = \theta_u(\kappa)$  such that if  $x, y \in \mathcal{X}$  and  $d(x, y) \geq \theta_u$ , then there exists  $V \in \mathfrak{S}$  such that  $d_V(x, y) \geq \kappa$ .

We will refer to the elements of the index set  $\mathfrak{S}$  as *domains* and use  $\mathfrak{S}$  to denote the entire HHS structure, including all the spaces, constants, projections and relations defined above. A quasigeodesic space  $\mathcal{X}$  is a *hierarchically hyperbolic space* (HHS) if it admits a hierarchically hyperbolic structure. We will use the pair  $(\mathcal{X}, \mathfrak{S})$  to denote  $\mathcal{X}$  equipped with the hierarchically hyperbolic structure  $\mathfrak{S}$ .

If  $(\mathcal{X}, \mathfrak{S})$  is a hierarchically hyperbolic space and  $f: \mathcal{Y} \rightarrow \mathcal{X}$  is a quasi-isometry, then  $\mathfrak{S}$  is also an HHS structure for  $\mathcal{Y}$  where the projections maps are defined by  $\pi_W \circ f$  for each  $W \in \mathfrak{S}$ .

Many of the key examples of hierarchically hyperbolic spaces are finitely generated groups where the Cayley graph admits an HHS structure. In the case where this structure is preserved by the group action, we will call those groups hierarchically hyperbolic groups.

**Definition 4.2** (hierarchically hyperbolic groups) Let  $G$  be a finitely generated group. We say  $G$  is a *hierarchically hyperbolic group* (HHG) if:

- (1)  $G$  with the word metric from a finite generating set admits an HHS structure  $\mathfrak{S}$ .
- (2) There is a  $\sqsubseteq, \perp$  and  $\pitchfork$  preserving action of  $G$  on  $\mathfrak{S}$  by bijections such that  $\mathfrak{S}$  contains finitely many  $G$  orbits.
- (3) For each  $W \in \mathfrak{S}$  and  $g \in G$ , there exists an isometry  $g_W : CW \rightarrow C(gW)$  satisfying the following for all  $V, W \in \mathfrak{S}$  and  $g, h \in G$ :
  - The map  $(gh)_W : CW \rightarrow C(ghW)$  is equal to  $g_{hW} \circ h_W : CW \rightarrow C(ghW)$ .
  - For each  $h \in G$ ,  $g_W(\pi_W(h))$  and  $\pi_{gW}(gh)$   $E$ -coarsely coincide.
  - If  $V \pitchfork W$  or  $V \sqsubseteq W$ , then  $g_W(\rho_W^V)$  and  $\rho_{gW}^{gV}$   $E$ -coarsely coincide.
  - If  $V \sqsubseteq W$  and  $p \in CW - N_E(\rho_W^V)$ , then  $g_W(\rho_W^V(p))$  and  $\rho_{gW}^{gV}(g_W(p))$   $E$ -coarsely coincide.

The HHS structure  $\mathfrak{S}$  satisfying (1)–(3) is called a hierarchically hyperbolic group (HHG) structure on  $G$  and we use  $(G, \mathfrak{S})$  to denote a group  $G$  equipped with a specific HHG structure  $\mathfrak{S}$ .

Being a hierarchically hyperbolic group is independent of choice of generating set by virtue of being able to pass the HHG structure through a  $G$ -equivariant quasi-isometry. The reader may find it helpful to note that the conditions in (3) above can be summarized by saying the diagrams

$$\begin{array}{ccc}
 G & \xrightarrow{g} & G \\
 \downarrow \pi_W & & \downarrow \pi_{gW} \\
 CW & \xrightarrow{g_W} & C(gW)
 \end{array}
 \quad \text{and} \quad
 \begin{array}{ccc}
 CV & \xrightarrow{g_V} & C(gV) \\
 \downarrow \rho_W^V & & \downarrow \rho_{gW}^{gV} \\
 CW & \xrightarrow{g_W} & C(gW)
 \end{array}$$

coarsely commute whenever  $V, U \in \mathfrak{S}$  are not orthogonal.

**Notation 4.3** When writing distances in  $CW$  for some  $W \in \mathfrak{S}$ , we often simplify the notation by suppressing the projection map  $\pi_W$ , that is, given  $x, y \in \mathcal{X}$  and  $p \in CW$  we write  $d_W(x, y)$  for  $d_W(\pi_W(x), \pi_W(y))$  and  $d_W(x, p)$  for  $d_W(\pi_W(x), p)$ . Note that when we measure distance between a pair of sets (typically both of bounded diameter) we are taking the minimum distance between the two sets. Given  $A \subseteq \mathcal{X}$  and  $W \in \mathfrak{S}$  we let  $\pi_W(A)$  denote  $\bigcup_{a \in A} \pi_W(a)$ .

The guiding philosophy of hierarchically hyperbolic spaces is that one can “pull back” the hyperbolic geometry of the various  $CW$ ’s to obtain features of negative curvature

in the original space. The most prominent example of this philosophy is the following distance formula which allows distances in the main space  $\mathcal{X}$  to be approximated by distances in the hyperbolic spaces.

**Theorem 4.4** (the distance formula; [10, Theorem 4.4]) *Let  $(\mathcal{X}, \mathfrak{S})$  be a hierarchically hyperbolic space. Then there exists  $\sigma_0$  such that, for all  $\sigma \geq \sigma_0$ , there exist  $K \geq 1$  and  $L \geq 0$  such that, for all  $x, y \in \mathcal{X}$ ,*

$$d_{\mathcal{X}}(x, y) \asymp_{K,L} \sum_{U \in \mathfrak{S}} \{ \{d_U(x, y)\} \}_\sigma,$$

where  $\{ \{N\} \}_\sigma = N$  if  $N \geq \sigma$  and 0 otherwise.

The distance formula can be “distributed” over a sum of distances in the hyperbolic spaces as described in the next lemma.

**Lemma 4.5** [48, Lemma 2.26] *Let  $(\mathcal{X}, \mathfrak{S})$  be an HHS and  $x_0, x_1, \dots, x_n$  be points in  $\mathcal{X}$ . If there exists  $C \geq 1$  such that  $\sum_{i=0}^{n-1} d_W(x_i, x_{i+1}) \asymp_{C,C} d_W(x_0, x_n)$  for all  $W \in \mathfrak{S}$ , then there exist  $K$  depending only on  $C, n$ , and  $(\mathcal{X}, \mathfrak{S})$  such that*

$$\sum_{i=0}^{n-1} d_{\mathcal{X}}(x_i, x_{i+1}) \asymp_{K,K} d_{\mathcal{X}}(x_0, x_n).$$

Part of the content of Theorem 4.4 is that for any pair of points in an HHS, there is only a finite number of domains where that pair of points can have a large projection. More precisely, if  $(\mathcal{X}, \mathfrak{S})$  is a hierarchically hyperbolic space, then a domain  $W \in \mathfrak{S}$  is said to be  $\sigma$ -relevant for  $x, y \in \mathcal{X}$  if  $d_W(x, y) > \sigma$ . We denote the set of all  $\sigma$ -relevant domains for  $x, y \in \mathcal{X}$  by  $\text{Rel}_\sigma(x, y)$ . By Theorem 4.4, for all  $\sigma \geq \sigma_0$ ,  $\text{Rel}_\sigma(x, y)$  has finite cardinality. The relevant facts about  $\text{Rel}_\sigma(x, y)$  that we will need are summarized in the following proposition.

**Proposition 4.6** [10, Lemma 2.2, Proposition 2.8, Lemma 2.14] *Let  $(\mathcal{X}, \mathfrak{S})$  be a hierarchically hyperbolic space and  $E \geq 0$  be the maximum of all the constants in the HHS structure for  $(\mathcal{X}, \mathfrak{S})$ .*

- (1) *There exists  $\chi > 0$  such that if  $\mathfrak{U} \subseteq \mathfrak{S}$  does not contain a pair of transverse domains, then  $|\mathfrak{U}| \leq \chi$ .*
- (2) *If  $\sigma \geq 100E$  and  $x, y \in \mathcal{X}$ , then the set  $\text{Rel}_\sigma(x, y)$  can be partially ordered by*

$$U \leq V \iff U = V \text{ or } U \pitchfork V \text{ and } d_V(\rho_V^U, y) \leq \kappa_0.$$

- (3) If  $\sigma \geq 100E$  and  $x, y \in \mathcal{X}$ , then there exists  $n \leq \chi$  such that  $\text{Rel}_\sigma(x, y)$  can be partitioned into  $n$  disjoint subsets  $\mathcal{U}_1, \dots, \mathcal{U}_n$  where, for each  $i$ ,  $\mathcal{U}_i$  is totally ordered with respect to the above ordering on  $\text{Rel}_\sigma(x, y)$ .

Hierarchically hyperbolic spaces contain a particularly nice class of quasigeodesics, called hierarchy paths. Even when considering a geodesic HHS, it is often preferable to work with hierarchy paths over geodesics.

**Definition 4.7** (hierarchy path) For  $\lambda \geq 1$ , a (not necessarily continuous) path  $\gamma: [a, b] \rightarrow \mathcal{X}$  is a  $\lambda$ -hierarchy path if

- (1)  $\gamma$  is a  $(\lambda, \lambda)$ -quasigeodesic,
- (2) for each  $W \in \mathfrak{S}$ , the path  $\pi_W \circ \gamma$  is an unparametrized  $(\lambda, \lambda)$ -quasigeodesic.

Recall that a map  $f: [a, b] \rightarrow X$  is an unparametrized  $(\lambda, \lambda)$ -quasigeodesic if there exists an increasing function  $g: [0, \ell] \rightarrow [a, b]$  such that  $g(0) = a$ ,  $g(\ell) = b$ , and  $f \circ g$  is a  $(\lambda, \lambda)$ -quasigeodesic of  $X$ .

While not every quasigeodesic in an HHS is a hierarchy path, every pair of points can be connected by a hierarchy path as the next theorem describes.

**Theorem 4.8** (existence of hierarchy paths; [10, Theorem 5.4]) Let  $(\mathcal{X}, \mathfrak{S})$  be a hierarchically hyperbolic space. Then there exists a  $\lambda_0$  such that any  $x, y \in \mathcal{X}$  are joined by a  $\lambda_0$ -hierarchy path.

## 4.1 Hierarchical quasiconvexity and gate maps

In [10], Behrstock, Hagen and Sisto introduced *hierarchical quasiconvexity*, a notion of convexity unique to hierarchically hyperbolic spaces.

**Definition 4.9** (hierarchical quasiconvexity; [10, Definition 5.1]) Let  $(\mathcal{X}, \mathfrak{S})$  be a hierarchically hyperbolic space and  $k: [0, \infty) \rightarrow [0, \infty)$ . A subset  $Y \subseteq \mathcal{X}$  is  $k$ -hierarchically quasiconvex if:

- (1) For all  $U \in \mathfrak{S}$ , the projection  $\pi_U(Y)$  is a  $k(0)$ -quasiconvex subspace of the  $\delta$ -hyperbolic space  $CU$ .
- (2) For every  $\kappa > 0$  and every point  $x \in \mathcal{X}$  satisfying  $d_U(x, Y) \leq \kappa$  for all  $U \in \mathfrak{S}$ , we have that  $d_{\mathcal{X}}(x, Y) \leq k(\kappa)$ .

While hierarchically quasiconvex subsets need not be strongly quasiconvex, they are “quasiconvex with respect to hierarchy paths”. That is, if  $Y \subseteq \mathcal{X}$  is  $k$ -hierarchically quasiconvex then any  $\lambda$ -hierarchy path with endpoints in  $Y$  must stay uniformly close to  $Y$ . The existence of hierarchy paths (Theorem 4.8) therefore ensures that if  $Y$  is equipped with the induced metric from  $\mathcal{X}$ , then  $Y$  is also a quasigeodesic metric space with constants depending only on  $(\mathcal{X}, \mathfrak{S})$  and  $k$ . In Section 5 we will prove that hierarchically quasiconvex subsets are actually characterized by this “quasiconvexity with respect to hierarchy paths”.

One of the key features of hierarchically quasiconvex subsets is that they are hierarchically hyperbolic spaces with the restriction of the HHS structure from the ambient space.

**Theorem 4.10** [10, Proposition 5.6] *Let  $(\mathcal{X}, \mathfrak{S})$  be a hierarchically hyperbolic space and  $Y \subseteq \mathcal{X}$  be  $k$ -hierarchically quasiconvex. Then  $(Y, \mathfrak{S})$  is a hierarchically hyperbolic space, where  $Y$  is equipped with the induced metric from  $\mathcal{X}$ .*

The following lemma is a special case of the powerful realization theorem for hierarchically hyperbolic spaces; see [10, Theorem 3.1]. It is often useful when verifying that a subset is hierarchically quasiconvex.

**Lemma 4.11** [10, Theorem 3.1, Lemma 5.3] *For each  $R \geq 0$  there is a  $\mu \geq 0$  such that the following holds. Let  $Y \subseteq \mathcal{X}$  be such that  $\pi_W(Y)$  is  $R$ -quasiconvex for each  $W \in \mathfrak{S}$ . Let  $x \in \mathcal{X}$  and for each  $W \in \mathfrak{S}$ , let  $p_W \in \pi_W(Y)$  satisfy  $d_V(x, p_W) \leq d_W(x, Y) + 1$ . Then there exists  $p \in \mathcal{X}$  such that  $d_W(p, p_W) \leq \mu$  for all  $W \in \mathfrak{S}$ .*

Given a subset  $Y \subseteq \mathcal{X}$ , there exists a *hierarchically quasiconvex hull* of  $Y$  which can be thought of as the coarsely smallest hierarchically quasiconvex subset of  $\mathcal{X}$  containing  $Y$ .

**Definition 4.12** (hierarchically quasiconvex hull) For each set  $Y \subseteq \mathcal{X}$  and  $W \in \mathfrak{S}$ , let  $\text{hull}_{CW}(Y)$  denote the convex hull of  $\pi_W(Y)$  in  $CW$ , ie the union of all  $CW$ -geodesics connecting pairs of points in  $\pi_W(Y)$ . Given  $\theta \geq 0$ , let  $H_\theta(Y)$  be the set of all  $p \in \mathcal{X}$  such that, for each  $W \in \mathfrak{S}$ , the set  $\pi_W(p)$  lies at distance at most  $\theta$  from  $\text{hull}_{CW}(Y)$ . Note that  $Y \subseteq H_\theta(Y)$ .

**Lemma 4.13** [10, Lemma 6.2] *Let  $(\mathcal{X}, \mathfrak{S})$  be an HHS. There exists  $\theta_0$  such that for each  $\theta \geq \theta_0$  there exists  $k: [0, \infty) \rightarrow [0, \infty)$  such that for each  $Y \subseteq \mathcal{X}$ , the hull  $H_\theta(Y)$  is  $k$ -hierarchically quasiconvex.*



In Section 5 we strengthen the analogy between hierarchically quasiconvex hulls and convex hulls in hyperbolic spaces, by showing that  $H_\theta(Y)$  can be constructed by iteratively connecting points in  $Y$  by hierarchy paths.

One of the important properties of hierarchically quasiconvex subsets is the existence of a *gate map* which retracts the entire space onto the hierarchically quasiconvex subset. The gate map is a generalization to hierarchically hyperbolic spaces of the closest point projection,  $\mathfrak{p}$ , defined at the end of Section 2.

**Lemma 4.14** (existence of coarse gates; [10, Lemma 5.5]) *If  $(\mathcal{X}, \mathfrak{S})$  is a hierarchically hyperbolic space and  $Y \subseteq \mathcal{X}$  is  $k$ -hierarchically quasiconvex and nonempty, then there exists a gate map  $\mathfrak{g}_Y : \mathcal{X} \rightarrow Y$  such that*

- (1)  $\mathfrak{g}_Y$  is  $(K, K)$ -coarsely Lipschitz;
- (2) for all  $y \in Y$ ,  $d_{\mathcal{X}}(y, \mathfrak{g}_Y(y)) \leq K$ ;
- (3) for all  $x \in \mathcal{X}$  and  $U \in \mathfrak{S}$ ,  $d_U(\mathfrak{g}_Y(x), \mathfrak{p}_{\pi_U(Y)}(\pi_U(x))) \leq K$ ;

where  $K$  depends only on  $k$  and  $\mathfrak{S}$ .

While the gate map need not be the closest point projection, it approximates the closest point projection with a multiplicative and additive error.

**Lemma 4.15** [11, Lemma 1.27] *Let  $Y$  be a  $k$ -hierarchically quasiconvex subset of the HHS  $(\mathcal{X}, \mathfrak{S})$  and  $x \in \mathcal{X}$ . If  $y \in Y$  is a point such that  $d_{\mathcal{X}}(x, y) \leq d_{\mathcal{X}}(x, Y) + 1$ , then  $d_{\mathcal{X}}(x, y) \asymp d_{\mathcal{X}}(x, \mathfrak{g}_Y(x))$  where the constants depend only on  $k$  and  $\mathfrak{S}$ .*

In the case of hierarchically hyperbolic groups, the gate is also coarsely equivariant.

**Lemma 4.16** (coarse equivariance of gate maps) *Let  $(G, \mathfrak{S})$  be a hierarchically hyperbolic group and let  $Y$  be a  $k$ -hierarchically quasiconvex subspace of  $G$ . There exists  $K$  depending on  $(G, \mathfrak{S})$  and  $k$  such that, for every  $g, x \in G$ ,*

$$d_G(g\mathfrak{g}_Y(x), \mathfrak{g}_Y(gx)) \leq K.$$

**Proof** Since  $G$  acts on the disjoint union of the  $CW$ 's by isometries, Lemma 4.14 and the definition of HHG provide a uniform bound on  $d_W(\pi_W(g\mathfrak{g}_Y(x)), \pi_W(\mathfrak{g}_Y(gx)))$  for all  $W \in \mathfrak{S}$ , which depends only on  $\mathfrak{S}$ ,  $k$ , and the choice of finite generating set for  $G$ . The result now follows from the distance formula (Theorem 4.4).  $\square$

The following lemma explains the nice behavior of the gates of hierarchically quasiconvex sets onto each other. The lemma is stated in slightly more generality than presented in [11], but the more general statement is implicit in the proof of [11, Lemma 1.20]. The following notation will simplify the exposition.

**Notation 4.17** If  $\mathfrak{S}$  is an HHS structure on a metric space  $\mathcal{X}$  and  $\mathcal{H} \subseteq \mathfrak{S}$  we use  $\mathcal{H}^\perp$  to denote the set  $\{W \in \mathfrak{S} \mid \forall H \in \mathcal{H}, H \perp W\}$ . In particular,  $\mathfrak{S}_U^\perp = \{W \in \mathfrak{S} \mid U \perp W\}$  for any  $U \in \mathfrak{S}$ . Note, if  $\mathcal{H} = \emptyset$ , then  $\mathcal{H}^\perp = \mathfrak{S}$  as every domain in  $\mathfrak{S}$  would vacuously satisfy the condition of the set.

**Theorem 4.18** (the bridge theorem; [11, Lemma 1.20]) *Let  $(\mathcal{X}, \mathfrak{S})$  be a hierarchically hyperbolic space and  $\theta_0$  be as in Lemma 4.13. For every  $k$  and  $\theta \geq \theta_0$ , there exist  $k' : [0, \infty) \rightarrow [0, \infty)$  and  $K_0 \geq 0$  such that, for any  $k$ -hierarchically quasiconvex sets  $A$  and  $B$ :*

- (1)  $\mathfrak{g}_A(B)$  is  $k'$ -hierarchically quasiconvex.
- (2) The composition  $\mathfrak{g}_A \circ \mathfrak{g}_B|_{\mathfrak{g}_A(B)}$  is bounded distance from the identity map  $\mathfrak{g}_A(B) \rightarrow \mathfrak{g}_A(B)$ .
- (3) For any  $a \in \mathfrak{g}_A(B)$  and  $b = \mathfrak{g}_B(a)$ , we have a  $(K_0, K_0)$ -quasi-isometric embedding  $f : \mathfrak{g}_A(B) \times H_\theta(a, b) \rightarrow \mathcal{X}$  with image  $H_\theta(\mathfrak{g}_A(B) \cup \mathfrak{g}_B(A))$  such that  $f(\mathfrak{g}_A(B) \times \{b\})$   $K_0$ -coarsely coincides with  $\mathfrak{g}_B(A)$ .

Let  $K \geq K_0$  and  $\mathcal{H} = \{U \in \mathfrak{S} : \text{diam}(\pi_U(\mathfrak{g}_A(B))) > K\}$ .

- (4) For each  $p, q \in \mathfrak{g}_A(B)$  and  $t \in H_\theta(a, b)$ ,

$$\text{Rel}_K(f(p, t), f(q, t)) \subseteq \mathcal{H}.$$

- (5) For each  $p \in \mathfrak{g}_A(B)$  and  $t_1, t_2 \in H_\theta(a, b)$ ,

$$\text{Rel}_K(f(p, t_1), f(p, t_2)) \subseteq \mathcal{H}^\perp.$$

- (6) For each  $p \in A, q \in B$ ,

$$d(p, q) \asymp_{K_0, K_0} d(p, \mathfrak{g}_A(B)) + d(q, \mathfrak{g}_B(A)) + d(A, B) + d(\mathfrak{g}_{\mathfrak{g}_B(A)}(p), \mathfrak{g}_{\mathfrak{g}_B(A)}(q)).$$

We name Theorem 4.18 the bridge theorem as one should think of the set

$$H_\theta(\mathfrak{g}_A(B) \cup \mathfrak{g}_B(A))$$

as a “bridge” between  $A$  and  $B$ : in order to efficiently travel between  $A$  and  $B$  one needs to always traverse this bridge. The bridge theorem, along with the construction of

the gate map and hulls produces the following fact about the set  $H_\theta(\mathfrak{g}_A(B) \cup \mathfrak{g}_B(A))$  which we will need in Section 8.

**Lemma 4.19** *For every  $k$  and  $\theta \geq \theta_0$ , there exists  $K$  such that for any  $k$ -hierarchically quasiconvex sets  $A$  and  $B$ , the sets  $\mathfrak{g}_B(H_\theta(\mathfrak{g}_A(B) \cup \mathfrak{g}_B(A)))$  and  $\mathfrak{g}_B(A)$   $K$ -coarsely coincide.*

We finish this section by recalling the construction of *standard product regions* introduced in [9, Section 13] and studied further in [10]. For what follows, fix a hierarchically hyperbolic space  $(\mathcal{X}, \mathfrak{S})$ .

**Definition 4.20** (nested partial tuple  $F_U$ ) Recall  $\mathfrak{S}_U = \{V \in \mathfrak{S} \mid V \sqsubseteq U\}$ . Define  $F_U$  to be the set of tuples in  $\prod_{V \in \mathfrak{S}_U} 2^{C^V}$  satisfying the conditions of Definition 4.1(5) for all  $V, W \in \mathfrak{S}_U$  with  $V \not\sqsubseteq W$ .

**Definition 4.21** (orthogonal partial tuple  $E_U$ ) Recall  $\mathfrak{S}_U^\perp = \{V \in \mathfrak{S} \mid V \perp U\}$ . Define  $E_U$  to be the set of tuples in  $\prod_{V \in \mathfrak{S}_U^\perp} 2^{C^V}$  satisfying the conditions of Definition 4.1(5) for all  $V, W \in \mathfrak{S}_U^\perp$  with  $V \not\sqsubseteq W$ .

**Definition 4.22** (product regions in  $\mathcal{X}$ ) Let  $U \in \mathfrak{S}$ . There exists  $\mu$  depending only on  $\mathfrak{S}$  such that for each  $(a_V)_{V \in \mathfrak{S}_U} \in F_U$  and  $(b_V)_{V \in \mathfrak{S}_U^\perp} \in E_U$ , there exists  $x \in \mathcal{X}$  such that for each  $V \in \mathfrak{S}$ :

- If  $V \sqsubseteq U$ , then  $d_V(x, a_V) \leq \mu$ .
- If  $V \perp U$ , then  $d_V(x, b_V) \leq \mu$ .
- If  $V \pitchfork U$  or  $U \sqsubseteq V$ , then  $d_V(x, \rho_V^U) \leq \mu$ .

Thus there is a map  $\phi_U: F_U \times E_U \rightarrow \mathcal{X}$ , whose image is  $k$ -hierarchically quasiconvex where  $k$  only depends on  $\mathfrak{S}$ . We call  $\phi_U(F_U \times E_U)$  the *product region for  $U$*  and denote it by  $P_U$ .

For any  $e \in E_U$  and  $f \in F_U$ , the sets  $\phi_U(F_U \times \{e\})$  and  $\phi_U(\{f\} \times E_U)$  will also be hierarchically quasiconvex; thus  $E_U$  and  $F_U$  are quasigeodesic metric spaces when equipped with the subspace metric from  $\phi_U(F_U \times \{e\})$  and  $\phi_U(\{f\} \times F_U)$ . While these metrics depend on the choice of  $e$  and  $f$ , the distance formula (Theorem 4.4) ensures that the different choices are all uniformly quasi-isometric.

The definition of the product regions ensure that they are not only uniformly hierarchically quasiconvex, but have easily described gate maps.

**Lemma 4.23** [11, Section 5] *Let  $(\mathcal{X}, \mathfrak{S})$  be an HHS. There exists  $k : [0, \infty) \rightarrow [0, \infty)$  such that for all  $U \in \mathfrak{S}$ , the product region  $\mathbf{P}_U$  is  $k$ -hierarchically quasiconvex. Moreover, there exists  $K \geq 0$  depending only on  $(\mathcal{X}, \mathfrak{S})$  such that for all  $x \in \mathcal{X}$ ,*

- $d_V(\mathbf{g}_{\mathbf{P}_U}(x), x) \leq K$  if  $V \sqsubseteq U$  or  $V \perp U$ ,
- $d_V(\mathbf{g}_{\mathbf{P}_U}(x), \rho_V^U) \leq K$  if  $V \pitchfork U$  or  $U \sqsubset V$ .

A version of our last result appeared as [10, Proposition 5.17]. However, that result contains an error in both its statement and its proof.<sup>3</sup> We provide a corrected statement and proof.

**Proposition 4.24** (active subpaths; corrected version of [10, Proposition 5.17]) *Let  $(\mathcal{X}, \mathfrak{S})$  be an HHS. There exist constants  $D, \nu, \lambda \geq 1$  such that for all  $x, y \in \mathcal{X}$ , if  $d_U(x, y) > D$  for some  $U \in \mathfrak{S}$ , then there exists a  $\lambda$ -hierarchy path  $\gamma : [a, b] \rightarrow \mathcal{X}$  joining  $x$  and  $y$  that has a subpath  $\alpha = \gamma|_{[a_1, b_1]}$  such that*

- (1)  $\alpha \subseteq N_\nu(\mathbf{P}_U)$ ;
- (2) the diameters of  $\pi_W(\gamma([a, a_1]))$  and  $\pi_W(\gamma([b_1, b]))$  are both bounded by  $\nu$ , for all  $W \in \mathfrak{S}_U \cup \mathfrak{S}_U^\perp$ ;
- (3) for any point  $p \in \gamma([a, a_1])$  or  $q \in \gamma([b_1, b])$ ,

$$d_{\mathcal{X}}(\mathbf{g}_{\mathbf{P}_U}(x), \mathbf{g}_{\mathbf{P}_U}(p)) \leq \nu \quad \text{and} \quad d_{\mathcal{X}}(\mathbf{g}_{\mathbf{P}_U}(y), \mathbf{g}_{\mathbf{P}_U}(q)) \leq \nu.$$

We call  $\alpha$  the **active subpath** of  $\gamma$  for  $U$ .

**Proof** Let  $\delta, E$ , and  $\kappa_0$  be the constants appearing in the HHS structure  $\mathfrak{S}$  for  $\mathcal{X}$ . Let  $x' = \mathbf{g}_{\mathbf{P}_U}(x)$  and  $y' = \mathbf{g}_{\mathbf{P}_U}(y)$ . Let  $\lambda_0 \geq 1$  be the constant such that every pair of points in  $\mathcal{X}$  can be joined by a  $\lambda_0$ -hierarchy path and  $\mu$  be the constant from Definition 4.22. Both  $\mu$  and  $\lambda_0$  depend only on  $(\mathcal{X}, \mathfrak{S})$ .

Let  $\gamma_0, \gamma_1$ , and  $\gamma_2$  be  $\lambda_0$ -hierarchy paths connecting the pairs  $(x, x')$ ,  $(x', y')$ , and  $(y', y)$  respectively. Let  $\gamma : [a, b] \rightarrow \mathcal{X}$  be the concatenation  $\gamma_0 * \gamma_1 * \gamma_2$ . We first verify that the path  $\gamma$  satisfies the requirements of the proposition with  $\alpha = \gamma_1$  and then verify that  $\gamma$  is in fact a hierarchy path with constant depending only on the HHS  $(\mathcal{X}, \mathfrak{S})$ .

---

<sup>3</sup>The error in the proof of [10, Proposition 5.17] is the incorrect claim that  $V \sqsubseteq U \implies \mathbf{P}_V \subseteq \mathbf{P}_U$ . The error in the statement is that *all* hierarchy paths have the stated properties instead of there existing at least one hierarchy path with the stated properties.

For the first item, let  $z \in \alpha = \gamma_1$ . By Lemma 4.23,  $\pi_W(\mathbf{g}_{P_U}(z))$  and  $\pi_W(z)$  are uniformly close for all  $W \in \mathfrak{S}_U \cup \mathfrak{S}_U^\perp$ . If  $W \notin \mathfrak{S}_U \cup \mathfrak{S}_U^\perp$ , then  $\pi_W(x')$ ,  $\pi_W(y')$ , and  $\pi_W(\mathbf{g}_{P_U}(z))$  are all  $\mu$ -close to  $\rho_W^U$  because  $x'$ ,  $y'$ , and  $\mathbf{g}_{P_U}(z)$  are all in  $P_U$ . Since  $\pi_W \circ \gamma_1$  is an unparametrized  $\lambda_0$ -quasigeodesic,  $\pi_W(z)$  must also be uniformly close to  $\rho_W^U$ . Therefore,  $d_W(\mathbf{g}_{P_U}(z), z)$  is uniformly bounded for all  $W \notin \mathfrak{S}_U \cup \mathfrak{S}_U^\perp$ . Since  $d_W(\mathbf{g}_{P_U}(z), z)$  is uniformly bounded for all  $W \in \mathfrak{S}$ , the distance formula (Theorem 4.4) provides  $\nu_1 \geq 0$  such that  $\gamma_1 \subseteq N_\nu(P_U)$ .

For the second item, if  $W \in \mathfrak{S}_U \cup \mathfrak{S}_U^\perp$ , then  $d_W(x, x')$  and  $d_W(y', y)$  are both uniformly bounded by Lemma 4.23. Since  $\pi_W \circ \gamma_0$  and  $\pi_W \circ \gamma_2$  are unparametrized  $(\lambda_0, \lambda_0)$ -quasigeodesics, there is a constant  $\nu_2 \geq 0$  satisfying the second item.

We prove the third item for  $p \in \gamma_0$  as the case  $q \in \gamma_2$  is identical. By the second item,  $d_W(x, p) \leq \nu_2$  for all  $W \in \mathfrak{S}_U \cup \mathfrak{S}_U^\perp$ . Since  $d_W(x, \mathbf{g}_{P_U}(x))$  and  $d_W(p, \mathbf{g}_{P_U}(p))$  are uniformly bounded for all  $W \in \mathfrak{S}_U \cup \mathfrak{S}_U^\perp$  as well (Lemma 4.23), we have that  $d_W(\mathbf{g}_{P_U}(x), \mathbf{g}_{P_U}(p))$  has a bound depending only on  $(\mathcal{X}, \mathfrak{S})$  for all  $W \in \mathfrak{S}_U \cup \mathfrak{S}_U^\perp$ . If instead  $U \sqsubset W$  or  $W \pitchfork U$ , then  $\pi_W(\mathbf{g}_{P_U}(x))$  and  $\pi_W(\mathbf{g}_{P_U}(p))$  are both uniformly close to  $\rho_W^U$  as they are points in the product region  $P_U$ . Hence  $d_W(\mathbf{g}_{P_U}(x), \mathbf{g}_{P_U}(p))$  is uniformly bounded for all  $W \in \mathfrak{S}$ . Thus, the distance formula provides  $\nu_3 \geq 0$  depending only on  $\mathfrak{S}$  such that  $d_{\mathcal{X}}(\mathbf{g}_{P_U}(x), \mathbf{g}_{P_U}(p)) \leq \nu_3$ .

Set  $\nu = \max\{\nu_1, \nu_2, \nu_3\}$ . This depends only on  $(\mathcal{X}, \mathfrak{S})$  since each of the  $\nu_i$  depend only on  $(\mathcal{X}, \mathfrak{S})$ . It remains to show that  $\gamma$  is a hierarchy path with constant depending only on  $(\mathcal{X}, \mathfrak{S})$ . For this we need to assume that  $d_U(x, y) > 10(E + \kappa_0)$ .

We first show that  $\pi_W \circ \gamma$  is a uniform unparametrized quasigeodesic for each  $W \in \mathfrak{S}$ .

- If  $W \in \mathfrak{S}_U \cup \mathfrak{S}_U^\perp$ , then  $\text{diam}(\pi_W(\gamma_0)) \leq \nu$ ,  $\text{diam}(\pi_W(\gamma_2)) \leq \nu$ , and  $\pi_W \circ \gamma_1$  is an unparametrized  $(\lambda_0, \lambda_0)$ -quasigeodesic. Hence  $\pi_W \circ \gamma$  is an unparametrized  $(\lambda_0, \lambda_0 + 2\nu)$ -quasigeodesic.
- If  $U \sqsubset W$ , then by the bounded geodesic image axiom (8) any  $CW$ -geodesic from  $\pi_W(x)$  to  $\pi_W(y)$  must intersect the  $E$ -neighborhood of  $\rho_W^U$ . Since all of  $\pi_W \circ \gamma_1$  is contained in  $N_{\lambda_0(E+\mu)+\lambda_0}(\rho_W^U)$ , the hyperbolicity of  $CW$  implies that both of the unparametrized quasigeodesics  $\pi_W \circ \gamma_0$  and  $\pi_W \circ \gamma_2$  are contained in a regular neighborhood of a  $CW$ -geodesic from  $\pi_W(x)$  to  $\pi_W(y)$ . Thus  $\pi_W \circ \gamma$  will be a unparametrized quasigeodesic with constants depending on  $\lambda_0, \mu, E$ , and  $\delta$ .
- If  $W \pitchfork U$ , then since  $d_U(x, y) > 10(E + \kappa_0)$ , the consistency axiom (5) ensures that at most one of  $d_W(x, \rho_W^U)$  and  $d_W(y, \rho_W^U)$  are larger than  $\kappa_0$ . Without loss of

generality, assume  $d_W(x, \rho_W^U) \leq \kappa_0$ . Since  $\pi_W(x')$  and  $\pi_W(y')$  are  $\mu$ -close to  $\rho_W^U$  and  $\gamma_0$  and  $\gamma_1$  are both  $\lambda_0$ -hierarchy paths, the diameter of  $\pi_W(\gamma_0) \cup \pi_W(\gamma_1)$  is at most

$$2\lambda_0(3E + \mu + \kappa_0) + 2\lambda_0.$$

This makes  $\pi_W \circ \gamma$  an unparametrized  $(\lambda_0, 2\lambda_0(3E + \mu + \kappa_0) + 3\lambda_0)$ -quasigeodesic.

The above shows that there exists  $\lambda' \geq 1$  depending only on  $(\mathcal{X}, \mathfrak{S})$  such that  $\pi_W \circ \gamma$  is an unparametrized  $(\lambda', \lambda')$ -quasigeodesic for all  $W \in \mathfrak{S}$ .

Finally we show that  $\gamma: [a, b] \rightarrow \mathcal{X}$  is a quasigeodesic with constants depending only on  $(\mathcal{X}, \mathfrak{S})$ . Let  $t, s \in [a, b]$  and let  $u = \gamma(t)$  and  $v = \gamma(s)$ . Since  $\gamma_0, \gamma_1$  and  $\gamma_2$  are all  $(\lambda_0, \lambda_0)$ -quasigeodesics, we can assume  $u$  and  $v$  do not lie in the same  $\gamma_i$ . Without loss of generality we have two cases.

In the first case,  $u \in \gamma_0$  and  $v \in \gamma_1$ . Since  $\pi_W \circ \gamma$  is a uniform unparametrized quasigeodesic, there exists  $C \geq 1$  such that

$$d_W(u, x') + d_W(x', v) \asymp_{C,C} d_W(u, v)$$

for all  $W \in \mathfrak{S}$ . By Lemma 4.5, there is a  $K \geq 1$  depending only on  $(\mathcal{X}, \mathfrak{S})$  such that

$$d_{\mathcal{X}}(u, x') + d_{\mathcal{X}}(x', v) \asymp_{K,K} d_{\mathcal{X}}(u, v),$$

which implies

$$\frac{1}{\lambda_0 K} |t - s| - \frac{2\lambda_0}{K} - K \leq d_{\mathcal{X}}(\gamma(t), \gamma(s)) \leq \lambda_0 |t - s| + 2\lambda_0$$

because  $\gamma_0$  and  $\gamma_1$  are  $(\lambda_0, \lambda_0)$ -quasigeodesics.

The second case is when  $u \in \gamma_0$  and  $v \in \gamma_2$ . The proof is the same as the first case using the fact that

$$d_W(u, x') + d_W(x', y') + d_W(y', v) \asymp d_W(u, v)$$

for all  $W \in \mathfrak{S}$  instead. Hence  $\gamma$  is a quasigeodesic with constants depending only on  $(\mathcal{X}, \mathfrak{S})$ , as desired. □

## 4.2 Summary of constants

Before continuing we summarize the constants associated to the hierarchically hyperbolic space  $(\mathcal{X}, \mathfrak{S})$  that we will utilize frequently.

- $\delta$  is the hyperbolicity constant of  $CW$  for each  $W \in \mathfrak{S}$ .

- $\kappa_0$  is the consistency constant from axiom (5).
- $E$  is the bound on projections in axioms (1), (5) and (8).
- $\sigma_0$  is the minimal threshold constant from the distance formula (Theorem 4.4).
- $\lambda_0$  is the constant such that any two points in  $\mathcal{X}$  can be joined by a  $\lambda_0$ -hierarchy path (Theorem 4.8).
- $\chi$  is the constant from Proposition 4.6 which bounds the cardinality of any subset of  $\mathfrak{S}$  that does not contain a pair of transverse domains.
- $\theta_0$  is the constant such that for all  $\theta \geq \theta_0$  and  $Y \subset \mathcal{X}$ ,  $H_\theta(Y)$  is hierarchically quasiconvex (Lemma 4.13).

We can and shall assume that  $E \geq \kappa_0$  and  $E \geq \delta$ . When we say that a quantity depends on  $\mathfrak{S}$ , we mean that it depends on any of the above constants.

## 5 Constructing hulls with hierarchy paths

In this section, we study hierarchically quasiconvex hulls in hierarchically hyperbolic spaces. The main result is Theorem 5.2 which says that the hierarchically quasiconvex hull can be constructed by iteratively connecting points with hierarchy paths. While our motivation for such a construction is to establish that strongly quasiconvex subsets are hierarchically quasiconvex (Proposition 5.7) we believe it will have many other applications. At the end of the section, we give an example of such an application by characterizing hierarchical quasiconvexity in terms of the coarse median structure on a hierarchically hyperbolic space.

**Definition 5.1** (hierarchy path hull) Let  $Y$  be a subset of the hierarchically hyperbolic space  $(\mathcal{X}, \mathfrak{S})$ . Define  $\mathcal{P}_\lambda^1(Y)$  to be the union of all  $\lambda$ -hierarchy paths between points in  $Y$ . Inductively define  $\mathcal{P}_\lambda^n(Y) = \mathcal{P}_\lambda^1(\mathcal{P}_\lambda^{n-1}(Y))$  for all integers  $n \geq 2$ . For all  $\lambda \geq \lambda_0$  and  $n \geq 1$ ,  $\mathcal{P}_\lambda^n(Y) \neq \emptyset$ .

**Theorem 5.2** (constructing hulls using hierarchy paths) Let  $(\mathcal{X}, \mathfrak{S})$  be a hierarchically hyperbolic space and  $N = 2\chi$ , where  $\chi$  is as in Proposition 4.6. There exist  $\bar{\theta} \geq \theta_0$  and  $\bar{\lambda} \geq \lambda_0$  depending only on  $\mathfrak{S}$  such that for all  $\theta \geq \bar{\theta}$ ,  $\lambda \geq \bar{\lambda}$  and  $Y \subseteq \mathcal{X}$ ,

$$d_{\text{Haus}}(\mathcal{P}_\lambda^N(Y), H_\theta(Y)) < D$$

where  $D$  depends only on  $\theta$ ,  $\lambda$ , and  $\mathfrak{S}$ .

In a recent paper, Bowditch [16] independently constructs hulls in coarse median spaces in a similar manner to the construction in Definition 5.1. Hierarchically hyperbolic spaces are one of the primary examples of coarse median spaces and [16, Lemma 7.3] establishes a version of Theorem 5.2 for finite subsets of hierarchically hyperbolic spaces. At the end of this section we show that Bowditch's coarse median hull is coarsely equal to the hierarchical quasiconvex hull for any subset of an HHS. This is achieved by using Theorem 5.2 to give a new characterization of the hierarchical quasiconvexity in terms of the coarse median structure on a hierarchically hyperbolic space.

The number of iterations of connecting pairs of points by hierarchy paths required by Theorem 5.2 is unlikely to be optimal. However, a simple example illustrates that the number of iteration required must increase with the maximal number of pairwise orthogonal domains. Consider the group  $\mathbb{Z}^n$  with the standard HHG structure. Let  $Y$  be the union of the positive halves of each of the coordinate axes. The hull  $H_\theta(Y)$  then coarsely coincides with the positive orthant of  $\mathbb{Z}^n$ , but  $\mathcal{P}_\lambda^m(Y)$  coarsely coincides with the set of points in the positive orthant where at most  $2^m$  coordinates are nonzero. Thus, the number of iterations of  $\mathcal{P}_\lambda^1(\cdot)$  required to achieve  $H_\theta(Y)$  will be approximately  $\log(n)$ .

For the remainder of this section, let  $(\mathcal{X}, \mathfrak{S})$  be a hierarchically hyperbolic space and  $Y \subseteq \mathcal{X}$ . Recall, there exist  $\theta_0$  and  $\lambda_0$  such that for all  $\theta \geq \theta_0$ ,  $H_\theta(Y)$  is hierarchically quasiconvex (Lemma 4.13) and any two points in  $\mathcal{X}$  can be joined by a  $\lambda_0$ -hierarchy path (Theorem 4.8).

The following lemma can be found in [10, Proposition 6.4.4] and says for sufficiently large  $\theta$ , all hierarchically quasiconvex hulls coarsely coincide. We record the proof for completeness.

**Lemma 5.3** [10, Proposition 6.4.4] *There exists  $\bar{\theta} \geq \theta_0$  depending only on  $\mathfrak{S}$ , such that for all  $\theta_1, \theta_2 \geq \bar{\theta}$ ,*

$$d_{\text{Haus}}(H_{\theta_1}(Y), H_{\theta_2}(Y)) \leq D,$$

where  $D$  depends on  $\theta_1$  and  $\theta_2$ .

**Proof** Without loss of generality, assume  $\theta_0 < \bar{\theta} \leq \theta_1 < \theta_2$  with  $\bar{\theta}$  to be determined below. By definition,  $H_{\theta_1}(Y) \subseteq H_{\theta_2}(Y)$ . Let  $x \in H_{\theta_2}(Y)$ . For each  $U \in \mathfrak{S}$ ,  $\pi_U(H_{\theta_0}(Y))$  is  $K$ -quasiconvex, where  $K$  depends on  $\theta_0$  and  $\delta$ . Let  $y_U$  be the closest point projection of  $\pi_U(x)$  onto  $\pi_U(H_{\theta_0}(Y))$ . By Lemma 4.11, there exist  $y \in \mathcal{X}$



and  $\theta'$  depending on  $\theta_0$  and  $\mathfrak{S}$  such that  $d_U(\pi_U(y), y_U) \leq \theta'$ . In particular, setting  $\bar{\theta} = \theta_0 + \theta'$ , we have  $y \in H_{\bar{\theta}}(Y) \subseteq H_{\theta_1}(Y)$ . To bound  $d_{\mathcal{X}}(x, y)$ , we will uniformly bound  $d_U(x, y_U)$  in terms of  $\theta_2$  for every  $U \in \mathfrak{S}$ ; the bound on  $d_{\mathcal{X}}(x, y)$  will then follow from the distance formula (Theorem 4.4). By the definition of  $y_U$  we have  $d_U(x, y_U) \leq d_U(x, \pi_U(H_{\theta_0}(Y))) + 1$ . Since  $\pi_U(H_{\theta_0}(Y))$  is quasiconvex, contains  $Y$ , and is contained in the  $\theta_0$ -neighborhood of  $\text{hull}_{CU}(Y)$ , there exists a  $D'$  depending only on  $\mathfrak{S}$  such that  $\text{hull}_{CU}(Y) \subseteq N_{D'}(\pi_U(H_{\theta_0}(Y)))$ . Since  $d_U(x, \text{hull}_{CU}(Y)) \leq \theta_2$ ,

$$d_U(x, y_U) \leq d_U(x, \pi_U(H_{\theta_0}(Y))) + 1 \leq \theta_2 + D' + 1,$$

providing the result. □

For the remainder of this section,  $\bar{\theta}$  will denote the constant from Lemma 5.3.

To prove Theorem 5.2 we will show for sufficiently large  $\theta$  and  $\lambda$ , we can find  $\theta' > \theta$  and  $\lambda' > \lambda$  such that

$$\mathcal{P}_{\lambda}^N(Y) \subseteq H_{\theta'}(Y) \quad \text{and} \quad H_{\theta}(Y) \subseteq \mathcal{P}_{\lambda'}^N(Y).$$

Theorem 5.2 will then follow by applying Lemma 5.3. The inclusion  $\mathcal{P}_{\lambda}^N(Y) \subseteq H_{\theta'}(Y)$  is the following direct consequence of hierarchical quasiconvexity.

**Lemma 5.4** *For each  $\lambda, n \geq 1$ , there exists  $\theta \geq \bar{\theta}$  such that for any  $Y \subseteq \mathcal{X}$ ,*

$$\mathcal{P}_{\lambda}^n(Y) \subseteq H_{\theta}(Y).$$

**Proof** The  $n = 1$  case follows directly from the definition of  $H_{\theta}(Y)$  and hierarchy paths. We can proceed by induction on  $n$  and assume there exists  $\theta' \geq \bar{\theta}$  such that  $\mathcal{P}_{\lambda}^{n-1}(Y) \subseteq H_{\theta'}(Y)$ . Let  $x \in \mathcal{P}_{\lambda}^n(Y)$ . There exist  $y_1, y_2 \in \mathcal{P}_{\lambda}^{n-1}(Y)$  such that  $x$  is on a  $\lambda$ -hierarchy path from  $y_1$  to  $y_2$ . For each  $U \in \mathfrak{S}$ ,  $\pi_U(y_i)$  is within  $\theta'$  of  $\text{hull}_{CU}(Y)$ . Therefore, quasiconvexity of  $\text{hull}_{CU}(Y)$  in  $CU$  guarantees there exists a  $\theta$  depending only on  $\lambda$  and  $\theta'$  (which in turn depends on  $n$ ) such that  $\pi_U(x)$  is within  $\theta$  of  $\text{hull}_{CU}(Y)$  and thus  $x \in H_{\theta}(Y)$ . □

The other inclusion,  $H_{\theta}(Y) \subseteq \mathcal{P}_{\lambda}^N(Y)$ , requires two main steps. First we prove that if  $x \in H_{\theta}(Y)$ , then there exists at most  $2\chi + 1$  points,  $x_1, \dots, x_n$ , in  $Y$  such that  $x \in H_{\theta'}(x_1, \dots, x_n)$  where  $\theta'$  depends only on  $\theta$  (Lemma 5.5). We then show that for any finite collection of points  $x_1, \dots, x_n \in \mathcal{X}$ ,  $H_{\theta'}(x_1, \dots, x_n) \subseteq \mathcal{P}_{\lambda}^{n-1}(x_1, \dots, x_n)$  where  $\lambda$  ultimately depends only on  $n$  and  $\theta$  (Proposition 5.6). Together, these imply  $H_{\theta}(Y) \subseteq \mathcal{P}_{\lambda}^{2\chi+1}(Y)$ .

We start with the first step, which can be thought of a version of Carathéodory's theorem for HHSs.

**Lemma 5.5** *Let  $Y \subseteq \mathcal{X}$ ,  $\theta \geq \bar{\theta}$ , and  $\chi$  be as in Proposition 4.6. For each  $x \in H_\theta(Y)$ , there exist  $x_1, \dots, x_{\ell+1} \in Y$ , where  $1 \leq \ell \leq 2\chi$ , and  $\theta'$  depending only on  $\theta$  such that  $x \in H_{\theta'}(x_1, \dots, x_{\ell+1})$ .*

**Proof** Let  $K = 100(E + 2\kappa_0 + \theta)$  and  $x \in H_\theta(Y)$ . If for all  $y \in Y$ ,  $\text{Rel}_K(x, y) = \emptyset$ , then  $x \in H_K(y)$  for each  $y \in Y$ . Thus we can assume there is  $y \in Y$  such that  $\text{Rel}_K(x, y) \neq \emptyset$ .

As in Proposition 4.6, we can partition  $\text{Rel}_K(x, y)$  in subsets  $\mathcal{U}_1, \dots, \mathcal{U}_n$  where  $n \leq \chi$ . Further, for each  $i$ , all the elements of  $\mathcal{U}_i$  are pairwise transverse and are totally ordered with respect to the order  $U \leq V$  if  $d_U(\rho_U^V, y) \leq \kappa_0$ . Let  $U_{i,1} < \dots < U_{i,k_i}$  be the distinct domains in  $\mathcal{U}_i$ . For each  $i$ , there exist  $a_i, b_i \in Y$  such that  $\pi_{U_{i,1}}(x)$  is within  $\theta$  of the  $CU_{i,1}$  geodesic between  $a_i$  and  $b_i$ . If  $a_i$  and  $b_i$  project  $2\kappa_0 + E$  close to  $y$  in  $CU_{i,1}$ , then  $d_{U_{i,1}}(x, y) \leq \theta + 4\kappa_0 + 3E$  which contradicts  $U_{i,1} \in \text{Rel}_K(x, y)$ . Thus without loss of generality,  $d_{U_{i,1}}(a_i, y) > 2\kappa_0 + E$  and in particular  $d_{U_{i,1}}(a_i, \rho_{U_{i,1}}^{U_{i,j}}) > \kappa_0$  for all  $j > 1$ . The total order on  $\mathcal{U}_i$  and the consistency axiom (5) ensure that  $d_{U_{i,j}}(x, a_i) \leq 2\kappa_0 + E$  for all  $1 < j \leq k_i$ . Thus for each  $U_{i,j}$ ,  $x$  projects  $\theta + 2\kappa_0 + E$  close to the  $CU_{i,j}$  geodesic between  $a_i$  and  $b_i$  and  $x \in H_K(y, a_1, \dots, a_n, b_1, \dots, b_n)$ .  $\square$

Armed with Lemma 5.5, the next step is to prove that for a finite set of points, the hierarchical hull is contained in the path hull.

**Proposition 5.6** *For each  $\theta \geq \bar{\theta}$  and  $n \geq 2$ , there exists  $\lambda \geq 1$  such that*

$$H_\theta(x_1, \dots, x_n) \subseteq \mathcal{P}_\lambda^{n-1}(x_1, \dots, x_n)$$

for any  $n$  distinct points  $x_1, \dots, x_n \in \mathcal{X}$ .

**Proof** We shall proceed by induction on  $n$ . First we will show the base case of  $n = 2$ .

**Claim 1** (base case) *For each  $\theta \geq \bar{\theta}$  there exists  $\lambda \geq 1$  such that*

$$H_\theta(x, y) \subseteq \mathcal{P}_\lambda^1(x, y)$$

for each  $x, y \in \mathcal{X}$ .

**Proof of Claim 1** Let  $z \in H_\theta(x, y)$ ,  $\gamma_0: [a, b] \rightarrow \mathcal{X}$  be a  $\lambda_0$ -hierarchy path from  $x$  to  $z$  and  $\gamma_1: [b, c] \rightarrow \mathcal{X}$  is a  $\lambda_0$ -hierarchy path from  $z$  to  $y$ . We will show that

$\gamma = \gamma_0 * \gamma_1 : [a, c] \rightarrow \mathcal{X}$  is a  $\lambda$ -hierarchy path from  $x$  to  $y$ , where  $\lambda$  depends only on  $\theta$ . By the definition of  $H_\theta(x, y)$  and hyperbolicity of the  $CU$ 's we have that  $\pi_U(\gamma)$  is an unparametrized  $(\lambda_1, \lambda_1)$ -quasigeodesic for each  $U \in \mathfrak{S}$ , where  $\lambda_1$  depends only on  $\theta$ . Therefore, it suffices to show that  $\gamma$  is a  $(\lambda, \lambda)$ -quasigeodesic in  $\mathcal{X}$ , where  $\lambda$  depends only on  $\theta$ . That is, we need to prove for each  $t, s \in [a, c]$ ,

$$|t - s| \asymp_{\lambda, \lambda} d_{\mathcal{X}}(\gamma(t), \gamma(s)).$$

Since  $\gamma_0$  and  $\gamma_1$  are both  $(\lambda_0, \lambda_0)$ -quasigeodesics, we can restrict ourselves to the case where  $t \in [a, b)$  and  $s \in (b, c]$ . Let  $u = \gamma(t)$  and  $v = \gamma(s)$ . Since  $\pi_U(\gamma)$  is a uniform unparametrized quasigeodesic for each  $U \in \mathfrak{S}$ ,

$$d_U(u, z) + d_U(z, v) \asymp_{C, C} d_U(u, v)$$

where  $C \geq 1$  depends only on  $\theta$ . Hence, Lemma 4.5 provides a constant  $K \geq 1$  depending only on  $\theta$  such that

$$d_{\mathcal{X}}(u, z) + d_{\mathcal{X}}(z, v) \asymp_{K, K} d_{\mathcal{X}}(u, v).$$

Since  $\gamma_0$  and  $\gamma_1$  are both  $(\lambda_0, \lambda_0)$ -quasigeodesics,

$$\frac{1}{\lambda_0 K} |t - s| - \frac{2\lambda_0}{K} - k \leq d_{\mathcal{X}}(\gamma(t), \gamma(s)) \leq \lambda_0 |t - s| + 2\lambda_0,$$

as desired. □

We now show the key fact for the inductive step, that the hull of  $n$  points can be obtained by taking the hull on  $n - 1$  points, and then considering all the hierarchy paths between this smaller hull and the remaining point.

**Claim 2** *Let  $x_1, \dots, x_n \in \mathcal{X}$ , for  $n \geq 2$ . If  $x \in H_\theta(x_1, \dots, x_n)$  where  $\theta \geq \bar{\theta}$ , then there exist  $\theta'$  and  $\lambda$  depending only on  $\theta$  and  $y \in H_{\theta'}(x_1, \dots, x_{n-1})$  such that  $x$  is on a  $\lambda$ -hierarchy path from  $x_n$  to  $y$ .*

**Proof of Claim 2** For  $1 \leq i \leq n$ , let  $A_i = \{x_1, \dots, x_i\}$ . For each  $U \in \mathfrak{S}$ ,  $\pi_U(H_\theta(A_{n-1}))$  is  $R$ -quasiconvex where  $R$  depends only on  $\theta$ . Let  $y_U$  be the closest point projection of  $\pi_U(x)$  to  $\pi_U(H_\theta(A_{n-1}))$ ,  $z_U$  be a point in  $\text{hull}_{CU}(A_n)$  within  $\theta$  of  $\pi_U(x)$ , and  $z'_U$  be the closest point projection of  $z_U$  to  $\pi_U(H_\theta(A_{n-1}))$ . By Lemma 4.11, there exist  $y \in \mathcal{X}$  and a constant  $\theta'$  depending on  $\theta$  and  $\delta$  such that  $d_U(\pi_U(y), y_U) \leq \theta'$ . Further, we can assume  $\theta'$  is large enough that

- (1)  $\theta' > \theta + \delta + R + 1$ ;
- (2)  $y \in H_{\theta'}(A_{n-1})$ ;

- (3) for all  $v, w \in CU$ , if  $d_U(v, w) < d_U(v, H_\theta(A_{n-1}))$ , then the closest point projection of  $v$  and  $w$  to  $\pi_U(H_\theta(A_{n-1}))$  are no more than  $\theta'$  apart.

For each  $U \in \mathfrak{S}$ , let  $\gamma_U$  be a  $CU$  geodesic from  $\pi_U(x_n)$  to  $\pi_U(y)$ . We will show that  $d_U(x_n, \gamma_U)$  is uniformly bounded for each  $U \in \mathfrak{S}$ . If  $d_U(y_U, z_U) \leq 5\theta'$ , then  $d_U(x, y_U) \leq 6\theta'$  which implies  $d_U(x, \gamma_U) \leq 7\theta'$ . Otherwise  $d_U(y_U, z_U) > 5\theta'$  implies that  $d_U(x, H_\theta(A_{n-1})) > d_U(x, z_U)$  and thus  $d_U(y_U, z'_U) \leq \theta'$  by (3). This implies that  $d_U(z_U, H_\theta(A_{n-1})) > 3\theta'$ . Since  $z_U \in \text{hull}_{CU}(A_n)$  and  $z_U \notin H_\theta(A_{n-1})$ , there exist  $D \geq 0$  depending only on  $\theta$  and  $x_U \in \pi_U(A_{n-1})$  such that  $z_U$  is within  $D$  of any  $CU$  geodesic from  $\pi_U(x_n)$  to  $x_U$ . Further, by increasing  $\theta'$ , we can assume  $D < \theta'$ . Take a geodesic triangle with endpoints  $\pi_U(x_n)$ ,  $y_U$  and  $x_U$ . Since  $d_U(z_U, H_\theta(A_{n-1})) > 3\theta'$ , it must be the case that  $z_U$  is within  $2\theta'$  of any  $CU$  geodesic from  $\pi_U(x_n)$  to  $y_U$ .

Thus there exists  $\theta''$  depending ultimately only on  $\theta$ , such that  $d_U(x, \gamma_U) \leq \theta''$  for all  $U \in \mathfrak{S}$ . Therefore  $x \in H_{\theta''}(x_n, y)$  and the statement in Claim 2 follows from Claim 1. □

We now finish the proof of Proposition 5.6. Let  $x \in H_\theta(x_1, \dots, x_n)$ . Claim 2 shows that there exist a  $\lambda' \geq 1$  and  $\theta' \geq \bar{\theta}$  such that  $x$  is on a  $\lambda'$ -hierarchy path from  $x_n$  to a point in  $H_{\theta'}(x_1, \dots, x_{n-1})$ . By induction, there exists  $\lambda \geq \lambda'$  such that  $H_{\theta'}(x_1, \dots, x_{n-1}) \subseteq \mathcal{P}_\lambda^{n-2}(x_1, \dots, x_{n-1})$  and therefore  $x \in \mathcal{P}_\lambda^{n-1}(x_1, \dots, x_n)$ . □

We can now finish the proof of Theorem 5.2.

**Proof of Theorem 5.2** Recall, we need to show that for all sufficiently large  $\theta$  and  $\lambda$ ,  $H_\theta(Y)$  coarsely coincides with  $\mathcal{P}_\lambda^N(Y)$  where  $N = 2\chi$ . First we will show that for all  $\theta \geq \bar{\theta}$ , there exists  $\lambda \geq 1$  such that  $H_\theta(Y) \subseteq \mathcal{P}_\lambda^N(Y)$ .

Let  $x \in H_\theta(Y)$  and let  $x_1, \dots, x_{\ell+1}$  be the finite number of points in  $Y$  provided by Lemma 5.5. By Proposition 5.6, there exists  $\lambda$  depending on  $\theta$  such that

$$x \in \mathcal{P}_\lambda^\ell(x_1, \dots, x_{\ell+1}) \subseteq \mathcal{P}_\lambda^\ell(Y) \subseteq \mathcal{P}_\lambda^N(Y).$$

Thus  $H_\theta(Y) \subseteq \mathcal{P}_\lambda^N(Y)$ .

Now, fix  $\bar{\lambda} \geq \lambda_0$  such that  $H_{\bar{\theta}}(Y) \subseteq \mathcal{P}_{\bar{\lambda}}^N(Y)$ . If  $\theta \geq \bar{\theta}$  and  $\lambda \geq \bar{\lambda}$ , then by Lemma 5.4 there exists  $\theta' > \bar{\theta}$  such that

$$H_{\bar{\theta}}(Y) \subseteq \mathcal{P}_{\bar{\lambda}}^N(Y) \subseteq H_{\theta'}(Y).$$

The conclusion now follows by Lemma 5.3. □

The primary use of Theorem 5.2 in this paper is the following proof that hierarchically quasiconvex subsets are exactly the subsets that are “quasiconvex with respect to hierarchy paths”. From this it immediately follows that all strongly quasiconvex subsets are hierarchically quasiconvex.

**Proposition 5.7** *Let  $(\mathcal{X}, \mathfrak{S})$  be a hierarchically hyperbolic space. A subset  $Y \subseteq \mathcal{X}$  is  $k$ -hierarchically quasiconvex if and only if there exists a function  $R: [1, \infty) \rightarrow [0, \infty)$  such that if  $\gamma$  is a  $\lambda$ -hierarchy path with endpoints on  $Y$ , then  $\gamma \subseteq N_{R(\lambda)}(Y)$  where  $k$  and  $R$  each determines the other. In particular, if  $Y$  is  $Q$ -strongly quasiconvex, then  $Y$  is  $k$ -hierarchically quasiconvex where  $k$  is determined by  $Q$ .*

**Proof** The proof of the forward implication follows directly from the definition of hierarchical quasiconvexity and hierarchy path. Assume there exists a function  $R: [1, \infty) \rightarrow [0, \infty)$  such that if  $\gamma$  is a  $\lambda$ -hierarchy path with endpoints in  $Y$ , then  $\gamma \subseteq N_{R(\lambda)}(Y)$ . The first condition of hierarchical quasiconvexity now follows from the existence of hierarchy paths (Theorem 4.8), the coarse Lipschitzness of the projection maps (axiom (1)), and the hyperbolicity of the  $CU$ 's. For the second condition, observe that the hypothesis implies there exists a bound on the Hausdorff distance between  $Y$  and  $\mathcal{P}_\lambda^n(Y)$  depending only on  $R$ ,  $n$ , and  $\lambda$ . Thus by Theorem 5.2, for each  $\theta \geq \bar{\theta}$ , there exists  $D_\theta$  such that  $d_{\text{Haus}}(H_\theta(Y), Y) \leq D_\theta$ . Let  $\kappa > 0$  and  $x \in \mathcal{X}$  such that  $d_U(x, Y) \leq \kappa$  for all  $U \in \mathfrak{S}$ . Thus  $x \in H_\theta(Y)$  for each  $\theta \geq \kappa + \bar{\theta}$ . Let  $k(\kappa) = D_{\bar{\theta} + \kappa}$ . Then  $d_{\mathcal{X}}(x, Y) \leq k(\kappa)$  and  $Y$  is hierarchically quasiconvex.  $\square$

**Remark 5.8** If  $\mathcal{X}$  is a hyperbolic space, there exist many HHS structures on  $\mathcal{X}$ ; see [51]. In this case, Proposition 5.7 recovers [51, Proposition 3.5], which states that a subset  $Y \subseteq \mathcal{X}$  is quasiconvex if and only if  $Y$  is hierarchically quasiconvex in any of the HHS structures on  $\mathcal{X}$ .

## 5.1 Hulls and coarse medians

We now take a small detour from the main thrust of the paper to highlight an application of Theorem 5.2 and discuss the relation of our work in this section to the hulls in coarse median spaces constructed in [16].

In [14], Bowditch axiomatized the notion of a coarse center of three points in a metric space and defined *coarse median spaces* as metric spaces where every triple of points has such a coarse center. Bowditch observed that all hierarchically hyperbolic spaces

are coarse median spaces; see also [10, Theorem 7.3]. The salient property of the coarse median structure of an HHS is the following fact.

**Lemma 5.9** (see proof of [10, Theorem 7.3]) *Let  $(\mathcal{X}, \mathfrak{S})$  be a hierarchically hyperbolic space. There exist  $\mu > 0$  and a map  $m: \mathcal{X} \times \mathcal{X} \times \mathcal{X} \rightarrow \mathcal{X}$  with the property that for every  $(x, y, z) \in \mathcal{X}^3$  and  $U \in \mathfrak{S}$ , the projection  $\pi_U(m(x, y, z))$  is within  $\mu$  of all three sides of any  $CU$  triangle with vertices  $\pi_U(x)$ ,  $\pi_U(y)$  and  $\pi_U(z)$ .*

We call the point  $m(x, y, z)$  the *coarse center* of  $x$ ,  $y$  and  $z$ . There is a natural notion of convexity for coarse median spaces, which we formulate in the hierarchically hyperbolic setting as follows.

**Definition 5.10** (coarse median quasiconvexity) *Let  $(\mathcal{X}, \mathfrak{S})$  be an HHS. A subset  $Y$  of  $\mathcal{X}$  is said to be  $Q$ -median quasiconvex if for every  $y, y' \in Y$  and  $x \in \mathcal{X}$  we have  $m(y, y', x) \in N_Q(Y)$ .*

Behrstock, Hagen and Sisto showed that a hierarchically quasiconvex subset is median quasiconvex in [10, Proposition 7.12]. Using Theorem 5.2, we establish the converse.

**Proposition 5.11** *Let  $(\mathcal{X}, \mathfrak{S})$  be an HHS and  $Y \subseteq \mathcal{X}$ .  $Y$  is  $k$ -hierarchically quasiconvex if and only if  $Y$  is  $Q$ -median quasiconvex where  $k$  and  $Q$  each determines the other.*

**Proof** Let  $Y$  be a  $Q$ -median quasiconvex subset of the HHS  $(\mathcal{X}, \mathfrak{S})$  and  $\gamma$  be a  $\lambda$ -hierarchy path with endpoints  $y_1, y_2 \in Y$ . If  $x \in \gamma$ , then  $d_U(x, m(y_1, y_2, x))$  is uniformly bounded in terms of  $\lambda$  and  $\mathfrak{S}$  for each  $U \in \mathfrak{S}$ . By the distance formula (Theorem 4.4),  $d_{\mathcal{X}}(x, m(y_1, y_2, x))$  is also uniformly bounded. Since  $Y$  is median quasiconvex, this implies that there exist  $R(\lambda)$  such that  $d_{\mathcal{X}}(x, Y) \leq R(\lambda)$ . In particular,  $\gamma \subseteq N_{R(\lambda)}(Y)$  and  $Y$  is  $k$ -hierarchically quasiconvex, with  $k$  determined by  $Q$ , by Proposition 5.7.  $\square$

If  $Y \subseteq \mathcal{X}$ , let  $M(Y)$  denote the *coarse median hull* defined in [16, Proposition 6.2]. Proposition 5.11 implies the following corollary that extends [16, Lemma 7.3] in the special case of hierarchically hyperbolic spaces.

**Corollary 5.12** *Let  $(\mathcal{X}, \mathfrak{S})$  be an HHS and  $Y \subseteq \mathcal{X}$ . For each  $\theta \geq \theta_0$ , there exists  $D$  depending only on  $\theta$  and  $\mathfrak{S}$  such that*

$$d_{\text{Haus}}(H_{\theta}(Y), M(Y)) \leq D.$$

**Proof** Let  $Y \subseteq \mathcal{X}$  and  $\theta \geq \theta_0$ . By Proposition 5.11,  $H_\theta(Y)$  is  $Q_1$ -median quasiconvex for some  $Q_1$  depending on  $\theta$  and  $\mathfrak{S}$ . By [16, Proposition 6.2]  $M(Y)$  is  $Q_2$ -median quasiconvex, where  $Q_2$  depends only on  $\mathfrak{S}$ , and there exists  $D_1$  depending on  $\theta$  such that  $M(Y) \subseteq N_{D_1}(H_\theta(Y))$ . By Proposition 5.11,  $M(Y)$  is  $k$ -hierarchically quasiconvex where  $k$  depends only on  $\mathfrak{S}$ . By the second condition in Definition 4.9, there exists  $D_2$  depending on  $\theta$  and  $\mathfrak{S}$  such that  $H_\theta(Y) \subseteq N_{D_2}(M(Y))$ .  $\square$

## 6 Characterization of strongly quasiconvex subsets in HHSs

We now turn our attention to the main objective of this paper, characterizing the strongly quasiconvex subsets of hierarchically hyperbolic spaces. From now on we shall restrict our attention to HHSs with the *bounded domain dichotomy*; a minor regularity condition satisfied by all HHGs as well as Teichmüller space with either the Weil–Peterson or Teichmüller metric and the fundamental groups of 3-manifolds without Nil or Sol components.

**Definition 6.1** (bounded domain dichotomy) A hierarchically hyperbolic space  $(\mathcal{X}, \mathfrak{S})$  has the *B-bounded domain dichotomy* if there exists  $B > 0$  such that for all  $U \in \mathfrak{S}$ , if  $\text{diam}(CU) > B$ , then  $\text{diam}(CU) = \infty$ .

The key to characterizing the strongly quasiconvex subsets of hierarchically hyperbolic spaces is to determine what the projection of a strongly quasiconvex subset to each of the associated hyperbolic spaces looks like. The property that characterizes the projection of strongly quasiconvex subsets is the following orthogonal projection dichotomy.

**Definition 6.2** (orthogonal projection dichotomy) For  $B \geq 0$ , a subset  $Y$  of an HHS  $(\mathcal{X}, \mathfrak{S})$  has the *B-orthogonal projection dichotomy* if for all  $U, V \in \mathfrak{S}$  with  $U \perp V$ , if  $\text{diam}(\pi_U(Y)) > B$  then  $CV \subseteq N_B(\pi_V(Y))$ .

From now on, when we consider an HHS with the  $B_0$ -bounded domain dichotomy and a subspace with the  $B$ -orthogonal projection dichotomy, we will assume that  $B \geq B_0$ .

We can now state our characterization of strongly quasiconvex subsets of hierarchically hyperbolic spaces with the bounded domain dichotomy.

**Theorem 6.3** (characterization of strong quasiconvexity) *Let  $(\mathcal{X}, \mathfrak{S})$  be a hierarchically hyperbolic space with the bounded domain dichotomy and  $Y \subseteq \mathcal{X}$ . Then the following are equivalent:*

- (1)  $Y$  is an  $(A, D)$ -contracting subset.
- (2) The lower relative divergence of  $\mathcal{X}$  with respect to  $Y$  is at least quadratic.
- (3) The lower relative divergence of  $\mathcal{X}$  with respect to  $Y$  is completely superlinear.
- (4)  $Y$  is  $Q$ -strongly quasiconvex.
- (5)  $Y$  is  $k$ -hierarchically quasiconvex and has the  $B$ -orthogonal projection dichotomy.

Moreover, the pair  $(A, D)$  in part (1), the convexity gauge  $Q$  in part (4), and the pair  $(k, B)$  in part (5) each determine the other two.

The work in Section 3 showed that the implications

$$(1) \implies (2) \implies (3) \implies (4)$$

hold in any quasigeodesic space and that the pair  $(A, D)$  determines  $Q$ . Further, Proposition 5.7 showed that every  $Q$ -strongly quasiconvex subset of a hierarchically hyperbolic space is  $k$ -hierarchically quasiconvex with  $Q$  determining  $k$ . Thus in the next two subsections, we only need to prove:

- If  $Y$  is  $Q$ -strongly quasiconvex, then there exists  $B > 0$  determined by  $Q$  such that  $Y$  has the  $B$ -orthogonal projection dichotomy (Section 6.1).
- If  $Y$  is  $k$ -hierarchically quasiconvex and has the  $B$ -orthogonal projection dichotomy, then  $Y$  is  $(A, D)$ -contracting where  $(A, D)$  is determined by  $(k, B)$  (Section 6.2).

Before beginning the proof, we record of the following corollary to Theorem 6.3 that allows us to characterize stable embeddings.

**Corollary 6.4** *Let  $(\mathcal{X}, \mathfrak{S})$  be an HHS with the bounded domain dichotomy and let  $i : Y \rightarrow \mathcal{X}$  be a quasi-isometric embedding from a uniform quasigeodesic space  $Y$  to  $\mathcal{X}$ . The following are equivalent:*

- (1)  $i$  is a stable embedding.
- (2)  $Z = i(Y)$  is hierarchically quasiconvex and there exists a  $B > 0$  such that for all  $U, V \in \mathfrak{S}$  with  $U \perp V$ , if  $\text{diam}(\pi_U(Z)) > B$ , then  $\text{diam}(\pi_V(Z)) < B$ .

**Proof** By [11, Corollary 2.16], an HHS  $(\mathcal{Z}, \mathfrak{T})$  is hyperbolic if and only if there exists  $B$  such that for all  $U, V \in \mathfrak{T}$  with  $U \perp V$ , either  $\text{diam}(\pi_U(\mathcal{Z})) < B$  or  $\text{diam}(\pi_V(\mathcal{Z})) < B$ .



By Proposition 2.8,  $i$  is a stable embedding if and only if the image  $Z = i(Y)$  is strongly quasiconvex in  $\mathcal{X}$  and hyperbolic. The equivalence follows from these observations and the fact that hierarchically quasiconvex subsets inherit the hierarchy structure from the ambient space as described in [10, Proposition 5.6].  $\square$

Corollary 6.4 should be compared with [1, Corollary 6.2]. If  $(\mathcal{X}, \mathfrak{S})$  has the extra assumption of unbounded products required in [1, Corollary 6.2], then Corollary 6.4 can be immediately improved to [1, Corollary 6.2]. However, Corollary 6.4 is a strict expansion of [1, Corollary 6.2] as many HHS structures do not have unbounded products. Naturally occurring HHS structures without unbounded products can be found in right angled Coxeter groups and the Weil–Petersson metric on Teichmüller space. We briefly describe these structures in Section 7.

## 6.1 Strongly quasiconvex subsets have orthogonal projection dichotomy

In this subsection, we provide the implication (4) to (5) in Theorem 6.3. Our focus will be on studying the following set of domains.

**Definition 6.5** Define  $\mathfrak{S}^*$  to be the set of domains  $U \in \mathfrak{S}$  such that  $\text{diam}(CU) = \infty$  and there exists  $V \in \mathfrak{S}_U^\perp$  such that  $\text{diam}(CV) = \infty$ .

For each  $U \in \mathfrak{S}^*$  we have that both factors of the product region  $P_U$  have infinite diameter. In particular, if  $\mathfrak{S}^* = \emptyset$  and  $\mathfrak{S}$  has the bounded domain dichotomy, then  $(\mathcal{X}, \mathfrak{S})$  is hyperbolic by [11, Corollary 2.16]. Thus the intuition for restricting our attention to these domains is that the domains in  $\mathfrak{S}^*$  are the source of nonhyperbolic behavior in  $(\mathcal{X}, \mathfrak{S})$ .

The crucial step to proving strongly quasiconvex subsets have the orthogonal projection dichotomy is the following proposition that establishes a sort of orthogonal projection dichotomy for the product regions of domains in  $\mathfrak{S}^*$ .

**Proposition 6.6** Let  $(\mathcal{X}, \mathfrak{S})$  be an HHS with the bounded domain dichotomy and  $Y \subseteq \mathcal{X}$  be a  $Q$ -strongly quasiconvex subset. There is a constant  $B_0 > 0$  depending on  $\mathfrak{S}$  and  $Q$  such that for all  $B \geq B_0$  and  $U \in \mathfrak{S}^*$ ,

$$\text{diam}(\pi_U(Y)) > B \implies P_U \subseteq N_B(\mathfrak{g}_{P_U}(Y)).$$

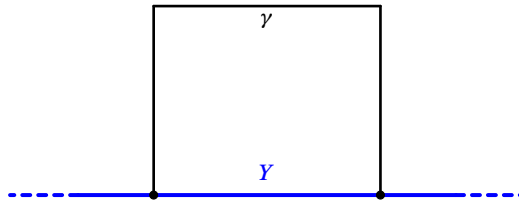


Figure 3: In  $\mathbb{R}^2$  (equipped with the  $\ell_1$ -metric) consider  $Y$  to be the  $x$ -axis. Let  $\gamma$  be the  $(3, 0)$ -quasigeodesic consisting of three sides of a square with the fourth side on  $Y$ . While the quasigeodesic constants do not change, increasing the distance between the endpoints of  $\gamma$  produces points of  $\gamma$  arbitrarily far away from  $Y$ .

Since  $U$  is in  $\mathfrak{S}^*$ , the product region  $\mathbf{P}_U$  coarsely coincides with the product of two infinite diameter metric spaces. The proof of Proposition 6.6 is therefore motivated by the situation described in Figure 3. Namely, if  $Y$  is a subset of the product of two infinite-diameter metric spaces, then either  $Y$  coarsely coincides with the whole product or there exists a quasigeodesic  $\gamma$  with endpoints on  $Y$  and fixed constants such that there are points of  $\gamma$  whose distance to  $Y$  is comparable to  $\text{diam}(Y)$ . Thus if  $Y$  is  $Q$ -strongly quasiconvex, then either  $Y$  has bounded diameter or it coarsely covers the entire product.

In Proposition 6.10, we prove that a similar situation holds for  $\mathbf{P}_U$ . We show if  $\text{diam}(\pi_U(Y))$  is sufficiently large and  $Y$  does not coarsely coincide with  $\mathbf{P}_U$ , then we can find a uniform constant quasigeodesic with endpoints on  $\mathfrak{g}_{\mathbf{P}_U}(Y)$  that contains points relatively far from  $\mathfrak{g}_{\mathbf{P}_U}(Y)$ . To utilize this to prove Proposition 6.6, we must promote this statement on  $\mathfrak{g}_{\mathbf{P}_U}(Y)$  to a statement on  $Y$ . Specifically, we show that we can realize every quasigeodesic of  $\mathbf{P}_U$  with endpoints on  $\mathfrak{g}_{\mathbf{P}_U}(Y)$  as a segment of a quasigeodesic with endpoints on  $Y$ , while maintaining uniform quasigeodesic constants (Lemma 6.11). This yields a quasigeodesic with endpoints on  $Y$  that contains a point  $x$  of  $\mathbf{P}_U$  such that  $d_{\mathcal{X}}(x, \mathfrak{g}_{\mathbf{P}_U}(Y))$  is comparable with  $\text{diam}(\mathfrak{g}_{\mathbf{P}_U}(Y))$ . If  $Y$  is strongly quasiconvex, the bridge theorem (Theorem 4.18) implies that  $d_{\mathcal{X}}(x, \mathfrak{g}_{\mathbf{P}_U}(Y))$  also provides a lower bound on the distance between  $x$  and  $Y$ . However, since  $Y$  is strongly quasiconvex, the distance between  $x$  and  $Y$  is uniformly bounded. Hence, if  $Y$  does not coarsely cover  $\mathbf{P}_U$ , we obtain that  $\mathfrak{g}_{\mathbf{P}_U}(Y)$  must have bounded diameter which contradicts the assumption on  $\text{diam}(\pi_U(Y))$ .

We begin by describing a particularly nice class of paths in product spaces and show that they are quasigeodesics (Lemma 6.8).

**Definition 6.7** (spiral path) Let  $X$  and  $Y$  be  $(K, L)$ -quasigeodesic metric spaces, and let  $Z = X \times Y$  be equipped with the  $\ell_1$ -metric. A *spiral path*  $\gamma$  in  $Z$  is the concatenation  $\gamma = \gamma_1 * \dots * \gamma_n$  of  $(K, L)$ -quasigeodesic of  $Z$  satisfying the following:

- Every  $\gamma_i$  is of the form  $\eta \times c_y$  or  $c_x \times \delta$  where  $\eta$  (resp.  $\delta$ ) is a  $(K, L)$ -quasigeodesic of  $X$  (resp.  $Y$ ) and  $c_{x_0}$  (resp.  $c_{y_0}$ ) is the constant function with value  $x_0 \in X$  (resp.  $y_0 \in Y$ ).
- For every  $i$ , if  $\gamma_i$  is constant on the  $X$  (resp.  $Y$ ) factor of  $Z = X \times Y$ , then  $\gamma_{i+1}$  is constant on the  $Y$  (resp.  $X$ ) component of  $Z = X \times Y$ .

A spiral path  $\gamma = \gamma_1 * \dots * \gamma_n$  has *slope*  $N$  if for every  $i \in \{1, \dots, n - 2\}$ ,

$$d(\gamma_{i+1}^+, \gamma_{i+1}^-) \geq Nd(\gamma_i^+, \gamma_i^-),$$

where  $\gamma_j^\pm$  are the endpoints of  $\gamma_j$ . Note that the distance between the endpoints of  $\gamma_n$  can be arbitrary.

**Lemma 6.8** (spiral paths are quasigeodesics) *For each  $K \geq 1$  and  $L \geq 0$  there are constants  $K'$  and  $L'$  such that the following holds. Let  $X$  and  $Y$  be  $(K, L)$ -quasigeodesic metric spaces. If  $\gamma = \gamma_1 * \dots * \gamma_n$  is a spiral path of slope  $N > 4K^2$  in  $Z = X \times Y$  such that the endpoints of  $\gamma_1$  are at least  $3K^2L + 1$  far apart, then  $\gamma$  is a  $(K', L')$ -quasigeodesic of  $X \times Y$ .*

The following proof is essentially the same as showing the logarithmic spiral in  $\mathbb{R}^2$  is a quasigeodesic. However, as we were not able to find a sufficient reference in the literature, we have included it in the interest of completeness.

**Proof** Let  $\gamma = \gamma_1 * \dots * \gamma_n: [a_0, a_n] \rightarrow Z$  be spiral path of slope  $N > 4K^2$  and let  $a_1 < \dots < a_n$  be points in  $[a_0, a_n]$  such that  $\gamma_i = \gamma|_{[a_{i-1}, a_i]}$ .

Let  $t_1, t_2 \in [a_0, a_n]$ . We claim that

$$(3) \quad d(\gamma(t_1), \gamma(t_2)) \leq (K + 1)|t_2 - t_1| + 2L.$$

Since each  $\gamma_i$  is a  $(K, L)$ -quasigeodesic of  $Z$  for each  $i$ , we only need to consider the case where  $t_1 \in [a_k, a_{k+1}]$  and  $t_2 \in [a_j, a_{j+1}]$  with  $j - k \geq 1$ . By the choice on the distance between endpoints of  $\gamma_1$  and the slope  $N$ ,

$$d(\gamma(a_{i-1}), \gamma(a_i)) > 3K^2L + 1,$$

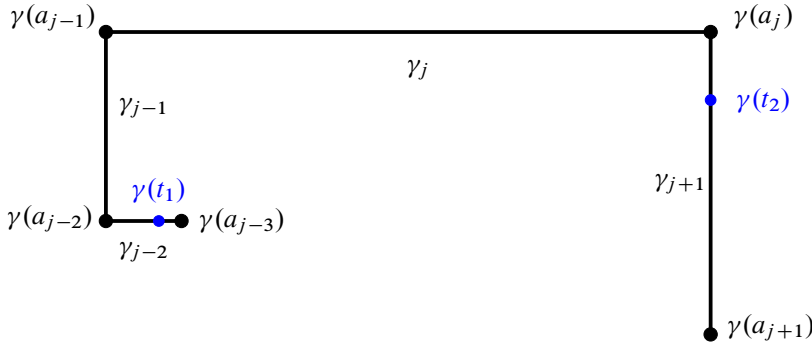


Figure 4

which implies  $|a_i - a_{i-1}| > L$ . Therefore,

$$|t_2 - t_1| \geq |a_j - a_{k+1}| \geq (j - k - 1)L.$$

Since each  $\gamma_i$  is  $(K, L)$ -quasigeodesic,

$$d(\gamma(t_1), \gamma(t_2)) \leq K|t_2 - t_1| + (j - k + 1)L \leq (K + 1)|t_2 - t_1| + 2L.$$

The remainder of the proof will show  $|t_2 - t_1| \leq d(\gamma(t_1), \gamma(t_2))$ .

For every  $i$ ,  $\gamma_i * \gamma_{i+1}$  is a  $(K, 2L)$ -quasigeodesic of  $Z$ , so we only need to consider the case where  $t_1 \in [a_k, a_{k+1}]$  and  $t_2 \in [a_j, a_{j+1}]$  with  $j - k \geq 2$  as in Figure 4.

We encourage the reader to refer to Figure 4 as they follow the remainder of the proof.

By the triangle inequality,

$$(4) \quad d(\gamma(t_2), \gamma(t_1)) \geq d(\gamma(t_2), \gamma(a_{j-1})) - d(\gamma(a_{j-1}), \gamma(t_1)).$$

The remainder of the proof has two parts. First we show that,  $d(\gamma(t_2), \gamma(a_{j-1}))$  is much larger than  $d(\gamma(a_{j-1}), \gamma(t_1))$ , so

$$d(\gamma(t_2), \gamma(t_1)) \geq d(\gamma(t_2), \gamma(a_{j-1})) \geq |t_2 - a_{j-1}|.$$

We then finish by showing that  $|t_2 - a_{j-1}| \geq |t_2 - t_1|$ .

To simplify notation let  $\ell(\gamma_i) = d(\gamma(a_{i-1}), \gamma(a_i))$ . The slope condition then says  $\frac{1}{N}\ell(\gamma_i) > \ell(\gamma_{i-1})$  for each  $1 \leq i \leq n - 1$ . Since  $N > 4K^2$ , we can iteratively apply the slope condition to get

$$(5) \quad \sum_{i=1}^{j-1} \ell(\gamma_i) \leq \left( \frac{1}{N^{j-2}} + \dots + \frac{1}{N} + 1 \right) \ell(\gamma_{j-1}) \leq 2\ell(\gamma_{j-1}) \leq \frac{2}{N}\ell(\gamma_j).$$

From the triangle inequality and the fact  $|a_{k+1} - t_1| \leq |a_{k+1} - a_k|$ ,

$$\begin{aligned} d(\gamma(t_1), \gamma(a_{j-1})) &\leq d(\gamma(t_1), \gamma(a_{k+1})) + \sum_{i=k+2}^{j-1} \ell(\gamma_i) \\ &\leq K|a_{k+1} - a_k| + L + \sum_{i=k+2}^{j-1} \ell(\gamma_i) \\ &\leq K(K\ell(\gamma_{k+1}) + KL) + L + \sum_{i=k+2}^{j-1} \ell(\gamma_i) \\ &\leq K^2 \left( \sum_{i=k+1}^{j-1} \ell(\gamma_i) \right) + 2K^2L. \end{aligned}$$

Then by applying inequality (5),

$$d(\gamma(t_1), \gamma(a_{j-1})) \leq \left( \frac{2K^2}{N} \right) \ell(\gamma_j) + 2K^2L \leq \frac{1}{2}d(\gamma(t_2), \gamma(a_{j-1})) + 2K^2L.$$

Substituting this into inequality (4) produces

$$d(\gamma(t_2), \gamma(t_1)) \geq \frac{1}{2}d(\gamma(t_2), \gamma(a_{j-1})) - 2K^2L.$$

We can then use the fact that  $\gamma_j * \gamma_{j+1}$  is a  $(K, 2L)$ -quasigeodesic to obtain

$$\begin{aligned} (6) \quad d(\gamma(t_2), \gamma(t_1)) &\geq \frac{1}{2}d(\gamma(t_2), \gamma(a_{j-1})) - 2K^2L \\ &\geq \frac{1}{2} \left( \frac{1}{K}|t_2 - a_{j-1}| - 2L \right) - 2K^2L \\ &\geq \frac{1}{2K}|t_2 - a_{j-1}| - 3K^2L. \end{aligned}$$

We now show that  $|t_2 - a_{j-1}| \geq |t_2 - t_1|$ , which completes the proof by inequality (6). Since we required that  $\ell(\gamma_1) > 3K^2L + 1$  and  $N > 4K^2$ , we have  $\frac{1}{K}|a_i - a_{i-1}| > 2L$  for each  $i$ . This implies

$$\ell(\gamma_i) \geq \frac{1}{K}|a_i - a_{i-1}| - L > \frac{1}{2K}|a_i - a_{i-1}|.$$

In particular, using inequality (5) we obtain

$$\frac{2}{N}(K|a_j - a_{j-1}| + L) \geq \frac{2}{N}\ell(\gamma_j) \geq \sum_{i=1}^{j-1} \ell(\gamma_i) \geq \sum_{i=1}^{j-1} \frac{1}{2K}|a_i - a_{i-1}| \geq \frac{1}{2K}|a_{j-1} - t_1|.$$

Hence,

$$|a_{j-1} - t_1| \leq \frac{4K^2}{N}|a_j - a_{j-1}| + \frac{4KL}{N} \leq |a_j - a_{j-1}| + L$$

and we can conclude

$$\begin{aligned} |t_2 - t_1| &= |t_2 - a_j| + |a_j - a_{j-1}| + |a_{j-1} - t_1| \\ &\leq |t_2 - a_j| + 2|a_j - a_{j-1}| + L \\ &\leq 3|t_2 - a_{j-1}| + L. \end{aligned}$$

Combining this with inequalities (3) and (6), we obtain that there are constants  $K'$  and  $L'$  depending on  $K$  and  $L$  such that

$$\frac{1}{K'}(t_2 - t_1) - L' \leq d(\gamma(t_2), \gamma(t_1)) \leq K'(t_2 - t_1) + L'. \quad \square$$

For the remainder of this section  $(\mathcal{X}, \mathfrak{S})$  will be an HHS with the bounded domain dichotomy and  $\mathfrak{S}^*$  is as in Definition 6.5. Recall, for each  $U \in \mathfrak{S}$ , the space  $F_U \times E_U$  consists of tuples  $a = (a_V)$ , where  $V \in \mathfrak{S}_U \cup \mathfrak{S}_U^\perp$ , and that  $P_U$  is defined as the image of  $\phi_U: F_U \times E_U \rightarrow \mathcal{X}$ . By restricting to a choice of factor, we can endow  $F_U$  and  $E_U$  with the subspace metric of their images under  $\phi_U$ . While this relies on the choice of factor, the distance formula (Theorem 4.4) says any two choices result in uniformly quasi-isometric metric spaces. Given  $a, b \in F_U \times E_U$  we use  $d_V(a, b)$  to denote  $d_V(a_V, b_V)$ , where  $V \in \mathfrak{S}_U \cup \mathfrak{S}_U^\perp$ . If  $U \in \mathfrak{S}^*$ , then both  $F_U$  and  $E_U$  are infinite diameter and so we can apply Proposition 6.9 to build the desired quasigeodesic in  $P_U$  based on  $\mathfrak{g}_{P_U}(Y)$ .

**Proposition 6.9** *Let  $Y \subseteq \mathcal{X}$ . There exist constants  $L'$  and  $r_0$ , and functions*

$$f, g, h: [r_0, \infty) \rightarrow [0, \infty),$$

*all depending only on  $\mathfrak{S}$ , such that  $f(r), g(r), h(r) \rightarrow \infty$  as  $r \rightarrow \infty$  and the following holds: for each  $U \in \mathfrak{S}^*$  and each  $r \geq r_0$ , if the  $r$ -neighborhood of  $\phi_U^{-1}(\mathfrak{g}_{P_U}(Y))$  does not cover  $F_U \times E_U$  and  $\text{diam}(\pi_U(Y)) > f(r)$ , then there exists a  $(L', L')$ -quasigeodesic  $\eta$  with endpoints  $a, b \in \phi_U^{-1}(\mathfrak{g}_{P_U}(Y))$  such that  $\eta$  is not contained in the  $g(r)$ -neighborhood of  $\phi_U^{-1}(\mathfrak{g}_{P_U}(Y))$  and  $d_U(a, b) > h(r)$ .*

**Proof** Our approach is to construct a spiral path of sufficient slope in  $F_U \times E_U$  and then apply Lemma 6.8 to conclude it is a quasigeodesic. Let  $d(\cdot, \cdot)$  denote the  $\ell_1$ -distance in  $F_U \times E_U$  and fix the following constants, which depend only on  $\mathfrak{S}$ :

- $L$  such that  $F_U$  and  $E_U$  are  $(L, L)$ -quasigeodesic spaces.
- $K$  such that  $\pi_U$  is  $(K, K)$ -coarsely Lipschitz.
- $N = 4L^2 + 1$  will be the slope of the spiral path we construct.

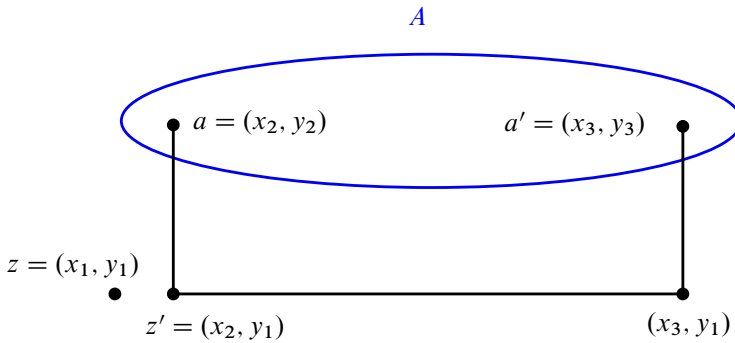


Figure 5: Spiral path constructed when  $d_{F_U}(x_1, x_2) \leq \frac{1}{2}(r + 2L + 1)$ .

Let  $r > 10L^3 + 6$  and  $A = \phi_U^{-1}(\mathfrak{g}_{P_U}(Y))$ . Suppose that the  $r$ -neighborhood of  $A$  does not cover  $F_U \times E_U$ . Thus there exists a point  $z = (x_1, y_1) \in F_U \times E_U$  such that  $r \leq d(z, A) \leq r + 2L$ . Let  $a = (x_2, y_2)$  be a point of  $A$  such that  $d(z, a) - 1 \leq d(z, A)$ . We have  $\min\{d_{F_U}(x_1, x_2), d_{E_U}(y_1, y_2)\} \leq \frac{1}{2}(r + 2L + 1)$ . There are two cases depending on which of the two factors realizes the minimum.

**Case 1** If  $d_{F_U}(x_1, x_2)$  realizes the minimum, let  $z' = (x_2, y_1)$  and  $D_r = \frac{1}{2}(r - 2L - 1)$ . Then  $d(z', A) \geq d(z, A) - d(z, z') \geq D_r$ , which implies  $d(z', a) > 3L^3 + 1$  because  $r > 10L^3 + 6$ .

There exists  $B_r > r$  such that for any pair of points  $u$  and  $v$  of  $F_U$ , if  $d_U(u, v) \geq B_r$  then

$$d_{F_U}(u, v) \geq 2(r + 2L + 1)N.$$

We shall assume  $\text{diam}(\pi_U(Y)) > 2B_r$ , so there is a point  $a' = (x_3, y_3)$  of  $A$  such that  $d_U(x_2, x_3) \geq B_r$  and  $d_{F_U}(x_2, x_3) > d_{E_U}(y_2, y_1)N$ . We can now form a spiral path  $\eta$  of slope  $N = 4L^2 + 1$  by connecting each sequential pair of points in the sequence

$$a = (x_2, y_2) - (x_2, y_1) - (x_3, y_1) - (x_3, y_3) = a'$$

by  $(L, L)$ -quasigeodesics. Since  $d_{E_U}(y_2, y_1) > 3L^3 + 1$ ,  $\eta$  satisfies the hypothesis of Lemma 6.8 and is therefore an  $(L', L')$ -quasigeodesic for some  $L'$  determined by  $L$ .

Since  $z' = (x_2, y_1)$  is at least  $D_r$  far from  $A$ ,  $\eta$  has endpoints in  $A$  and is not contained in the  $D_r$ -neighborhood of  $A$ . Moreover,  $d_U(a, a') \geq B_r$  and we get the claim with  $f(r) = 2B_r$ ,  $g(r) = D_r$ , and  $h(r) = B_r$ .

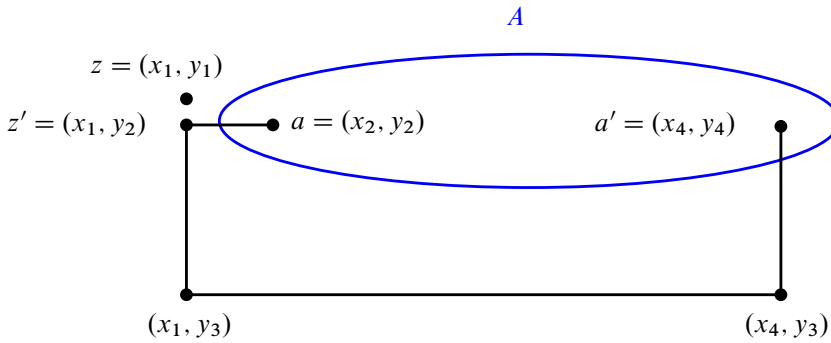


Figure 6: Spiral path constructed when  $d_{E_U}(y_1, y_2) \leq \frac{1}{2}(r + 2L + 1)$ .

**Case 2** If  $d_{E_U}(y_1, y_2)$  realizes the minimum, let  $z' = (x_1, y_2)$ . As before we have that  $d(z', A) \geq D_r = \frac{1}{2}(r - 2L - 1)$ , which implies  $d(z', a) > 3L^3 + 1$ . Let  $y_3$  be a point of  $E_U$  such that

$$(r + 2L + 1)N \leq d_{E_U}(y_2, y_3) \leq 2(r + 2L + 1)N.$$

There exists  $C_r > r$  such that for any pair of points  $u$  and  $v$  of  $F_U$ , if  $d_U(u, v) \geq C_r$  then

$$d_{F_U}(u, v) \geq 2(r + 2L + 1)N^2.$$

We shall assume  $\text{diam}(\pi_U(Y)) > 2C_r$ , so there exists  $a' = (x_4, y_4) \in A$  such that  $d_U(x_1, x_4) > C_r$ . This implies  $d_{F_U}(x_1, x_4) > 2(r + 2L + 1)N^2$  and we can now form a spiral path  $\eta$  of slope  $N = 4L^2 + 1$  by connecting each sequential pair of points in the sequence

$$a = (x_2, y_2) - (x_1, y_2) - (x_1, y_3) - (x_4, y_3) - (x_4, y_4) = a'$$

by an  $(L, L)$ -quasi-geodesics.

As before,  $\eta$  satisfies the hypothesis of Lemma 6.8 and is therefore an  $(L', L')$ -quasi-geodesic for some  $L'$  determined by  $L$ . The remaining claims follow as in the preceding case. □

The distance formula makes the map  $\phi_U : F_U \times E_U \rightarrow \mathcal{X}$  a uniform quasi-isometric embedding. Thus  $\mathfrak{g}_{P_U}(Y)$  coarsely covers  $P_U$  if and only if  $\phi_U^{-1}(\mathfrak{g}_{P_U}(Y))$  coarsely covers  $F_U \times E_U$ , Proposition 6.9 therefore allows us to immediately deduce the following result for  $P_U \subseteq \mathcal{X}$ .



**Proposition 6.10** *Let  $Y \subseteq \mathcal{X}$ . There exist constants  $L'$  and  $r_0$ , and functions*

$$f, g, h: [r_0, \infty) \rightarrow [0, \infty),$$

*all depending only on  $\mathfrak{S}$ , such that  $f(r), g(r), h(r) \rightarrow \infty$  as  $r \rightarrow \infty$  and the following holds: for each  $U \in \mathfrak{S}^*$  and each  $r \geq r_0$ , if the  $r$ -neighborhood of  $\mathfrak{g}_{\mathbf{P}_U}(Y)$  does not cover  $\mathbf{P}_U$  and  $\text{diam}(\pi_U(Y)) > f(r)$ , then there exists an  $(L', L')$ -quasigeodesic  $\eta$  with endpoints  $a, b \in \mathfrak{g}_{\mathbf{P}_U}(Y)$  such that*

- (1)  $\eta \subseteq \mathbf{P}_U$ ,
- (2)  $\eta$  is not contained in the  $g(r)$ -neighborhood of  $\mathfrak{g}_{\mathbf{P}_U}(Y)$ ,
- (3)  $d_U(a, b) > h(r)$ .

Proposition 6.10 furnishes a quasigeodesic  $\eta$  with endpoints in  $\mathfrak{g}_{\mathbf{P}_U}(Y)$  that can be made as far from  $\mathfrak{g}_{\mathbf{P}_U}(Y)$  as desired by increasing  $\text{diam}(\pi_U(Y))$ . However, to exploit the fact that  $Y$  is a strongly quasiconvex subset, we need the next lemma, which “promotes”  $\eta$  to a quasigeodesic with endpoints in  $Y$ .

**Lemma 6.11** *There exists  $D > 0$  such that if  $x, y \in \mathcal{X}$  and  $U \in \mathfrak{S}$ , with  $d_U(x, y) > D$  and  $\eta$  is a  $(k, c)$ -quasigeodesic contained in  $\mathbf{P}_U$  with endpoints  $\mathfrak{g}_{\mathbf{P}_U}(x)$  and  $\mathfrak{g}_{\mathbf{P}_U}(y)$ , then there exists a  $(k', c')$ -quasigeodesic containing  $\eta$  and with endpoints  $x$  and  $y$ , where  $k'$  and  $c'$  depend only on  $\lambda$  and  $\epsilon$ .*

**Proof** Let  $D$  and  $\lambda$  be as in Proposition 4.24. We further assume  $\lambda$  is large enough that every pair of points in  $\mathcal{X}$  can be joined by a  $\lambda$ -hierarchy path (Theorem 4.8).

Assume  $d_U(x, y) > D$  and let  $\tilde{\gamma}$  be the  $\lambda$ -hierarchy path connecting  $x$  and  $y$  provided by Proposition 4.24. Let  $\alpha$  be the active subpath of  $\tilde{\gamma}$  corresponding to  $U$ . Define  $x'$  (resp.  $y'$ ) to be the endpoint of  $\alpha$  closest to  $x$  (resp.  $y$ ) and  $x'' = \mathfrak{g}_{\mathbf{P}_U}(x)$  (resp.  $y'' = \mathfrak{g}_{\mathbf{P}_U}(y)$ ). If  $\eta: [b, c] \rightarrow \mathbf{P}_U$  is any  $(k, c)$ -quasigeodesic in  $\mathbf{P}_U$  connecting  $x''$  and  $y''$ , let  $\gamma$  be the concatenation of  $\tilde{\gamma} - \alpha$ , any  $\lambda$ -hierarchy path from  $x'$  to  $x''$ ,  $\eta$ , and any  $\lambda$ -hierarchy path from  $y'$  to  $y''$ . We will show that this path  $\gamma$  is a  $(k', c')$ -quasigeodesic where the constants depend only on  $k$  and  $c$ .

The distances  $d_{\mathcal{X}}(x', \mathbf{P}_U)$  and  $d_{\mathcal{X}}(y', \mathbf{P}_U)$  are uniformly bounded by Proposition 4.24. By Lemma 4.15, the distances  $d_{\mathcal{X}}(x', \mathfrak{g}_{\mathbf{P}_U}(x'))$  and  $d_{\mathcal{X}}(y', \mathfrak{g}_{\mathbf{P}_U}(y'))$  are uniformly bounded as well. Again by Proposition 4.24,  $\mathfrak{g}_{\mathbf{P}_U}(x)$  coarsely coincides with  $\mathfrak{g}_{\mathbf{P}_U}(x')$  and  $\mathfrak{g}_{\mathbf{P}_U}(y)$  coarsely coincides with  $\mathfrak{g}_{\mathbf{P}_U}(y')$ . Thus there exists  $\mu$  depending only on  $\mathfrak{S}$  such that  $d_{\mathcal{X}}(x', x''), d_{\mathcal{X}}(y', y'') \leq \mu$ .

Now, let  $\gamma_x$  (resp.  $\gamma_y$ ) be the subset of  $\gamma$  from  $x$  to  $x''$  (resp.  $y$  to  $y''$ ). Since  $d_{\mathcal{X}}(x', x'')$  and  $d_{\mathcal{X}}(y', y'')$  are uniformly bounded by  $\mu$ ,  $\gamma_x$  and  $\gamma_y$  are both uniform quasigeodesics. By Lemma 4.15 and Proposition 4.24, there exists  $K \geq 1$  depending on  $k, c$ , and  $\mathfrak{G}$  such that

- $d_{\mathcal{X}}(x', x''), d_{\mathcal{X}}(y', y'') \leq K$ ;
- $\text{diam}(\mathfrak{g}_{\mathbf{P}_U}(\gamma_x)), \text{diam}(\mathfrak{g}_{\mathbf{P}_U}(\gamma_y)) \leq K$ ;
- $\gamma_x, \gamma_y$  and  $\eta$  are all  $(K, K)$ -quasigeodesics;
- for all  $p \in \mathbf{P}_U$  and  $q \in \mathcal{X}$ ,  $d_{\mathcal{X}}(q, \mathfrak{g}_{\mathbf{P}_U}(q)) \leq Kd_{\mathcal{X}}(p, q) + K$ .

Let  $\gamma = \gamma_x * \eta * \gamma_y : [a, d] \rightarrow \mathcal{X}$  and  $a < b < c < d$  such that  $\gamma|_{[a,b]} = \gamma_x, \gamma|_{[b,c]} = \eta$  and  $\gamma|_{[c,d]} = \gamma_y$ . For  $t, s \in [a, d]$ , let  $u = \gamma(t), v = \gamma(s)$ . We want to show  $|t - s| \asymp d_{\mathcal{X}}(u, v)$  for some constants depending only on  $K$ . The only interesting cases are when  $u$  and  $v$  are in different components of  $\gamma = \gamma_x * \eta * \gamma_y$ , so without loss of generality, we have the following two cases.

**Case 1** Assume  $t \in [a, b]$  and  $s \in [b, c]$ . Thus  $u \in \gamma_x$  and  $v \in \eta$ , and

$$d_{\mathcal{X}}(u, v) \leq d_{\mathcal{X}}(u, x'') + d_{\mathcal{X}}(x'', v) \leq K|t - b| + K|b - s| + 2K \leq K|t - s| + 2K.$$

For the inequality  $|t - s| \leq d_{\mathcal{X}}(u, v)$ , our choice of  $K$  provides

$$d_{\mathcal{X}}(u, x'') \leq d_{\mathcal{X}}(u, \mathfrak{g}_{\mathbf{P}_U}(u)) + K \leq Kd_{\mathcal{X}}(u, v) + 2K.$$

By the triangle inequality  $d_{\mathcal{X}}(v, x'') \leq d_{\mathcal{X}}(v, u) + d_{\mathcal{X}}(u, x'')$  and we derive the desired inequality as

$$\begin{aligned} |t - s| &= |t - b| + |b - s| \\ &\leq Kd_{\mathcal{X}}(u, x'') + Kd_{\mathcal{X}}(v, x'') + 2K \\ &\leq K^2d_{\mathcal{X}}(u, v) + K(d_{\mathcal{X}}(u, v) + d_{\mathcal{X}}(u, x'')) + 2K^2 + 2K \\ &\leq K^2d_{\mathcal{X}}(u, v) + Kd_{\mathcal{X}}(u, v) + K^2d_{\mathcal{X}}(u, v) + 4K^2 + 2K \\ &\leq 3K^2d_{\mathcal{X}}(u, v) + 6K^2. \end{aligned}$$

**Case 2** Assume  $t \in [a, b]$  and  $s \in [c, d]$  so that  $u \in \gamma_x$  and  $v \in \gamma_y$ . Further we can assume  $u, v \in \tilde{\gamma}$ , since otherwise the above proof holds by increasing the constants by  $4K$ . The inequality  $d_{\mathcal{X}}(u, v) \leq |t - s|$  can be established by a nearly identical argument to the previous case. For the inequality  $|t - s| \leq d_{\mathcal{X}}(u, v)$  we need to utilize the fact that  $\tilde{\gamma}$  is a  $(\lambda_0, \lambda_0)$ -quasigeodesic. Thus, by increasing  $K$ , we can ensure that

- $d_{\mathcal{X}}(u, v) \asymp_{K,K} d_{\mathcal{X}}(u, x') + d_{\mathcal{X}}(x', y') + d_{\mathcal{X}}(y', v)$ ,

- $d_{\mathcal{X}}(x', y') \asymp_{1,2K} d_{\mathcal{X}}(x'', y'') \asymp_{K,K} |b - c|$ ,
- $d_{\mathcal{X}}(u, x') \asymp_{1,K} d_{\mathcal{X}}(u, x'') \asymp_{K,K} |t - b|$ ,
- $d_{\mathcal{X}}(v, y') \asymp_{1,K} d_{\mathcal{X}}(v, y'') \asymp_{K,K} |c - s|$ .

We then have the calculation

$$\begin{aligned}
 |t - s| &= |t - b| + |b - c| + |c - s| \\
 &\leq Kd_{\mathcal{X}}(u, x'') + Kd_{\mathcal{X}}(x'', y'') + Kd_{\mathcal{X}}(y'', v) + 3K \\
 &\leq Kd_{\mathcal{X}}(u, x') + Kd_{\mathcal{X}}(x', y') + Kd_{\mathcal{X}}(y', v) + 7K^2 \\
 &\leq K^2 d_{\mathcal{X}}(u, v) + 8K^2.
 \end{aligned}
 \tag*{$\square$}$$

We can now provide the proof of Proposition 6.6.

**Proof of Proposition 6.6** Let  $Y \subseteq \mathcal{X}$  be  $Q$ -strongly quasiconvex and  $U \in \mathfrak{S}$  such that  $\text{diam}(CU) = \infty$  and there exists  $V \in \mathfrak{S}^{\perp}_U$  with  $\text{diam}(CV) = \infty$ . Recall our goal is to show that there exists  $B$  depending on  $\mathfrak{S}$  and  $Q$  such that if  $\text{diam}(\pi_U(Y)) > B$ , then  $\mathbf{P}_U \subseteq N_B(\mathfrak{g}_{\mathbf{P}_U}(Y))$ . Begin by fixing the following constants that all depend only on  $\mathfrak{S}$  and  $Q$ :

- $\mu$  such that for all  $x \in \mathcal{X}$ ,  $d_U(x, \mathfrak{g}_{\mathbf{P}_U}(x)) < \mu$ .
- $D$ , the constant from Lemma 6.11.
- $L'$ , the quasigeodesic constant from Proposition 6.10.
- $k'$ , the quasigeodesic constant obtained by applying Lemma 6.11 to a  $(L', L')$ -quasigeodesic.
- $K$ , the constant from the bridge theorem (Theorem 4.18) for  $Y$  and  $\mathbf{P}_U$  (recall  $Y$  is hierarchically quasiconvex by Proposition 5.7).

Let  $f, g$  and  $h$  be as in Proposition 6.10 and fix  $r$  be large enough that

$$g(r) > 2KQ(k', k') + K^2 + K \quad \text{and} \quad h(r) > D + 2\mu.$$

If  $\mathbf{P}_U \subseteq N_r(\mathfrak{g}_{\mathbf{P}_U}(Y))$ , then we are done. So for the purposes of contradiction, suppose that  $\mathbf{P}_U \not\subseteq N_r(\mathfrak{g}_{\mathbf{P}_U}(Y))$  and that  $\text{diam}(\pi_U(Y)) > f(r)$ . Let  $\eta$  be the  $(L', L')$ -quasigeodesic provided by Proposition 6.10 and let  $a_1, b_1 \in \mathfrak{g}_{\mathbf{P}_U}(Y)$  be the endpoints of  $\eta$ . Let  $a_0, b_0 \in Y$  such that  $\mathfrak{g}_{\mathbf{P}_U}(a_0) = a_1$  and  $\mathfrak{g}_{\mathbf{P}_U}(b_0) = b_1$ . Since

$$d_U(a_0, b_0) > d_U(a_1, b_1) - 2\mu > h(r) - 2\mu > D,$$

Lemma 6.11 produces a  $(k', k')$ -quasigeodesic  $\gamma$  with endpoints  $a_0$  and  $b_0$  and containing  $\eta$  where  $k'$  depending ultimately only on  $\mathfrak{S}$ . Since  $Y$  is  $Q$ -strongly quasiconvex,  $\gamma \subseteq N_{Q(k', k')}(Y)$ . By Proposition 6.10, there exists  $x \in \eta$  such that  $d_{\mathcal{X}}(x, \mathfrak{g}_{P_U}(Y)) > g(r)$ . Let  $y \in Y$  be such that  $d_{\mathcal{X}}(x, y) - 1 \leq d_{\mathcal{X}}(x, Y)$ . Then by the bridge theorem (Theorem 4.18) we have a contradiction,

$$Q(k', k') \geq d_{\mathcal{X}}(x, y) - 1 \geq \frac{1}{K}d_{\mathcal{X}}(x, \mathfrak{g}_{P_U}(Y)) - K - 1 > 2Q(k', k'). \quad \square$$

The following proposition uses Proposition 6.6 to finish the proof of the implication from (4) to (5) in Theorem 6.3.

**Proposition 6.12** *If  $(\mathcal{X}, \mathfrak{S})$  is an HHS with the bounded domain dichotomy and  $Y$  is a  $Q$ -strongly quasiconvex subset of  $\mathcal{X}$ , then there exists  $B > 0$  depending only on  $Q$  and  $\mathfrak{S}$  such that  $Y$  has the  $B$ -orthogonal projection dichotomy.*

**Proof** Let  $Y \subseteq \mathcal{X}$  be  $Q$ -strongly quasiconvex and  $B' > 0$  be larger than the bounded domain dichotomy constant for  $\mathfrak{S}$  and the constant  $B_0$  from Proposition 6.6. Let  $U \in \mathfrak{S}$ . If  $U \notin \mathfrak{S}^*$ , then by the bounded domain dichotomy, either  $\text{diam}(CU) < B'$  or for all  $V \in \mathfrak{S}_U^\perp$ ,  $\text{diam}(CV) < B'$ . In either case, the  $B'$ -orthogonal projection dichotomy is satisfied for  $U$ . Thus we can assume that  $U \in \mathfrak{S}^*$ , so  $\text{diam}(CU) = \infty$  and there exists  $V \in \mathfrak{S}_U^\perp$  with  $\text{diam}(CV) = \infty$ . Suppose  $\text{diam}(\pi_U(Y)) > B'$ . By Proposition 6.6,  $P_U \subseteq N_{B'}(\mathfrak{g}_{P_U}(Y))$ . For all  $V \in \mathfrak{S}_U^\perp$ ,  $\pi_V(P_U)$  uniformly coarsely covers  $CV$ , thus there exists  $B \geq B'$  depending only on  $Q$  and  $\mathfrak{S}$  such that  $CV \subseteq N_B(\pi_V(Y))$ .  $\square$

## 6.2 Contracting subsets in HHSs

We now finish the proof of Theorem 6.3 by showing that for hierarchically quasiconvex subsets, the orthogonal projection dichotomy implies that the gate map  $\mathfrak{g}_Y$  is contracting.

**Proposition 6.13** *Let  $(\mathcal{X}, \mathfrak{S})$  be a hierarchically hyperbolic space with the bounded domain dichotomy and  $Y \subseteq \mathcal{X}$  be  $k$ -hierarchically quasiconvex. If  $Y$  has the  $B$ -orthogonal projection dichotomy, then the gate map  $\mathfrak{g}_Y : \mathcal{X} \rightarrow Y$  is  $(A, D)$ -contracting, where  $A$  and  $D$  depend only on  $k, B$ , and  $\mathfrak{S}$ .*

**Proof** The gate map satisfies the first two condition in the definition of a contracting map by Lemma 4.14. It only remains to prove: *there exist  $0 < A < 1$  and  $D \geq 1$  depending only on  $k, B$ , and  $\mathfrak{S}$ , such that for all  $x \in \mathcal{X}$ ,  $\text{diam}(\mathfrak{g}_Y(B_R(x))) \leq D$  where  $R = Ad(x, Y)$ .*

Fix a point  $x_0 \in \mathcal{X}$  with  $d_{\mathcal{X}}(x_0, Y) \geq C_0$  and let  $x \in \mathcal{X}$  be any point with

$$d_{\mathcal{X}}(x_0, x) < C_1 d_{\mathcal{X}}(x_0, Y)$$

for constants  $C_0$  and  $C_1$  to be determined below. We will prove that for each domain  $U \in \mathfrak{S}$  the distance  $d_U(\mathfrak{g}_Y(x_0), \mathfrak{g}_Y(x))$  is uniformly bounded, then the above will follow from the distance formula (Theorem 4.4).

We choose a “large” number  $L$  (we will clarify how large  $L$  is later). Let  $K \geq 1$  be the coarse equality constant from the distance formula with thresholds  $L$  and  $2L$ . Take  $C_0 > (2K + 1)K$  sufficiently large so there is  $W \in \mathfrak{S}$  such that  $d_W(x_0, \mathfrak{g}_Y(x_0)) > 2L$ . Choose  $C_1 < 1/(2K^2 + 1)$ , ensuring that  $d_{\mathcal{X}}(x_0, \mathfrak{g}_Y(x_0)) > (2K^2 + 1)d_{\mathcal{X}}(x_0, x)$ . If  $d_{\mathcal{X}}(x_0, x) \leq C_0$ , then by the coarse Lipschitzness of the projections  $d_U(\mathfrak{g}_Y(x_0), \mathfrak{g}_Y(x))$  is uniformly bounded by a number depending on  $C_0$  for each  $U \in \mathfrak{S}$ . Therefore, we can assume that  $d_{\mathcal{X}}(x_0, x) > C_0$ . We claim that there is a  $V \in \mathfrak{S}$  such that  $d_V(x_0, \mathfrak{g}_Y(x_0)) > d_V(x_0, x) + L$ .

Assume for the purpose of contradiction that  $d_W(x_0, \mathfrak{g}_Y(x_0)) \leq d_W(x_0, x) + L$  for all  $W \in \mathfrak{S}$ . Therefore,  $d_W(x_0, \mathfrak{g}_Y(x_0)) \geq 2L \implies d_W(x_0, x) \geq L$  and this implies

$$\{\{d_W(x_0, \mathfrak{g}_Y(x_0))\}\}_{2L} \leq 2\{\{d_W(x_0, x)\}\}_L$$

for all  $W \in \mathfrak{S}$ . Thus,

$$\begin{aligned} d_{\mathcal{X}}(x_0, \mathfrak{g}_Y(x_0)) &\leq K \sum_{W \in \mathfrak{S}} \{\{d_W(x_0, \mathfrak{g}_Y(x_0))\}\}_{2L} + K \\ &\leq 2K \sum_{W \in \mathfrak{S}} \{\{d_W(x_0, x)\}\}_L + K \\ &\leq 2K(Kd_{\mathcal{X}}(x_0, x) + K) + K \\ &\leq 2K^2 d_{\mathcal{X}}(x_0, x) + (2K + 1)K \\ &\leq 2K^2 d_{\mathcal{X}}(x_0, x) + C_0 \\ &\leq (2K^2 + 1)d_{\mathcal{X}}(x_0, x) \end{aligned}$$

which contradicts  $C_1 < 1/(2K^2 + 1)$ . Therefore, we can fix  $V \in \mathfrak{S}$  such that

$$d_V(x_0, \mathfrak{g}_Y(x_0)) > d_V(x_0, x) + L.$$

The construction of the gate map and the hyperbolicity of  $CV$  ensure that, after enlarging  $L$  and shrinking  $C_1$  if necessary,  $d_V(\mathfrak{g}_Y(x_0), \mathfrak{g}_Y(x)) < r$  where  $r$  depends only on  $k$  and  $\mathfrak{S}$ . The triangle inequality then gives us

$$d_V(x, \mathfrak{g}_Y(x_0)) > L \quad \text{and} \quad d_V(x, \mathfrak{g}_Y(x)) > L - r.$$

Now let  $U \in \mathfrak{S}$ . If  $\text{diam}(\pi_U(Y)) \leq B$ , then  $d_U(\mathfrak{g}_Y(x_0), \mathfrak{g}_Y(x)) \leq B$  and we are done. Thus we can assume that  $\text{diam}(\pi_U(Y)) > B$ . If  $U = V$ , then the distance  $d_U(\mathfrak{g}_Y(x_0), \mathfrak{g}_Y(x))$  is uniformly bounded above by the number  $r$  and we are done. We now consider the other possible cases depending on the relation between  $U$  and  $V$ .

**Case 1** Suppose  $V \sqsubseteq U$ . If we choose  $L$  greater than  $E + r$ , then

$$d_V(x_0, \mathfrak{g}_Y(x_0)) > E \quad \text{and} \quad d_V(x, \mathfrak{g}_Y(x)) > E.$$

Thus by the bounded geodesic image axiom (8), the  $CU$  geodesics from  $\pi_U(x_0)$  to  $\pi_U(\mathfrak{g}_Y(x_0))$  and from  $\pi_U(x)$  to  $\pi_U(\mathfrak{g}_Y(x))$  must intersect  $N_E(\rho_U^V)$ . Therefore, the distance  $d_U(\mathfrak{g}_Y(x_0), \mathfrak{g}_Y(x))$  is uniformly bounded due to the hyperbolicity of  $CU$  and the properties of the gate map (Lemma 4.14).

**Case 2** Suppose  $U \sqsubseteq V$ . If some  $CV$  geodesic from  $\pi_V(\mathfrak{g}_Y(x_0))$  to  $\pi_V(\mathfrak{g}_Y(x))$  stays  $E$ -far from  $\rho_V^U$ , then by the bounded geodesic image axiom (8),  $d_U(\mathfrak{g}_Y(x_0), \mathfrak{g}_Y(x)) \leq E$  and we are done. Therefore, we assume that all  $CV$  geodesics from  $\pi_V(\mathfrak{g}_Y(x_0))$  to  $\pi_V(\mathfrak{g}_Y(x))$  intersect  $N_E(\rho_V^U)$ . Since  $d_V(x_0, \mathfrak{g}_Y(x_0)) > d_V(x_0, x) + L$ , if there was also a  $CV$  geodesic from  $\pi_V(x_0)$  to  $\pi_V(x)$  that intersected  $N_E(\rho_V^U)$  we would have

$$\begin{aligned} d_V(\mathfrak{g}_Y(x_0), \rho_V^U) &\geq d_V(\mathfrak{g}_Y(x_0), x_0) - d_V(x_0, \rho_V^U) \\ &> d_V(\mathfrak{g}_Y(x_0), x_0) - d_V(x_0, x) - 2E \\ &\geq L - 2E. \end{aligned}$$

However,  $d_V(\mathfrak{g}_Y(x_0), \mathfrak{g}_Y(x)) \leq r$  which implies  $\pi_V(\mathfrak{g}_Y(x_0))$  lies in  $N_{E+r}(\rho_V^U)$ . Therefore, by assuming  $L > 4E + r$  we can ensure that no  $CV$  geodesic from  $\pi_V(x_0)$  to  $\pi_V(x)$  intersects  $N_E(\rho_V^U)$ . Thus  $d_U(x_0, x) < E$  by the bounded geodesic image axiom and it immediately follows that  $d_U(\mathfrak{g}_Y(x_0), \mathfrak{g}_Y(x))$  is bounded by a constant depending on  $k$  and  $\mathfrak{S}$ .

**Case 3** Suppose  $U \not\sqsubseteq V$  and  $V \not\sqsubseteq U$ . Recall that we can assume  $\text{diam}(\pi_U(Y)) > B$ . Thus if  $U \perp V$ , we have  $CV \subseteq N_B(\pi_V(Y))$  by the orthogonal projection dichotomy. However  $d_V(x_0, \mathfrak{g}_Y(x_0)) > L$ , so by Lemma 4.15 we can choose  $L$  large enough such that  $\pi_V(x_0)$  does not lie in the  $B$ -neighborhood of  $\pi_V(Y)$ . Thus  $U$  and  $V$  cannot be orthogonal and hence  $U \pitchfork V$ .

Now assume  $L > 2\kappa_0 + 3r + 2E + 1$ . Then if  $d_V(\mathfrak{g}_Y(x_0), \rho_V^U) \leq \kappa_0 + r + E$ ,

$$d_V(x_0, \rho_V^U) \geq d_V(x_0, \mathfrak{g}_Y(x_0)) - d_V(\mathfrak{g}_Y(x_0), \rho_V^U) - E \geq L - (\kappa_0 + r + E) - E > \kappa_0$$

and

$$d_V(x, \rho_V^U) \geq d_V(x, \mathfrak{g}_Y(x_0)) - d_V(\mathfrak{g}_Y(x_0), \rho_V^U) - E > L - (\kappa_0 + r + E) - E > \kappa_0.$$

Therefore,  $d_U(x_0, \rho_U^V) < \kappa_0$  and  $d_U(x, \rho_U^V) < \kappa_0$  by consistency (axiom (5)). This implies that  $d_U(x_0, x) \leq 2\kappa_0 + E$  and thus  $d_U(\mathfrak{g}_Y(x_0), \mathfrak{g}_Y(x))$  is bounded by a constant depending on  $k$  and  $\mathfrak{S}$ .

If instead  $d_V(\mathfrak{g}_Y(x_0), \rho_V^U) > \kappa_0 + r + E$ , then  $d_V(\mathfrak{g}_Y(x), \rho_V^U) > \kappa_0$  since

$$d_V(\mathfrak{g}_Y(x_0), \mathfrak{g}_Y(x)) < r.$$

By consistency,  $d_U(\mathfrak{g}_Y(x_0), \rho_U^V) < \kappa_0$  and  $d_U(\mathfrak{g}_Y(x), \rho_U^V) < \kappa_0$ , which implies that

$$d_U(\mathfrak{g}_Y(x_0), \mathfrak{g}_Y(x)) \leq 2\kappa_0 + E. \quad \square$$

**Remark 6.14** Both hypotheses on the subspace in Proposition 6.13 are in fact required. In the standard HHG structure of  $\mathbb{Z}^2$ , the subgroup  $\langle(1, 0)\rangle$  is hierarchically quasiconvex, but does not satisfy the orthogonal projection dichotomy. On the other hand, the subgroup  $\langle(1, 1)\rangle$  has the orthogonal projection dichotomy, but is not hierarchically quasiconvex. Neither of these subsets are strongly quasiconvex and thus neither are contracting. Both of the above examples can even be made to be (nonstrongly) quasiconvex by choosing  $\{(1,0), (1,1), (0,1)\}$  to be the generating set for  $\mathbb{Z}^2$ .

### 6.3 A generalization of the bounded geodesic image property

As a first application of Theorem 6.3 — our characterization of strongly quasiconvex subsets — we show that strongly quasiconvex subspaces of HHSs satisfy a version of the bounded geodesic image property. First recall the bounded geodesic image property for quasiconvex subsets of hyperbolic spaces (not to be confused with the bounded geodesic image axiom of an HHS).

**Proposition 6.15** (bounded geodesic image property for hyperbolic spaces) *Let  $Y$  be a  $K$ -quasiconvex subset of a geodesic  $\delta$ -hyperbolic space  $X$ . Then there exists  $r > 0$  (depending on  $\delta$  and  $K$ ) such that if  $d(p_Y(x), p_Y(y)) > r$ , then every geodesic connecting  $x$  and  $y$  must intersect the  $r$ -neighborhood of  $Y$ .*

In the case of strongly quasiconvex subsets of hierarchically hyperbolic space, we replace the closest point projection with the gate map and geodesics with hierarchy paths. Theorem 1.5 from the introduction will follow as a result of the following proposition, which is a version of the active subpath theorem (Proposition 4.24) for strongly quasiconvex subsets.

**Proposition 6.16** *Let  $(\mathcal{X}, \mathfrak{S})$  be an HHS with the bounded domain dichotomy and  $Y \subseteq \mathcal{X}$  be a  $Q$ -strongly quasiconvex. For all  $\lambda \geq 1$ , there exist constants  $\nu$  and  $D$ , depending on  $\lambda$  and  $Q$ , such that for all  $x, y \in \mathcal{X}$ , if  $d_{\mathcal{X}}(\mathfrak{g}_Y(x), \mathfrak{g}_Y(y)) > D$  and  $\gamma: [a, b] \rightarrow \mathcal{X}$  is a  $\lambda$ -hierarchy path joining  $x$  and  $y$ , then there is a subpath  $\alpha = \gamma|_{[a_1, b_1]}$  of  $\gamma$  with*

- (1)  $\alpha \subseteq N_{\nu}(Y)$ ,
- (2) the diameters of  $\mathfrak{g}_Y(\gamma([a, a_1]))$  and  $\mathfrak{g}_Y(\gamma([b_1, b]))$  both bounded by  $\nu$ .

**Proof** By Theorem 6.3,  $Y$  is hierarchically quasiconvex and has the orthogonal domain dichotomy. In particular,  $\pi_U(Y)$  is uniformly quasiconvex in  $CU$  for all  $U \in \mathfrak{S}$ . Let  $x, y \in \mathcal{X}$  and  $\gamma$  be a  $\lambda$ -hierarchy path connecting  $x$  and  $y$ . Since  $\gamma$  is a  $(\lambda, \lambda)$ -quasigeodesic, we can choose

$$x = x_0, x_1, x_2, \dots, x_n = y$$

on  $\gamma$  such that the distances between  $x_i$  and  $x_{i+1}$  are all bounded by  $2\lambda$ . We will show that there exist  $0 \leq i_0 \leq j_0 \leq n$  such that:

- For  $i = i_0$  or  $i = j_0$ ,  $d_{\mathcal{X}}(x_i, \mathfrak{g}_Y(x_i))$  is bounded by a constant depending only on  $Q$ ,  $\lambda$ , and  $\mathfrak{S}$ .
- If  $s < t < i_0$  or  $j_0 < s < t$ , then  $d_{\mathcal{X}}(\mathfrak{g}_Y(x_s), \mathfrak{g}_Y(x_t))$  is bounded by a constant depending only on  $Q$ ,  $\lambda$ , and  $\mathfrak{S}$ .

Since  $Y$  is strongly quasiconvex, once we have shown the above, the proposition will follow with  $\alpha$  as the subsegment of  $\gamma$  between  $x_{i_0}$  and  $x_{j_0}$ .

For each  $U \in \mathfrak{S}$ , the projection  $\pi_U$  is uniformly coarsely Lipschitz, thus there is a  $\lambda'$  depending on  $(\mathcal{X}, \mathfrak{S})$  and  $\lambda$  such that the distances  $d_U(x_i, x_{i+1})$  are all bounded above by  $\lambda'$ .

By the hyperbolicity of each  $CU$  and the properties of gate map (Lemma 4.14), there are constants  $B$  and  $\mu$  depending only on  $\mathfrak{S}$ ,  $Q$ , and  $\lambda$  such that for each  $V \in \mathfrak{S}$  satisfying  $d_V(\mathfrak{g}_Y(x), \mathfrak{g}_Y(y)) > B$  there are  $0 \leq I_V < J_V \leq n$  with the following properties:

- (1)  $d_V(x_i, \mathfrak{g}_Y(x_i)) \leq \mu$  for  $I_V \leq i \leq J_V$ .
- (2) If  $s < t < I_V$  or  $J_V < s < t$ , then  $d_V(\mathfrak{g}_Y(x_s), \mathfrak{g}_Y(x_t)) < \mu$ .
- (3)  $d_V(x_{I_V}, x_{J_V}) \geq 10D$ , where  $D = 3(E + \mu + \kappa_0 + \lambda')$ .



For future convenience, we can and shall assume  $B$  is large enough that  $B > E$ ,  $(\mathcal{X}, \mathfrak{S})$  has the  $B$ -bounded domain dichotomy, and  $Y$  has the  $B$ -orthogonal projection dichotomy. By the uniqueness axiom (10), there is a constant  $K$  depending on  $B$  and  $(\mathcal{X}, \mathfrak{S})$  such that if  $d_{\mathcal{X}}(\mathfrak{g}_Y(x), \mathfrak{g}_Y(y)) > K$ , then the set  $\mathcal{R} = \text{Rel}_B(\mathfrak{g}_Y(x), \mathfrak{g}_Y(y))$  is nonempty. Since for each  $V \in \mathcal{R}$  we have  $d_V(x_{I_V}, x_{J_V}) \geq 10D$  and each distance  $d_V(x_i, x_{i+1})$  is bounded above by  $\lambda' < D$ , there are  $I_V < i_V < j_V < J_V$  such that

$$(*) \quad D \leq d_V(x_{i_V}, x_{I_V}) \leq 2D \quad \text{and} \quad D \leq d_V(x_{j_V}, x_{J_V}) \leq 2D.$$

Let  $i_0 = \min_{V \in \mathcal{R}} i_V$  and  $j_0 = \max_{V \in \mathcal{R}} j_V$ .

We first prove that for each  $s$  and  $t$  that are both less than  $i_0$  or both greater than  $j_0$  the distance  $d_{\mathcal{X}}(\mathfrak{g}_Y(x_s), \mathfrak{g}_Y(x_t))$  is uniformly bounded by some constant depending only on  $\mathfrak{S}$ ,  $Q$  and  $\lambda$ . We will provide the proof for the case  $s$  and  $t$  are both less than  $i_0$  and the proof for the other case is essentially identical. Let  $V \in \mathfrak{S}$ . If  $V \notin \mathcal{R}$ , then  $d_V(\mathfrak{g}_Y(x), \mathfrak{g}_Y(y)) \leq B$  which implies  $\text{diam}(\pi_V(\mathfrak{g}_Y(\gamma)))$  is bounded by a constant that depends only on  $B$ ,  $\lambda$ ,  $Q$  and  $\mathfrak{S}$ . In particular,  $d_V(\mathfrak{g}_Y(x_s), \mathfrak{g}_Y(x_t))$  is also uniformly bounded by this constant. When  $V \in \mathcal{R}$ , then  $s$  and  $t$  are both less than  $i_V$ . Therefore by item (2) above and  $(*)$  we have that  $d_V(\mathfrak{g}_Y(x_s), \mathfrak{g}_Y(x_t))$  is bounded by a constant depending only on  $\mathfrak{S}$ ,  $Q$ , and  $\lambda$ . By the distance formula (Theorem 4.4) the distance  $d_{\mathcal{X}}(\mathfrak{g}_Y(x_s), \mathfrak{g}_Y(x_t))$  is therefore bounded by a constant that ultimately depends only on  $\mathfrak{S}$ ,  $Q$  and  $\lambda$ .

We now prove that there exists  $v'$  depending on  $\mathfrak{S}$ ,  $Q$  and  $\lambda$  such that for  $i = i_0$  or  $i = j_0$ ,

$$(**) \quad d_{\mathcal{X}}(x_i, \mathfrak{g}_Y(x_i)) \leq v'.$$

Again we only give the proof for the case of  $i = i_0$  and the argument for the case  $i = j_0$  is almost identical. By the distance formula, it is sufficient to check that we can uniformly bound  $d_U(x_i, \mathfrak{g}_Y(x_i))$  for each  $U \in \mathfrak{S}$ .

Fix a domain  $V \in \mathcal{R}$  such that  $i = i_0 = i_V$ . We shall show  $d_U(x_i, \mathfrak{g}_Y(x_i))$  for all  $U \in \mathfrak{S}$  by examining the four cases for how  $U$  can be related to  $V$ .

**Case 1** Suppose  $V \perp U$ . Since  $Y$  has the  $B$ -orthogonal domain dichotomy,

$$V \in \mathcal{R} \implies CU \subseteq N_B(\pi_U(Y)).$$

Therefore by the properties of the gate map (Lemma 4.14), we have that  $d_{\mathcal{X}}(x_i, \mathfrak{g}_Y(x_i))$  is uniformly bounded.

**Case 2** Suppose  $V \pitchfork U$ . If  $d_V(x_i, \rho_V^U) > \kappa_0 + \mu + E$ , then

$$d_V(\mathfrak{g}_Y(x_i), \rho_V^U) > \kappa_0$$

and by the consistency axiom (5) and triangle inequality,

$$d_U(x_i, \mathfrak{g}_Y(x_i)) \leq 2\kappa_0 + E.$$

Now assume that  $d_V(x_i, \rho_V^U) < \kappa_0 + \mu + E$ . Since  $D > \mu + E + \kappa_0$ ,  $d_V(x_i, x_{I_V}) \geq D$ , and  $d_V(x_i, x_{J_V}) \geq D$ , we have that  $x_{I_V}$ ,  $\mathfrak{g}_Y(x_{I_V})$ ,  $x_{J_V}$  and  $\mathfrak{g}_Y(x_{J_V})$  all project at least  $\kappa_0$  far from  $\rho_V^U$  in  $CV$ . Therefore, by the consistency axiom and triangle inequality,

$$d_U(x_{I_V}, \mathfrak{g}_Y(x_{I_V})) \leq 2\kappa_0 + E \quad \text{and} \quad d_U(x_{J_V}, \mathfrak{g}_Y(x_{J_V})) \leq 2\kappa_0 + E.$$

Thus, by the quasiconvexity of  $\pi_U(Y)$  in  $CU$  and the properties of the gate map, the distance  $d_U(x_i, \mathfrak{g}_Y(x_i))$  is bounded by a uniform constant determined by  $\mathfrak{S}$ ,  $Q$  and  $\lambda$ .

**Case 3** Suppose  $U \sqsubseteq V$ . Consider geodesics in  $CV$  connecting the projections of the pairs of points  $(x_{I_V}, \mathfrak{g}_Y(x_{I_V}))$ ,  $(x_i, \mathfrak{g}_Y(x_i))$  and  $(x_{J_V}, \mathfrak{g}_Y(x_{J_V}))$ . By the assumptions on  $I_V$ ,  $i$  and  $J_V$ , at most one of these geodesics intersects  $N_E(\rho_V^U)$ . If such a geodesic is not the one connecting  $\pi_V(x_i)$  and  $\pi_V(\mathfrak{g}_Y(x_i))$ , then we are done by the bounded geodesic image axiom (8). Otherwise, the bounded geodesic image axiom requires that  $\pi_V(x_{I_V})$  and  $\pi_V(x_{J_V})$  are contained in the  $3E$ -neighborhood of  $\pi_U(Y)$  in  $CU$ . By the quasiconvexity of  $\pi_U(Y)$  in  $CU$  and the properties of the gate map, the distance  $d_U(x_i, \mathfrak{g}_Y(x_i))$  is thus bounded by a uniform constant determined by  $\mathfrak{S}$ ,  $Q$  and  $\lambda$ .

**Case 4** Suppose  $V \sqsubseteq U$ . Recall that  $\pi_U(\gamma)$  is a unparametrized quasigeodesic in  $CU$ , and let  $\gamma_0$  be the subsegment of  $\pi_U(\gamma)$  from  $x_{I_V}$  to  $x_i$  and  $\gamma_1$  be the subsegment from  $x_i$  to  $x_{J_V}$ . By the bounded geodesic image axiom and  $(*)$ , there exists  $E' \geq E$  determined by  $\mathfrak{S}$ , such that both  $\gamma_0$  and  $\gamma_1$  intersect the  $E'$ -neighborhood of  $\rho_U^V$ . Since  $\pi_U(\gamma)$  is an unparametrized  $(\lambda, \lambda)$ -quasigeodesic, there exists  $R$  depending on  $E'$  and  $\lambda$  such that  $d_U(x_i, \rho_U^V) \leq R$ . If  $\alpha$  is some  $CU$  geodesic connecting  $\mathfrak{g}_Y(x)$  and  $\mathfrak{g}_Y(y)$ , then  $\alpha$  also intersects the  $E$ -neighborhood of  $\rho_U^V$  by the bounded geodesic image axiom. Therefore, by the quasiconvexity of  $\pi_U(Y)$  in  $CU$  and the properties of the gate map, the distance  $d_U(x_i, \mathfrak{g}_Y(x_i))$  is bounded by a uniform constant determined by  $\mathfrak{S}$ ,  $Q$  and  $\lambda$ . □

**Remark 6.17** The hypotheses of Proposition 6.16 cannot be relaxed by taking  $Y$  to be hierarchically quasiconvex instead of strongly quasiconvex. As a counterexample, one can consider  $\mathbb{Z}^2$  with the standard HHG structure and let  $Y$  be the  $x$ -axis. As any horizontal line in  $\mathbb{Z}^2$  is a hierarchy path, for any  $D > 0$ , there exists a hierarchy path  $\gamma$  where both  $d_{\mathcal{X}}(\gamma, Y) > D$  and  $\text{diam}(\mathfrak{g}_Y(\gamma)) > D$ .

## 7 Strongly quasiconvex subsets in familiar examples

In this section, we utilize Theorem 6.3 to give descriptions of the strongly quasiconvex subsets in well studied examples of hierarchically hyperbolic spaces. We will begin by briefly discussing the HHS structure for the mapping class group, Teichmüller space, right-angled Artin and Coxeter groups, and graph manifolds. The descriptions are not complete as we only describe the parts of the HHS structure that we shall need to be able to apply the results from the general case. We direct the reader to the references provided alongside each example for complete details.

**The mapping class group and Teichmüller space** For the mapping class group, see [10; 41]; for the Teichmüller metric, see [23; 26; 47]; and for the Weil–Petersson metric, see [17].

Let  $S$  be an oriented, connected, finite-type surface with genus  $g$  and  $p$  punctures. The *complexity of  $S$*  is  $\xi(S) = 3g - 3 + p$ . Assume  $\xi(S) \geq 1$  and let  $\mathcal{X}$  be the marking complex of  $S$ .

- **Index set**  $\mathfrak{S}$  will be the collection of isotopy classes of (not necessarily connected) essential subsurfaces of  $S$  excluding 3-punctured sphere, but including annuli.
- **Hyperbolic spaces** For each  $U \in \mathfrak{S}$ ,  $CU$  will be the curve graph of  $U$ . The space  $CU$  will be infinite diameter if and only if  $U$  is connected. The projection maps,  $\pi_U$ , are the well studied subsurface projections of Masur and Minsky.
- **Relations**  $U \perp V$  if  $U$  and  $V$  are disjoint and  $U \sqsubseteq V$  if  $U$  is nested into  $V$ . If  $U \sqsubseteq V$ , then  $\rho_V^U$  will be the subset of curves in  $CV$  corresponding to  $\partial U$ .

The HHS structure for Teichmüller space with either metric is identical except for the annular domains of  $\mathfrak{S}$ . For the Teichmüller metric, modify the curve graphs of the annular domains by attaching a horoball. For the Weil–Petersson metric, the index set  $\mathfrak{S}$  simply excludes annuli. This difference in the treatment of annular domains accounts for all of the differences in the coarse geometry of the these three spaces.

**RAAGs and RACGs** [9] Let  $\Gamma$  be a finite simplicial graph and  $G_\Gamma$  be the associated right-angled Artin or right-angled Coxeter group equipped with a word metric from a finite generating set. For an induced subgraph  $\Lambda \subseteq \Gamma$ ,  $\text{link}(\Lambda)$  is the subgraph of  $\Gamma - \Lambda$  induced by the vertices adjacent to every vertex in  $\Lambda$  and  $\text{star}(\Lambda)$  be the induced

subgraph of  $\text{link}(\Lambda) \cup \Lambda$ . If  $\Lambda$  is an induced subgraph of  $\Gamma$ , then  $G_\Lambda$  is a subgroup of  $G_\Gamma$ . We call subgroups of this form the *special subgroups* of  $G_\Gamma$ . The following is an HHG structure on  $G_\Gamma$ .

- **Index set** For  $g, h \in G_\Gamma$  and  $\Lambda$  a nonempty, induced subgraph of  $\Gamma$ , define the equivalence relation  $gG_\Lambda \sim hG_\Lambda$  if  $g^{-1}h \in G_{\text{star}(\Lambda)}$ . Let  $\mathfrak{S}$  be defined as  $\{gG_\Lambda\}/\sim$ .
- **Hyperbolic spaces**  $C[gG_\Lambda]$  can be obtained by starting with the coset  $gG_\Lambda$  and coning off each left coset of the special subgroups contained in  $gG_\Lambda$ .  $C[gG_\Lambda]$  is infinite diameter if and only if  $G_\Lambda$  is infinite and  $\Lambda$  does not split as a join.
- **Relations**  $[gG_{\Lambda'}] \sqsubseteq [gG_\Lambda]$  if  $\Lambda' \subseteq \Lambda$  and  $[gG_{\Lambda'}] \perp [gG_\Lambda]$  if  $\Lambda \subseteq \text{link}(\Lambda')$  (and hence  $\Lambda' \subseteq \text{link}(\Lambda)$ ). If  $[gG_{\Lambda'}] \sqsubseteq [gG_\Lambda]$ , then  $\rho_{[gG_\Lambda]}^{[gG_{\Lambda}]}$  will be the subset  $gG_{\Lambda'}$  in  $C[gG_\Lambda]$ .

**Graph manifolds** [10] Let  $M$  be a nongeometric graph manifold and  $\mathcal{X}$  be the universal cover of  $M$ . Since the fundamental group of every graph manifold is quasi-isometric to the fundamental group of a flip graph manifold, we will assume  $M$  is flip. Let  $T$  be Bass–Serre tree for  $M$  and  $X_v$  be the subspace of  $\mathcal{X}$  corresponding to a vertex  $v \in T$ . Each  $X_v$  is bi-Lipschitz to the product  $R_v \times H_v$  where  $R_v$  is a copy of the real line and  $H_v$  is the universal cover of a hyperbolic surface with totally geodesic boundary. If  $v, w$  are adjacent vertices in  $T$ , then let  $\partial_w H_v$  and  $\partial_v H_w$  denote the boundary components of  $H_v$  and  $H_w$  such that  $R_v \times \partial_w H_v$  is identified with  $R_w \times \partial_v H_w$  in  $\mathcal{X}$ . Since  $M$  is flip,  $R_v$  is identified with  $\partial_v H_w$ . For each  $v \in T$ , let  $\hat{H}_v$  denote the spaced obtained from  $H_v$  after coning off each copy of  $\partial_w H_v$  for each vertex  $w$  adjacent to  $v$ . The following is an HHS structure on  $\mathcal{X}$ .

- **Index set** For adjacent vertices  $v, w \in T$ , define  $R_v \sim \partial_v H_w$  and then let  $\mathfrak{S} = \{T, R_v, \partial_v H_w, \hat{H}_w\}/\sim$ .
- **Hyperbolic spaces** Every element of  $\mathfrak{S}$  is a hyperbolic space, so we have  $CU = U$  for all  $U \in \mathfrak{S}$ . The diameter of  $CU$  is infinite for all  $U \in \mathfrak{S}$ .
- **Relations**  $T$  is the  $\sqsubseteq$ -maximal domain and  $[\partial_w H_v] \sqsubseteq \hat{H}_v$  for all  $w$  and  $v$  adjacent in  $T$ . For adjacent vertices  $v, w \in T$ ,  $\rho_T^{[R_v]} = \rho_T^{[\partial_w H_v]} = \{v, w\} \subset T$  and  $\rho_{\hat{H}_v}^{[\partial_w H_v]}$  is the cone point for  $\partial_w H_v$  in  $\hat{H}_v$ . For  $v$  and  $w$  adjacent in  $T$ , we have  $[R_v] \perp \hat{H}_v$  and  $[R_v] \perp [R_w]$  (recall  $[\partial_w H_v] = [R_w]$ ).

**Remark 7.1** When the manifold  $M$  is flip, the above describes an HHG structure on  $\pi_1(M)$ . However, if  $M$  is not flip, then the quasi-isometry from  $\pi_1(M)$  to the

fundamental group of the flip graph manifold need not be equivariant and the above will be an HHS, but not an HHG structure on  $\pi_1(M)$ . See [10, Remark 10.2] for a discussion of the existence of HHG structures on 3-manifold groups.

In the case of right-angled Artin groups with connected defining graphs, Tran and Genevois independently showed that strongly quasiconvex subgroups are either finite-index or hyperbolic (and are actually free when they are hyperbolic) [28; 54]. The same result is shown for the mapping class group in [38] and for certain  $\mathcal{CFS}$  right-angled Coxeter groups in [43]. Based on these examples, one may conjecture that the strongly quasiconvex subsets of any not relatively hyperbolic, hierarchically hyperbolic space are either hyperbolic or coarsely cover the entire space. While [54] provides a counterexample to this conjecture in right-angled Coxeter groups, it nevertheless holds in many of the examples described above. In Proposition 7.2, we give sufficient conditions for every strongly quasiconvex subset of an HHS to be either hyperbolic or coarsely covering. We then unite and expand the work of Genevois, Kim, Nguyen and Tran by applying Proposition 7.2 to the mapping class group, Teichmüller space, right-angled Artin and Coxeter groups, and graph manifolds in Corollary 7.4.

**Proposition 7.2** *Let  $(\mathcal{X}, \mathfrak{S})$  be an HHS with the bounded domain dichotomy and let  $\mathfrak{S}^*$  be as defined in Definition 6.5. Assume the following two conditions hold:*

- (1) *For all  $W \in \mathfrak{S} - \mathfrak{S}^*$  either  $CW$  has bounded diameter or the set*

$$\{\rho_W^V \mid V \in \mathfrak{S}^* \text{ with } V \pitchfork W \text{ or } V \sqsubseteq W\}$$

*uniformly coarsely covers  $CW$ .*

- (2) *For every  $U, V \in \mathfrak{S}^*$  there exists a sequence  $U = U_1, \dots, U_n = V$  of domains in  $\mathfrak{S}^*$  with  $U_i \perp U_{i+1}$  for all  $1 \leq i \leq n - 1$ .*

*Then, if  $Y \subseteq \mathcal{X}$  is strongly quasiconvex, either  $Y$  is hyperbolic or some finite neighborhood of  $Y$  covers all of  $\mathcal{X}$ .*

**Proof** Let  $Y \subseteq \mathcal{X}$  be  $Q$ -strongly quasiconvex. By Theorem 6.3 there exists  $B$ , depending only on  $Q$  and  $\mathfrak{S}$ , such that  $Y$  has the  $B$ -orthogonal projection dichotomy. Further, we can assume  $B$  is large enough such that  $(\mathcal{X}, \mathfrak{S})$  satisfies the  $B$ -bounded domain dichotomy. We will show that if  $Y$  is not hyperbolic, then for all  $W \in \mathfrak{S}$  we have that  $CW$  is uniformly coarsely covered by  $\pi_W(Y)$ . Thus for all  $x \in \mathcal{X}$  we will have that  $d_W(x, \mathfrak{g}_Y(x))$  is uniformly bounded and therefore  $Y$  will coarsely cover  $\mathcal{X}$  by the distance formula (Theorem 4.4).

Suppose that  $Y$  is not hyperbolic. By Proposition 2.8, the inclusion map  $i: Y \hookrightarrow X$  cannot be a stable embedding. Therefore by Corollary 6.4, there exists a domain  $U \in \mathfrak{S}^*$  such that  $\text{diam}(\pi_U(Y)) > B$ . First we will show that for any domain  $W \in \mathfrak{S}^*$ ,  $CW \subseteq N_B(\pi_W(Y))$ .

Let  $W \in \mathfrak{S}^*$ . By hypothesis, there exists a sequence  $U = U_1, \dots, U_n = W$  of domains in  $\mathfrak{S}^*$  with  $U_i \perp U_{i+1}$  for all  $1 \leq i \leq n-1$ . Since  $Y$  has the  $B$ -orthogonal projection dichotomy and  $\text{diam}(CU_i) = \infty$  for each  $1 \leq i \leq n$ , we have  $CU_i \subseteq N_B(\pi_{U_i}(Y))$  for all  $1 \leq i \leq n$ . In particular,  $CW \subseteq N_B(\pi_W(Y))$ .

Now let  $W \in \mathfrak{S} - \mathfrak{S}^*$  such that  $\text{diam}(CW) = \infty$ . We will show that  $\pi_W(Y)$  uniformly coarsely covers  $CW$  by showing that for all  $V \in \mathfrak{S}^*$  such that  $\rho_W^V$  is defined there exists  $y \in Y$  such that  $\pi_W(y)$  is uniformly close to  $\rho_W^V$ . First suppose  $V \in \mathfrak{S}^*$  with  $V \sqsubseteq W$ . By the preceding paragraph, there exist  $x, x' \in Y$  such that  $d_V(x, x') > 100E$ . If  $\gamma$  is a hierarchy path connecting  $x$  and  $x'$ , then  $\pi_W(\gamma)$  is uniformly close to  $\rho_W^V$  by the bounded geodesic image axiom (8). Further, since  $Y$  is strongly quasiconvex there exists  $y \in Y$  such that  $d_W(\rho_W^V, \pi_W(y)) < B'$  where  $B'$  depends only on  $Q$  and  $\mathfrak{S}$ . If instead  $V \in \mathfrak{S}^*$  and  $V \pitchfork W$ , then there exists  $y \in Y$  such that  $d_V(y, \rho_V^W) > \kappa_0$ . Thus  $d_W(y, \rho_W^V) \leq \kappa_0$  by the consistency axiom (5). Since the set

$$\{\rho_W^V \mid V \in \mathfrak{S}^* \text{ with } V \pitchfork W \text{ or } V \sqsubseteq W\}$$

uniformly coarsely covers  $CW$  by hypothesis, we have that  $\pi_W(Y)$  uniformly coarsely covers all of  $CW$  as well.

Hence we have that for all  $W \in \mathfrak{S}$ ,  $CW$  is uniformly coarsely covered by  $\pi_W(Y)$  and so  $Y$  coarsely covers  $\mathcal{X}$  by the distance formula.  $\square$

Before continuing, we will take a brief detour to define a property of graphs that will be relevant to our study of right-angled Coxeter groups. Given a graph  $\Gamma$ , define  $\Gamma^4$  as the graph whose vertices are induced 4-cycles of  $\Gamma$ . Two vertices in  $\Gamma^4$  are adjacent if and only if the corresponding induced 4-cycles in  $\Gamma$  have two nonadjacent vertices in common. Given graphs  $\Lambda_1$  and  $\Lambda_2$ , recall that the join  $\Lambda_1 * \Lambda_2$  is the graph obtained from  $\Lambda_1 \sqcup \Lambda_2$  by adding an edge between each vertex of  $\Lambda_1$  and each vertex of  $\Lambda_2$ .

**Definition 7.3** (constructed from squares) A graph  $\Gamma$  is  $\mathcal{CFS}$  if  $\Gamma = \Omega * K$ , where  $K$  is a (possibly empty) clique and  $\Omega$  is a nonempty subgraph such that  $\Omega^4$  has a connected component  $T$  where every vertex of  $\Omega$  is contained in a 4-cycle that is a vertex of  $T$ . If  $\Gamma$  is  $\mathcal{CFS}$  and  $\Omega^4$  is connected, then we say  $\Gamma$  is *strongly  $\mathcal{CFS}$* . If  $\Gamma$  is

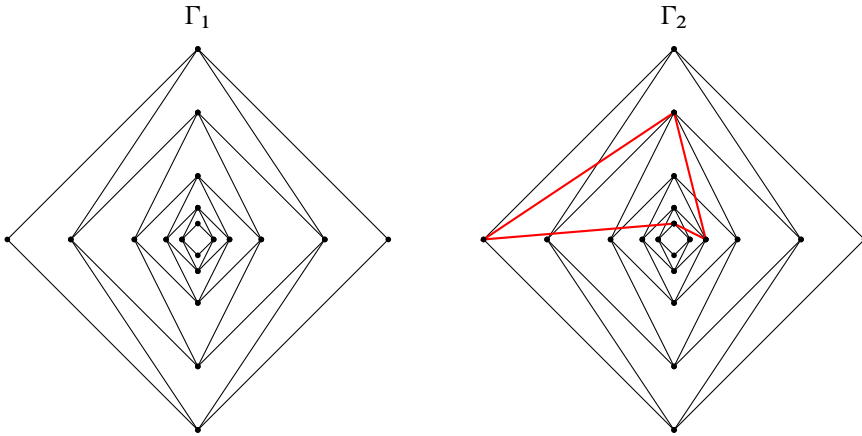


Figure 7: Two graphs  $\Gamma_1$  and  $\Gamma_2$  are both  $\mathcal{CFS}$ . However, graph  $\Gamma_1$  is graph strongly  $\mathcal{CFS}$  but  $\Gamma_2$  is not since the red induced 4-cycle in  $\Gamma_2$  is not “connected” to any other induced 4-cycle in the graph.

(strongly)  $\mathcal{CFS}$ , then by abuse of language we will say that the right-angled Coxeter group  $G_\Gamma$  is (strongly)  $\mathcal{CFS}$ . See Figure 7 for examples of  $\mathcal{CFS}$  and strongly  $\mathcal{CFS}$  graphs.

**Corollary 7.4** *The following HHSs have the property that every strongly quasiconvex subset is either hyperbolic or coarsely covers the entire space:*

- (a) *The Teichmüller space of a finite-type, oriented surface with the Teichmüller metric.*
- (b) *The Teichmüller space of a finite-type, oriented surface of complexity at least 6 with the Weil–Petersson metric.*
- (c) *The mapping class group of a finite-type, oriented surface.*
- (d) *A right-angled Artin group with connected defining graph.*
- (e) *A right-angled Coxeter group with strongly  $\mathcal{CFS}$  defining graph.*
- (f) *The fundamental group of a nongeometric graph manifold.*

*In particular, if  $H$  is a strongly quasiconvex subgroup in any of the groups (c)–(f), then  $H$  is either stable or finite-index.*

**Proof** All of the above examples have the bounded domain dichotomy. We shall show they satisfy the two hypotheses of Proposition 7.2.

**Mapping class group and Teichmüller metric** If  $\xi(S) \leq 1$ , then the mapping class group and Teichmüller space will both be hyperbolic; thus we can assume  $\xi(S) \geq 2$ . In this case,  $\mathfrak{S}^*$  is the set of all connected proper subsurfaces. Thus hypothesis (1) follows from the fact that every curve on the surface corresponds to the boundary curve of some connected subsurface. Given two subsurfaces  $U$  and  $V$ , a sequence satisfying hypothesis (2) is found by taking a path in  $CS$  connecting  $\partial U$  and  $\partial V$ .

**Weil–Peterson**  $\mathfrak{S}^*$  is the collection of all connected proper subsurfaces whose complement contains a subsurface of complexity at least 1. In particular, since the complexity is at least 6,  $\mathfrak{S}^*$  contains every subsurface of complexity 1. For every connected subsurface  $W \notin \mathfrak{S}^*$ , every curve on  $W$  corresponds to the boundary curve of some complexity 1 subsurface providing hypothesis (1). Hypothesis (2) follows from the observations that if  $U \subset S$  is a subsurface of complexity 1 and  $\alpha$  is a curve disjoint from  $U$ , then there exists  $V \subseteq S$ , a subsurface of complexity 1, such that  $\alpha \subseteq \partial V$  and  $U$  is disjoint from  $V$ . Thus any path in  $CS$  can be promoted to a sequence of sequentially disjoint subsurfaces in  $\mathfrak{S}^*$ .

**RAAGs**  $\mathfrak{S}^*$  is the collection of  $[gG_\Lambda]$  such that there exists  $\Delta \subseteq \text{link}(\Lambda)$  where  $\Lambda$  and  $\Delta$  are both nonempty and not joins. In particular, since  $\Gamma$  is connected,  $\mathfrak{S}^*$  contains all of the  $[gG_\Lambda]$  where  $\Lambda$  is a single vertex. Hypothesis (1) follows from the fact that  $G_\Lambda$  acts cocompactly on its Cayley graph and the construction of  $C[gG_\Lambda]$ . For hypothesis (2), let  $[g_1G_{\Lambda_1}], [g_2G_{\Lambda_2}] \in \mathfrak{S}^*$  and  $m = |g_1^{-1}g_2|$ . We shall proceed by induction on  $m$ . If  $m = 0$ , then  $g_1 = g_2 = g$  and since  $\Gamma$  is connected, there is a sequence of vertices  $v_1, v_2, \dots, v_n$  such that  $v_i$  and  $v_{i+1}$  are adjacent for all  $1 \leq i \leq n-1$  and  $v_1 \in \text{link}(\Lambda_1), v_n \in \text{link}(\Lambda_2)$ . Thus  $[gG_{\Lambda_1}], [gG_{v_1}], \dots, [gG_{v_n}], [gG_{\Lambda_2}]$  is the required sequence.

If  $m > 0$ , then there exists  $g_3 \in G_\Gamma$  such that  $|g_1^{-1}g_3| = m - 1$  and  $|g_3^{-1}g_2| = 1$ . Let  $v$  be the vertex of  $\Gamma$  such that  $g_3^{-1}g_2$  is either  $v$  or  $v^{-1}$ . By induction, there exist two sequences of elements of  $\mathfrak{S}^*$ ,

$$[g_1G_{\Lambda_1}] = U_1, U_2, \dots, U_n = [g_3G_v] \quad \text{and} \quad [g_2G_v] = V_1, V_2, \dots, V_k = [g_2G_{\Lambda_2}],$$

such that  $U_i \perp U_{i+1}$  for  $1 \leq i \leq n - 1$  and  $V_i \perp V_{i+1}$  for all  $1 \leq i \leq k - 1$ . Since  $[g_3G_v] = [g_2G_v]$ ,

$$[g_1G_{\Lambda_1}] = U_1, U_2, \dots, U_n, V_2, \dots, V_n = [g_2G_{\Lambda_2}]$$

is the required sequence.



**RACGs** Since  $\Gamma$  is strongly  $\mathcal{CFS}$ , we can write  $\Gamma = \Omega * K$  where  $K$  is a clique (possibly empty) and  $\Omega$  is a nonempty graph such that  $\Omega^4$  is connected and every vertex of  $\Omega$  is contained a 4-cycle that is a vertex of  $\Omega^4$ . Since  $G_\Omega$  is a finite-index subgroup of  $G_\Gamma$ , it suffices to prove that every strongly quasiconvex subset of  $G_\Omega$  is either hyperbolic or coarsely covers  $G_\Omega$ . We now prove that the standard HHG structure,  $\mathfrak{S}$ , on  $G_\Omega$  satisfies satisfy the two hypotheses of Proposition 7.2. The argument will be similar to the case of right-angled Artin groups above.

We first observe that  $\mathfrak{S}^*$  is the collection of  $[gG_\Lambda]$  such that there exists  $\Delta \subseteq \text{link}(\Lambda)$  where  $\Lambda$  and  $\Delta$  both have at least two points and they are not joins. In particular,  $\mathfrak{S}^*$  contains all domains  $[gG_{\{a,b\}}]$  where  $a$  and  $b$  are two nonadjacent vertices of an induced 4-cycle. Hypothesis (1) follows from the fact that  $G_\Lambda$  acts cocompactly on its Cayley graph and the construction of  $C[gG_\Lambda]$ .

For hypothesis (2), let  $[g_1G_{\Lambda_1}], [g_2G_{\Lambda_2}] \in \mathfrak{S}^*$  and  $m = |g_1^{-1}g_2|$ . We shall proceed by induction on  $m$ . We first assume that  $m = 0$ . Therefore,  $g_1 = g_2 = g$ . We note that for  $i = 0$  or  $1$  there exists  $\Delta_i \subseteq \text{link}(\Lambda_i)$  where  $\Lambda_i$  and  $\Delta_i$  both contain at least two vertices and are not joins. Therefore,  $\text{link}(\Lambda_i)$  contains a pair  $(u_i, v_i)$  of two nonadjacent vertices of some induced 4-cycle. Since  $\Omega^4$  is connected, there is a sequence of pairs of nonadjacent vertices  $(u_1, v_1) = (a_1, b_1), (a_2, b_2), \dots, (a_n, b_n) = (u_2, v_2)$  such that  $a_i$  and  $b_i$  are both adjacent to  $a_{i+1}$  and  $b_{i+1}$  for all  $1 \leq i \leq n - 1$ . Thus  $[gG_{\Lambda_1}], [gG_{\{a_1,b_1\}}], \dots, [gG_{\{a_n,b_n\}}], [gG_{\Lambda_2}]$  is the required sequence.

If  $m > 0$ , then there exists  $g_3 \in G_\Omega$  such that  $|g_1^{-1}g_3| = m - 1$  and  $|g_3^{-1}g_2| = 1$ . Let  $v$  be the vertex of  $\Omega$  such that  $g_3^{-1}g_2 = v$ . Since every vertex of  $\Omega$  is contained in a four cycle that is a vertex of  $\Omega^4$ , there is a vertex  $w$  such that  $v$  and  $w$  are two nonadjacent vertices of an induced 4-cycle. By induction, there exist two sequences of elements of  $\mathfrak{S}^*$ ,

$$[g_1G_{\Lambda_1}] = U_1, U_2, \dots, U_n = [g_3G_{\{v,w\}}]$$

and

$$[g_2G_{\{v,w\}}] = V_1, V_2, \dots, V_k = [g_2G_{\Lambda_2}],$$

such that  $U_i \perp U_{i+1}$  for  $1 \leq i \leq n - 1$  and  $V_i \perp V_{i+1}$  for all  $1 \leq i \leq k - 1$ . Since  $[g_3G_{\{v,w\}}] = [g_2G_{\{v,w\}}]$ ,

$$[g_1G_{\Lambda_1}] = U_1, U_2, \dots, U_n, V_2, \dots, V_n = [g_2G_{\Lambda_2}]$$

is the required sequence.

**Graph manifolds** In this case,  $\mathfrak{S}^* = \mathfrak{S} - \{T\}$  and hypothesis (1) is immediate from the facts that for every vertex  $v \in T$  is an element of  $\rho_T^{[R_v]}$  and every point in  $H_v$  is uniformly close to some boundary component  $\partial_w H_v$ . For hypothesis (2), consider  $U, W \in \mathfrak{S}^*$ . If  $U = [R_u]$  and  $W = [R_w]$ , let  $v_1, \dots, v_n$  be a sequence of adjacent vertices in  $T$  such that  $v_1$  is adjacent to  $u$  and  $v_n$  is adjacent to  $w$ . In this case the sequence  $[R_u], [R_{v_1}], \dots, [R_{v_n}], [R_w]$  satisfies the hypothesis. If  $U = [\hat{H}_u]$  or  $W = [\hat{H}_w]$ , the hypothesis is satisfied by adding  $[\hat{H}_u]$  before  $[R_u]$  or  $[\hat{H}_w]$  after  $[R_w]$  to  $[R_u], [R_{v_1}], \dots, [R_{v_n}], [R_w]$  as needed.  $\square$

In the setting of 2-dimensional right-angled Coxeter groups, Tran provided a characterization of the special strongly quasiconvex subgroups [54]. This characterization was expanded by Genevois to include all right-angled Coxeter groups in [28]. We provide a new proof of this characterization using Theorem 6.3.

**Theorem 7.5** [28; 54] *Let  $\Gamma$  be a simplicial graph and  $\Delta$  an induced subgraph of  $\Gamma$ . If  $G_\Gamma$  is the right-angled Coxeter group corresponding to  $\Gamma$  and  $G_\Delta$  is the subgroup generated by the vertices of  $\Delta$ , then the following conditions are equivalent.*

- (1) *The subgroup  $G_\Delta$  is strongly quasiconvex in  $G_\Gamma$ .*
- (2) *If  $\Delta$  contains two nonadjacent vertices of an induced 4-cycle  $\sigma$ , then  $\Delta$  contains all vertices of  $\sigma$ .*

**Proof** Before we begin, we document a few additional facts we will need about the HHG structure on a right-angled Coxeter group. For any induced subgraph  $\Lambda$ ,  $\mathbf{P}_{[G_\Lambda]}$  is coarsely equal to the subgroup  $G_\Lambda \times G_{\text{link}(\Lambda)}$  and  $G_\Lambda$  can be coarsely identified with  $\mathbf{F}_{[G_\Lambda]}$ . In particular,  $G_\Lambda$  is hierarchically quasiconvex,  $\pi_U(G_\Lambda)$  uniformly coarsely covers  $CU$  for  $U \subseteq [G_\Lambda]$ , and  $\pi_V(G_\Lambda)$  is uniformly bounded for all  $V \not\subseteq [G_\Lambda]$ . See [9] for full details on the HHG structure on right-angled Coxeter groups.

(1)  $\implies$  (2) Assume for a contradiction that  $G_\Delta$  is strongly quasiconvex, but there is a 4-cycle  $\sigma$  with two pairs of nonadjacent vertices  $\{a_1, a_2\}$  and  $\{b_1, b_2\}$  such that  $\{a_1, a_2\}$  is a subset of  $\Delta$  and  $\{b_1, b_2\}$  is not. We know that  $U = [G_{\{a_1, a_2\}}]$  and  $[G_{\{b_1, b_2\}}] = V$  are orthogonal domains. However,  $\pi_U(G_\Delta)$  coarsely covers  $CU$ , but  $\pi_V(G_\Delta)$  has uniformly bounded diameter which contradicts Theorem 6.3.

(2)  $\implies$  (1) As  $G_\Delta$  is hierarchically quasiconvex, we only need to demonstrate that  $G_\Delta$  satisfies the orthogonal projection dichotomy. Let  $B$  be a positive number such that  $(G_\Gamma, \mathfrak{S})$  has the  $B$ -bounded domain dichotomy,  $CW \subseteq N_B(\pi_W(G_\Delta))$  for all  $W \subseteq [G_\Delta]$ , and  $\text{diam}(\pi_W(G_\Delta)) < B$  for all  $W \not\subseteq [G_\Delta]$ . If  $\text{diam}(\pi_U(G_\Delta)) > B$ , then it

must be the case that  $U = [G_\Lambda]$  where  $\Lambda \subseteq \Delta$  and  $\Lambda$  contains two nonadjacent vertices  $s$  and  $t$ . If  $V \in \mathfrak{S}_V^\perp$ , then  $V = [G_{\Lambda'}]$  where  $\Lambda' \subseteq \text{link}(\Lambda)$  and  $\Lambda \subseteq \text{link}(\Lambda')$ . If  $\Lambda'$  is a join or  $\Lambda' = \{v\}$ , then  $\text{diam}(CV) \leq B$  and  $CV \subseteq N_{2B}(\pi_V(G_\Delta))$ . In the other case, we will show  $\Lambda' \subseteq \Delta$ .

If  $\Lambda'$  is not a join and contains at least two vertices, then for each vertex  $v \in \Lambda'$  there exists a vertex  $w \in \Lambda'$  that is not adjacent to  $v$ . Since  $\Lambda \subseteq \text{link}(\Lambda')$ , the vertices  $v, s, w$  and  $t$  form a 4-cycle. However, (2) then requires  $v, w \in \Delta$ . Hence,  $\Lambda' \subseteq \Delta$  and  $V = [G_{\Lambda'}] \subseteq [G_\Delta]$  implying  $CV \subseteq N_B(\pi_V(G_\Delta))$ . Thus  $G_\Delta$  has the  $2B$ -orthogonal projection dichotomy and we are finished by Theorem 6.3. □

### 7.1 CFS right-angled Coxeter groups

Recently, Behrstock proposed the program of classifying all  $\mathcal{CFS}$  right-angled Coxeter groups up to quasi-isometry and commensurability. This was motivated by the genericity of  $\mathcal{CFS}$  right-angled Coxeter groups among random right-angled Coxeter groups as well as the fact that being  $\mathcal{CFS}$  is a necessary (but not sufficient) condition for a right-angled Coxeter group to be quasi-isometric to a right-angled Artin group; see [7].

In [7], Behrstock presented the first example of a  $\mathcal{CFS}$  right-angled Coxeter group that contains a one-ended stable subgroup answering outstanding questions about stable subgroups and quasi-isometries between right-angled Artin groups and right-angled Coxeter groups. Using Theorem 7.5, we can expand Behrstock’s construction to produce  $\mathcal{CFS}$  right-angled Coxeter groups that contain any other right-angled Coxeter group as a strongly quasiconvex subgroup. This shows that there is incredible diversity among the quasi-isometry types of  $\mathcal{CFS}$  right-angled Coxeter groups.

**Proposition 7.6** *Any right-angled Coxeter group (resp. hyperbolic right-angled Coxeter group) is an infinite-index strongly quasiconvex subgroup (resp. stable subgroup) of a  $\mathcal{CFS}$  right-angled Coxeter group.*

**Proof** To prove the proposition we shall utilize a construction of certain  $\mathcal{CFS}$  graphs described in [7]. Let  $\Omega_n$  be a graph with  $2n$  vertices built in the following inductive way. Let  $\Omega_1$  be a pair of vertices  $a_1, b_1$  with no edge between them. Given the graph  $\Omega_{n-1}$ , we obtain the graph  $\Omega_n$  by adding a new pair of vertices  $a_n$  and  $b_n$  to the graph  $\Omega_{n-1}$  and adding four new edges, one connecting each of  $\{a_{n-1}, b_{n-1}\}$  to each of  $\{a_n, b_n\}$ . In Figure 8, graph  $\Gamma_1$  is exactly  $\Omega_{13}$ . For each integer  $m \geq 2$  there is a sufficiently large  $n$  such that the graph  $\Omega_n$  contains  $m$  vertices whose pairwise distances are at least 3.

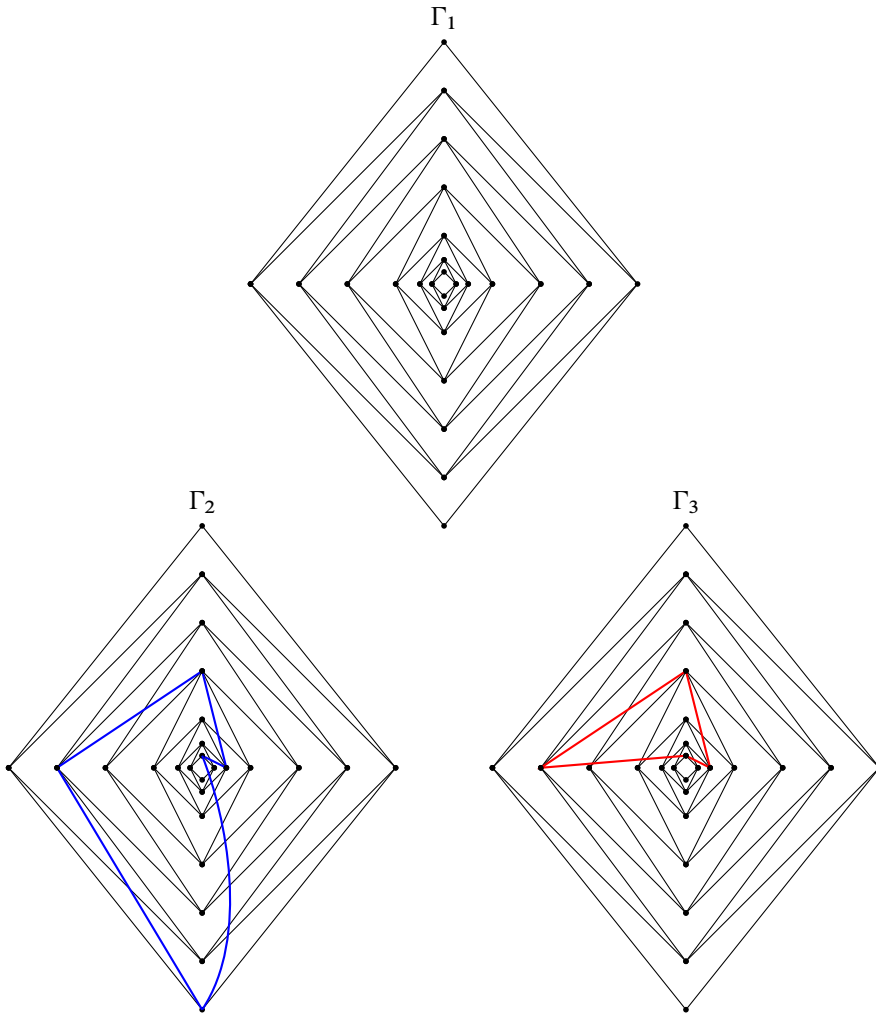


Figure 8: Three graphs  $\Gamma_1$ ,  $\Gamma_2$ , and  $\Gamma_3$  are all  $\mathcal{CFS}$ , but no pair of them are quasi-isometric.

Let  $G_\Gamma$  be an arbitrary right-angled Coxeter group. We will construct a  $\mathcal{CFS}$  right-angled Coxeter group  $G_\Omega$  that contains  $G_\Gamma$  as a strongly quasiconvex subgroup. Let  $m$  be a number of vertices of  $\Gamma$ . Choose a positive integer  $n$  sufficient large so the graph  $\Omega_n$  contains a set  $S$  of  $m$  vertices whose pairwise distance is at least 3. We glue the graphs  $\Gamma$  and  $\Omega_n$  by identifying the vertex set of  $\Gamma$  to  $S$ . Let  $\Omega$  be the resulting graph. In Figure 8, graph  $\Gamma_2$  is an example of graph  $\Omega$  when  $\Gamma$  is the 5-cycle graph and graph  $\Gamma_3$  is another example of graph  $\Omega$  when  $\Gamma$  is the 4-cycle graph.

	$G_{\Gamma_1}$	$G_{\Gamma_2}$	$G_{\Gamma_3}$
strongly $\mathcal{CFS}$	yes	yes	no
noncoarsely covering strongly quasiconvex subsets	all quasitrees	all hyperbolic; contains a one-ended stable subgroup	contains a strongly quasiconvex virtually $\mathbb{Z}^2$ subgroup
Morse boundary	totally disconnected	contains a circle	connectivity unknown
quasi-isometric to an RAAG	yes	no	no

Table 1: Note that Karrer has since shown that the Morse boundary of  $G_{\Gamma_3}$  is totally disconnected [37].

The graphs  $\Omega$  and  $\Omega_n$  have the same vertex set and  $\Omega_n^4 \subset \Omega^4$ . Thus  $\Omega$  is a  $\mathcal{CFS}$  graph as  $\Omega_n$  is a  $\mathcal{CFS}$  graph. Since the distance in  $\Omega_n$  between any distinct vertices of  $S$  is at least 3,  $\Gamma$  is an induced subgraph of  $\Omega$  with the property that if  $\Gamma$  contains two nonadjacent vertices of an induced 4-cycle  $\sigma$ , then  $\Gamma$  contains all vertices of  $\sigma$ . Therefore,  $G_\Gamma$  is a strongly quasiconvex subgroup of  $G_\Omega$  by Theorem 7.5. If  $G_\Gamma$  is a hyperbolic group, then it is a stable subgroup of  $G_\Omega$ .  $\square$

In light of Proposition 7.6, we believe that strongly quasiconvex subgroups will play an important role in understanding the quasi-isometry classification of  $\mathcal{CFS}$  right-angled Coxeter groups. In particular, it suggests that the quasi-isometry classification of  $\mathcal{CFS}$  right-angled Coxeter groups may be no simpler than the quasi-isometry classification of all right-angled Coxeter groups.

We finish this section by illustrating the results of this section with three  $\mathcal{CFS}$  right-angled Coxeter groups whose quasi-isometry types can be distinguished utilizing their strongly quasiconvex subsets.

**Example 7.7** Let  $\Gamma_1, \Gamma_2,$  and  $\Gamma_3$  be the graphs in Figure 8. All of the right-angled Coxeter groups  $G_{\Gamma_1}, G_{\Gamma_2},$  and  $G_{\Gamma_3}$  are  $\mathcal{CFS}$ , but no pair of them are quasi-isometric. By [43],  $G_{\Gamma_1}$  is quasi-isometric to a right-angled Artin group with connected defining graph. Thus, all of  $G_{\Gamma_1}$ 's noncoarsely covering strongly quasiconvex subsets are quasitrees. However,  $G_{\Gamma_2}$  contains a one-ended hyperbolic strongly quasiconvex subgroup (induced by the blue 5-cycle) and  $G_{\Gamma_3}$  contain a virtually  $\mathbb{Z}^2$  strongly quasiconvex subgroup (induced by the red 4-cycle). Table 1 summarizes some of the differences between  $G_{\Gamma_1}, G_{\Gamma_2},$  and  $G_{\Gamma_3}$ .

## 8 Hyperbolically embedded subgroups of HHGs

In this section, we utilize Theorem 6.3 to prove the following classification of hyperbolically embedded subgroups of hierarchically hyperbolic groups. As our proof does not directly utilize the definition of hyperbolically embedded, we shall omit the definition here and direct the curious reader to [20].

**Theorem 8.1** *Let  $G$  be a hierarchically hyperbolic group and let  $\{H_i\}$  be a finite collection of subgroups. Then the following are equivalent:*

- (1) *The collection  $\{H_i\}$  is hyperbolically embedded in  $G$ .*
- (2) *The collection  $\{H_i\}$  is almost malnormal and each  $H_i$  is strongly quasiconvex.*

Combining work of Dahmani, Guirardel and Osin [20] and Sisto [50], the implication (1)  $\implies$  (2) holds for all finitely generated groups. To see that the converse does not hold in general, consider a nonvirtually cyclic lacunary hyperbolic group  $G$  where every proper subgroup is infinite cyclic and strongly quasiconvex — the existence of such a group is shown in [44, Theorem 1.12]. If  $I$  is a proper subgroup of  $G$ , then by [54, Theorem 1.2],  $I$  has finite index in its commensurator  $H$ . Thus  $H$  is a proper, infinite, almost malnormal, strongly quasiconvex subgroup of  $G$ . However,  $H$  cannot be hyperbolically embedded as  $G$  does not contain any nonabelian free subgroups and thus fails to be acylindrically hyperbolic; see [20; 45].

Despite this failure in general, Genevois showed that in the setting of CAT(0) cubical groups, (2) does imply (1) [28, Theorem 6.31]. Genevois employed a combination of the Bestvina–Bromberg–Fujiwara construction [12, Theorems A and B] with some work of Sisto [49, Theorems 6.3 and 6.4] that is summarized in the following sufficient conditions for a collection of subgroups to be hyperbolically embedded.

**Theorem 8.2** [12; 49] *Let  $G$  be a finitely generated group and  $\mathcal{Z}$  be the collection of all (left) cosets of a finite collection of finitely generated subgroups  $\{H_i\}$  in  $G$ . Fix a finite generating set  $S$  for  $G$  such that  $H_i = \langle H_i \cap S \rangle$  for all  $i$ . Suppose for every  $Z_1 \neq Z_2 \in \mathcal{Z}$  we are given a subset  $\tau_{Z_1}(Z_2) \subseteq Z_1$  and for  $Z_1, Z_2, Z_3 \in \mathcal{Z}$  define  $d_{Z_3}^\tau(Z_1, Z_2) = \text{diam}_{Z_3}(\tau_{Z_3}(Z_1) \cup \tau_{Z_3}(Z_2))$ . The collection  $\{H_i\}$  is hyperbolically embedded in  $G$  if there exists  $C > 0$  such that:*

- (P0) *For all  $Z_1 \neq Z_2$ ,  $\text{diam}(\tau_{Z_1}(Z_2)) \leq C$ .*
- (P1) *For any triple  $Z_1, Z_2, Z_3 \in \mathcal{Z}$  of distinct elements, at most one of the three numbers  $d_{Z_1}^\tau(Z_2, Z_3)$ ,  $d_{Z_2}^\tau(Z_1, Z_3)$  and  $d_{Z_3}^\tau(Z_1, Z_2)$  is greater than  $C$ .*

(P2) For any  $Z_1, Z_2 \in \mathcal{Z}$ , the set

$$\{Z \in \mathcal{Z} \mid d_Z^\tau(Z_1, Z_2) > C\}$$

is finite.

(P3) For all  $g \in G$ ,  $d_{gZ_1}^\tau(gZ_2, gZ_3) = d_{Z_1}^\tau(Z_2, Z_3)$  for any  $Z_1, Z_2, Z_3 \in \mathcal{Z}$ .

As Genevois does in the cubical case, we shall show that an almost malnormal collection of strongly quasiconvex subgroups of an HHG satisfies (P0)–(P3) of Theorem 8.2. The bulk of that work is in Proposition 8.6, which we will state and prove after collecting a few preliminary lemmas.

**Lemma 8.3** *Let  $\{H_1, \dots, H_n\}$  be an almost malnormal collection of subgroups of a finitely generated group  $G$  and  $B \geq 0$ . For all  $g_1, g_2 \in G$ , if  $g_1H_i \neq g_2H_j$ , then  $\text{diam}(N_B(g_1H_i) \cap N_B(g_2H_j))$  is finite.*

**Proof** The conclusion follows directly from [36, Proposition 9.4] and the definition of almost malnormal. □

The next two lemmas tell us that a hierarchically quasiconvex subset coarsely intersects a strongly quasiconvex subset whenever the image under the gate map is large. Further, the diameter of this coarse intersection is proportional to the diameter of the gate. In addition to being key components in our proof of Theorem 8.1, these lemmas can also be interpreted as additional generalizations of the bounded geodesic image property of strongly quasiconvex subsets of hyperbolic spaces.

**Lemma 8.4** *Let  $(\mathcal{X}, \mathfrak{S})$  be an HHS with the bounded domain dichotomy,  $A \subseteq \mathcal{X}$  be  $k$ -hierarchically quasiconvex subset, and  $Y \subseteq \mathcal{X}$  be  $Q$ -strongly quasiconvex. There exists  $r > 1$  depending on  $Q$  and  $k$  such that if  $\text{diam}_{\mathcal{X}}(g_Y(A)) > r$ , then  $d_{\mathcal{X}}(a, g_Y(a)) < r$  for each  $a \in g_A(Y)$ .*

**Proof** By Proposition 5.7, there exists  $k'$  such that both  $A$  and  $Y$  are  $k'$ -hierarchically quasiconvex. Recall that for each point  $x \in \mathcal{X}$  and  $U \in \mathfrak{S}$ , the distance in  $CU$  between  $g_Y(x)$  and the closest point projection of  $\pi_U(x)$  onto  $\pi_U(Y)$  is uniformly bounded by some  $\epsilon > 1$ . Let  $K \geq \epsilon$  be such that  $Y$  has the  $K$ -orthogonal projection dichotomy and that  $K$  is larger than the constant from the bridge theorem (Theorem 4.18) determined by  $k'$ . Define  $\mathcal{H} = \{U \in \mathfrak{S} : \text{diam}(\pi_U(g_Y(A))) > 2K\}$ . By the uniqueness axiom (10), there exists  $C$  such that if  $\text{diam}(g_Y(A)) > C$ , then  $\mathcal{H} \neq \emptyset$ . Assume  $\text{diam}(g_Y(A)) > C$  and let  $a \in g_A(Y)$ . By (5) of the bridge theorem,  $\text{Rel}_{2K}(a, g_Y(a)) \subseteq \mathcal{H}^\perp$ . Suppose for the purposes of contradiction that  $V \in \text{Rel}_{2K}(a, g_Y(a))$ . Thus, there must exist  $H \in \mathcal{H}$

with  $V \perp H$ . By Theorem 6.3,  $CH \subseteq N_K(\pi_H(Y))$  and  $CV \subseteq N_K(\pi_V(Y))$  which implies that  $d_V(a, g_Y(a)) < K + \epsilon < 2K$ . However, this contradicts  $V \in \text{Rel}_{2K}(a, g_Y(a))$ . Hence,  $\text{Rel}_{2K}(a, g_Y(a)) = \emptyset$ , and by the distance formula (Theorem 4.4), there exists  $K'$  depending only on  $K$  (and thus only on  $Q$  and  $\kappa_1$ ) such that  $d_X(a, g_Y(a)) < K'$ . The conclusion follows by choosing  $r = \max\{K', C\}$ .  $\square$

**Lemma 8.5** *Let  $(X, \mathfrak{S})$  be an HHS with the bounded domain dichotomy,  $A \subseteq X$  be a  $k$ -hierarchically quasiconvex subset and  $Y \subseteq X$  be  $Q$ -strongly quasiconvex. There exists  $r > 1$  depending on  $k$  and  $Q$  such that for all  $D \geq r$  if  $\text{diam}(g_Y(A)) > r$ , then there exists  $K \geq 1$  depending on  $k, D$  and  $Q$  such that*

$$\text{diam}(N_D(A) \cap N_D(Y)) \asymp_{1,K} \text{diam}(g_Y(A)).$$

**Proof** Let  $r$  be the constant given by Lemma 8.4 and suppose  $\text{diam}(g_Y(A)) > r$ . Thus, for  $D \geq r$ ,  $\text{diam}(N_D(A) \cap N_D(Y)) \neq \emptyset$ . First consider  $x, y \in N_D(A) \cap N_D(Y)$ . Let  $x', y' \in A$  be points such that  $d_X(x, x') \leq D$  and  $d_X(y, y') \leq D$ . By Lemma 4.15 and the fact that  $x, y \in N_D(Y)$ , there exists  $K'$  depending on  $Q$  such that

$$d_X(x, g_Y(x')) \leq 4DK' \quad \text{and} \quad d_X(y, g_Y(y')) \leq 4DK'.$$

Hence,

$$d_X(x, y) \leq d_X(g_Y(x'), g_Y(y')) + 8DK',$$

which shows

$$\text{diam}(N_D(A) \cap N_D(Y)) \leq \text{diam}(g_Y(A)) + 8DK'.$$

For the inequality  $\text{diam}(g_Y(A)) \leq \text{diam}(N_D(A) \cap N_D(Y))$ , Lemma 8.4 provides  $g_Y(g_A(Y)) \subseteq N_D(A) \cap N_D(Y)$  and the bridge theorem (Theorem 4.18) says there exists  $K''$  depending on  $k$  and  $Q$  such that  $g_Y(A) \subseteq N_{K''}(g_Y(g_A(Y)))$ . Thus,

$$\text{diam}(g_Y(A)) \leq \text{diam}(g_Y(g_A(Y))) + 2K'' \leq \text{diam}(N_D(A) \cap N_D(Y)) + 2K''$$

and we are finished by choosing  $K = \max\{2K'', 6DK' + 3K'\}$ .  $\square$

We now prove that the cosets of a collection of almost malnormal, strongly quasiconvex subgroups of an HHG satisfy (P0)–(P2) of Theorem 8.2 when  $\tau_{Z_1}(Z_2)$  is defined by the gate map. This is the main tool for the proof of Theorem 8.1.

**Proposition 8.6** *Let  $(G, \mathfrak{S})$  be an HHG and  $d(\cdot, \cdot)$  denote the distance in the word metric on  $G$  with respect to some fixed finite generating set. If  $\{H_1, \dots, H_n\}$  is a collection of  $Q$ -strongly quasiconvex, almost malnormal subgroups of  $G$  and  $\mathcal{Z}$  is the collection of all left cosets of the  $H_i$ , then there exists  $C > 0$  such that, for all distinct  $Z_1, Z_2, Z_3 \in \mathcal{Z}$ :*



- (1)  $\text{diam}(\mathfrak{g}_{Z_1}(Z_2)) \leq C$ .
- (2) If  $d(\mathfrak{g}_{Z_3}(Z_1), \mathfrak{g}_{Z_3}(Z_2)) > C$ , then
 
$$d(\mathfrak{g}_{Z_2}(Z_1), \mathfrak{g}_{Z_2}(Z_3)) < C \quad \text{and} \quad d(\mathfrak{g}_{Z_1}(Z_2), \mathfrak{g}_{Z_1}(Z_3)) < C.$$
- (3)  $\{Z \in \mathcal{Z} \mid d(\mathfrak{g}_Z(Z_1), \mathfrak{g}_Z(Z_2)) > C\}$  has only a finite number of elements.

**Proof** We will prove each of the three assertions individually. Before beginning, we remind the reader that all hierarchically hyperbolic groups satisfy the bounded domain dichotomy and that every element of  $\mathcal{Z}$  is  $k$ -hierarchically quasiconvex for some  $k$  depending only on  $Q$ .

**Assertion (1)** There exists  $C_1 > 0$  such that  $\text{diam}(\mathfrak{g}_{Z_1}(Z_2)) \leq C_1$  for all  $Z_1, Z_2 \in \mathcal{Z}$ .

**Proof** Let  $r > 1$  be the constant from Lemma 8.5 for  $Q$  and define

$$F = \{gH_i \in \mathcal{Z} \mid gH_i \cap B_r(e) \neq \emptyset\}$$

where  $B_r(e)$  is the ball of radius  $r$  around the identity in  $G$ . Since  $F$  is a finite set, Lemma 8.3 provides a uniform number  $D_1$  such that  $\text{diam}(N_r(gH_i) \cap N_r(H_j)) \leq D_1$  for any distinct  $gH_i, H_j \in F$ . By Lemma 8.5, there exists  $D_2$  depending on  $Q$  such that  $\text{diam}(\mathfrak{g}_{H_j}(gH_i)) \leq D_2$  where  $gH_i \neq H_j$  are elements in  $F$ .

We now prove that there is a uniform constant  $C_1$  such that for each pair of distinct cosets  $g_1H_i$  and  $g_2H_j$  we have

$$\text{diam}(\mathfrak{g}_{g_1H_i}(g_2H_j)) \leq C_1.$$

If  $\text{diam}(\mathfrak{g}_{g_1H_i}(g_2H_j)) \leq r$ , then we are done. Otherwise, by Lemma 8.4, there are elements  $h_i \in H_i$  and  $h_j \in H_j$  such that  $d_G(g_1h_i, g_2h_j) < r$ . This implies that  $h_i^{-1}g_1^{-1}g_2h_j$  is an element in  $F$  and  $h_i^{-1}g_1^{-1}g_2h_j \neq H_i$ . Therefore,

$$\text{diam}(\mathfrak{g}_{H_i}(h_i^{-1}g_1^{-1}g_2h_j)) \leq D_2.$$

Thus, by the coarse equivariance of the gate maps (Lemma 4.16), the diameter of  $\mathfrak{g}_{g_1H_i}(g_2H_j)$  is bounded above by a uniform number  $C_1$ . □

**Assertion (2)** There exists  $C_2 > 0$  such that for all  $Z_1, Z_2, Z_3 \in \mathcal{Z}$ , if

$$d(\mathfrak{g}_{Z_3}(Z_1), \mathfrak{g}_{Z_3}(Z_2)) > C_2,$$

then

$$d(\mathfrak{g}_{Z_2}(Z_1), \mathfrak{g}_{Z_2}(Z_3)) < C_2 \quad \text{and} \quad d(\mathfrak{g}_{Z_1}(Z_2), \mathfrak{g}_{Z_1}(Z_3)) < C_2.$$

**Proof** Fix  $\theta \geq \theta_0$ . Let  $Z_1, Z_2, Z_3 \in \mathcal{Z}$  and  $B = H_\theta(\mathfrak{g}_{Z_2}(Z_1) \cup \mathfrak{g}_{Z_1}(Z_2))$ . We remind the reader that they should view  $B$  as a bridge between  $Z_1$  and  $Z_2$ . Our goal is to show that there exists  $b \in B$  such that  $d(b, \mathfrak{g}_{Z_3}(b))$  is uniformly bounded. From this our conclusion will follow from the coarse Lipschitzness of the gate map.

By assertion (1),  $\mathfrak{g}_{Z_3}(Z_1)$  and  $\mathfrak{g}_{Z_3}(Z_2)$  are uniformly coarsely contained in  $\mathfrak{g}_{Z_3}(B)$ . Since the gate map is coarsely Lipschitz,

$$\text{diam}(\mathfrak{g}_{Z_3}(B)) \geq d(\mathfrak{g}_{Z_3}(Z_1), \mathfrak{g}_{Z_3}(Z_2))$$

with constants depending only on  $Q$ . Let  $r$  be the constant from Lemma 8.4 with  $A = B$  and  $Y = Z_3$  and suppose  $d(\mathfrak{g}_{Z_3}(Z_1), \mathfrak{g}_{Z_3}(Z_2))$  is large enough that  $\text{diam}(\mathfrak{g}_{Z_3}(B)) > r$ . By Lemma 8.4, there exists  $b \in B$  such that  $d(b, Z_3) < r$ .

By Lemma 4.19, we have that  $\mathfrak{g}_{Z_2}(Z_1)$  is uniformly coarsely equal to  $\mathfrak{g}_{Z_2}(B)$  in particular  $\mathfrak{g}_{Z_2}(b)$  is uniformly coarsely contained in  $\mathfrak{g}_{Z_2}(Z_1)$ . Since the gate maps are uniformly coarsely Lipschitz and  $d(b, Z_3) < r$ , we have that  $d(\mathfrak{g}_{Z_2}(Z_3), \mathfrak{g}_{Z_2}(Z_1)) < C_2$ . By switching the roles of  $Z_1$  and  $Z_2$ , we get  $d(\mathfrak{g}_{Z_1}(Z_3), \mathfrak{g}_{Z_1}(Z_2)) < C_2$ . □

**Assertion (3)** *There exists  $C_3 > 0$  such that for all  $Z_1, Z_2 \in \mathcal{Z}$ , the set*

$$\{Z \in \mathcal{Z} \mid d_X(\mathfrak{g}_Z(Z_1), \mathfrak{g}_Z(Z_2)) > C_3\}$$

*has only a finite number of elements.*

**Proof** Let  $Z_1, Z_2 \in \mathcal{Z}$ . Fix  $\theta \geq \theta_0$  and let  $B = H_\theta(\mathfrak{g}_{Z_2}(Z_1) \cup \mathfrak{g}_{Z_1}(Z_2))$ . By the bridge theorem, we have that  $B$  is coarsely equals to the product of  $\mathfrak{g}_{Z_1}(Z_2) \times H_\theta(a, b)$ , where  $a \in \mathfrak{g}_{Z_1}(Z_2)$  and  $b = \mathfrak{g}_{Z_2}(a)$ . By assertion (1), the gate  $\mathfrak{g}_{Z_1}(Z_2)$  has uniformly bounded diameter. By Proposition 5.6, there exists  $\lambda \geq \lambda_0$  such that  $H_\theta(a, b)$  is contained in  $\mathcal{P}_\lambda^1(a, b)$ , the set of  $\lambda$ -hierarchy paths between  $a$  and  $b$ . Since the distance between  $a$  and  $b$  is finite, so is the diameter of  $\mathcal{P}_\lambda^1(a, b)$ . Therefore  $H_\theta(a, b)$  has bounded diameter and so does the set  $B = H_\theta(\mathfrak{g}_{Z_2}(Z_1) \cup \mathfrak{g}_{Z_1}(Z_2))$ . Since  $G$  is locally finite,  $B$  can contain only a finite number of elements of  $G$ .

Let  $r$  be as in Lemma 8.4. Since  $\mathfrak{g}_{Z_2}(Z_1), \mathfrak{g}_{Z_1}(Z_2) \subseteq B$ , for any  $Z \in \mathcal{Z}$  with  $d(\mathfrak{g}_Z(Z_1), \mathfrak{g}_Z(Z_2))$  larger than  $r$  we have  $\text{diam}(\mathfrak{g}_Z(B)) > r$ . Thus every such  $Z$  intersects the  $r$ -neighborhood of  $B$ . By locally finiteness of  $G$ , we obtain that  $N_r(B)$  contains a finite number of element of  $G$ . Since the elements of  $\mathcal{Z}$  are cosets of finitely many subgroups, every point of  $N_r(B)$  can belong to uniformly finitely many elements of  $\mathcal{Z}$ , which concludes the proof of assertion (3). □

Proposition 8.6 now holds by taking  $C = \max\{C_1, C_2, C_3\}$ . □

We now have all the ingredients needed to give the proof of Theorem 8.1.

**Proof of Theorem 8.1** Recall, we need to show that if  $G$  is a hierarchically hyperbolic group and  $\{H_i\}$  a finite almost malnormal collection of strongly quasiconvex subgroups, then  $\{H_i\}$  is hyperbolically embedded in  $G$ . In particular, we shall show that the left cosets of the  $H_i$ 's satisfy the requirements of Theorem 8.2. Since each  $H_i$  is a strongly quasiconvex subgroup of  $G$ , by [54, Theorem 1.2] we have that they are all finitely generated. Let  $S$  be a finite generating set for  $G$  such that for each  $i$ ,  $H_i \cap S$  generates  $H_i$ . As before, let  $\mathcal{Z}$  be the set of all left cosets of  $\{H_i\}$ . For every pair of distinct  $Z_1, Z_2 \in \mathcal{Z}$  we want to define a set  $\tau_{Z_1}(Z_2)$  that satisfies (P0)–(P3) of Theorem 8.2. If we define  $\tau_{Z_1}(Z_2)$  as  $\mathfrak{g}_{Z_1}(Z_2)$ , Proposition 8.6 provides that (P0)–(P2) will be satisfied. However, since the gate maps are only coarsely equivariant, condition (P3) may not hold.

Thus, for  $Z_1 \neq Z_2$  define

$$\tau_{Z_1}(Z_2) = \bigcup_{g \in G} g^{-1} \mathfrak{g}_{gZ_1}(gZ_2).$$

By construction we have that  $\tau_{gZ_1}(gZ_2) = g(\tau_{Z_1}(Z_2))$  and thus (P3) holds. Since  $\tau_{Z_1}(Z_2)$  and  $\mathfrak{g}_{Z_1}(Z_2)$  uniformly coarsely coincide by the coarse equivariance of the gates maps (Lemma 4.16), (P0)–(P2) are satisfied as a corollary of Proposition 8.6. Hence, the collection  $\{H_i\}$  is hyperbolically embedded in  $G$  by Theorem 8.2.  $\square$

Our method of proof for Theorem 8.1 relies in a fundamental way upon the coarse equivariance of the gate map. If the group  $G$  has an HHS structure, but not an HHG structure, then the gate map need not be coarsely equivariant. In particular, Theorem 8.1 does not (currently) apply to the fundamental groups of nonflip graph manifolds and thus we have the following interesting case of Question 4.

**Question 6** *If  $M$  is a nonflip graph manifold and  $\{H_i\}$  is a finite, almost malnormal collection of strongly quasiconvex subgroups of  $\pi_1(M)$ , is  $\{H_i\}$  hyperbolically embedded in  $\pi_1(M)$ ?*

## Appendix Subsets with arbitrary reasonable lower relative divergence

The proposition in this appendix utilizes the notion of asymptotic equivalence between families of functions. We will present the definition in the specific case we need and direct the reader to [53, Section 2] for the more general case.

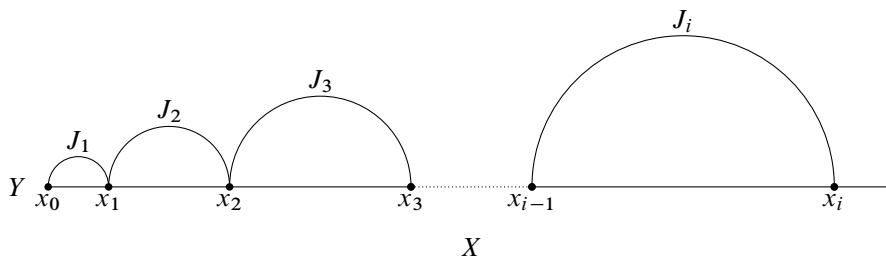


Figure 9: By controlling the length of each arc  $J_i$  we can get the desired lower relative divergence of the geodesic space  $X$  with respect to the subspace  $Y$ .

**Definition A.1** Let  $f$  and  $g$  be two functions from  $[0, \infty)$  to  $[0, \infty)$ . The function  $f$  is *dominated by the function*  $g$  if there are positive constants  $A, B, C$  and  $D$  such that  $f(r) \leq Ag(Br) + Cr$  for all  $r > D$ . Two functions  $f$  and  $g$  are *equivalent* if  $f$  is dominated by  $g$  and vice versa.

Let  $X$  be a geodesic metric space and  $\{\sigma_\rho^n\} = \text{div}(X, Y)$  be the lower relative divergence of  $X$  with respect to some subset  $Y \subseteq X$ . We say  $\text{div}(X, Y)$  is *equivalent* to a function  $f : [0, \infty) \rightarrow [0, \infty)$  if there exists  $L \in (0, 1]$  and a positive integer  $M$  such that  $\sigma_{L\rho}^{Mn}$  is equivalent to  $f$  for all  $\rho \in (0, 1]$  and  $n \geq 2$ .

**Proposition A.2** Let  $f : [0, \infty) \rightarrow [0, \infty)$  be a nondecreasing function, and assume that there is a positive integer  $r_0$  such that  $f(r) \geq r$  for each  $r > r_0$ . There is a geodesic space  $X$  with a subspace  $Y$  such that the lower relative divergence  $\text{div}(X, Y)$  is equivalent to  $f$ .

**Proof** Let  $Y$  be a ray with initial point  $x_0$ . Let  $(x_i)$  be the sequence of points along  $Y$  such that for each  $i \geq 1$  the distance  $d_Y(x_{i-1}, x_i) = i$  and we connect each pair  $(x_{i-1}, x_i)$  by a segment  $J_i$  of length  $f(i)$ ; see Figure 9. Let  $X$  be the resulting geodesic space and  $\text{div}(X, Y) = \{\sigma_\rho^n\}$ . We shall show that  $\text{div}(X, Y)$  is equivalent to  $f$ .

We first prove that for all  $n \geq 3$  and  $\rho \in (0, 1]$ ,  $f$  dominates  $\sigma_\rho^n$  by showing that  $\sigma_\rho^n(r) \leq f((n + 3)r)$  for each  $r > r_0$ . Let  $i_0$  be a smallest integer that is greater or equal to  $(n + 2)r$ . Let  $x$  and  $y$  be two points in the segment  $J_{i_0}$  such that

$$d(x_{i_0-1}, x) = d(x_{i_0}, y) = r.$$

Both  $x$  and  $y$  belong to  $\partial N_r(Y)$ . Moreover, the subpath  $\alpha$  of  $J_{i_0}$  connecting  $x$  and  $y$  lies outside the  $r$ -neighborhood of  $Y$ , and the length of  $\alpha$  is exactly is  $f(i) - 2r$ .

Therefore,  $d(x, y) = \min\{i_0 + 2r, f(i_0) - 2r\}$ . Hence  $d(x, y) \geq nr$  as

$$f(i_0) - 2r \geq f((n+2)r) - 2r \geq (n+2)r - 2r = nr$$

and

$$i_0 + 2r \geq (n+4)r \geq nr.$$

Since  $\alpha$  is the unique path outside the  $\rho r$ -neighborhood of  $Y$  connecting  $x$  and  $y$ ,

$$\sigma_\rho^n(r) \leq d_{\rho r}(x, y) = f(i_0) - 2r \leq f(i_0).$$

Since  $i_0 \leq (n+2)r + 1 \leq (n+3)r$  and  $f$  is nondecreasing,  $f(i_0) \leq f((n+3)r)$ . Thus,  $\sigma_\rho^n(r) \leq f((n+3)r)$ , which implies that  $\sigma_\rho^n$  is dominated by  $f$ .

Now we prove that for all  $n \geq 3$  and  $\rho \in (0, 1]$ ,  $\sigma_\rho^n$  dominates  $f$  by showing that  $\sigma_\rho^n(r) \geq f(r) - 2r$  for each  $r > r_0$ . Let  $u$  and  $v$  be an arbitrary points in  $\partial N_r(Y)$  such that  $d(u, v) \geq nr$  and there is a path outside the  $r$ -neighborhood of  $Y$  connecting  $u$  and  $v$ . Therefore,  $u$  and  $v$  must lie in some segment  $J_{i_1}$ . We can assume that  $d(u, x_{i_1-1}) = d(v, x_{i_1}) = r$ . Therefore,

$$i_1 \geq d(x_{i_1-1}, x_{i_1}) \geq d(u, v) - 2r \geq nr - 2r \geq r.$$

This implies that  $f(i_1) \geq f(r)$  since  $f$  is nondecreasing. Since the subpath  $\beta$  of  $J_{i_1}$  connecting  $u$  and  $v$  is the unique path outside the  $\rho r$ -neighborhood of  $Y$  connecting these points,

$$d_{\rho r}(u, v) = f(i_1) - 2r \geq f(r) - 2r.$$

Therefore,  $\sigma_\rho^n(r) \geq f(r) - 2r$  which implies that  $\sigma_\rho^n$  dominates  $f$ . Thus, the lower relative divergence  $\text{div}(X, Y)$  is equivalent to  $f$ .  $\square$

## References

- [1] **C Abbott, J Behrstock, M G Durham**, *Largest acylindrical actions and stability in hierarchically hyperbolic groups*, Trans. Amer. Math. Soc. Ser. B 8 (2021) 66–104 MR Zbl
- [2] **I Agol**, *The virtual Haken conjecture*, Doc. Math. 18 (2013) 1045–1087 MR Zbl
- [3] **Y Antolín, M Mj, A Sisto, S J Taylor**, *Intersection properties of stable subgroups and bounded cohomology*, Indiana Univ. Math. J. 68 (2019) 179–199 MR Zbl
- [4] **T Aougab, M G Durham, S J Taylor**, *Pulling back stability with applications to  $\text{Out}(F_n)$  and relatively hyperbolic groups*, J. Lond. Math. Soc. 96 (2017) 565–583 MR Zbl
- [5] **G N Arzhantseva**, *On quasiconvex subgroups of word hyperbolic groups*, Geom. Dedicata 87 (2001) 191–208 MR Zbl

- [6] **G N Arzhantseva, C H Cashen, D Gruber, D Hume**, *Characterizations of Morse quasi-geodesics via superlinear divergence and sublinear contraction*, Doc. Math. 22 (2017) 1193–1224 MR Zbl
- [7] **J Behrstock**, *A counterexample to questions about boundaries, stability, and commensurability*, from “Beyond hyperbolicity” (M Hagen, R Webb, H Wilton, editors), London Math. Soc. Lecture Note Ser. 454, Cambridge Univ. Press (2019) 151–159 MR Zbl
- [8] **J Behrstock, M F Hagen, A Sisto**, *Asymptotic dimension and small-cancellation for hierarchically hyperbolic spaces and groups*, Proc. Lond. Math. Soc. 114 (2017) 890–926 MR Zbl
- [9] **J Behrstock, M F Hagen, A Sisto**, *Hierarchically hyperbolic spaces, I: Curve complexes for cubical groups*, Geom. Topol. 21 (2017) 1731–1804 MR Zbl
- [10] **J Behrstock, M Hagen, A Sisto**, *Hierarchically hyperbolic spaces II: Combination theorems and the distance formula*, Pacific J. Math. 299 (2019) 257–338 MR Zbl
- [11] **J Behrstock, M F Hagen, A Sisto**, *Quasiflats in hierarchically hyperbolic spaces*, Duke Math. J. 170 (2021) 909–996 MR Zbl
- [12] **M Bestvina, K Bromberg, K Fujiwara**, *Constructing group actions on quasi-trees and applications to mapping class groups*, Publ. Math. Inst. Hautes Études Sci. 122 (2015) 1–64 MR Zbl
- [13] **B H Bowditch**, *Relatively hyperbolic groups*, Internat. J. Algebra Comput. 22 (2012) art. id. 1250016 MR Zbl
- [14] **B H Bowditch**, *Coarse median spaces and groups*, Pacific J. Math. 261 (2013) 53–93 MR Zbl
- [15] **B H Bowditch**, *Uniform hyperbolicity of the curve graphs*, Pacific J. Math. 269 (2014) 269–280 MR Zbl
- [16] **B H Bowditch**, *Convex hulls in coarse median spaces*, preprint (2018) Available at <http://homepages.warwick.ac.uk/~masgak/papers/hulls-cms.pdf>
- [17] **J F Brock**, *The Weil–Petersson metric and volumes of 3–dimensional hyperbolic convex cores*, J. Amer. Math. Soc. 16 (2003) 495–535 MR Zbl
- [18] **C H Cashen**, *Morse subsets of CAT(0) spaces are strongly contracting*, Geom. Dedicata 204 (2020) 311–314 MR Zbl
- [19] **R Charney, H Sultan**, *Contracting boundaries of CAT(0) spaces*, J. Topol. 8 (2015) 93–117 MR Zbl
- [20] **F Dahmani, V Guirardel, D Osin**, *Hyperbolically embedded subgroups and rotating families in groups acting on hyperbolic spaces*, Mem. Amer. Math. Soc. 1156, Amer. Math. Soc., Providence, RI (2017) MR Zbl
- [21] **C Druţu, M Kapovich**, *Geometric group theory*, American Mathematical Society Colloquium Publications 63, Amer. Math. Soc., Providence, RI (2018) MR Zbl
- [22] **C Druţu, S Mozes, M Sapir**, *Divergence in lattices in semisimple Lie groups and graphs of groups*, Trans. Amer. Math. Soc. 362 (2010) 2451–2505 MR Zbl

- [23] **M G Durham**, *The augmented marking complex of a surface*, J. Lond. Math. Soc. 94 (2016) 933–969 MR Zbl
- [24] **M G Durham, M F Hagen, A Sisto**, *Boundaries and automorphisms of hierarchically hyperbolic spaces*, Geom. Topol. 21 (2017) 3659–3758 MR Zbl
- [25] **M G Durham, S J Taylor**, *Convex cocompactness and stability in mapping class groups*, Algebr. Geom. Topol. 15 (2015) 2839–2859 MR Zbl
- [26] **A Eskin, H Masur, K Rafi**, *Large-scale rank of Teichmüller space*, Duke Math. J. 166 (2017) 1517–1572 MR Zbl
- [27] **B Farb**, *Some problems on mapping class groups and moduli space*, from “Problems on mapping class groups and related topics” (B Farb, editor), Proc. Sympos. Pure Math. 74, Amer. Math. Soc., Providence, RI (2006) 11–55 MR Zbl
- [28] **A Genevois**, *Hyperbolicities in CAT(0) cube complexes*, Enseign. Math. 65 (2019) 33–100 MR Zbl
- [29] **A Genevois**, *Quasi-isometrically rigid subgroups in right-angled Coxeter groups*, Algebr. Geom. Topol. 22 (2022) 657–708 MR Zbl
- [30] **R Gitik**, *Ping-pong on negatively curved groups*, J. Algebra 217 (1999) 65–72 MR Zbl
- [31] **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
- [32] **M Gromov**, *Asymptotic invariants of infinite groups*, from “Geometric group theory, II” (G A Niblo, M A Roller, editors), London Math. Soc. Lecture Note Ser. 182, Cambridge Univ. Press (1993) 1–295 MR Zbl
- [33] **D Groves, J F Manning**, *Hyperbolic groups acting improperly*, preprint (2018) arXiv 1808.02325
- [34] **T Haettel**, *Hyperbolic rigidity of higher rank lattices*, Ann. Sci. Éc. Norm. Supér. 53 (2020) 439–468 MR Zbl
- [35] **M F Hagen, H Petyt**, *Projection complexes and quasimedial maps*, Algebr. Geom. Topol. 22 (2022) 3277–3304 MR Zbl
- [36] **G C Hruska**, *Relative hyperbolicity and relative quasiconvexity for countable groups*, Algebr. Geom. Topol. 10 (2010) 1807–1856 MR Zbl
- [37] **A Karrer**, *Right-angled Coxeter groups with totally disconnected Morse boundaries*, preprint (2021) arXiv 2105.04029
- [38] **H Kim**, *Stable subgroups and Morse subgroups in mapping class groups*, Internat. J. Algebra Comput. 29 (2019) 893–903 MR Zbl
- [39] **T Koberda, J Mangahas, S J Taylor**, *The geometry of purely loxodromic subgroups of right-angled Artin groups*, Trans. Amer. Math. Soc. 369 (2017) 8179–8208 MR Zbl
- [40] **H A Masur, Y N Minsky**, *Geometry of the complex of curves, I: Hyperbolicity*, Invent. Math. 138 (1999) 103–149 MR Zbl
- [41] **H A Masur, Y N Minsky**, *Geometry of the complex of curves, II: Hierarchical structure*, Geom. Funct. Anal. 10 (2000) 902–974 MR Zbl

- [42] **H Masur, S Schleimer**, *The geometry of the disk complex*, J. Amer. Math. Soc. 26 (2013) 1–62 MR Zbl
- [43] **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
- [44] **A Y Ol’shanskii, D V Osin, M V Sapir**, *Lacunary hyperbolic groups*, Geom. Topol. 13 (2009) 2051–2140 MR Zbl
- [45] **D Osin**, *Acylindrically hyperbolic groups*, Trans. Amer. Math. Soc. 368 (2016) 851–888 MR Zbl
- [46] **N Petrosyan**, *Decomposing groups by codimension-1 subgroups*, Proc. Amer. Math. Soc. 150 (2022) 4587–4601 MR Zbl
- [47] **K Rafi**, *A combinatorial model for the Teichmüller metric*, Geom. Funct. Anal. 17 (2007) 936–959 MR Zbl
- [48] **J Russell**, *From hierarchical to relative hyperbolicity*, Int. Math. Res. Not. 2022 (2022) 575–624 MR Zbl
- [49] **A Sisto**, *On metric relative hyperbolicity*, preprint (2012) arXiv 1210.8081
- [50] **A Sisto**, *Quasi-convexity of hyperbolically embedded subgroups*, Math. Z. 283 (2016) 649–658 MR Zbl
- [51] **D Spriano**, *Hyperbolic HHS, II: Graphs of hierarchically hyperbolic groups*, preprint (2018) arXiv 1801.01850
- [52] **H Sultan**, *Hyperbolic quasi-geodesics in CAT(0) spaces*, Geom. Dedicata 169 (2014) 209–224 MR Zbl
- [53] **H C Tran**, *Relative divergence of finitely generated groups*, Algebr. Geom. Topol. 15 (2015) 1717–1769 MR Zbl
- [54] **H C Tran**, *On strongly quasiconvex subgroups*, Geom. Topol. 23 (2019) 1173–1235 MR Zbl
- [55] **D T Wise**, *From riches to raags: 3-manifolds, right-angled Artin groups, and cubical geometry*, CBMS Regional Conference Series in Mathematics 117, Amer. Math. Soc., Providence, RI (2012) MR Zbl

*Math Department, Rice University*

*Houston, TX, United States*

*Mathematical Institute, University of Oxford*

*Oxford, United Kingdom*

*Department of Mathematics, The University of Oklahoma*

*Norman, OK, United States*

`jacob.russell@rice.edu`, `davide.spriano@chch.ox.ac.uk`,  
`hungtran280687@gmail.com`

Received: 16 September 2020

Revised: 17 June 2021



# Finite presentations for stated skein algebras and lattice gauge field theory

JULIEN KORINMAN

We provide finite presentations for stated skein algebras and deduce that those algebras are Koszul and that they are isomorphic to the quantum moduli algebras appearing in lattice gauge field theory, generalizing previous results of Bullock, Frohman, Kania-Bartoszyńska and Faitg.

57R56; 57K31

1. Introduction	1249
2. Finite presentations for stated skein algebras	1253
3. Proof of Theorems 1.1 and 1.2	1273
4. Lattice gauge field theory	1284
5. Concluding remarks	1296
References	1299

## 1 Introduction

**Stated skein algebras and lattice gauge field theory** A *punctured surface* is a pair  $\Sigma = (\Sigma, \mathcal{P})$ , where  $\Sigma$  is a compact oriented surface and  $\mathcal{P}$  is a (possibly empty) finite subset of  $\Sigma$  which nontrivially intersects each boundary component. We write  $\Sigma_{\mathcal{P}} := \Sigma \setminus \mathcal{P}$ . The set  $\partial\Sigma \setminus \mathcal{P}$  consists of a disjoint union of open arcs, which we call *boundary arcs*.

**Warning** In this paper, the punctured surface  $\Sigma$  will be called open if the surface  $\Sigma$  has nonempty boundary and closed otherwise. This convention differs from the traditional one, where some authors refer to an open surface as a punctured surface  $\Sigma = (\Sigma, \mathcal{P})$  with  $\Sigma$  closed and  $\mathcal{P} \neq \emptyset$  (in which case  $\Sigma_{\mathcal{P}}$  is not closed).

The *Kauffman-bracket skein algebras* were introduced by Bullock and Turaev as a tool to study the  $SU(2)$  Witten–Reshetikhin–Turaev topological quantum field theories [45; 51]. They are associative unitary algebras  $\mathcal{S}_\omega(\Sigma)$  indexed by a closed punctured surface  $\Sigma$  and an invertible element  $\omega \in \mathbb{k}^\times$  in some commutative unital ring  $\mathbb{k}$ . Bonahon and Wong [12] and Lê [40] generalized the notion of Kauffman-bracket skein algebras to open punctured surfaces, where in addition to closed curves the algebras are generated by arcs whose endpoints are endowed with a sign,  $\pm$  (a state). The motivation for the introduction of these so-called *stated skein algebras* is their good behavior for the operation of gluing two boundary arcs together. This property permitted the authors of [12] to define an embedding of the skein algebra into a quantum torus, named the quantum trace, and offers new tools to study the representation theory of skein algebras.

Except for genus 0 and 1 surfaces (see Bullock and Przytycki [21]), no finite presentation for the Kauffman-bracket skein algebras is known, though a conjecture in that direction was formulated in Santharoubane [46, Conjecture 1.2]. However, it is well known that they are finitely generated; see Abdiel and Frohman [1], Bullock [18], Frohman and Kania-Bartoszyńska [30] and Santharoubane [46]. The corresponding problem for stated skein algebras of open punctured surfaces is easier. Finite presentations of stated skein algebras were given for a disc with two punctures on its boundary (for the bigon) and for the disc with three punctures on its boundary (for the triangle) in [40], for the disc with two punctures on its boundary and one inner puncture in Korinman [35] and for any connected punctured surface having exactly one boundary component, one puncture on the boundary and possibly inner punctures in Faitg [27].

Our first purpose is to provide explicit finite presentations for stated skein algebras of an arbitrary connected open punctured surface  $\Sigma$ . Let us briefly sketch their construction; we refer to Section 2.2 for details.

The finite presentations we will define depend on the choice of a finite presentation  $\mathbb{P}$  of some groupoid  $\Pi_1(\Sigma_{\mathcal{P}}, \mathbb{V})$ . In brief, for each boundary arc  $a$  of  $\Sigma$ , choose a point  $v_a \in a$  and let  $\mathbb{V}$  be the set of such points. The groupoid  $\Pi_1(\Sigma_{\mathcal{P}}, \mathbb{V})$  is the full subcategory of the fundamental groupoid of  $\Sigma_{\mathcal{P}}$  whose set of objects is  $\mathbb{V}$ . A finite presentation  $\mathbb{P} = (\mathbb{G}, \mathbb{RL})$  for  $\Pi_1(\Sigma_{\mathcal{P}}, \mathbb{V})$  will consist in a finite set  $\mathbb{G}$  of generating paths relating points of  $\mathbb{V}$  and a finite set  $\mathbb{RL}$  of relations among those paths which satisfy some axioms (see Section 2.2 for details). For instance for the triangle  $\mathbb{T}$  (the disc with three punctures on its boundary), the groupoid  $\Pi_1(\mathbb{T}, \mathbb{V})$  admits the presentation with generators  $\mathbb{G} = \{\alpha, \beta, \gamma\}$ , drawn in Figure 1, and the unique relation  $\alpha\beta\gamma = 1$ .

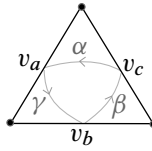


Figure 1: The triangle and some paths.

A path  $\alpha \in \mathbb{G}$  can be seen as an arc in  $\Sigma_{\mathcal{P}}$  and, after choosing some states  $\varepsilon, \varepsilon' \in \{-, +\}$  for its endpoints, we get an element  $\alpha_{\varepsilon\varepsilon'} \in \mathcal{S}_{\omega}(\Sigma)$  in the stated skein algebra. We denote by  $\mathcal{A}^{\mathbb{G}} \subset \mathcal{S}_{\omega}(\Sigma)$  the (finite) set of such elements. It was proved in Korinman [38] that  $\mathcal{A}^{\mathbb{G}}$  generates  $\mathcal{S}_{\omega}(\Sigma)$  and its elements will be the generators of our presentations.

Concerning the relations, first for each  $\alpha \in \mathbb{G}$ , one has a  $q$ -determinant relation between the elements  $\alpha_{\varepsilon\varepsilon'}$ . For each pair  $(\alpha, \beta) \in \mathbb{G}^2$  we will associate a finite set of arc exchange relations permitting us to express an element of the form  $\alpha_{\varepsilon\varepsilon'}\beta_{\mu\mu'} \in \mathcal{S}_{\omega}(\Sigma)$  as a linear combination of elements of the form  $\beta_{ab}\alpha_{cd}$ . Finally, to each relation  $R \in \mathbb{RL}$  in the finite presentation  $\mathbb{P}$ , we will associate a finite set of so-called trivial loop relations.

**Theorem 1.1** *Let  $\Sigma$  be a connected open punctured surface and  $\mathbb{P}$  a finite presentation of  $\Pi_1(\Sigma_{\mathcal{P}}, \mathbb{V})$ . Then the stated skein algebra  $\mathcal{S}_{\omega}(\Sigma)$  is presented by the set of generators  $\mathcal{A}^{\mathbb{G}}$  and by the  $q$ -determinant, arc exchange and trivial loop relations.*

For every open punctured surface, we can choose a finite presentation  $\mathbb{P}$  of  $\Pi_1(\Sigma_{\mathcal{P}}, \mathbb{V})$  such that the set of relations is empty (for instance for the triangle of Figure 1, one might choose the presentation with generators  $\mathbb{G} = \{\alpha, \beta\}$  and no relations). In this case, the presentation of  $\mathcal{S}_{\omega}(\Sigma)$  is quadratic inhomogeneous and, by using the diamond lemma, we prove:

**Theorem 1.2** *For  $\Sigma$  a connected open punctured surface, the quadratic inhomogeneous algebra  $\mathcal{S}_{\omega}(\Sigma)$  is Koszul and admits a Poincaré–Birkhoff–Witt (PBW) basis.*

Theorem 1.2 implies that  $\mathcal{S}_{\omega}(\Sigma)$  has an explicit minimal projective resolution (the so-called Koszul resolution), which permits us to effectively compute its cohomology (see Loday and Vallette [42] for details).

Let  $(\Gamma, c)$  be a ciliated graph, that is a finite graph with the data for each vertex of a linear ordering of its adjacent half-edges. Inspired by Fock and Rosly’s original work in [29] on the Poisson structure of character varieties, Alekseev, Grosse and

Schomerus [2; 3; 5] and Buffenoir and Roche [15; 16] independently defined the so-called *quantum moduli algebras*  $\mathcal{L}_\omega(\Gamma, c)$ , which are combinatorial quantizations of relative character varieties (see Section 4.2 for details). Those algebras arise with some right comodule map  $\Delta^{\mathcal{G}}: \mathcal{L}_\omega(\Gamma, c) \rightarrow \mathcal{L}_\omega(\Gamma, c) \otimes \mathcal{O}_q[\mathcal{G}]$ , where  $\mathcal{O}_q[\mathcal{G}] = \mathcal{O}_q[\mathrm{SL}_2] \otimes^{\mathring{V}(\Gamma)}$  is the so-called quantum gauge group Hopf algebra and  $q := \omega^{-4}$ . The subalgebra  $\mathcal{L}_\omega^{\mathrm{inv}}(\Gamma) \subset \mathcal{L}_\omega(\Gamma, c)$  of coinvariant vectors plays an important role in combinatorial quantization. More precisely, as reviewed in Section 4.1, we associate to each ciliated graph  $(\Gamma, c)$  two punctured surfaces: an open one  $\Sigma^0(\Gamma, c)$  and a closed one  $\Sigma(\Gamma)$ , such that the algebras  $\mathcal{L}_\omega(\Gamma, c)$  and  $\mathcal{L}_\omega^{\mathrm{inv}}(\Gamma)$  are quantizations of the  $\mathrm{SL}_2(\mathbb{C})$  (relative) character varieties of  $\Sigma^0(\Gamma, c)$  and  $\Sigma(\Gamma)$ , respectively, with their Fock–Rosly Poisson structures. We deduce from Theorem 1.1:

**Theorem 1.3** *There exist isomorphisms of algebras  $\mathcal{S}_\omega(\Sigma^0(\Gamma, c)) \cong \mathcal{L}_\omega(\Gamma, c)$  and  $\mathcal{S}_\omega(\Sigma(\Gamma)) \cong \mathcal{L}_\omega^{\mathrm{inv}}(\Gamma)$ .*

Theorem 1.3 is not surprising and was already proved in some cases. First it is well known that (stated) skein algebras also induce deformation quantizations of (relative) character varieties: it follows from the work in Bullock [17], Przytycki and Sikora [44] and Turaev [50] for closed punctured surfaces and is proved in Korinman and Quesney [39, Theorem 1.3] and Costantino and Lê [26, Theorem 8.12] for open punctured surfaces. So Theorem 1.3 was expected; for instance its statement was conjectured in [26]. Next the skein origin of the defining relations of quantum moduli algebra was discovered by Bullock, Frohman and Kania-Bartoszyńska in [19] where the authors already proved that  $\mathcal{S}_\omega(\Sigma(\Gamma))$  and  $\mathcal{L}_\omega^{\mathrm{inv}}(\Gamma)$  are isomorphic in the particular case where  $\mathbb{k} = \mathbb{C}[[\hbar]]$  and  $q := \omega^{-4} = \exp \hbar$ . However, their proof does not extend to arbitrary ring (see item (vi) of Section 5). Finally, in the special case where  $(\Gamma, c)$  is the so-called daisy graph (it has only one vertex, so  $\Sigma^0(\Gamma, c)$  has exactly one boundary component with one puncture on it), Theorem 1.3 was proved by Faitg in [27] in the case where  $\omega$  is not a root of unity. A detailed comparison between Faitg’s isomorphism and ours is made in Section 4.4. Faitg’s result can also be derived indirectly from the works in Ben-Zvi, Brochier and Jordan [9] and Gunningham, Jordan and Safronov [31], as detailed in Section 4.4. As pointed out to us by the anonymous referee, there is an important difference between our definition of quantum moduli algebras and the original one. In the original approaches, the algebra  $\mathcal{L}_\omega(\Gamma, c)$  is seen as a  $U_q \mathfrak{sl}_2^{\otimes n}$ -module, where  $n$  is the number of external vertices of  $\Gamma$ , and  $\mathcal{L}_\omega^{\mathrm{inv}}(\Gamma)$  is then defined as the subalgebra of invariant vectors for this action. Here,  $\mathcal{L}_\omega(\Gamma, c)$  is rather seen as an  $\mathcal{O}_q[\mathrm{SL}_2]^{\otimes n}$ -comodule and  $\mathcal{L}_\omega^{\mathrm{inv}}(\Gamma)$  is defined as the subalgebra of coinvariant vectors

instead. When  $q$  is generic both definitions coincide, however when  $q$  is a root of unity they differ in general (see Section 5 for details). In particular, the isomorphism in Theorem 1.3 holds with our definition of quantum moduli algebra and might fail with the original one, at roots of unity.

**Acknowledgments** The author thanks S Baseilhac, F Costantino, M Faitg, L Funar, D Jordan, A Quesney, P Roche and P Safronov for useful discussions and the anonymous referees for interesting suggestions and corrections and for pointing out to us the difference between  $U_q\mathfrak{sl}_2$ -invariant and  $\mathcal{O}_q[\mathrm{SL}_2]$ -coinvariants at roots of unity. He acknowledges support from the Japanese Society for Promotion of Science (JSPS) and the Centre National de la Recherche Scientifique (CNRS).

## 2 Finite presentations for stated skein algebras

### 2.1 Definitions and first properties of stated skein algebras

**Definition 2.1** A *punctured surface* is a pair  $\Sigma = (\Sigma, \mathcal{P})$  where  $\Sigma$  is a compact oriented surface and  $\mathcal{P}$  is a finite subset of  $\Sigma$  which nontrivially intersects each boundary component. A *boundary arc* is a connected component of  $\partial\Sigma \setminus \mathcal{P}$ . We write  $\Sigma_{\mathcal{P}} := \Sigma \setminus \mathcal{P}$ .

**Definition of stated skein algebras** Before precisely stating the definition of stated skein algebras, let us sketch it informally. Given a punctured surface  $\Sigma$  and an invertible element  $\omega \in \mathbb{k}^\times$  in some commutative unital ring  $\mathbb{k}$ , the stated skein algebra  $\mathcal{S}_\omega(\Sigma)$  is the quotient of the  $\mathbb{k}$ -module freely spanned by isotopy classes of stated tangles in  $\Sigma_{\mathcal{P}} \times (0, 1)$  by some local skein relations. Figure 2, left, illustrates such a stated tangle: each point of  $\partial T \subset \partial\Sigma_{\mathcal{P}}$  is equipped with a sign  $+$  or  $-$  (the state). Here the stated tangle is the union of three stated arcs and one closed curve. In order to work with two-dimensional pictures, we will consider the projection of tangles in  $\Sigma_{\mathcal{P}}$  as in Figure 2, right; such a projection will be referred to as a diagram.

A *tangle* in  $\Sigma_{\mathcal{P}} \times (0, 1)$  is a compact framed, properly embedded one-dimensional manifold  $T \subset \Sigma_{\mathcal{P}} \times (0, 1)$  such that for every point of  $\partial T \subset \partial\Sigma_{\mathcal{P}} \times (0, 1)$  the framing is parallel to the  $(0, 1)$  factor and points in the direction of 1. Here, by framing, we refer to a thickening of  $T$  to an oriented surface. The *height* of  $(v, h) \in \Sigma_{\mathcal{P}} \times (0, 1)$  is  $h$ . If  $b$  is a boundary arc and  $T$  a tangle, we impose that no two points in  $\partial_b T := \partial T \cap b \times (0, 1)$  have the same heights, hence the set  $\partial_b T$  is totally ordered by the heights. Two tangles are isotopic if they are isotopic through the class of tangles that preserve the boundary height orders. By convention, the empty set is a tangle only isotopic to itself.

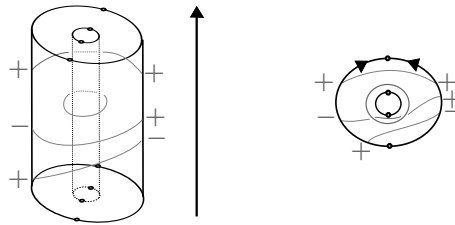


Figure 2: A stated tangle (left) and its associated diagram (right). The arrows represent the height orders.

Let  $\pi: \Sigma_{\mathcal{P}} \times (0, 1) \rightarrow \Sigma_{\mathcal{P}}$  be the projection with  $\pi(v, h) = v$ . A tangle  $T$  is in *generic position* if, for each of its points, the framing is parallel to the  $(0, 1)$  factor, points in the direction of 1 and is such that  $\pi|_T: T \rightarrow \Sigma_{\mathcal{P}}$  is an immersion with at most transversal double points in the interior of  $\Sigma_{\mathcal{P}}$ . Every tangle is isotopic to a tangle in generic position. A *diagram* is the image  $D = \pi(T)$  of a tangle in generic position, together with the over/undercrossing information at each double point. An isotopy class of diagram  $D$  together with a total order of  $\partial_b D := \partial D \cap b$  for each boundary arc  $b$  uniquely define an isotopy class of a tangle. When choosing an orientation  $\sigma(b)$  of a boundary arc  $b$  and a diagram  $D$ , the set  $\partial_b D$  receives a natural order by setting that the points are increasing when going in the direction of  $\sigma(b)$ . We will represent tangles by drawing a diagram and an orientation (an arrow) for each boundary arc, as in Figure 2. When a boundary arc  $b$  is oriented we assume that the order of the heights of the points of  $\partial_b D$  coincides with the order induced by the orientation of the boundary arc. A *state* of a tangle is a map  $s: \partial T \rightarrow \{-, +\}$ . A pair  $(T, s)$  is called a *stated tangle*. We define a *stated diagram*  $(D, s)$  in a similar manner.

Let  $\omega \in \mathbb{k}^\times$  be an invertible element and write  $A := \omega^{-2}$ .

**Definition 2.2** [40] The *stated skein algebra*  $\mathcal{S}_\omega(\Sigma)$  is the free  $\mathbb{k}$ -module generated by isotopy classes of stated tangles in  $\Sigma_{\mathcal{P}} \times (0, 1)$  modulo the relations (1) and (2):

$$\begin{aligned}
 (1) \quad & \begin{array}{c} \diagup \diagdown \\ \diagdown \diagup \end{array} = A \begin{array}{c} \diagdown \diagup \\ \diagup \diagdown \end{array} + A^{-1} \begin{array}{c} \diagdown \diagdown \\ \diagup \diagup \end{array} \quad \text{and} \quad \bigcirc = -(A^2 + A^{-2}) \begin{array}{c} \diagup \diagdown \\ \diagdown \diagup \end{array}, \\
 (2) \quad & \begin{array}{c} \uparrow \\ \square \\ \uparrow \end{array} = \begin{array}{c} \uparrow \\ \square \\ \downarrow \end{array} = 0, \quad \begin{array}{c} \uparrow \\ \square \\ \uparrow \end{array} = \omega \begin{array}{c} \square \\ \square \\ \square \end{array} \quad \text{and} \quad \omega^{-1} \begin{array}{c} \uparrow \\ \square \\ \uparrow \end{array} - \omega^{-5} \begin{array}{c} \uparrow \\ \square \\ \downarrow \end{array} = \begin{array}{c} \square \\ \square \\ \square \end{array}.
 \end{aligned}$$

The product of two classes of stated tangles  $[T_1, s_1]$  and  $[T_2, s_2]$  is defined by isotoping  $T_1$  and  $T_2$  in  $\Sigma_{\mathcal{P}} \times (\frac{1}{2}, 1)$  and  $\Sigma_{\mathcal{P}} \times (0, \frac{1}{2})$ , respectively, and then setting  $[T_1, s_1] \cdot [T_2, s_2]$  equal to  $[T_1 \cup T_2, s_1 \cup s_2]$ . Figure 3 illustrates this product.

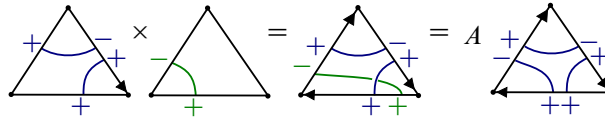


Figure 3: An illustration of the product in stated skein algebras.

For a closed punctured surface,  $\mathcal{S}_\omega(\Sigma)$  coincides with the classical (Turaev’s) Kauffman-bracket skein algebra.

**Reflexion anti-involution** Suppose  $\mathbb{k} = \mathbb{Z}[\omega^{\pm 1}]$  and consider the  $\mathbb{Z}$ -linear involution  $x \mapsto x^*$  on  $\mathbb{k}$  sending  $\omega$  to  $\omega^{-1}$ . Let  $r : \Sigma_{\mathcal{P}} \times (0, 1) \xrightarrow{\cong} \Sigma_{\mathcal{P}}$  be the homeomorphism defined by  $r(x, t) = (x, 1 - t)$ . Define an antilinear map  $\theta : \mathcal{S}_\omega(\Sigma) \xrightarrow{\cong} \mathcal{S}_\omega(\Sigma)$  by

$$\theta \left( \sum_i x_i [T_i, s_i] \right) := \sum_i x_i^* [r(T_i), s_i \circ r].$$

**Proposition 2.3** [40, Proposition 2.7] *The map  $\theta$  is an antimorphism of algebras, ie  $\theta(xy) = \theta(y)\theta(x)$ .*

**Bases for stated skein algebras** A closed component of a diagram  $D$  is trivial if it bounds an embedded disc in  $\Sigma_{\mathcal{P}}$ . An open component of  $D$  is trivial if it can be isotoped, relatively to its boundary, inside some boundary arc. A diagram is *simple* if it has neither double point nor trivial component. By convention, the empty set is a simple diagram. Let  $\sigma$  denote an arbitrary orientation of the boundary arcs of  $\Sigma$ . For each boundary arc  $b$  we denote by  $<_\sigma$  the induced total order on  $\partial_b D$ . A state  $s : \partial D \rightarrow \{-, +\}$  is  $\sigma$ -increasing if, for any boundary arc  $b$  and any two points  $x, y \in \partial_b D$ , then  $x <_\sigma y$  implies  $s(x) < s(y)$ , with the convention  $- < +$ .

**Definition 2.4** We denote by  $\mathcal{B}^\sigma \subset \mathcal{S}_\omega(\Sigma)$  the set of classes of stated diagrams  $(D, s)$  such that  $D$  is simple and  $s$  is  $\sigma$ -increasing.

**Theorem 2.5** [40, Theorem 2.11] *The set  $\mathcal{B}^\sigma$  is a basis of  $\mathcal{S}_\omega(\Sigma)$ .*

**Remark 2.6** The basis  $\mathcal{B}^\sigma$  is independent of the choice of the ground ring  $\mathbb{k}$  and of  $\omega \in \mathbb{k}^\times$ . This fact has the following useful consequence: Let  $\mathbb{k} := \mathbb{Z}[\omega^{\pm 1}]$  and  $\mathbb{k}'$  be any other commutative unital ring with an invertible element  $\omega' \in \mathbb{k}'^\times$ . There is a unique morphism of rings  $\mu : \mathbb{k} \rightarrow \mathbb{k}'$  sending  $\omega$  to  $\omega'$  and the two  $\mathbb{k}'$  algebras  $\mathcal{S}_\omega(\Sigma) \otimes_{\mathbb{k}} \mathbb{k}'$  and  $\mathcal{S}_{\omega'}(\Sigma)$  are canonically isomorphic through the isomorphism preserving the basis  $\mathcal{B}^\sigma$ . This fact permits us to prove formulas in  $\mathbb{k}$  using the reflexion anti-involution  $\theta$  and then apply them to any ring  $\mathbb{k}'$  by changing the coefficients.

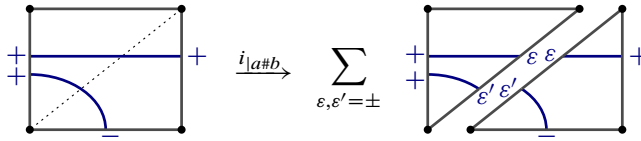


Figure 4: An illustration of the gluing map  $i_{|a\#b}$ .

**Gluing maps** Let  $a$  and  $b$  be two distinct boundary arcs of  $\Sigma$  and let  $\Sigma_{|a\#b}$  be the punctured surface obtained from  $\Sigma$  by gluing  $a$  and  $b$ . Denote by  $\pi : \Sigma_{\mathcal{P}} \rightarrow (\Sigma_{|a\#b})_{\mathcal{P}_{|a\#b}}$  the projection and  $c := \pi(a) = \pi(b)$ . Let  $(T_0, s_0)$  be a stated framed tangle of  $\Sigma_{|a\#b} \times (0, 1)$  transverse to  $c \times (0, 1)$  and such that the heights of the points of  $T_0 \cap c \times (0, 1)$  are pairwise distinct and the framing of the points of  $T_0 \cap c \times (0, 1)$  is vertical. Let  $T \subset \Sigma_{\mathcal{P}} \times (0, 1)$  be the framed tangle obtained by cutting  $T_0$  along  $c$ . Any two states  $s_a : \partial_a T \rightarrow \{-, +\}$  and  $s_b : \partial_b T \rightarrow \{-, +\}$  give rise to a state  $(s_a, s, s_b)$  on  $T$ . Both the sets  $\partial_a T$  and  $\partial_b T$  are in canonical bijection with the set  $T_0 \cap c$  by the map  $\pi$ . Hence the two sets of states  $s_a$  and  $s_b$  are both in canonical bijection with the set  $\text{St}(c) := \{s : c \cap T_0 \rightarrow \{-, +\}\}$ .

**Definition 2.7** Let  $i_{|a\#b} : \mathcal{S}_\omega(\Sigma_{|a\#b}) \rightarrow \mathcal{S}_\omega(\Sigma)$  be the linear map given, for any  $(T_0, s_0)$  as above, by

$$i_{|a\#b}([T_0, s_0]) := \sum_{s \in \text{St}(c)} [T, (s, s_0, s)].$$

**Theorem 2.8** [40, Theorem 3.1] *The linear map  $i_{|a\#b} : \mathcal{S}_\omega(\Sigma_{|a\#b}) \rightarrow \mathcal{S}_\omega(\Sigma)$  is an injective morphism of algebras. Moreover the gluing operation is coassociative in the sense that if  $a, b, c$  and  $d$  are four distinct boundary arcs, then  $i_{|a\#b} \circ i_{|c\#d} = i_{|c\#d} \circ i_{|a\#b}$ .*

**Relation with  $U_q \mathfrak{sl}_2$  and  $\mathcal{O}_q[\text{SL}_2]$**  Recall that  $A = \omega^{-2}$  and write  $q := A^2$ . The stated skein algebra has deep relations with the quantum enveloping algebra  $U_q \mathfrak{sl}_2$  and the quantum group  $\mathcal{O}_q(\text{SL}_2)$ , explored in [26; 27; 32; 39; 40], that we briefly reproduce here for later use by using the notation of [22; 33; 47]. Suppose that  $q$  is generic (not a root of unity) and let  $\rho : U_q \mathfrak{sl}_2 \rightarrow \text{End}(V)$  be the standard representation of  $U_q \mathfrak{sl}_2$ , where  $V$  is two-dimensional with basis  $(v_+, v_-)$  and

$$\rho(E) = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \quad \rho(F) = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} \quad \text{and} \quad \rho(K) = \begin{pmatrix} q & 0 \\ 0 & q^{-1} \end{pmatrix}.$$

When  $q$  is a generic parameter,  $U_q \mathfrak{sl}_2$  has the structure of topological half-ribbon Hopf algebra in the sense of [47], that is, it admits an  $R$ -matrix

$$R = q^{\frac{1}{2}(H \otimes H)} \exp_q((q - q^{-1})E \otimes F) \in \widetilde{U_q \mathfrak{sl}_2^{\otimes 2}}$$



(see [22] for details) and a half-ribbon element  $\Omega \in \widetilde{U_q \mathfrak{sl}_2}$  (defined by Kirillov and Reshetikhin in [34], where  $\Omega$  is denoted by  $w^{-1}$ ) such that  $\Delta(\Omega) = (\Omega \otimes \Omega)R$  and such that  $(U_q \mathfrak{sl}_2, R, \Omega^{-2})$  is a topological ribbon Hopf algebra. Note that the ribbon element  $v := \Omega^{-2}$  is not the usual one (see [47; 49] for details) but the Kauffman-bracket one (the one for which  $\text{qdim}(V) = -q - q^{-1}$  instead of  $q + q^{-1}$ ).

In the standard basis  $(v_+, v_-)$  of  $V$ , the matrix  $C = \text{Mat}_{(v_+, v_-)}(\Omega^{-1})$  is written

$$C = \begin{pmatrix} C_+^+ & C_+^- \\ C_-^+ & C_-^- \end{pmatrix} := \begin{pmatrix} 0 & \omega \\ -\omega^5 & 0 \end{pmatrix}.$$

Therefore

$$C^{-1} = -A^3 C = \begin{pmatrix} 0 & -\omega^{-5} \\ \omega^{-1} & 0 \end{pmatrix}.$$

Define the operators  $\tau, q^{\frac{1}{2}(H \otimes H)} \in \text{End}(V \otimes V)$  by

$$\tau(v_i \otimes v_j) := v_j \otimes v_i \quad \text{and} \quad q^{\frac{1}{2}(H \otimes H)}(v_i \otimes v_j) = A^{ij} v_i \otimes v_j$$

for  $i, j \in \{+, -\}$  (we identified  $-$  with  $-1$  and  $+$  with  $+1$ ). The braiding associated to the  $R$ -matrix is

$$\begin{aligned} \mathcal{R} = c_{V,V} &:= \tau \circ q^{\frac{1}{2}(H \otimes H)} \circ \exp_q((q - q^{-1})\rho(E) \otimes \rho(F)) \\ &= \tau \circ q^{\frac{1}{2}(H \otimes H)} \circ (\mathbb{1}_2 + (q - q^{-1})\rho(E) \otimes \rho(F)). \end{aligned}$$

In the basis  $(v_+ \otimes v_+, v_+ \otimes v_-, v_- \otimes v_+, v_- \otimes v_-)$ , it is written

$$\mathcal{R} = \begin{pmatrix} \mathcal{R}_{++}^{++} & \mathcal{R}_{+-}^{++} & \mathcal{R}_{-+}^{++} & \mathcal{R}_{--}^{++} \\ \mathcal{R}_{++}^{+-} & \mathcal{R}_{+-}^{+-} & \mathcal{R}_{-+}^{+-} & \mathcal{R}_{--}^{+-} \\ \mathcal{R}_{++}^{-+} & \mathcal{R}_{+-}^{-+} & \mathcal{R}_{-+}^{-+} & \mathcal{R}_{--}^{-+} \\ \mathcal{R}_{++}^{--} & \mathcal{R}_{+-}^{--} & \mathcal{R}_{-+}^{--} & \mathcal{R}_{--}^{--} \end{pmatrix} := \begin{pmatrix} A & 0 & 0 & 0 \\ 0 & 0 & A^{-1} & 0 \\ 0 & A^{-1} & A - A^{-3} & 0 \\ 0 & 0 & 0 & A \end{pmatrix},$$

so

$$\mathcal{R}^{-1} = \begin{pmatrix} A^{-1} & 0 & 0 & 0 \\ 0 & A^{-1} - A^3 & A & 0 \\ 0 & A & 0 & 0 \\ 0 & 0 & 0 & A^{-1} \end{pmatrix}.$$

We now list three families of skein relations, which are straightforward consequences of the definition, work regardless whether  $q$  is generic or a root of unity, and will be used later. Let  $i, j \in \{-, +\}$ .

- The *trivial arc relations*, which are given by

$$(3) \quad \begin{array}{|c|} \hline \uparrow \\ \hline \square_j^i \\ \hline \end{array} = C_j^i \begin{array}{|c|} \hline \square_j^i \\ \hline \end{array} \quad \text{and} \quad \begin{array}{|c|} \hline \downarrow \\ \hline \square_j^i \\ \hline \end{array} = (C^{-1})_j^i \begin{array}{|c|} \hline \square_j^i \\ \hline \end{array}.$$

- The *cutting arc relations*, which are given by

$$(4) \quad \boxed{C} = \sum_{i,j=\pm} C_j^i \begin{array}{c} \uparrow \\ \boxed{i} \\ \boxed{j} \end{array} \quad \text{and} \quad \boxed{C^{-1}} = \sum_{i,j=\pm} (C^{-1})_j^i \begin{array}{c} \boxed{i} \\ \boxed{j} \\ \downarrow \end{array}.$$

- The *height exchange relations*, which are given by

$$(5) \quad \begin{array}{c} \uparrow \\ \boxed{i} \\ \boxed{j} \end{array} = \begin{array}{c} \boxed{j} \\ \times \\ \boxed{i} \end{array} = \sum_{k,l=\pm} \mathcal{R}_{ij}^{kl} \begin{array}{c} \boxed{l} \\ \boxed{k} \\ \downarrow \end{array} \quad \text{and} \quad \begin{array}{c} \boxed{j} \\ \boxed{i} \\ \downarrow \end{array} = \begin{array}{c} \boxed{i} \\ \times \\ \boxed{j} \end{array} = \sum_{k,l=\pm} (\mathcal{R}^{-1})_{ij}^{kl} \begin{array}{c} \uparrow \\ \boxed{k} \\ \boxed{l} \end{array}.$$

We refer to [40] for proofs.

The algebra  $\mathcal{O}_q[\text{SL}_2]$  is the algebra presented by generators  $x_{\varepsilon\varepsilon'}, \varepsilon, \varepsilon' \in \{-, +\}$  and relations

$$\begin{aligned} x_{++}x_{+-} &= q^{-1}x_{+-}x_{++}, & x_{++}x_{-+} &= q^{-1}x_{-+}x_{++}, \\ x_{--}x_{+-} &= qx_{+-}x_{--}, & x_{--}x_{-+} &= qx_{-+}x_{--}, \\ x_{++}x_{--} &= 1 + q^{-1}x_{+-}x_{-+}, & x_{--}x_{++} &= 1 + qx_{+-}x_{-+}, \\ x_{-+}x_{+-} &= x_{+-}x_{-+}, \end{aligned}$$

It has a Hopf algebra structure characterized by

$$\begin{aligned} \begin{pmatrix} \Delta(x_{++}) & \Delta(x_{+-}) \\ \Delta(x_{-+}) & \Delta(x_{--}) \end{pmatrix} &= \begin{pmatrix} x_{++} & x_{+-} \\ x_{-+} & x_{--} \end{pmatrix} \otimes \begin{pmatrix} x_{++} & x_{+-} \\ x_{-+} & x_{--} \end{pmatrix}, \\ \begin{pmatrix} \epsilon(x_{++}) & \epsilon(x_{+-}) \\ \epsilon(x_{-+}) & \epsilon(x_{--}) \end{pmatrix} &= \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \\ \begin{pmatrix} S(x_{++}) & S(x_{+-}) \\ S(x_{-+}) & S(x_{--}) \end{pmatrix} &= \begin{pmatrix} x_{--} & -qx_{+-} \\ -q^{-1}x_{-+} & x_{++} \end{pmatrix}. \end{aligned}$$

When  $q \in \mathbb{C}^*$  is generic (not a root of unity),  $\mathcal{O}_q[\text{SL}_2]$  is the subalgebra of the restricted dual of  $U_q\mathfrak{sl}_2$  generated by the matrix elements of the integrable modules; see [14; 22]. The *bigon*  $\mathbb{B}$  is the punctured surface made of a disc with two punctures on its boundary. It has two boundary arcs  $a$  and  $b$  and is generated by the stated arcs  $\alpha_{\varepsilon\varepsilon'}, \varepsilon, \varepsilon' = \pm$  made of an arc  $\alpha$  linking  $a$  to  $b$  with state  $\varepsilon$  on  $\alpha \cap a$  and  $\varepsilon'$  on  $\alpha \cap b$ . Consider a disjoint union  $\mathbb{B} \sqcup \mathbb{B}$  of two bigons; by gluing together the boundary arc  $b_1$  of the first bigon with the boundary arc  $a_2$  of the second, one obtains a morphism  $\Delta := i|_{b_1 \# a_2} : \mathcal{S}_\omega(\mathbb{B}) \rightarrow \mathcal{S}_\omega(\mathbb{B})^{\otimes 2}$  which endows  $\mathcal{S}_\omega(\mathbb{B})$  with the structure of Hopf algebra where  $\Delta$  is the coproduct.

**Theorem 2.9** [26; 39; 40] *There is an isomorphism  $\varphi : \mathcal{O}_q[\text{SL}_2] \cong \mathcal{S}_\omega(\mathbb{B})$  of Hopf algebras sending the generator  $x_{\varepsilon\varepsilon'} \in \mathcal{O}_q[\text{SL}_2]$  to the element  $\alpha_{\varepsilon\varepsilon'} \in \mathcal{S}_\omega(\mathbb{B})$ .*

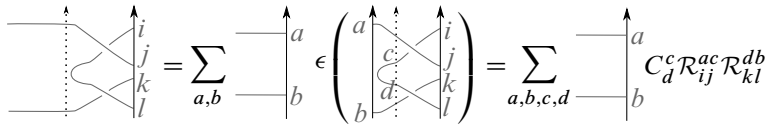


Figure 5: An example of the boundary skein relation.

More precisely, the fact that  $\varphi$  is an isomorphism of algebras is proved in [40] and the fact that it preserves the coproduct was noticed independently in [26; 39]. Throughout, we will (abusively) identify the Hopf algebras  $\mathcal{O}_q[\mathrm{SL}_2]$  and  $\mathcal{S}_\omega(\mathbb{B})$  using  $\varphi$ . Note that the definition of  $\varphi$  depends on an indexing by  $a$  and  $b$  of the boundary arcs of  $\mathbb{B}$ .

Now consider a punctured surface  $\Sigma$  and a boundary arc  $c$ . By gluing a bigon  $\mathbb{B}$  along  $\Sigma$  while gluing  $b$  with  $c$ , one obtains a punctured surface isomorphic to  $\Sigma$ , hence a map  $\Delta_c^L := i_{|b\#c} : \mathcal{S}_\omega(\Sigma) \rightarrow \mathcal{O}_q[\mathrm{SL}_2] \otimes \mathcal{S}_\omega(\Sigma)$  which endows  $\mathcal{S}_\omega(\Sigma)$  with the structure of left  $\mathcal{O}_q[\mathrm{SL}_2]$  comodule. Similarly, gluing  $c$  with  $a$  induces a right comodule morphism  $\Delta_c^R := i_{|c\#a} : \mathcal{S}_\omega(\Sigma) \rightarrow \mathcal{S}_\omega(\Sigma) \otimes \mathcal{O}_q[\mathrm{SL}_2]$ . The following theorem characterizes the image of the gluing map and was proved independently in [26; 39].

**Theorem 2.10** [26, Theorem 4.7; 39, Theorem 1.1] *Let  $\Sigma$  be a punctured surface, and  $a$  and  $b$  two boundary arcs. The sequence*

$$0 \rightarrow \mathcal{S}_\omega(\Sigma_{|a\#b}) \xrightarrow{i_{|a\#b}} \mathcal{S}_\omega(\Sigma) \xrightarrow{\Delta_a^L - \sigma \circ \Delta_b^R} \mathcal{O}_q[\mathrm{SL}_2] \otimes \mathcal{S}_\omega(\Sigma)$$

is exact, where  $\sigma(x \otimes y) := y \otimes x$ .

An easy but very important consequence of the fact that  $\Delta_a^L$  and  $\Delta_a^R$  are comodule maps are the *boundary skein relations*

$$(6) \quad (\epsilon \otimes \mathrm{id}) \circ \Delta_a^L = \mathrm{id} \quad \text{and} \quad (\mathrm{id} \otimes \epsilon) \circ \Delta_a^R = \mathrm{id}.$$

The image through the counit  $\epsilon$  of a stated diagram in  $\mathbb{B}$  can be computed using

$$(7) \quad \begin{aligned} \epsilon\left(\begin{array}{c} \uparrow \\ \boxed{\text{C}} \\ \uparrow \\ i \\ j \end{array}\right) &= C_j^i, & \epsilon\left(\begin{array}{c} \uparrow \\ \boxed{\text{D}} \\ \uparrow \\ i \\ j \end{array}\right) &= (C^{-1})_j^i, \\ \epsilon\left(\begin{array}{c} \uparrow \\ \boxed{\text{X}} \\ \uparrow \\ i \\ j \end{array} \begin{array}{c} \uparrow \\ \boxed{\text{Y}} \\ \uparrow \\ k \\ l \end{array}\right) &= \mathcal{R}_{kl}^{ij}, & \epsilon\left(\begin{array}{c} \uparrow \\ \boxed{\text{Z}} \\ \uparrow \\ i \\ j \end{array} \begin{array}{c} \uparrow \\ \boxed{\text{W}} \\ \uparrow \\ k \\ l \end{array}\right) &= (\mathcal{R}^{-1})_{kl}^{ij}. \end{aligned}$$

Figure 5 illustrates an instance of boundary skein relation (6). Here we draw a dotted arrow to illustrate where we cut the bigon. Note that all the trivial arc (3), cutting arc (4) and height exchange (5) relations are particular cases of (6).

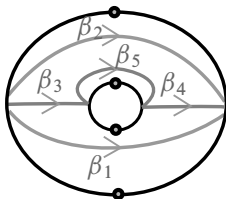


Figure 6: A punctured surface and a set of generators for its small fundamental groupoid.

### 2.2 The small fundamental groupoid and its finite presentations

In this section we fix a punctured surface  $\Sigma = (\Sigma, \mathcal{P})$  such that  $\Sigma$  is connected and has nonempty boundary. For each boundary arc  $a$  of  $\Sigma$ , fix a point  $v_a \in a$  and denote by  $\mathbb{V}$  the set  $\{v_a\}_a$ .

**Definition 2.11** The *small fundamental groupoid*  $\Pi_1(\Sigma_{\mathcal{P}}, \mathbb{V})$  is the full subcategory of the fundamental groupoid  $\Pi_1(\Sigma_{\mathcal{P}})$  generated by  $\mathbb{V}$ .

Said differently,  $\Pi_1(\Sigma_{\mathcal{P}}, \mathbb{V})$  is the small groupoid whose set of objects is  $\mathbb{V}$  and such that a morphism (called a path)  $\alpha: v_1 \rightarrow v_2$  is a homotopy class of continuous maps  $\varphi_\alpha: [0, 1] \rightarrow \Sigma_{\mathcal{P}}$  with  $\varphi_\alpha(0) = v_1$  and  $\varphi_\alpha(1) = v_2$ . The map  $\varphi_\alpha$  will be referred to as a *geometric representative* of  $\alpha$ . The composition is the concatenation of paths. For a path  $\alpha: v_1 \rightarrow v_2$  we write  $s(\alpha) = v_1$  (the source point) and  $t(\alpha) = v_2$  (the target point), and  $\alpha^{-1}: v_2 \rightarrow v_1$  is the path with opposite orientation ( $\varphi_{\alpha^{-1}}(t) = \varphi_\alpha(1 - t)$ ).

We will define the notion of *finite presentation*  $\mathbb{P}$  of the groupoid  $\Pi_1(\Sigma_{\mathcal{P}}, \mathbb{V})$  and attach to each such  $\mathbb{P}$  a finite presentation of  $\mathcal{S}_\omega(\Sigma)$ . In order to get some intuition, consider the punctured surface in Figure 6: it is an annulus with two punctures per boundary component, so it has four boundary arcs. The figure shows some paths  $\beta_1, \dots, \beta_5$  and we will say that  $\Pi_1(\Sigma_{\mathcal{P}}, \mathbb{V})$  is finitely presented by the set of generators  $\{\beta_1, \dots, \beta_5\}$  together with the relation  $\beta_2^{-1} \beta_4 \beta_5 \beta_3 = 1$ . We will deduce that  $\mathcal{S}_\omega(\Sigma)$  is generated by the stated arcs  $(\beta_i)_{i \in \mathcal{E}}$  and that the relation  $\beta_2^{-1} \beta_4 \beta_5 \beta_3 = 1$  induces a relation among them. Alternatively, the fundamental groupoid of the same punctured surface has a presentation with the smaller set of generators  $\{\beta_1, \dots, \beta_4\}$  and no relation. The induced finite presentation of  $\mathcal{S}_\omega(\Sigma)$  will be simpler.

**Definition 2.12** (i) A *set of generators* for  $\Pi_1(\Sigma_{\mathcal{P}}, \mathbb{V})$  is a set  $\mathbb{G}$  consisting of paths in  $\Pi_1(\Sigma_{\mathcal{P}}, \mathbb{V})$  such that any path  $\alpha \in \Pi_1(\Sigma_{\mathcal{P}}, \mathbb{V})$  decomposes as  $\alpha = \alpha_1^{\varepsilon_1} \dots \alpha_n^{\varepsilon_n}$  with  $\varepsilon_i = \pm 1$  and  $\alpha_i \in \mathbb{G}$ . We also require that each path  $\alpha \in \mathbb{G}$  is the homotopy class of

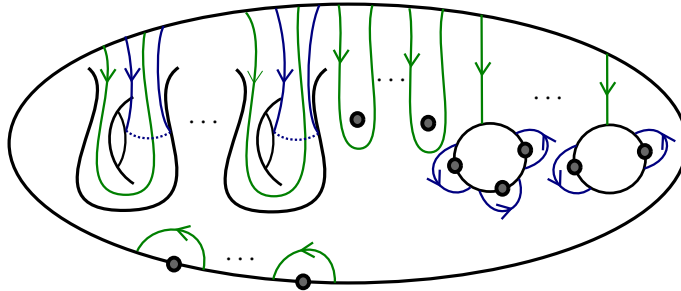


Figure 7: The geometric representatives of a set of generators for  $\Pi_1(\Sigma_{\mathcal{P}}, \mathbb{V})$ .

some embedding  $\varphi_\alpha : [0, 1] \rightarrow \Sigma_{\mathcal{P}}$  such that the images of the  $\varphi_\alpha$  do not intersect outside  $\mathbb{V}$  and possibly intersect transversally at  $\mathbb{V}$ . The *generating graph* is the oriented ribbon graph  $\Gamma \subset \Sigma_{\mathcal{P}}$  whose set of vertices is  $\mathbb{V}$  and edges are the images of the  $\varphi_\alpha$ . We will always assume implicitly that the geometric representatives  $\varphi_\alpha$  are part of the data defining a set of generators. Moreover, when  $\alpha \in \mathbb{G}$  is a path such that  $s(\alpha) = t(\alpha)$  (ie  $\alpha$  is a loop) we add the additional datum of a “height order” for its endpoints, that is we specify whether  $h(s(\alpha)) < h(t(\alpha))$  or  $h(t(\alpha)) < h(s(\alpha))$ .

(ii) For a path  $\alpha : v_1 \rightarrow v_2$  and  $\varepsilon, \varepsilon' \in \{-, +\}$ , we denote by  $\alpha_{\varepsilon\varepsilon'} \in \mathcal{S}_\omega(\Sigma)$  the class of the stated arc  $(\alpha, \sigma)$ , where the state  $\sigma$  is given by  $\sigma(v_1) = \varepsilon$  and  $\sigma(v_2) = \varepsilon'$ . When both endpoints lie in the same boundary arc (when  $s(\alpha) = t(\alpha)$ ) we use the chosen height order to specify which endpoint lies on the top. Set

$$\mathcal{A}^{\mathbb{G}} := \{\alpha_{\varepsilon\varepsilon'} \mid \alpha \in \mathbb{G} \text{ and } \varepsilon, \varepsilon' \in \{-, +\}\} \subset \mathcal{S}_\omega(\Sigma).$$

**Example 2.13** For any connected open punctured surface  $\Sigma$ , the groupoid  $\Pi_1(\Sigma_{\mathcal{P}}, \mathbb{V})$  admits a finite set of generators depicted in Figure 7 and defined as follows. Denote by  $a_0, \dots, a_n$  the boundary arcs, by  $\partial_0, \dots, \partial_r$  the boundary components of  $\Sigma$  with  $a_0 \subset \partial_0$ , and write  $v_i := a_i \cap \mathbb{V}$ . Let  $\bar{\Sigma}$  be the surface obtained from  $\Sigma$  by gluing a disc along each boundary component  $\partial_i$  for  $1 \leq i \leq r$ , and choose  $\alpha_1, \beta_1, \dots, \alpha_g, \beta_g$  some paths in  $\pi_1(\Sigma_{\mathcal{P}}, v_0)$  (which equals  $\text{End}_{\Pi_1(\Sigma_{\mathcal{P}}, \mathbb{V})}(v_0)$ ) such that their images in  $\bar{\Sigma}$  generate the free group  $\pi_1(\bar{\Sigma}, v_0)$  (said differently, the  $\alpha_i$  and  $\beta_i$  are longitudes and meridians of  $\Sigma$ ). For each inner puncture  $p$  choose a peripheral curve  $\gamma_p \in \pi_1(\Sigma_{\mathcal{P}}, v_0)$  encircling  $p$  once and for each boundary puncture  $p_\partial$  between two boundary arcs  $a_i$  and  $a_j$ , consider the path  $\alpha_{p_\partial} : v_i \rightarrow v_j$  represented by the corner arc in  $p_\partial$ . Finally, for each boundary component  $\partial_j$ , with  $1 \leq j \leq r$ , containing a boundary arc  $a_{k_j} \subset \partial_j$ , choose a path  $\delta_{\partial_j} : v_0 \rightarrow v_{k_j}$ . The set

$$\mathbb{G}' := \{\alpha_i, \beta_i, \alpha_p, \delta_{\partial_j} \mid 1 \leq i \leq g, p \in \mathcal{P} \text{ and } 1 \leq j \leq r\}$$

Figure 8: How an application of the cutting arc relations permits us to express any simple stated diagram in terms of the elements of  $\mathcal{A}^{\mathbb{G}}$ . Here  $\mathbb{G} = \{\beta_1, \beta_2, \beta_3, \beta_4\}$  is the set of generators of Figure 6. We draw dotted arrows to exhibit where we perform the cutting arc relations.

is a generating set for  $\Pi_1(\Sigma_{\mathcal{P}}, \mathbb{V})$  and Figure 7 represents a set of geometric representatives for  $\mathbb{G}'$ . Moreover each of its generators which is not one of the  $\delta_{\partial_j}$  can be expressed as a composition of the other ones (we will soon say that there is a relation among those generators), therefore a set  $\mathbb{G}$  obtained from  $\mathbb{G}'$  by removing one of the element of the form  $\alpha_i, \beta_i$  or  $\gamma_p$  is still a generating set for  $\Pi_1(\Sigma_{\mathcal{P}}, \mathbb{V})$ . The height orders can be chosen arbitrarily. Note that  $\mathbb{G}$  has cardinality  $2g - 2 + s + n_{\partial}$ , where  $g$  is the genus of  $\Sigma$ ,  $s := |\mathcal{P}|$  is the number of punctures and  $n_{\partial} := |\pi_0(\partial\Sigma)|$  is the number of boundary components.

In the particular case where  $\Sigma$  has exactly one boundary component with one puncture on it (and possibly inner punctures), the generating graph of  $\mathbb{G}$  is called the *daisy graph* (see Figure 9). The daisy graph was first considered in [4] in the context of classical lattice gauge field theory and in [5; 8; 27; 28] in the quantum case.

**Proposition 2.14** [38, Proposition 3.4] *If  $\mathbb{G}$  is a set of generators of  $\Pi_1(\Sigma_{\mathcal{P}}, \mathbb{V})$ , then the set  $\mathcal{A}^{\mathbb{G}}$  generates  $\mathcal{S}_{\omega}(\Sigma)$  as an algebra.*

The proof of Proposition 2.14 is an easy consequence of the cutting arc relations illustrated in Figure 8.

We now define the notion of relations for a generating set  $\mathbb{G}$ . Let  $\mathcal{F}(\mathbb{G})$  denote the free semigroup generated by the elements of  $\mathbb{G}$  and let  $\text{Rel}_{\mathbb{G}}$  denote the subset of  $\mathcal{F}(\mathbb{G})$  of elements of the form  $R = \beta_1 \star \dots \star \beta_n$  such that  $s(\beta_i) = t(\beta_{i+1})$  and such that the path  $\beta_1 \dots \beta_n$  is trivial. We write  $R^{-1} := \beta_n^{-1} \star \dots \star \beta_1^{-1}$ . A relation  $R = \beta_1 \star \dots \star \beta_n \in \text{Rel}_{\mathbb{G}}$  is called *simple* if the  $\beta_i$  admit as representatives embedded curves whose concatenation forms a contractible simple closed curve  $\gamma$  in  $\Sigma_{\mathcal{P}}$  whose orientation coincides with the orientation of the disc bounded by  $\gamma$ . Note that “being simple” depends on the choice of geometric representatives of the generators.

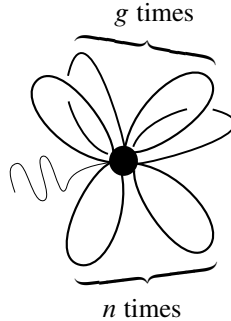


Figure 9: A daisy graph.

**Definition 2.15** A finite subset  $\mathbb{RL} \subset \text{Rel}_{\mathbb{G}}$  is called a *finite set of relations* if its elements are simple and every word  $R \in \text{Rel}_{\mathbb{G}}$  can be decomposed as

$$R = \beta \star R_1^{\varepsilon_1} \star \cdots \star R_m^{\varepsilon_m} \star \beta^{-1},$$

where  $R_i \in \mathbb{RL}$ ,  $\varepsilon_i \in \{\pm 1\}$  and  $\beta = \beta_1 \star \cdots \star \beta_n \in \mathcal{F}(\mathbb{G})$  is such that  $s(\beta_i) = t(\beta_{i+1})$ . The pair  $\mathbb{P} := (\mathbb{G}, \mathbb{RL})$  is called a *finite presentation* of  $\Pi_1(\Sigma_{\mathcal{P}}, \mathbb{V})$ .

As illustrated in the introduction, the small fundamental groupoid of the triangle  $\mathbb{T}$  admits the finite presentation with generating set  $\mathbb{G} = \{\alpha, \beta, \gamma\}$  and unique relation  $\mathbb{RL} = \{\alpha \star \beta \star \gamma\}$ .

For a general connected open punctured surface  $\Sigma$ , the set  $\mathbb{G}$  of Example 2.13 is the generating set of a presentation of  $\Pi_1(\Sigma_{\mathcal{P}}, \mathbb{V})$  with no relations.

### 2.3 Relations among the generators of the stated skein algebras

We fix a connected open punctured surface  $\Sigma$ , a finite presentation  $\mathbb{P} = (\mathbb{G}, \mathbb{RL})$  of  $\Pi_1(\Sigma_{\mathcal{P}}, \mathbb{V})$ , and look for relations in  $\mathcal{S}_{\omega}(\Sigma)$  among the elements of  $\mathcal{A}^{\mathbb{G}}$ .

**Definition 2.16** An *oriented arc*  $\beta$  is a nonclosed connected simple diagram of  $\Sigma_{\mathcal{P}}$  together with an orientation plus a possible height order of its endpoints in the case where they both lie in the same boundary arc. We will denote by  $s(\beta)$  and  $t(\beta)$  its endpoints so that  $\beta$  is oriented from  $s(\beta)$  towards  $t(\beta)$ . For  $\varepsilon, \varepsilon' \in \{-, +\}$ , we denote by  $\beta_{\varepsilon\varepsilon'} \in \mathcal{S}_{\omega}(\Sigma)$  the class of the stated diagram  $(\beta, \sigma)$  where  $\sigma(s(\beta)) = \varepsilon$  and  $\sigma(t(\beta)) = \varepsilon'$ .

Note that to each oriented arc one can associate a path in  $\Pi_1(\Sigma_{\mathcal{P}}, \mathbb{V})$  by first isotoping its endpoints to  $\mathbb{V}$  and then taking its homotopy class. However a path in  $\Pi_1(\Sigma_{\mathcal{P}}, \mathbb{V})$  can be associated to several distinct oriented arcs, so an oriented arc contains more information than a path in the small fundamental groupoid.

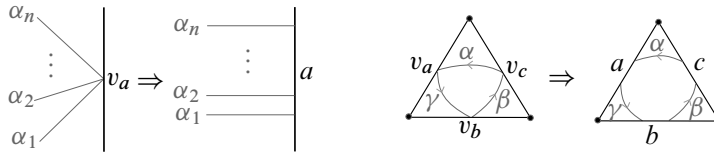


Figure 10: Left: an illustration of the local isotopy we perform to turn the set of edges of a (ribbon) presenting graph into a set of pairwise nonintersecting oriented arcs. Right: an example in the case of the triangle.

We want to see the elements of  $\mathbb{G}$  as pairwise nonintersecting oriented arcs as illustrated in Figure 10. Recall that by Definition 2.12, any path  $\alpha \in \mathbb{G}$  is endowed with a geometric representative  $\varphi_\alpha$  whose image is an oriented arc  $\underline{\alpha} \subset \Sigma_{\mathcal{P}}$  such that the  $\underline{\alpha}$  pairwise do not intersect outside of  $\mathbb{V}$  and they intersect transversally in  $\mathbb{V}$ . So each point  $v_a \in \mathbb{V}$  is endowed with a total order  $<_{v_a}$  on the set of its adjacent arcs (so the presenting graph has a ciliated ribbon graph structure).

The orientation of  $\Sigma_{\mathcal{P}}$  induces an orientation of its boundary arcs, which, in turn, induces a total order  $<_a$  on each boundary arc  $a$ , where  $v_1 <_a v_2$  if  $a$  is oriented from  $v_1$  towards  $v_2$ . After isotoping the  $\underline{\alpha}$  in a small neighborhood of each  $v_a$  in such a way that the vertex order  $<_{v_a}$  matches with the boundary arc order  $<_a$  as illustrated in Figure 10, we get a family of pairwise nonintersecting oriented arcs representing the elements of  $\mathbb{G}$ .

**Convention 2.17** From now on we consider the elements of  $\mathbb{G}$  as pairwise nonintersecting oriented arcs.

**Definition 2.18** Let  $\alpha$  be an oriented arc, set  $v_1 := s(\alpha)$  and  $v_2 := t(\alpha)$  and denote by  $u$  and  $v$  the boundary arcs containing  $v_1$  and  $v_2$ , respectively. The arc  $\alpha$  is

- of type *a* if  $u \neq v$ ,
- of type *b* if  $u = v$ ,  $h(v_1) < h(v_2)$  and  $v_2 <_u v_1$ ,
- of type *c* if  $u = v$ ,  $h(v_2) < h(v_1)$  and  $v_1 <_u v_2$ ,
- of type *d* if  $u = v$ ,  $h(v_1) < h(v_2)$  and  $v_1 <_u v_2$ ,
- of type *e* if  $u = v$ ,  $h(v_2) < h(v_1)$  and  $v_2 <_u v_1$ .

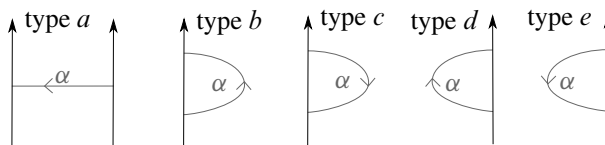


Figure 11: An illustration of the five types of oriented arcs.



Here  $h(v)$  represents the height of  $v$  ( $h$  is the second projection  $\Sigma_{\mathcal{P}} \times (0, 1) \rightarrow (0, 1)$ ). Figure 11 illustrates the five types of oriented arcs.

**Notation 2.19** (i) For  $\alpha$  an oriented arc, write  $M(\alpha) := \begin{pmatrix} \alpha_{++} & \alpha_{+-} \\ \alpha_{-+} & \alpha_{--} \end{pmatrix}$ , the  $2 \times 2$  matrix with coefficients in  $\mathcal{S}_\omega(\Sigma)$ . The relations among the generators of  $\mathcal{S}_\omega(\Sigma)$  that we will soon define are much more elegant when written using the matrix

$$N(\alpha) := \begin{cases} M(\alpha) & \text{if } \alpha \text{ is of type } a, \\ M(\alpha)C & \text{if } \alpha \text{ is of type } b, \\ M(\alpha)^t C & \text{if } \alpha \text{ is of type } c, \\ C^{-1}M(\alpha) & \text{if } \alpha \text{ is of type } d, \\ {}^t C^{-1}M(\alpha) & \text{if } \alpha \text{ is of type } e, \end{cases}$$

where  ${}^t M$  denotes the transpose of  $M$ .

(ii) Let  $M_{a,b}(R)$  be the ring of  $a \times b$  matrices with coefficients in some ring  $R$  (here  $R$  will be  $\mathcal{S}_\omega(\Sigma)$ ). The Kronecker product  $\odot: M_{a,b}(R) \otimes M_{c,d}(R) \rightarrow M_{ac,bd}(R)$  is defined by  $(A \odot B)_{j,l}^{i,k} = A_j^i B_l^k$ . For instance,

$$M(\alpha) \odot M(\beta) = \begin{pmatrix} \alpha_{++}\beta_{++} & \alpha_{++}\beta_{+-} & \alpha_{+-}\beta_{++} & \alpha_{+-}\beta_{+-} \\ \alpha_{++}\beta_{-+} & \alpha_{++}\beta_{--} & \alpha_{+-}\beta_{-+} & \alpha_{+-}\beta_{--} \\ \alpha_{-+}\beta_{++} & \alpha_{-+}\beta_{+-} & \alpha_{--}\beta_{++} & \alpha_{--}\beta_{+-} \\ \alpha_{-+}\beta_{-+} & \alpha_{-+}\beta_{--} & \alpha_{--}\beta_{-+} & \alpha_{--}\beta_{--} \end{pmatrix}.$$

(iii) By abuse of notation  $\tau$  also denotes the matrix of the flip map  $\tau: V^{\otimes 2} \rightarrow V^{\otimes 2}$  given by  $v_i \otimes v_j \mapsto v_j \otimes v_i$ :

$$\tau = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

(iv) For a  $4 \times 4$  matrix  $X = (X_{kl}^{ij})_{i,j,k,l=\pm}$ , we define the  $2 \times 2$  matrices  $\text{tr}_L(X)$  and  $\text{tr}_R(X)$  by

$$\text{tr}_L(X)_a^b := \sum_{i=\pm} X_{ia}^{ib} \quad \text{and} \quad \text{tr}_R(X)_a^b := \sum_{i=\pm} X_{ai}^{bi}.$$

(v) For  $M = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ , we set  $\det_q(M) := ad - q^{-1}bc$  and  $\det_{q^2}(M) := ad - q^{-2}bc$ .

**Lemma 2.20** (orientation-reversing formulas) *Let  $\alpha$  be an oriented arc and  $\alpha^{-1}$  be the same arc with opposite orientation. Then one has*

$$M(\alpha^{-1}) = {}^t M(\alpha).$$

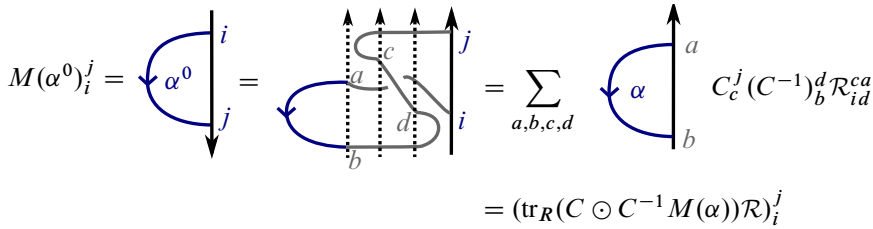


Figure 12: An illustration of the proof of (9) in the case where  $\alpha$  is of type  $e$ .

Therefore,

$$(8) \quad N(\alpha^{-1}) = \begin{cases} {}^t N(\alpha) & \text{if } \alpha \text{ is of type } a, \\ {}^t C^{-1} {}^t N(\alpha) {}^t C & \text{if } \alpha \text{ is of type } b \text{ or } d, \\ C^{-1} {}^t N(\alpha) C & \text{if } \alpha \text{ is of type } c \text{ or } e. \end{cases}$$

**Proof** This is a straightforward consequence of the definitions. □

**Lemma 2.21** (height-reversing formulas) *Let  $\alpha$  be an oriented arc with both endpoints in the same boundary arc and let  $\alpha^0$  be the same arc with reversed height order for its endpoints. Then one has*

$$(9) \quad M(\alpha^0) = \begin{cases} \text{tr}_R(\mathcal{R}^{-1}({}^t C^{-1} \odot M(\alpha) {}^t C)) & \text{if } \alpha \text{ is of type } b, \\ \text{tr}_L(\mathcal{R}^{-1}(M(\alpha) C \odot C^{-1})) & \text{if } \alpha \text{ is of type } c, \\ \text{tr}_L({}^t C^{-1} M(\alpha) \odot {}^t C) \mathcal{R} & \text{if } \alpha \text{ is of type } d, \\ \text{tr}_R((C \odot C^{-1} M(\alpha)) \mathcal{R}) & \text{if } \alpha \text{ is of type } e. \end{cases}$$

**Proof** Equation (9) is obtained by using the boundary skein relations (6). Figure 12 illustrates the proof in the case where  $\alpha$  is of type  $e$ . The other cases are similar and left to the reader.

In Figure 12, we represent the curve  $\alpha$  in blue to emphasize that, despite what the picture suggests, the curve can be arbitrarily complicated. Since the boundary arc relation only involves the intersection of  $\alpha$  with a small neighborhood (a bigon) of the boundary arc (colored in gray), the exact structure of the blue part of the figure does not matter. □

**Remark 2.22** Reversing the orientation of an arc exchanges type  $b$  with type  $c$  and type  $d$  with type  $e$ , whereas reversing the height order exchanges type  $b$  with type  $e$  and type  $c$  with type  $d$ . Therefore (8) and (9) permit us to switch between the types  $b, c, d$  and  $e$ ; this will permit us to write the arc exchange and trivial loop relations in a simpler form by specifying the type of arc.

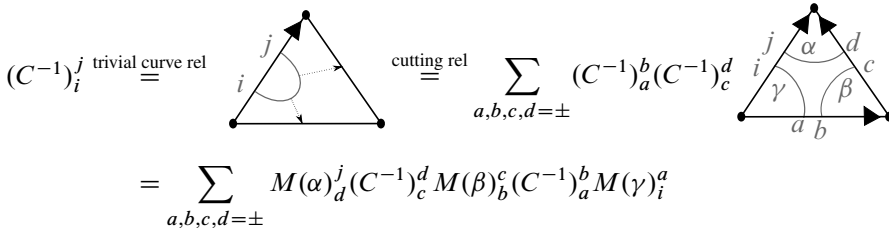


Figure 13: An illustration of the proof of (10) in the case of the triangle.

**Lemma 2.23** (trivial loop relations) *Let  $R = \beta_k \star \dots \star \beta_1$  be a simple relation. Suppose that all arcs  $\beta_i$  are either of type  $a$  or  $d$ . Then*

$$(10) \quad \mathbb{1}_2 = CM(\beta_k)C^{-1}M(\beta_{k-1})C^{-1} \dots C^{-1}M(\beta_1).$$

**Proof** Equation (10) is a consequence of the trivial arc and cutting arc relations illustrated in Figure 13 in the case of the triangle with presentation whose generators are the arcs  $\{\alpha, \beta, \gamma\}$  drawn in Figure 1 and where the relation is  $\alpha \star \beta \star \gamma = 1$ . Figure 13 shows the equality between the matrix coefficients of  $C^{-1}$  and  $M(\alpha)C^{-1}M(\beta)C^{-1}M(\gamma)$ .

Let us detail the proof in the general case. Since  $\beta_i$  is either of type  $a$  or  $d$ , it can be represented by a tangle  $T(\beta_i)$  such that the height of the source endpoint of  $\beta_i$  (say  $v_i$ ) is smaller than the height of its target endpoint (say  $w_i$ ); said differently  $h(v_i) < h(w_i)$ . One can further choose the  $T(\beta_i)$  so that  $T(\beta_{i+1})$  lies on the top of  $T(\beta_i)$  (so  $h(v_1) < h(w_1) < h(v_2) < \dots < h(w_k)$ ). Let  $T$  be the tangle made of the disjoint union of the  $T(\beta_i)$ . By the assumption that  $R$  is a simple relation, we can suppose that  $T$  is in generic position (in the sense of Section 2.1) and that its projection diagram is simple. Fix  $i, j \in \{-, +\}$  and let  $\alpha^0$  be a trivial arc with endpoints  $s(\alpha^0) = v_1$  and  $t(\alpha^0) = w_k$  such that  $\alpha^0$  can be isotoped (relative to its boundary) to an arc inside  $\partial\Sigma_{\mathcal{P}}$ . On the one hand, the trivial arc relation (3) gives the equality  $\alpha_{ij}^0 = (C^{-1})_i^j$ . On the other hand, the cutting arc relation (4) gives the equality

$$\begin{aligned} (C^{-1})_i^j &= \alpha_{ij}^0 = \sum_{\substack{s \in \text{St}(T) \\ s(v_1)=i \\ s(w_k)=j}} [T, s](C^{-1})_{s(w_1)}^{s(v_2)}(C^{-1})_{s(w_2)}^{s(v_3)} \dots (C^{-1})_{s(w_{k-1})}^{v_k} \\ &= \sum_{\mu_1, \dots, \mu_{2k-2}=\pm} M(\beta_k)_{\mu_1}^j (C^{-1})_{\mu_2}^{\mu_1} M(\beta_{k-1})_{\mu_2}^{\mu_3} \dots M(\beta_1)_{\mu_1}^{\mu_{2k-2}} \\ &= (M(\beta_k)C^{-1}M(\beta_{k-1})C^{-1} \dots M(\beta_1))_i^j. \quad \square \end{aligned}$$

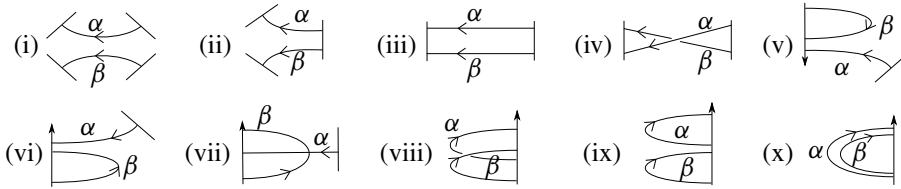


Figure 14: Ten configurations for two nonintersecting oriented arcs.

Let  $\alpha$  and  $\beta$  be two nonintersecting oriented arcs. Denote by  $a, b, c$  and  $d$  the boundary arcs containing  $s(\alpha), t(\alpha), s(\beta)$  and  $t(\beta)$ , respectively. Reversing the orientation and the height order of  $\alpha$  or  $\beta$  if necessary, we have ten different possibilities illustrated in Figure 14. The proof of the following lemma is very similar to the computations made by Faitg in [27].

**Lemma 2.24** (i) *If the elements of  $\{a, b, c, d\}$  are pairwise distinct, one has*

$$(11) \quad N(\alpha) \odot N(\beta) = \tau(N(\beta) \odot N(\alpha))\tau.$$

(ii) *When  $a = c$ ,  $\{a, b, d\}$  has cardinality 3 and  $s(\beta) <_a s(\alpha)$ , one has*

$$(12) \quad N(\alpha) \odot N(\beta) = \tau(N(\beta) \odot N(\alpha))\mathcal{R}.$$

(iii) *When  $a = c \neq b = d$ ,  $s(\beta) <_a s(\alpha)$  and  $t(\alpha) <_b t(\beta)$ , one has*

$$(13) \quad N(\alpha) \odot N(\beta) = \mathcal{R}^{-1}(N(\beta) \odot N(\alpha))\mathcal{R}.$$

(iv) *When  $a = c \neq b = d$ ,  $s(\beta) <_a s(\alpha)$  and  $t(\beta) <_b t(\alpha)$ , one has*

$$(14) \quad N(\alpha) \odot N(\beta) = \mathcal{R}(N(\beta) \odot N(\alpha))\mathcal{R}.$$

(v) *When  $b = c = d \neq a$ ,  $s(\beta) <_a t(\beta) <_a t(\alpha)$  and  $h(s(\beta)) < h(t(\beta))$ , one has*

$$(15) \quad N(\alpha) \odot N(\beta) = \mathcal{R}^{-1}(N(\beta) \odot \mathbb{1}_2)\mathcal{R}(N(\alpha) \odot \mathbb{1}_2).$$

(vi) *When  $b = c = d \neq a$ ,  $t(\alpha) <_a t(\beta) <_a s(\beta)$  and  $h(s(\beta)) < h(t(\beta)) < h(t(\alpha))$ , one has*

$$(16) \quad N(\alpha) \odot N(\beta) = \mathcal{R}^{-1}(N(\beta) \odot \mathbb{1}_2)\mathcal{R}(N(\alpha) \odot \mathbb{1}_2).$$

(vii) *When  $b = c = d \neq a$ ,  $t(\beta) <_a t(\alpha) <_a s(\beta)$  and  $h(s(\beta)) < h(t(\alpha)) < h(t(\beta))$ , one has*

$$(17) \quad N(\alpha) \odot N(\beta) = \mathcal{R}(N(\beta) \odot \mathbb{1}_2)\mathcal{R}(N(\alpha) \odot \mathbb{1}_2).$$

(viii) When  $a = b = c = d$ ,  $s(\beta) <_a s(\alpha) <_a t(\beta) <_a t(\alpha)$  and

$$h(s(\beta)) < h(s(\alpha)) < h(t(\beta)) < h(t(\alpha)),$$

one has

$$(18) \quad (\mathbb{1}_2 \odot N(\alpha))\mathcal{R}^{-1}(\mathbb{1}_2 \odot N(\beta))\mathcal{R}^{-1} = \mathcal{R}(\mathbb{1}_2 \odot N(\beta))\mathcal{R}^{-1}(\mathbb{1}_2 \odot N(\alpha)).$$

(ix) When  $a = b = c = d$ ,  $s(\beta) <_a t(\beta) <_a s(\alpha) <_a t(\alpha)$  and

$$h(s(\beta)) < h(t(\beta)) < h(s(\alpha)) < h(t(\alpha)),$$

one has

$$(19) \quad \mathcal{R}^{-1}(\mathbb{1}_2 \odot N(\alpha))\mathcal{R}(\mathbb{1}_2 \odot N(\beta)) = (\mathbb{1}_2 \odot N(\beta))\mathcal{R}^{-1}(\mathbb{1}_2 \odot N(\alpha))\mathcal{R}.$$

(x) When  $a = b = c = d$ ,  $s(\alpha) <_a s(\beta) <_a t(\beta) <_a t(\alpha)$  and

$$h(s(\alpha)) < h(s(\beta)) < h(t(\beta)) < h(t(\alpha)),$$

one has

$$(20) \quad (\mathbb{1}_2 \odot N(\alpha))\mathcal{R}^{-1}(\mathbb{1}_2 \odot N(\beta))\mathcal{R} = \mathcal{R}(\mathbb{1}_2 \odot N(\beta))\mathcal{R}^{-1}(\mathbb{1}_2 \odot N(\alpha)).$$

**Proof** Equation (11) says that in case (i) any  $\alpha_{ij}$  commutes with any  $\beta_{kl}$ , which is obvious. Equations (12), (13) and (14) in cases (ii), (iii) and (iv) are straightforward consequences of the height exchange relation (5). All other cases will be derived using the boundary skein relations (6). As in the proof of Lemma 2.21, we will color the arcs  $\alpha$  and  $\beta$  in red and blue to remind the reader that they might be much more complicated than they look in the picture: in the computations we perform while using the boundary skein relation we only care about the restriction of the diagrams (depicted in gray) in a small bigon in the neighborhood of the boundary arc  $a$  and not the actual shape of the blue and red parts.

Equations (15) and (16) in cases (v) and (vi) are proved in a very similar way; we detail the proof of (16) and leave (17) to the reader. In case (vi), one has

$$\begin{aligned} (M(\alpha) \odot M(\beta))_{kl}^{ij} &= \alpha_{ki} \beta_{lj} = \begin{array}{c} i \nearrow \alpha \searrow k \\ j \text{---} \beta \text{---} l \end{array} = \begin{array}{c} i \nearrow \alpha \searrow k \\ j \text{---} \beta \text{---} l \\ \text{---} a \text{---} b \text{---} c \text{---} d \text{---} e \text{---} f \end{array} \\ &= \sum_{a,b,c,d,e,f=\pm} (\mathcal{R}^{-1})_{fd}^{ij} M(\beta)_e^f C_c^e \mathcal{R}_{ab}^{cd} M(\alpha)_k^a (C^{-1})_l^b \\ &= (\mathcal{R}^{-1}(M(\beta)C \odot \mathbb{1}_2)\mathcal{R}(M(\alpha) \odot C^{-1}))_{kl}^{ij}. \end{aligned}$$

To handle cases (vii)–(x), we introduce the  $4 \times 4$  matrix  $V = (V_{kl}^{ij})_{i,j,k,l \in \{-,+\}}$ , where  $V_{kl}^{ij} = [\alpha \cup \beta, \sigma_{ijkl}] \in \mathcal{S}_\omega(\Sigma)$  is the class of the simple diagram  $\alpha \cup \beta$  with state  $\sigma_{ijkl}$  sending  $t(\alpha), t(\beta), s(\alpha)$  and  $s(\beta)$  to  $i, j, k$  and  $l$ , respectively. Here the height order of the points of  $\partial(\alpha \cup \beta)$  is given by the boundary arc orientation drawn in Figure 14. The trick is to compute  $V$  in two different ways and then equate the two obtained formulas.

In case (vii), on the one hand, we first prove  $V = \tau(M(\beta)C \odot \mathbb{1}_2)\mathcal{R}(M(\alpha) \odot C^{-1})$ :

$$V_{kl}^{ij} = \begin{array}{c} j \\ \beta \\ i \\ \alpha \\ l \end{array} \Big|_k = \begin{array}{c} j \\ \beta \\ i \\ \alpha \\ l \end{array} \Big|_k = ((M(\beta)C \odot \mathbb{1}_2)\mathcal{R}(M(\alpha) \odot C^{-1}))_{kl}^{ji}.$$

On the other hand, we prove  $V = \tau\mathcal{R}^{-1}(M(\alpha) \odot M(\beta))$ :

$$V_{kl}^{ij} = \begin{array}{c} j \\ \beta \\ i \\ \alpha \\ l \end{array} \Big|_k = \begin{array}{c} j \\ \beta \\ i \\ \alpha \\ l \end{array} \Big|_k = (\mathcal{R}^{-1}(M(\alpha) \odot M(\beta)))_{kl}^{ji}.$$

So we get the equality  $\mathcal{R}^{-1}(M(\alpha) \odot M(\beta)) = (M(\beta)C \odot \mathbb{1}_2)\mathcal{R}(M(\alpha) \odot C^{-1})$  (which equals  $\tau V$ ) and (17) follows.

In case (viii), on the one hand, we first prove  $V = \tau(C \odot M(\alpha))\mathcal{R}^{-1}(\mathbb{1}_2 \odot C^{-1}M(\beta))$ :

$$V_{kl}^{ij} = \begin{array}{c} i \\ \alpha \\ j \\ \beta \\ k \end{array} \Big|_l = \begin{array}{c} i \\ \alpha \\ j \\ \beta \\ k \end{array} \Big|_l = ((C \odot M(\alpha))\mathcal{R}^{-1}(\mathbb{1}_2 \odot C^{-1}M(\beta)))_{kl}^{ji}.$$

On the other hand, we prove  $V = \tau(C \odot C)\mathcal{R}(\mathbb{1}_2 \odot C^{-1}M(\beta))\mathcal{R}^{-1}(\mathbb{1}_2 \odot C^{-1}M(\alpha))\mathcal{R}$ :

$$V_{kl}^{ij} = \begin{array}{c} i \\ \alpha \\ j \\ \beta \\ k \end{array} \Big|_l = \begin{array}{c} i \\ \alpha \\ j \\ \beta \\ k \end{array} \Big|_l = ((C \odot C)\mathcal{R}(\mathbb{1}_2 \odot C^{-1}M(\beta))\mathcal{R}^{-1}(\mathbb{1}_2 \odot C^{-1}M(\alpha))\mathcal{R})_{kl}^{ji}.$$

Equation (18) follows by equating the two obtained expressions for  $V$ .

In case (x), on the one hand, we first prove  $V = (C \odot M(\alpha))\mathcal{R}^{-1}(\mathbb{1}_2 \odot C^{-1}M(\beta))\mathcal{R}$ :

$$V_{kl}^{ij} = \begin{array}{c} i \\ \alpha \\ j \\ \beta \\ k \end{array} \Big|_l = \begin{array}{c} i \\ \alpha \\ j \\ \beta \\ k \end{array} \Big|_l = ((C \odot M(\alpha))\mathcal{R}^{-1}(\mathbb{1}_2 \odot C^{-1}M(\beta))\mathcal{R})_{kl}^{ij}.$$

On the other hand, we prove  $V = (C \odot C)\mathcal{R}(\mathbb{1}_2 \odot C^{-1}M(\beta))\mathcal{R}^{-1}(\mathbb{1}_2 \odot C^{-1}M(\alpha))$ :

$$\begin{aligned}
 V_{kl}^{ij} &= \alpha \begin{array}{c} \text{---} i \\ \text{---} j \\ \text{---} k \\ \text{---} l \end{array} \beta = \alpha \begin{array}{c} \text{---} i \\ \text{---} j \\ \text{---} k \\ \text{---} l \end{array} \beta \\
 &= ((C \odot C)\mathcal{R}(\mathbb{1}_2 \odot C^{-1}M(\beta))\mathcal{R}^{-1}(\mathbb{1}_2 \odot C^{-1}M(\alpha)))_{kl}^{ij}.
 \end{aligned}$$

Therefore, we obtain the following equality that will be used in the proof of Lemma 2.25:

$$\begin{aligned}
 (21) \quad V &= (C \odot M(\alpha))\mathcal{R}^{-1}(\mathbb{1}_2 \odot C^{-1}M(\beta))\mathcal{R} \\
 &= (C \odot C)\mathcal{R}(\mathbb{1}_2 \odot C^{-1}M(\beta))\mathcal{R}^{-1}(\mathbb{1}_2 \odot C^{-1}M(\alpha)).
 \end{aligned}$$

Equation (20) follows.

In (ix) we slightly change strategy. Define the  $4 \times 4$  matrix  $W = (W_{kl}^{ij})_{i,j,k,l \in \{-,+\}}$  by

$$W_{kl}^{ij} := \alpha \begin{array}{c} \text{---} i \\ \text{---} j \\ \text{---} k \\ \text{---} l \end{array} \beta$$

We first prove  $W = (C \odot M(\beta))\mathcal{R}^{-1}(\mathbb{1}_2 \odot C^{-1}M(\alpha))$ :

$$W_{kl}^{ij} = \alpha \begin{array}{c} \text{---} i \\ \text{---} j \\ \text{---} k \\ \text{---} l \end{array} \beta = \alpha \begin{array}{c} \text{---} i \\ \text{---} j \\ \text{---} k \\ \text{---} l \end{array} \beta = ((C \odot M(\beta))\mathcal{R}^{-1}(\mathbb{1}_2 \odot C^{-1}M(\alpha)))_{kl}^{ij}.$$

Next, we prove  $W = (C \odot C)\mathcal{R}^{-1}(\mathbb{1}_2 \odot C^{-1}M(\alpha))\mathcal{R}(\mathbb{1}_2 \odot C^{-1}M(\beta))\mathcal{R}^{-1}$ :

$$\begin{aligned}
 W_{kl}^{ij} &= \alpha \begin{array}{c} \text{---} i \\ \text{---} j \\ \text{---} k \\ \text{---} l \end{array} \beta = \alpha \begin{array}{c} \text{---} i \\ \text{---} j \\ \text{---} k \\ \text{---} l \end{array} \beta \\
 &= ((C \odot C)\mathcal{R}^{-1}(\mathbb{1}_2 \odot C^{-1}M(\alpha))\mathcal{R}(\mathbb{1}_2 \odot C^{-1}M(\beta))\mathcal{R}^{-1})_{kl}^{ij}.
 \end{aligned}$$

Equation (19) follows by equating the two obtained expressions for  $W$ . □

**Lemma 2.25** ( $q$ -determinant relations) *Let  $\alpha$  be an oriented arc. Then*

$$(22) \quad \det_q(N(\alpha)) = 1 \text{ if } \alpha \text{ is of type } a, \text{ and } \det_{q^2}(N(\alpha)) = 1 \text{ otherwise.}$$

**Proof** First suppose that  $\alpha$  is of type  $a$ . Applying the trivial arc and cutting arc relation, we obtain

$$(C^{-1})_{+}^{-} = \begin{array}{c} \uparrow \\ \text{---} \\ \text{---} \\ \text{---} \\ \downarrow \end{array} = (C^{-1})_{-}^{+} \begin{array}{c} \alpha \\ \text{---} \\ \alpha \\ \text{---} \end{array} + (C^{-1})_{+}^{-} \begin{array}{c} \alpha \\ \text{---} \\ \alpha \\ \text{---} \end{array},$$

which is equivalent to the equation  $\alpha_{+++}\alpha_{--} - q^{-1}\alpha_{+-}\alpha_{-+} = 1$  as claimed. Next we suppose that  $\alpha$  is of type  $d$ . Let  $\beta$  be an arc isotopic to and disjoint from  $\alpha$ , placed as in Figure 14(x) (so  $\beta_{ij} = \alpha_{ij}$ ). Consider the matrix  $V = (V_{kl}^{ij})_{i,j,k,l \in \{-,+\}}$ , where  $V_{kl}^{ij} = [\alpha \cup \beta, \sigma_{ijkl}] \in \mathcal{S}_\omega(\Sigma)$  is the class of the simple diagram  $\alpha \cup \beta$  with state  $\sigma_{ijkl}$  sending  $t(\alpha)$ ,  $t(\beta)$ ,  $s(\alpha)$  and  $s(\beta)$  to  $i$ ,  $j$ ,  $k$  and  $l$  respectively, like in the proof of Lemma 2.24, ie

$$V_{kl}^{ij} = \begin{array}{c} \uparrow i \\ \text{---} \\ \alpha \\ \text{---} \\ \beta \\ \text{---} \\ \downarrow k \end{array}$$

Again, using the trivial arc and cutting arc relation, we obtain

$$(23) \quad C_{+}^{-} = \begin{array}{c} \uparrow \\ \text{---} \\ \text{---} \\ \downarrow \end{array} = (C^{-1})_{-}^{+} \begin{array}{c} \alpha \\ \text{---} \\ \beta \\ \text{---} \end{array} + (C^{-1})_{+}^{-} \begin{array}{c} \alpha \\ \text{---} \\ \beta \\ \text{---} \end{array} \\ \iff A^5 V_{+-}^{-+} - A^3 V_{-+}^{-+} = 1.$$

To develop the elements  $V_{kl}^{ij}$  as linear combinations of the  $\alpha_{ab}\alpha_{cd}$  we can either consider the matrix coefficients of the equality  $V = (C \odot M(\alpha))\mathcal{R}^{-1}(\mathbb{1}_2 \odot C^{-1}M(\beta))\mathcal{R}$  proved in the proof of Lemma 2.24, or we can perform the skein computation

$$\begin{aligned} \alpha_{ij}\alpha_{kl} &= \begin{array}{c} \uparrow j \\ \text{---} \\ \alpha \\ \text{---} \\ \beta \\ \text{---} \\ \downarrow k \end{array} = q \begin{array}{c} \uparrow j \\ \text{---} \\ \alpha \\ \text{---} \\ \beta \\ \text{---} \\ \downarrow k \end{array} + q^{-1} \begin{array}{c} \uparrow j \\ \text{---} \\ \alpha \\ \text{---} \\ \beta \\ \text{---} \\ \downarrow k \end{array} + \begin{array}{c} \uparrow j \\ \text{---} \\ \alpha \\ \text{---} \\ \beta \\ \text{---} \\ \downarrow k \end{array} + \begin{array}{c} \uparrow j \\ \text{---} \\ \alpha \\ \text{---} \\ \beta \\ \text{---} \\ \downarrow k \end{array} + \begin{array}{c} \uparrow j \\ \text{---} \\ \alpha \\ \text{---} \\ \beta \\ \text{---} \\ \downarrow k \end{array} \\ &= q C_k^j C_l^i + C_i^j C_k^l + q^{-1} V_{lk}^{ji} + C_l^i (C^{-1})_{+}^{-} V_{+k}^{j-} + C_l^i (C^{-1})_{-}^{+} V_{-k}^{j+}, \end{aligned}$$

from which we deduce the equalities

$$V_{+-}^{-+} = q\alpha_{+-}\alpha_{-+} + A^{-1}, \quad V_{-+}^{+-} = q\alpha_{-+}\alpha_{+-} + A^{-1}, \quad V_{-+}^{-+} = \alpha_{--}\alpha_{++} - A^{-3}.$$

Now, using the skein relation (2), we find

$$V_{+-}^{-+} = qV_{-+}^{+-} + A^{-1} = V_{-+}^{-+},$$

so  $V_{+-}^{-+} = V_{-+}^{-+}$ , which implies that  $\alpha_{+-}\alpha_{-+} = \alpha_{-+}\alpha_{+-}$ .

Next, replacing the elements  $V_{+-}^{-+}$  and  $V_{-+}^{-+}$  in (23) by their expressions in terms of the  $\alpha_{ij}\alpha_{kl}$ , we find

$$(24) \quad \alpha_{--}\alpha_{++} - q^2\alpha_{+-}\alpha_{-+} = A.$$



Using  $\alpha_{+-}\alpha_{-+} = \alpha_{-+}\alpha_{+-}$  we obtain the desired equality:

$$\det_{q^2}(N(\alpha)) = \det_{q^2} \begin{pmatrix} -\omega^{-5}\alpha_{-+} & -\omega^{-5}\alpha_{--} \\ \omega^{-1}\alpha_{++} & \omega^{-1}\alpha_{+-} \end{pmatrix} = -A^3\alpha_{-+}\alpha_{+-} + A^{-1}\alpha_{--}\alpha_{++} = 1.$$

Now, if  $\alpha$  is of type  $e$ , then  $\alpha^{-1}$  is of type  $d$ . A simple computation shows that if  $M = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$  is such that  $ad = da$  then  $\det_{q^2}(M) = \det_{q^2}(C^{-1t}MC)$ , so we deduce the  $q$ -determinant formula for  $\alpha$  of type  $e$  from the fact that it holds for  $\alpha^{-1}$ , from the orientation-reversing formula in Lemma 2.20 and from  $\alpha_{+-}\alpha_{-+} = \alpha_{-+}\alpha_{+-}$ .

Suppose that  $\alpha$  is of type  $c$  and choose  $\mathbb{k} = \mathbb{Z}[\omega^{\pm 1}]$ . Recall from Section 2.1 the reflexion anti-involution  $\theta$ . The image  $\theta(\alpha)$  is of type  $d$ , so applying  $\theta$  to (24),

$$(25) \quad \alpha_{++}\alpha_{--} - q^{-2}\alpha_{-+}\alpha_{+-} = A^{-1}.$$

By Remark 2.6, since (25) holds for  $\mathbb{k} = \mathbb{Z}[\omega^{\pm 1}]$ , it also holds for any other ring. Also using  $\theta$ , we find that  $\alpha_{+-}\alpha_{-+} = \alpha_{-+}\alpha_{+-}$  and the equation  $\det_{q^2}(N(\alpha)) = 1$  follows. Finally, when  $\alpha$  is of type  $b$ , we deduce the  $q$ -determinant relation from the fact that it holds for  $\alpha^{-1}$  (of type  $c$ ), from the orientation-reversing formulas of Lemma 2.20 and from the identity  $\alpha_{+-}\alpha_{-+} = \alpha_{-+}\alpha_{+-}$ .  $\square$

**Definition 2.26** Let  $\mathbb{P} = (\mathbb{G}, \mathbb{RL})$  be a finite presentation of  $\Pi_1(\Sigma_{\mathcal{P}}, \mathbb{V})$ . The set  $\mathcal{A}^{\mathbb{G}}$  generates  $\mathcal{S}_{\omega}(\Sigma)$  by Proposition 2.14, and we have found three families of relations:

- (i) For each  $\alpha \in \mathbb{G}$  we have either the relation  $\det_q(N(\alpha)) = 1$  or  $\det_{q^2}(N(\alpha)) = 1$  by (22) in Lemma 2.25; we call these the *q-determinant relations*.
- (ii) For each  $R \in \mathbb{RL}$ , we have four relations obtained by considering the matrix coefficients in (10) in Lemma 2.23; we call these *trivial loop relations*.
- (iii) For each pair  $(\alpha, \beta)$  of elements in  $\mathbb{G}$ , we have 16 relations obtained by considering the matrix coefficients in one of (11)–(20) of Lemma 2.24 after having possibly replaced  $\alpha$  or  $\beta$  by  $\alpha^{-1}$  or  $\beta^{-1}$ , if necessary, and using the inversion formula (8); we call these *arc exchange relations*.

### 3 Proof of Theorems 1.1 and 1.2

**Definition 3.1** Let  $\mathcal{L}_{\omega}(\mathbb{P})$  be the algebra generated by the elements of  $\mathbb{G}$  modulo the  $q$ -determinant, trivial loops and arc exchange relations, and write  $\Psi: \mathcal{L}_{\omega}(\mathbb{P}) \rightarrow \mathcal{S}_{\omega}(\Sigma)$  the obvious algebra morphism.

By Proposition 2.14,  $\Psi$  is surjective and we need to show that  $\Psi$  is injective to prove Theorem 1.1. We cut the proof of Theorem 1.1 in three steps: In step 1, we show that

it is sufficient to consider the case where  $\mathbb{P}$  has no relations (as in Example 2.13); in this particular case, the finite presentation defining  $\mathcal{L}_\omega(\mathbb{P})$  is inhomogeneous quadratic and we will use the diamond lemma to extract PBW bases of  $\mathcal{L}_\omega(\mathbb{P})$  and to prove it is Koszul. In step 2 we extract the rewritten rules and their leading terms from the  $q$ -determinant and arc exchange relations, and exhibit the associated spanning family  $\underline{\mathcal{B}}^{\mathbb{G}} \subset \mathcal{L}_\omega(\mathbb{P})$ . Finally in step 3, we show that the image by  $\Psi$  of  $\underline{\mathcal{B}}^{\mathbb{G}}$  is a basis; this will prove both the injectivity of  $\Psi$  and the fact that  $\underline{\mathcal{B}}^{\mathbb{G}}$  is a Poincaré–Birkhoff–Witt basis, and conclude the proofs of Theorems 1.1 and 1.2.

### 3.1 Step 1: reduction to the case where $\mathbb{P}$ has no relations

Let  $\Gamma$  be the presenting graph of  $\mathbb{P}$  and consider its fundamental groupoid  $\Pi_1(\Gamma)$ : the objects of  $\Pi_1(\Gamma)$  are the vertices of  $\Gamma$  (ie the set  $\mathbb{V}$ ) and the morphisms are compositions  $\alpha_k^{\varepsilon_k} \cdots \alpha_1^{\varepsilon_1}$  where  $\alpha_i \in \mathbb{G}$ . The inclusion  $\Gamma \subset \Sigma_{\mathcal{P}}$  induces a functor  $F: \Pi_1(\Gamma) \rightarrow \Pi_1(\Sigma_{\mathcal{P}}, \mathbb{V})$ , which is the identity on the objects. The fact that  $\mathbb{G}$  is a set of generators implies that  $F$  is full and  $\mathbb{P}$  has no relations if and only if  $F$  is faithful. Fix  $v_0 \in \mathbb{V}$ . For a relation  $R \in \mathbb{RL}$  of the form  $R = \beta_k \star \cdots \star \beta_1$ , the *basepoint* of  $R$  is  $s(\beta_1) = t(\beta_k)$ . By inspecting the trivial loop relation (10), we see that changing a relation  $R$  by a relation  $\beta \star R \star \beta^{-1}$  does not change the algebra  $\mathcal{L}_\omega(\mathbb{P})$ . Since  $\Sigma_{\mathcal{P}}$  is assumed to be connected, we can suppose that all relations in  $\mathbb{RL}$  have the same basepoint  $v_0$ , so each relation  $R = \beta_k \star \cdots \star \beta_1$  induces an element  $[R] = \beta_k \cdots \beta_1 \in \pi_1(\Gamma, v_0)$ . The functor  $F$  induces a surjective group morphism  $F_{v_0}: \pi_1(\Gamma, v_0) \rightarrow \pi_1(\Sigma_{\mathcal{P}}, v_0)$  and the fact that  $\mathbb{RL}$  is a set of relations implies that  $\{[R] \mid R \in \mathbb{RL}\}$  generates  $\ker(F_{v_0})$ . Since  $\pi_1(\Gamma, v_0)$  is a free group, so is  $\ker(F_{v_0})$ . Let  $R_1, \dots, R_m \in \mathbb{RL}$  be such that  $\{[R_1], \dots, [R_m]\}$  is a minimal set of generators for the free group  $\ker(F_{v_0})$ . For each  $R_i$ , choose an element  $\beta_i \in \mathbb{G}$  such that either  $\beta_i$  or  $\beta_i^{-1}$  appears in the expression of  $R_i$  and such that the set  $\mathbb{G}'$  obtained from  $\mathbb{G}$  by removing the  $\beta_i$  is a generating set. So if  $\Gamma'$  is the presenting graph of  $\mathbb{G}'$ , the morphism  $F'_{v_0}: \pi_1(\Gamma', v_0) \rightarrow \pi_1(\Sigma_{\mathcal{P}}, v_0)$  is injective, thus the functor  $F': \Pi_1(\Gamma') \rightarrow \Pi_1(\Sigma_{\mathcal{P}}, \mathbb{V})$  is faithful and  $\mathbb{P}' := (\mathbb{G}', \emptyset)$  is a finite presentation of  $\Pi_1(\Sigma_{\mathcal{P}}, \mathbb{V})$  with no relations.

The inclusion  $\mathbb{G}' \subset \mathbb{G}$  induces an algebra morphism  $\tilde{\varphi}: \mathcal{T}[\mathbb{G}'] \hookrightarrow \mathcal{T}[\mathbb{G}]$  on the free tensor algebras generated by  $\mathbb{G}'$  and  $\mathbb{G}$ , respectively, and  $\tilde{\varphi}$  sends  $q$ -determinant and arc exchange relations to  $q$ -determinant and arc exchange relations, so it induces an algebra morphism

$$\varphi: \mathcal{L}_\omega(\mathbb{P}') \rightarrow \mathcal{L}_\omega(\mathbb{P}).$$

**Lemma 3.2** *The morphism  $\varphi$  is an isomorphism.*

**Proof** To prove the surjectivity we need to show that, for each removed path  $\beta_i \in \mathbb{G} \setminus \mathbb{G}'$ , the stated arcs  $(\beta_i)_{\varepsilon\varepsilon'}$  can be expressed as a polynomial in the stated arcs  $(\alpha^{\pm 1})_{\mu\mu'}$  for  $\alpha \in \mathbb{G}'$ . This follows from the trivial loop relation (10) associated to the relation  $R_i \in \mathbb{R}\mathbb{L}$  containing  $\beta_i^{\pm 1}$ . Injectivity of  $\varphi$  is a straightforward consequence of the definition.  $\square$

### 3.2 Step 2: Poincaré–Birkhoff–Witt bases and Koszulness

**Convention 3.3** In the rest of the section, we suppose that  $\mathbb{P} = (\mathbb{G}, \emptyset)$  is a presentation with no relations and that every arc in  $\mathbb{G}$  is either of type  $a$ ,  $c$  or  $d$ .

Note that the convention on the type of the generators is not restrictive but purely conventional since we can always replace a generator  $\alpha$  by  $\alpha^{-1}$  without changing the set  $\mathcal{A}^{\mathbb{G}}$  of generators of  $\mathcal{S}_{\omega}(\Sigma)$ .

Since  $\mathbb{P}$  has no relations, the defining presentation of  $\mathcal{L}_{\omega}(\mathbb{P})$  contains only  $q$ -determinant and arc exchange relations. All these relations are quadratic (inhomogeneous) in the generators  $\mathcal{A}^{\mathbb{G}}$  and we want to apply the diamond lemma to prove that  $\mathcal{L}_{\omega}(\mathbb{P})$  is Koszul.

**Reminder of the diamond lemma for PBW bases** Following the exposition in Section 4 of [42], we briefly recall the statement of the diamond lemma for PBW bases:

Let  $V$  be a free finite rank  $\mathbb{k}$ -module, denote by  $T(V) := \bigoplus_{n \geq 0} V^{\otimes n}$  the tensor algebra and fix  $R \subset V^{\otimes 2}$  a finite subset. The quotient algebra  $\mathcal{A} := T(V)/(R)$  is called a *quadratic algebra*. Let  $\{v_i\}_{i \in I}$  be a totally ordered basis of  $V$  and write  $I = \{1, \dots, k\}$  so that  $v_i < v_{i+1}$ . Then the set  $J := \bigsqcup_{n \geq 0} I^n$  (where  $I^0 = \{0\}$ ) is totally ordered by the lexicographic order and the set of elements  $v_{\mathbf{i}} = v_{i_1} \cdots v_{i_n}$ , for  $\mathbf{i} = (i_1, \dots, i_k)$ , forms a basis of  $T(V)$ . We suppose that the elements  $r \in R$  (named *relators*) have the form

$$r = v_i v_j - \sum_{(k,l) < (i,j)} \lambda_{kl}^{ij} v_k v_l.$$

The term  $v_i v_j$  is called the *leading term* of  $r$ . We assume that two distinct relators have distinct leading terms. Define the family

$$(26) \quad \mathcal{B} := \{v_{i_1} \cdots v_{i_n} \mid v_{i_k} v_{i_{k+1}} \text{ is not a leading term for all } 1 \leq k \leq n-1\},$$

and denote by  $\mathcal{B}^{(3)} \subset \mathcal{B}$  the subset of elements of length 3 (of the form  $v_{i_1} v_{i_2} v_{i_3}$ ). Obviously the set  $\mathcal{B}$  spans  $\mathcal{A}$ .

**Theorem 3.4** (diamond lemma for PBW bases, Bergman [10]; see also Loday and Vallette [42, Theorem 4.3.10]) *If  $\mathcal{B}^{(3)}$  is free, then  $\mathcal{B}$  is a (Poincaré–Birkhoff–Witt) basis and  $\mathcal{A}$  is Koszul.*

The arc exchange relations defining  $\mathcal{L}_\omega(\mathbb{P})$  are quadratic, however the  $q$ -determinant relations are not (because of the 1 in  $\det_q(N(\alpha)) = 1$ ), so  $\mathcal{L}_\omega(\mathbb{P})$  is not quadratic but rather inhomogeneous quadratic. An *inhomogeneous quadratic algebra* is an algebra of the form  $\mathcal{A} := T(V)/(R)$ , where  $R \subset V^{\otimes 2} \oplus V \oplus \mathbb{k} \subset T(V)$ . We further assume

$$(ql_1) \quad R \cap V = \{0\}$$

and

$$(ql_2) \quad (R \otimes V + V \otimes R) \cap V^{\otimes 2} \subset R \cap V^{\otimes 2}.$$

The hypothesis  $(ql_2)$  says that one cannot create new relations by adding an element to  $R$ , so it is not restrictive. Like before, we fix an ordered basis  $\{v_i\}_{i \in I}$  of  $V$  and suppose that the relators of  $R$  have the form

$$(27) \quad r = v_i v_j - \sum_{(k,l) < (i,j)} \lambda_{kl}^{ij} v_k v_l - c_{i,j},$$

where  $c_{i,j}$  are some scalars and we suppose that two distinct relators have distinct leading terms. The associated quadratic algebra  $q\mathcal{A}$  is the algebra with same generators  $v_i$  but where the relators have been changed by replacing the scalars  $c_{i,j}$  by 0. Let  $q\mathcal{B} \subset q\mathcal{A}$  and  $\mathcal{B} \subset \mathcal{A}$  be the two generating families defined by (26).

**Theorem 3.5** [42, Theorem 4.3.18] *Suppose that  $q\mathcal{B}^{(3)} \subset q\mathcal{A}$  is free. Then both  $q\mathcal{B}$  and  $\mathcal{B}$  are (PBW) bases of  $q\mathcal{A}$  and  $\mathcal{A}$ , respectively, and both  $q\mathcal{A}$  and  $\mathcal{A}$  are Koszul.*

There exists a linear surjective morphism  $\varphi: q\mathcal{A} \rightarrow \mathcal{A}$  sending the generating family  $q\mathcal{B}$  to  $\mathcal{B}$ ; see [42, Section 4.2.9]. So, if  $\mathcal{B}$  is a basis of  $\mathcal{A}$ , then  $q\mathcal{B}$  is free, therefore Theorem 3.5 implies that  $\mathcal{A}$  is Koszul. Therefore:

**Theorem 3.6** *If  $\mathcal{B}$  is a basis of  $\mathcal{A}$ , then  $\mathcal{A}$  is Koszul.*

**The relators of the stated skein presentations and PBW bases** For  $\alpha \in \mathbb{G}$ , define  $\mathcal{B}(\alpha)$  as

$$\{(\alpha_{++})^a (\alpha_{+-})^b (\alpha_{--})^c \mid a, b, c \geq 0\} \cup \{(\alpha_{++})^a (\alpha_{-+})^b (\alpha_{--})^c \mid a, b, c \geq 0\} \subset \mathcal{L}_\omega(\mathbb{P}).$$

Fix a total order  $<$  on the set  $\mathbb{G}$  of generators and index its elements as  $\mathbb{G} = \{\alpha_1, \dots, \alpha_n\}$ , where  $\alpha_i < \alpha_{i+1}$ . Let

$$\underline{\mathcal{B}}^\mathbb{G} := \{m_1 m_2 \cdots m_n \mid m_i \in \mathcal{B}(\alpha_i)\} \subset \mathcal{L}_\omega(\mathbb{P}).$$

We want to apply Theorem 3.6 to prove that  $\mathcal{L}_\omega(\mathbb{P})$  is Koszul. By definition,  $\mathcal{L}_\omega(\mathbb{P})$  is an inhomogeneous quadratic algebra with generators  $\mathcal{A}^\mathbb{G} = \{\alpha_{ij} \mid \alpha \in \mathbb{G} \text{ and } i, j = \pm\}$  and whose relations are the arc exchange and  $q$ -determinant relations.

We first define a total order  $<$  on  $\mathcal{A}^{\mathbb{G}}$  by imposing that  $\alpha_{ab} < \beta_{cd}$  if  $\alpha < \beta$  and that  $\alpha_{++} < \alpha_{+-} < \alpha_{-+} < \alpha_{--}$ .

The goal of this subsection is to rewrite the  $q$ -determinant and arc exchange relations so that they define a set of relators of the form (27) whose leading terms are pairwise distinct, satisfying  $(ql_1)$  and  $(ql_2)$  and such that the set of leading terms is

$$(28) \quad \text{LT} := \{\alpha_{ab}\beta_{cd} \mid \text{either } \alpha > \beta, \text{ or } \alpha = \beta \text{ and either } a < c \text{ or } b < d\}.$$

The set  $\underline{\mathcal{B}}^{\mathbb{G}}$  is the generating set defined by (26) with this set of leading terms (ie  $\underline{\mathcal{B}}^{\mathbb{G}}$  is the set of elements  $v_1 \cdots v_n$  where  $v_i \in \mathcal{A}^{\mathbb{G}}$  and  $v_i v_{i+1}$  is not in LT). At this stage, it will become clear that  $\underline{\mathcal{B}}^{\mathbb{G}}$  spans  $\mathcal{L}_{\omega}(\mathbb{P})$ . Once we perform this task, we will prove in step 3 that  $\underline{\mathcal{B}}^{\mathbb{G}}$  is free by showing that its image through  $\Psi: \mathcal{L}_{\omega}(\mathbb{P}) \rightarrow \mathcal{S}_{\omega}(\Sigma)$  is a basis of  $\mathcal{S}_{\omega}(\Sigma)$ . This will imply that  $\Psi$  is an isomorphism (proving Theorem 1.1) and Theorem 3.6 will imply that  $\mathcal{L}_{\omega}(\mathbb{P})$  is Koszul (proving Theorem 1.2).

Consider two distinct generators  $\alpha, \beta \in \mathbb{G}$  such that  $\alpha > \beta$ . For each  $a, b, c, d \in \{\pm\}$ , we have an arc exchange relation of the form

$$\alpha_{ab}\beta_{cd} = \sum_{ijkl=\pm} c_{a,b,c,d}^{i,j,k,l} \beta_{ij}\alpha_{kl},$$

where  $c_{a,b,c,d}^{i,j,k,l}$  are some scalars. We associate the relator

$$r = \alpha_{ab}\beta_{cd} - \sum_{ijkl=\pm} c_{a,b,c,d}^{i,j,k,l} \beta_{ij}\alpha_{kl},$$

whose leading term is  $\alpha_{ab}\beta_{cd}$  (because  $\alpha > \beta$  implies that  $\alpha_{ab}\beta_{cd} > \beta_{ij}\alpha_{kl}$ ) and denote by  $R_{\alpha,\beta}$  the set (of cardinality 16) of such relators.

Now suppose that  $\alpha \in \mathbb{G}$  is of type  $a$ . The set of relations between the generators  $\alpha_{ij}$  is given by

$$M(\alpha) \odot M(\alpha) = \mathcal{R}^{-1}(M(\alpha) \odot M(\alpha))\mathcal{R} \quad \text{and} \quad \det_q(M(\alpha)) = 1.$$

Note that in this case, the subalgebra of  $\mathcal{L}_{\omega}(\mathbb{P})$  generated by the  $\alpha_{ij}$  is isomorphic to  $\mathcal{O}_q[\text{SL}_2] \cong \mathcal{S}_{\omega}(\mathbb{B})$ . We rewrite those relations:

$$\begin{aligned} \text{(Ra)} \quad \alpha_{+-}\alpha_{++} &= q\alpha_{++}\alpha_{+-}, & \alpha_{-+}\alpha_{++} &= q\alpha_{++}\alpha_{-+}, \\ \alpha_{--}\alpha_{+-} &= q\alpha_{+-}\alpha_{--}, & \alpha_{--}\alpha_{-+} &= q\alpha_{-+}\alpha_{--}, \\ \alpha_{+-}\alpha_{-+} &= q\alpha_{++}\alpha_{--} - q, & \alpha_{-+}\alpha_{+-} &= q\alpha_{++}\alpha_{--} - q, \\ \alpha_{--}\alpha_{++} &= q^2\alpha_{++}\alpha_{--} + 1 - q^2. \end{aligned}$$

The associated set of relators  $R_\alpha$  is defined by assigning, to each of the seven equalities of the form  $x = y$  in the system (Ra), the relator  $r := x - y$  with leading term  $x$ . Note that the set of leading terms of the elements of  $R_\alpha$  is the set of elements  $\alpha_{ab}\alpha_{cd}$  such that either  $a < c$  or  $b < d$ .

Now suppose that  $\alpha \in \mathbb{G}$  is of type  $d$ . The set of relations between the generators  $\alpha_{ij}$  are given by

$$(\mathbb{1}_2 \odot N(\alpha))\mathcal{R}^{-1}(\mathbb{1}_2 \odot N(\alpha))\mathcal{R} = \mathcal{R}(\mathbb{1}_2 \odot N(\alpha))\mathcal{R}^{-1}(\mathbb{1}_2 \odot N(\alpha)), \quad \det_{q^2}(N(\alpha)) = 1,$$

where  $N(\alpha) = C^{-1}M(\alpha)$ . These relations generate the same ideal as the set of relations

$$\begin{aligned} \text{(Rd)} \quad & \alpha_{-+}\alpha_{++} = \alpha_{++}\alpha_{-+} + (q - q^{-1})q^2\alpha_{+-}\alpha_{--}, & \alpha_{+-}\alpha_{++} &= q^2\alpha_{++}\alpha_{+-}, \\ & \alpha_{--}\alpha_{-+} = \alpha_{-+}\alpha_{--} + (q - q^{-1})q^2\alpha_{+-}\alpha_{--}, & \alpha_{--}\alpha_{+-} &= q^2\alpha_{+-}\alpha_{--}, \\ & \alpha_{+-}\alpha_{-+} = \alpha_{++}\alpha_{--} - (q - q^{-1})^2\alpha_{+-}^2 - A, \\ & \alpha_{-+}\alpha_{+-} = \alpha_{++}\alpha_{--} - (q - q^{-1})^2\alpha_{+-}^2 - A, \\ & \alpha_{--}\alpha_{++} = q^2\alpha_{++}\alpha_{--} - q^2(q - q^{-1})^2\alpha_{+-}^2 + A(1 - q^2). \end{aligned}$$

As before, we denote by  $R_\alpha$  the set of relators obtained from system (Rd) by assigning, to each of the seven equalities of the form  $x = y$  in the system (Rd), the relator  $r := x - y$  with leading term  $x$ . Again, the set of leading terms of the elements of  $R_\alpha$  is the set of elements  $\alpha_{ab}\alpha_{cd}$  such that either  $a < c$  or  $b < d$ .

For  $\alpha \in \mathbb{G}$  of type  $c$ , the set of relations between the elements  $\alpha_{ij}$  can be obtained from the system (Rd) using the reflection anti-involution. Rearranging the terms, we get the system of relations

$$\begin{aligned} \text{(Rc)} \quad & \alpha_{-+}\alpha_{++} = \alpha_{++}\alpha_{-+} + (q - q^{-1})\alpha_{+-}\alpha_{--}, & \alpha_{+-}\alpha_{++} &= q^2\alpha_{++}\alpha_{+-}, \\ & \alpha_{--}\alpha_{-+} = \alpha_{-+}\alpha_{--} + (q - q^{-1})\alpha_{+-}\alpha_{--}, & \alpha_{--}\alpha_{+-} &= q^2\alpha_{+-}\alpha_{--}, \\ & \alpha_{+-}\alpha_{-+} = q^2\alpha_{++}\alpha_{--} - A^3, & \alpha_{-+}\alpha_{+-} &= q^2\alpha_{++}\alpha_{--} - A^3, \\ & \alpha_{--}\alpha_{++} = q^2\alpha_{++}\alpha_{--} + (q - q^{-1})^2\alpha_{+-}^2 + A^{-1}(1 - q^2). \end{aligned}$$

Like previously, we denote by  $R_\alpha$  the associated set of relators and note that the set of leading terms is the set of elements  $\alpha_{ab}\alpha_{cd}$  such that either  $a < c$  or  $b < d$ .

Let  $V$  be the free  $\mathbb{k}$ -module with basis  $\mathcal{A}^\mathbb{G}$  and  $R \subset \mathbb{k} \oplus V^{\otimes 2} \subset T(V)$  be the union of the sets of relators  $R_{\alpha,\beta}$  and  $R_\alpha$ , where  $\alpha, \beta \in \mathbb{G}$  and  $\alpha > \beta$ . Then  $\mathcal{L}_\omega(\mathbb{P}) = T(V)/(R)$ , the leading terms of  $R$  are pairwise distinct and they form the set LT of (28), and the hypotheses  $(ql_1)$  and  $(ql_2)$  are obviously satisfied. Therefore, if we prove that  $\underline{\mathcal{B}}^\mathbb{G}$  is a basis of  $\mathcal{L}_\omega(\mathbb{P})$  then Theorem 3.6 would imply that  $\mathcal{L}_\omega(\mathbb{P})$  is Koszul.

### 3.3 Step 3: injectivity of $\Psi$

Denote by  $\mathcal{B}^{\mathbb{G}} \subset \mathcal{S}_{\omega}(\mathbb{P})$  the image of  $\underline{\mathcal{B}}^{\mathbb{G}}$  under  $\Psi: \mathcal{L}_{\omega}(\mathbb{P}) \rightarrow \mathcal{S}_{\omega}(\Sigma)$ .

**Theorem 3.7** *The set  $\mathcal{B}^{\mathbb{G}}$  is a basis of  $\mathcal{S}_{\omega}(\Sigma)$ .*

**Corollary 3.8** (i) *The morphism  $\Psi: \mathcal{L}_{\omega}(\mathbb{P}) \rightarrow \mathcal{S}_{\omega}(\Sigma)$  is an isomorphism.*

(ii) *The family  $\mathcal{B}^{\mathbb{G}}$  is a PBW basis and  $\mathcal{S}_{\omega}(\Sigma)$  is Koszul.*

The fact that  $\mathcal{B}^{\mathbb{G}}$  linearly spans  $\mathcal{S}_{\omega}(\Sigma)$  follows from the surjectivity of  $\Psi$  (so follows from Proposition 2.14), however we will reprove this fact. The proof of Theorem 3.7 is divided into two steps. First we introduce another family  $\mathcal{B}_{+}^{\mathbb{G}} \subset \mathcal{S}_{\omega}(\Sigma)$  and prove that  $\mathcal{B}_{+}^{\mathbb{G}}$  is free by relating it to the basis  $\mathcal{B}$ . Next we use a filtration of  $\mathcal{S}_{\omega}(\Sigma)$  to deduce that  $\mathcal{B}^{\mathbb{G}}$  is free from the fact that  $\mathcal{B}_{+}^{\mathbb{G}}$  is free.

For  $\alpha \in \mathbb{G}$  and  $n \geq 0$ , we denote by  $\alpha^{(n)}$  the simple diagram made of  $n$  pairwise nonintersecting copies of  $\alpha$ . For  $\mathbf{n} \in \mathbb{N}^{\mathbb{G}}$ , we denote by  $D(\mathbf{n})$  the simple diagram  $\bigsqcup_{\alpha \in \mathbb{G}} \alpha^{(n(\alpha))}$ . Denote by  $v$  and  $w$  the two endpoints of  $\alpha$ , and by  $a$  and  $b$  the (not necessarily distinct) boundary arcs containing  $v$  and  $w$ , respectively. Write  $v_1, \dots, v_n$  and  $w_1, \dots, w_n$  the endpoints of  $\alpha^{(n)}$  so that  $v_i <_a v_{i+1}$  and  $w_i <_b w_{i+1}$  (so  $v_i$  and  $w_i$  are not necessarily the boundary points of the same component of  $\alpha^{(n)}$ ). A state  $s \in \text{St}(D(\mathbf{n}))$  is *positive* if for all  $\alpha \in \mathbb{G}$  and for all  $i \leq j$  one has  $s(v_i) \leq s(v_j)$  and  $s(w_i) \leq s(w_j)$ ; we let  $\text{St}^{+}(D(\mathbf{n}))$  denote the set of positive states.

**Definition 3.9** We denote by  $\mathcal{B}_{+}^{\mathbb{G}} \subset \mathcal{S}_{\omega}(\Sigma)$  the set of classes  $[D(\mathbf{n}), s]$  for  $\mathbf{n} \in \mathbb{N}^{\mathbb{G}}$  and  $s \in \text{St}^{+}(D(\mathbf{n}))$ .

**Proposition 3.10** *The family  $\mathcal{B}_{+}^{\mathbb{G}}$  is a basis of  $\mathcal{S}_{\omega}(\Sigma)$ .*

The fact that  $\mathcal{B}_{+}^{\mathbb{G}}$  is free will follow from this elementary lemma, which basically says that an upper triangular matrix with invertible diagonal elements is invertible:

**Lemma 3.11** *Let  $V$  be a free  $\mathbb{k}$ -module,  $\mathcal{B}$  a basis of  $V$  equipped with a partial order  $\leq$ , and  $\mathcal{B}' \subset V$  a family such that there exist two maps  $m: \mathcal{B}' \rightarrow \mathcal{B}$  and  $c: \mathcal{B}' \rightarrow \mathbb{k}^{\times}$  such that*

- (i)  *$m$  is injective, and*
- (ii) *every element  $b' \in \mathcal{B}'$  decomposes as*

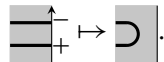
$$b' = c(b')m(b') + \sum_{b > m(b')} \alpha_{b,b'} b.$$

*Then  $\mathcal{B}'$  is free.*

**Proof** Consider a vanishing linear combination  $\sum_{b' \in \mathcal{B}'} x_{b'} b' = 0$ , where  $x_{b'} \in \mathbb{K}$ . Set  $S := \{m(b') \mid x_{b'} \neq 0\}$ . For contradiction, suppose that  $S \neq \emptyset$  and choose  $b_0$  a minimum for  $S$ . Let  $b'_0 \in \mathcal{B}'$  be the unique element such that  $m(b'_0) = b_0$ . Then the equality  $\sum_{b' \in \mathcal{B}'} x_{b'} b' = 0$  together with the decomposition hypothesis imply that  $c(b'_0)x_{b'_0} = 0$ . Since  $c(b'_0) \in \mathbb{K}^\times$  is invertible by hypothesis,  $x_{b'_0} = 0$ , so we have a contradiction.  $\square$

**Notation 3.12** (i) Let  $(D, s)$  be a stated diagram and  $a$  a boundary arc. We denote by  $d_a([D, s]) \in \mathbb{N}$  the number of pairs  $(v, w)$  in  $\partial_a D$  such that  $v <_a w$  and  $(s(v), s(w)) = (+, -)$ ; recall that the orientation of  $\Sigma_{\mathcal{P}}$  induces an orientation of  $a$  which in turn induces the order  $<_a$ . We also write  $d([D, s]) = \sum_a d_a([D, s])$ . Note that  $[D, s] \in \mathcal{B}$  if and only if  $d([D, s]) = 0$ .

(ii) Let  $\mathcal{D}$  denote the set of stated diagrams  $(D, s)$  with  $D$  simple, so both  $\mathcal{B}$  and  $\mathcal{B}_+^{\mathbb{G}}$  are subsets of  $\mathcal{D}$ . Define a binary operation  $\mapsto_o$  on  $\mathcal{D}$  as follows. If  $(D, s)$  contains a pair  $(v, w)$  in  $\partial_a D$  of consecutive points for the height ordering (there is no  $z \in \partial_a D$  such that  $v <_a z <_a w$ ) with  $v <_a w$  and such that  $(s(v), s(w)) = (+, -)$ , let  $(D', s')$  be the stated diagram obtained by joining  $v$  and  $w$  to a single point and then pushing it to the interior of  $\Sigma$ , that is  $(D', s')$  is obtained from  $(D, s)$  by the local move



$(D'', s'')$  be obtained from  $(D', s')$  by removing the possible trivial component if any. In this case, we write  $(D, s) \mapsto_o (D'', s'')$ . Since  $d([D'', s'']) < d([D, s])$ , the relation  $\mapsto_o$  is terminal, with  $\mathcal{B}$  as the set of terminal objects. Define a partial order  $\leq_o$  by setting  $(D, s) >_o (D', s')$  if there exists a sequence  $(D, s) \mapsto_o (D_2, s_2) \mapsto_o \dots \mapsto_o (D', s')$ . Clearly,  $\leq_o$  is filtrant, ie if  $(D_1, s_2) \leq_o (D, s)$  and  $(D_2, s_2) \leq_o (D, s)$  there exists  $(D_0, s_0)$  such that  $(D_0, s_0) \leq_o (D_i, s_i)$  for  $i = 1, 2$ .

(iii) Let  $\alpha$  be an oriented arc. Since  $\mathbb{G}$  is a generating set, the associated path in  $\Pi_1(\Sigma_{\mathcal{P}}, \mathbb{V})$  decomposes as  $\alpha = \beta_1^{\epsilon_1} \dots \beta_k^{\epsilon_k}$  and, since  $(\mathbb{G}, \emptyset)$  is a presentation with no relation, this decomposition is unique. We denote by  $\text{lw}(\alpha) := k$  its length. For  $(D, s) \in \mathcal{D}$ , where  $D = \alpha_1 \cup \dots \cup \alpha_n$  with  $\alpha_i$  connected, we set

$$l(D, s) := \sum_{i=1}^n (\text{lw}(\alpha_i) - 1).$$

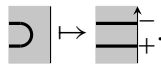
Note that  $\mathcal{B}_+^{\mathbb{G}}$  is the subset of elements  $(D, s) \in \mathcal{D}$  such that  $l(D, s) = 0$ .





Figure 15: An element  $b \in \mathcal{B}_+^{\mathbb{G}}$  (left) and its associated element  $m(b) \in \mathcal{B}$  (right). Here  $\mathbb{G} = \{\beta_1, \beta_2, \beta_3, \beta_4\}$  is the set of generators of Figure 6.

(iv) We define a binary operation  $\mapsto_{\mathbb{B}}$  on  $\mathcal{D}$  as follows. Let  $(D, s) \in \mathcal{D}$  and  $\alpha$  a connected component of  $D$  with  $lw(\alpha) > 1$ . Choose a decomposition  $\alpha = \alpha_1\alpha_2$  where  $lw(\alpha_i) < lw(\alpha)$ , set  $D' := (D \setminus \alpha) \cup \alpha_1 \cup \alpha_2$  and fix the height orders and the state  $s'$  such that  $(D', s')$  is obtained from  $(D, s)$  by the local move



In this case, we write  $(D, s) \mapsto_{\mathbb{G}} (D', s')$ . Since  $l(D', s') < l(D, s)$ , the relation  $\mapsto_{\mathbb{G}}$  is terminal with  $\mathcal{B}_+^{\mathbb{G}}$  as the set of terminal objects. Define a partial order  $\leq_{\mathbb{G}}$  on  $\mathcal{D}$  by setting  $(D, s) >_{\mathbb{G}} (D', s')$  if there exists a sequence  $(D, s) \mapsto_{\mathbb{G}} (D_2, s_2) \mapsto_{\mathbb{G}} \dots \mapsto_{\mathbb{G}} (D', s')$ . It follows from the unicity of the decomposition of a path in  $\mathbb{G}$  (so from the fact that  $(\mathbb{G}, \emptyset)$  is a presentation with no relation) that  $\leq_{\mathbb{G}}$  is filtrant.

(v) Let  $m: \mathcal{B}_+^{\mathbb{G}} \rightarrow \mathcal{B}$  be the map sending a class  $[D, s]$  to the class of the unique minimum for  $\leq_0$  of the successors of  $(D, s)$  (the existence and unicity are guaranteed by the fact that  $\leq_0$  is terminal and filtrant). Similarly, let  $m': \mathcal{B} \rightarrow \mathcal{B}_+^{\mathbb{G}}$  be the map sending a class  $[D, s]$  to the class of the unique minimum for  $\leq_{\mathbb{G}}$  of the set of successors of  $(D, s)$ ; see Figure 15 for an example.

**Proof of Proposition 3.10** We will apply Lemma 3.11 to the map  $m: \mathcal{B} \rightarrow \mathcal{B}_+^{\mathbb{G}}$ , where we equip  $\mathcal{B}$  with the partial order  $<$  where  $[D, s] < [D', s']$  if  $|\partial D| < |\partial D'|$ .

**The map  $m$  is injective** By definition, if  $(D, s) \mapsto_{\mathbb{G}} (D', s')$  then  $(D', s') \mapsto_0 (D, s)$  (the converse is false in general). Therefore, for  $[D, s] \in \mathcal{B}$ , given a sequence

$$(D, s) \mapsto_{\mathbb{G}} (D_2, s_2) \mapsto_{\mathbb{G}} \dots \mapsto_{\mathbb{G}} (D_n, s_n) \mapsto_{\mathbb{G}} m(D, s)$$

one has a sequence

$$m(D, s) \mapsto_0 (D_n, s_n) \mapsto_0 \dots \mapsto_0 (D_2, s_2) \mapsto_0 D.$$

This implies that  $m'(m(D)) = D$  so  $m' \circ m = \text{id}$  and  $m$  is injective.

$\mathcal{B}_+^{\mathbb{G}}$  is upper triangular Suppose that  $(D, s) \mapsto_o (D', s')$ . The skein relation

$$\begin{array}{c} \uparrow \\ \hline \hline \\ + \end{array} = q \begin{array}{c} \uparrow \\ \hline \hline \\ - \end{array} + \omega \begin{array}{c} \cap \\ \hline \hline \end{array}$$

shows that  $[D, s] = \omega[D', s'] + q[D'', s'']$ , where  $|\partial D'| < |\partial D''|$ . So for  $[D, s] \in \mathcal{B}_+^{\mathbb{G}}$  with  $m([D, s]) = [D_0, s_0]$  and given  $(D, s) \mapsto_o (D_2, s_2) \mapsto_o \dots \mapsto_o (D_n, s_n) \mapsto_o (D_0, s_0)$ , we have

$$[D, s] = \omega^n m(D, s) + \text{higher terms,}$$

where “higher terms” is a linear combination of elements  $(D', s')$  with  $|\partial D'| > |\partial D_0|$ . Since  $\mathcal{B}$  is free, Lemma 3.11 implies that  $\mathcal{B}_+^{\mathbb{G}}$  is free. To prove that it spans  $\mathcal{S}_\omega(\Sigma)$  we note that if  $(D, s) \mapsto_{\mathbb{G}} (D', s')$ , the same skein relation

$$\begin{array}{c} \uparrow \\ \hline \hline \\ + \end{array} = q \begin{array}{c} \uparrow \\ \hline \hline \\ - \end{array} + \omega \begin{array}{c} \cap \\ \hline \hline \end{array}$$

implies that

$$[D, s] = \omega^{-1}[D', s'] - \omega^{-5}[D'', s'']$$

for another element  $(D'', s'') \in \mathcal{D}$  such that  $l(D', s') < l(D, s)$  and  $l(D'', s'') < l(D, s)$ . We then prove that any element of  $\mathcal{B}$  is a linear combination of elements of  $\mathcal{B}_+^{\mathbb{G}}$  by induction on  $l(D, s)$ . □

We now want to deduce that  $\mathcal{B}^{\mathbb{G}}$  is a basis from the fact that  $\mathcal{B}_+^{\mathbb{G}}$  is a basis. The argument is based on the use of an algebra filtration of  $\mathcal{S}_\omega(\Sigma)$  that we now introduce:

**Definition 3.13** For  $\mathbf{n} \in \mathbb{N}^{\mathbb{G}}$ , we let  $|\mathbf{n}| := \sum_{\alpha \in \mathbb{G}} n(\alpha)$ . For a class  $[D(\mathbf{n}), s]$ , we set  $\|[D(\mathbf{n}), s]\| := (|\mathbf{n}|, -d([D(\mathbf{n}), s])) \in \mathbb{N} \times \mathbb{Z}$ . Denote by  $<$  the lexicographic order on  $\mathbb{N} \times \mathbb{Z}$ , ie  $(k_1, k_2) < (k'_1, k'_2)$  if either  $k_1 < k'_1$ , or  $k_1 = k'_1$  and  $k_2 < k'_2$ . Finally, to  $\mathbb{k} = (k_1, k_2) \in \mathbb{N} \times \mathbb{Z}$  we associate the submodule

$$\mathcal{F}_{\mathbb{k}} := \text{Span}([D(\mathbf{n}), s] : \|[D(\mathbf{n}), s]\| \leq \mathbb{k}).$$

In order to prove that  $\{\mathcal{F}_{\mathbb{k}}\}$  forms an algebra filtration, the following elementary observation will be quite useful:

**Lemma 3.14** Let  $T$  and  $T'$  be two tangles in  $\Sigma_{\mathcal{P}} \times (0, 1)$  which are isotopic through an isotopy that does not preserves the height orders. Let  $s \in \text{St}(T)$  and  $s' \in \text{St}(T')$  be two states such that for a boundary arc  $a$ , if  $\partial_a T = \{v_1, \dots, v_n\}$  and  $\partial_a T' = \{w_1, \dots, w_n\}$

are ordered so that  $h(v_i) < h(v_{i+1})$  and  $h(w_i) < h(w_{i+1})$ , then  $s(v_i) = s'(w_i)$  for all  $i \in \{1, \dots, n\}$ . Then one has

$$(29) \quad [T, s] = \omega^n [T', s'] + \sum_{\sigma \in \text{St}(T'), d([T', \sigma]) < d([T', s'])} x_\sigma [T', \sigma],$$

where  $n \in \mathbb{Z}$ ,  $x_\sigma \in \mathbb{k}$  and the sum in the right-hand side is over states  $\sigma$  of  $T'$  such that  $d([T', \sigma]) < d([T', s'])$ .

**Proof** We say that a tangle  $T_i$  is obtained from a tangle  $T_{i+1}$  by an elementary height exchange if there exists a boundary arc  $a$  and two consecutive points  $v$  and  $w$  in  $\partial_a T_i$  with  $h(v) < h(w)$  (“consecutive” means that there does not exist any  $p \in \partial_a T_i$  such that  $h(v) < h(p) < h(w)$ ) such that  $T_{i+1}$  is the tangle obtained from  $T_i$  by exchanging the heights of  $v$  and  $w$ . Since  $T$  and  $T'$  are isotopic, through an isotopy that does not preserve the height orders, we can obtain  $T'$  from  $T$  by a finite sequence  $T = T_1 \mapsto T_2 \mapsto \dots \mapsto T_n = T'$  of elementary height exchanges. It is clear that if one has a development (29) when the pair  $(T, T')$  is equal to a pair  $(T_i, T_{i+1})$  and a pair  $(T_{i+1}, T_{i+2})$ , then it holds for the pair  $(T_i, T_{i+2})$ . So by induction on the size  $n$  of the finite sequence, it is sufficient to prove the lemma in the particular case where  $T$  and  $T'$  differ by an elementary height exchange. In this case, (29) follows from the height exchange relations

$$\begin{aligned} \begin{array}{c} \diagup \diagdown \\ \diagdown \diagup \end{array} \begin{array}{c} \uparrow \\ \downarrow \end{array} &= A \begin{array}{c} \uparrow \\ \downarrow \end{array} \begin{array}{c} \uparrow \\ \downarrow \end{array}, & \begin{array}{c} \diagup \diagdown \\ \diagdown \diagup \end{array} \begin{array}{c} \downarrow \\ \uparrow \end{array} &= A \begin{array}{c} \downarrow \\ \uparrow \end{array} \begin{array}{c} \downarrow \\ \uparrow \end{array}, & \begin{array}{c} \diagup \diagdown \\ \diagdown \diagup \end{array} \begin{array}{c} \uparrow \\ \downarrow \end{array} &= A^{-1} \begin{array}{c} \downarrow \\ \uparrow \end{array} \begin{array}{c} \downarrow \\ \uparrow \end{array}, \\ \begin{array}{c} \diagup \diagdown \\ \diagdown \diagup \end{array} \begin{array}{c} \downarrow \\ \uparrow \end{array} &= A^{-1} \begin{array}{c} \uparrow \\ \downarrow \end{array} \begin{array}{c} \downarrow \\ \uparrow \end{array} + (A - A^{-3}) \begin{array}{c} \downarrow \\ \uparrow \end{array} \begin{array}{c} \downarrow \\ \uparrow \end{array}. & & \square \end{aligned}$$

**Notation 3.15** Let  $b \in \mathcal{B}^{\mathbb{G}}$ , so by definition  $b = b_{\alpha_1} \cdots b_{\alpha_n}$ , where  $b_{\alpha_i} \in \mathcal{B}(\alpha_i)$ . That is, one has either  $b_{\alpha_i} = \alpha_{++}^{a_i} \alpha_{+-}^{b_i} \alpha_{--}^{c_i}$  or  $b_{\alpha_i} = \alpha_{++}^{a_i} \alpha_{-+}^{b_i} \alpha_{--}^{c_i}$  for some  $a_i, b_i, c_i \geq 0$ . Let  $\mathbf{n} \in \mathbb{N}^{\mathbb{G}}$  be defined by  $\mathbf{n}(\alpha_i) := a_i + b_i + c_i$ . Let  $T(\mathbf{n})$  be the tangle underlying  $D(\mathbf{n})$ . Let  $(T, s)$  be a stated tangle (unique up to isotopy) such that  $b = [T, s]$ , so that  $T(\mathbf{n})$  is obtained from  $T$  by an isotopy that does not necessarily preserve the height order. Finally we define the element  $b^+ := [T(\mathbf{n}), s^+] \in \mathcal{B}_+^{\mathbb{G}}$ , where  $s^+ \in \text{St}^+(T(\mathbf{n}))$  is the unique state such that  $(T, s)$  and  $(T(\mathbf{n}), s^+)$  satisfy the assumptions of (29). Note that the induced map  $(\cdot)^+ : \mathcal{B}^{\mathbb{G}} \rightarrow \mathcal{B}_+^{\mathbb{G}}$ , sending  $b$  to  $b^+$ , is a bijection.

**Lemma 3.16** (i) For  $\mathbb{k}, \mathbb{k}' \in \mathbb{N} \times \mathbb{Z}$ , one has  $\mathcal{F}_{\mathbb{k}} \cdot \mathcal{F}_{\mathbb{k}'} \subset \mathcal{F}_{\mathbb{k} + \mathbb{k}'}$ .  
 (ii) For  $b \in \mathcal{B}^{\mathbb{G}}$ , one has

$$(30) \quad b = \omega^n b^+ + \text{lower terms},$$

where  $n \in \mathbb{Z}$  and “lower terms” is a linear combination of basis elements  $b_i^+ \in \mathcal{B}_+^{\mathbb{G}}$  such that  $\|b_i^+\| < \|b^+\|$ .

Note that the second assertion of Lemma 3.16 implies that  $\mathcal{B}^{\mathbb{G}}$  spans  $\mathcal{S}_\omega(\Sigma)$ , so reproves Proposition 2.14.

**Proof** (i) Let  $x := [T(\mathbf{n}), s]$  and  $y := [T(\mathbf{n}'), s']$ , and denote by  $(T(\mathbf{n}) \cup T(\mathbf{n}'), s \cup s')$  the stated tangle obtained by stacking  $(T(\mathbf{n}), s)$  on top of  $(T(\mathbf{n}'), s')$ , so that

$$x \cdot y = [T(\mathbf{n}) \cup T(\mathbf{n}'), s \cup s'].$$

The tangles  $T(\mathbf{n}) \cup T(\mathbf{n}')$  and  $T(\mathbf{n} + \mathbf{n}')$  differ by an isotopy that does not necessarily preserve the height orders, so Lemma 3.14 implies that  $x \cdot y$  is a linear combination of elements of the form  $[D(\mathbf{n} + \mathbf{n}'), \sigma]$  such that  $\|[D(\mathbf{n} + \mathbf{n}'), \sigma]\| \leq \|x\| + \|y\|$ . This proves the first assertion.

(ii) Using Notation 3.15, we apply Lemma 3.14 to  $b = [T, s]$  and  $b^+ = [T(\mathbf{n}), s^+]$ , and (30) is just a rewriting of (29).  $\square$

**Proof of Theorem 3.7** Both Proposition 2.14 and the second assertion of Lemma 3.16 imply that  $\mathcal{B}^{\mathbb{G}}$  generates  $\mathcal{S}_\omega(\Sigma)$ . To prove that  $\mathcal{B}^{\mathbb{G}}$  is free, we apply Lemma 3.11 to the injective map  $(\cdot)^+ : \mathcal{B}^{\mathbb{G}} \rightarrow \mathcal{B}_+^{\mathbb{G}}$  where we equip  $\mathcal{B}_+^{\mathbb{G}}$  with the ordering  $[D, s] < [D', s']$  if  $\|[D, s]\| > \|[D', s']\|$ . The hypotheses of Lemma 3.11 are satisfied by virtue of Proposition 3.10 and Lemma 3.16, so  $\mathcal{B}$  is free.  $\square$

## 4 Lattice gauge field theory

### 4.1 Ciliated graphs and quantum gauge group coaction

Since the pioneering work of Fock and Rosly [29], constructions in lattice gauge field theory are based on ciliated graphs. As we now explain, to a ciliated graph  $(\Gamma, c)$  one can associate a punctured surface  $\Sigma^0$  together with a finite presentation  $\mathbb{P}$  of its associated groupoid.

**Definition 4.1** (i) A *ribbon graph*  $\Gamma$  is a finite graph together with the data, for each vertex, of a cyclic ordering of its adjacent half-edges. An *orientation* for a ribbon graph is the choice of an orientation for each of its edges.

(ii) A *ciliated ribbon graph*  $(\Gamma, c)$  is a ribbon graph  $\Gamma$  together with a lift, for each vertex, of the cyclic ordering of the adjacent half-edges to a linear ordering. If the half-edges adjacent to a vertex have the cyclic ordering  $e_1 < e_2 < \dots < e_n < e_1$  that we lift to the linear ordering  $e_1 < e_2 < \dots < e_n$ , we draw a *cilium* between  $e_n$  and  $e_1$ .

(iii) We associate surfaces to ribbon graphs as follows.

(a) Place a disc  $D_v$  on top of each vertex  $v$  and a band  $B_e$  on top of each edge  $e$ , then glue the discs to the bands using the cyclic ordering. We thus get a surface  $S(\Gamma)$  named the *fattening* of  $\Gamma$ .

(b) The *closed punctured surface*  $\Sigma(\Gamma) = (\Sigma(\Gamma), \mathcal{P})$  associated to  $\Gamma$  is the closed punctured surface obtained from  $S(\Gamma)$  by gluing a disc to each boundary component and placing a puncture inside each added disc. So  $S(\Sigma)$  deformation retracts to  $\Sigma_{\mathcal{P}}(\Gamma)$ .

(c) The *open punctured surface*  $\Sigma^0(\Gamma, c) = (\Sigma^0(\Gamma, c), \mathcal{P}^0)$  associated to  $(\Gamma, c)$  is obtained from  $S(\Gamma)$  by first pushing each vertex  $v$  to the boundary of  $S(\Gamma)$  in the direction of the associated cilium. Said differently, if the ordered half-edges adjacent to  $v$  are  $e_1 < e_2 < \dots < e_n$ , we push  $v$  in the boundary of  $D_v$  so that it lies between the band  $B_{e_n}$  and the band  $B_{e_1}$ . Next place a puncture  $p_v$  next to  $v$  (in the counterclockwise direction) on the same boundary component as  $v$ . Finally, to each boundary component of  $S(\Gamma)$  which does not contain any puncture  $p_v$ , glue a disc and place a puncture inside the disc. In the so-obtained punctured surface  $\Sigma^0(\Gamma, c)$ , each boundary arc contains exactly one vertex  $v$  of  $\Gamma$ , so we denote by  $a_v$  the boundary arc containing  $v$ . Suppose that  $\Gamma$  is oriented. Then the oriented edges of  $\Gamma$  form a set  $\mathbb{G}$  of generators of  $\Pi_1(\Sigma_{\mathcal{P}}^0, \mathbb{V})$  such that  $\mathbb{P}(\Gamma, c) := (\mathbb{G}, \emptyset)$  is a finite presentation without relations.

(iv) For  $v_1$  and  $v_2$  two distinct vertices of  $(\Gamma, c)$ , the ciliated graph  $(\Gamma_{v_1\#v_2}, c_{v_1\#v_2})$  is obtained by gluing the vertices  $v_1$  and  $v_2$  to a vertex  $v$  in such a way that if  $e_1 < \dots < e_n$  and  $f_1 < \dots < f_m$  are the ordered half-edges adjacent to  $v_1$  and  $v_2$ , respectively, then the linear order of the half-edges adjacent to  $v$  is  $e_1 < \dots < e_n < f_1 < \dots < f_m$ . Note that  $c_{v_1\#v_2} \neq c_{v_2\#v_1}$ .

Figure 16 illustrates two examples having the same ribbon graph but different ciliated structures: the punctured surface  $\Sigma^0(\Gamma, c)$  is a disc with two inner punctures and two boundary punctures whereas  $\Sigma^0(\Gamma, c')$  is an annulus with one puncture per boundary component and one inner puncture.

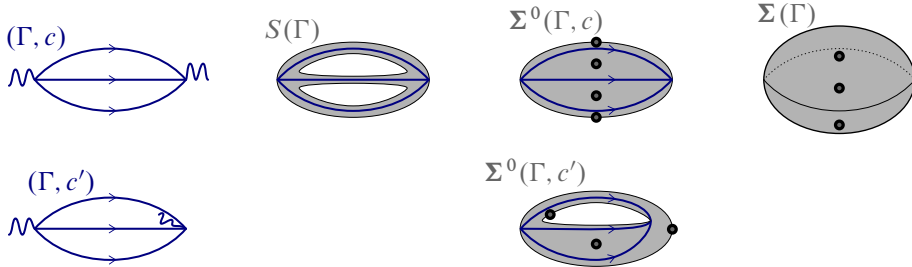


Figure 16: Top, from left to right: a ciliated graph  $(\Gamma, c)$ , its fattening  $S(\Gamma)$ , its open punctured surface  $\Sigma^0(\Gamma, c)$  and its closed punctured surface  $\Sigma(\Gamma)$ . Bottom: the same ribbon graph with a different ciliated structure  $c'$  (left) and the associated open punctured surface  $\Sigma^0(\Gamma, c')$  (right).

**Remark 4.2** In [26] Costantino and Lê made the following important remark: the punctured surface  $\Sigma^0(\Gamma_{v_1 \# v_2}, c_{v_1 \# v_2})$  is obtained from  $\Sigma^0(\Gamma, c) \sqcup \mathbb{T}$  by gluing the boundary arcs  $a_{v_1}$  and  $a_{v_2}$  to two faces of the triangle  $\mathbb{T}$ . In particular, when  $\Gamma = \Gamma_1 \sqcup \Gamma_2$  with  $v_1 \in \Gamma_1$  and  $v_2 \in \Gamma_2$ , this property, together with Theorem 2.10, permitted the authors of [26] to prove that  $\mathcal{S}_\omega(\Sigma^0(\Gamma_{v_1 \# v_2}, c_{v_1 \# v_2}))$  is the cobraided tensor product of  $\mathcal{S}_\omega(\Sigma^0(\Gamma_1, c_1))$  with  $\mathcal{S}_\omega(\Sigma^0(\Gamma_2, c_2))$ . The same gluing property was first discovered by Alekseev, Grosse and Schomerus in [2; 3] for the quantum moduli spaces (see [37] for a survey on the classical and quantum versions of the fusion operation).

For an oriented ciliated graph  $(\Gamma, c)$ , we denote by  $V(\Gamma)$  its set of vertices and  $\mathcal{E}(\Gamma)$  its set of (oriented) edges. Like in the previous section, we see the elements of  $\mathcal{E}(\Gamma)$  as oriented arcs. Denote by  $\mathbb{D}_0$  the punctured surface made of a disc with a single puncture on its boundary. The closed punctured surface  $\Sigma(\Gamma)$  is obtained from the open one  $\Sigma^0(\Gamma, c)$  by gluing a copy  $\mathbb{D}_0$  along each boundary arc  $a_v$ . Therefore, writing  $\widehat{\mathbb{D}} := \bigsqcup_{v \in V(\Gamma)} \mathbb{D}_0$ , by Theorem 2.10 one has the exact sequence

$$(31) \quad 0 \rightarrow \mathcal{S}_\omega(\Sigma(\Gamma)) \xrightarrow{i} \mathcal{S}_\omega(\Sigma^0(\Gamma, c) \sqcup \widehat{\mathbb{D}}) \xrightarrow{\Delta^R - \sigma \circ \Delta^L} \mathcal{S}_\omega(\Sigma^0(\Gamma, c) \sqcup \widehat{\mathbb{D}}) \otimes \mathcal{O}_q[\mathrm{SL}_2]^{\otimes V(\Gamma)},$$

where  $i$  represents the gluing map.

Using the isomorphism  $\mathcal{S}_\omega(\mathbb{D}_0) \cong \mathbb{k}$  sending the class of the empty stated tangle to the neutral element  $1 \in \mathbb{k}$ , we define an isomorphism

$$\kappa: \mathcal{S}_\omega(\Sigma^0(\Gamma, c) \sqcup \widehat{\mathbb{D}}) \cong \mathcal{S}_\omega(\Sigma^0(\Gamma, c)) \otimes \bigotimes_{v \in V(\Gamma)} \mathcal{S}_\omega(\mathbb{D}_0) \cong \mathcal{S}_\omega(\Sigma^0(\Gamma, c)).$$

Denote by  $\iota: \mathcal{S}_\omega(\Sigma(\Gamma)) \hookrightarrow \mathcal{S}_\omega(\Sigma^0(\Gamma, c))$  the injective morphism  $\iota := \kappa \circ i$ . Also denote by  $\Delta^{\mathcal{G}}: \mathcal{S}_\omega(\Sigma^0(\Gamma, c)) \rightarrow \mathcal{S}_\omega(\Sigma^0(\Gamma, c)) \otimes \mathcal{O}_q[\mathrm{SL}_2]^{\otimes V(\Gamma)}$  the (unique) morphism making the following diagram commute:

$$\begin{array}{ccc} \mathcal{S}_\omega(\Sigma^0(\Gamma, c) \sqcup \widehat{\mathbb{D}}) & \xrightarrow{\Delta^R} & \mathcal{S}_\omega(\Sigma^0(\Gamma, c) \sqcup \widehat{\mathbb{D}}) \otimes \mathcal{O}_q[\mathrm{SL}_2]^{\otimes V(\Gamma)} \\ \cong \downarrow \kappa & & \cong \downarrow \kappa \otimes \mathrm{id} \\ \mathcal{S}_\omega(\Sigma^0(\Gamma, c)) & \xrightarrow{\Delta^{\mathcal{G}}} & \mathcal{S}_\omega(\Sigma^0(\Gamma, c)) \otimes \mathcal{O}_q[\mathrm{SL}_2]^{\otimes V(\Gamma)} \end{array}$$

**Definition 4.3** The quantum gauge group is the Hopf algebra  $\mathcal{O}_q[\mathcal{G}] := \mathcal{O}_q[\mathrm{SL}_2]^{\otimes V(\Gamma)}$ . The (right) Hopf-comodule map  $\Delta^{\mathcal{G}}: \mathcal{S}_\omega(\Sigma^0(\Gamma, c)) \rightarrow \mathcal{S}_\omega(\Sigma^0(\Gamma, c)) \otimes \mathcal{O}_q[\mathcal{G}]$  is called the quantum gauge group coaction.

Note that, by definition, the following diagram commutes:

$$\begin{array}{ccc} \mathcal{S}_\omega(\Sigma^0(\Gamma, c) \sqcup \widehat{\mathbb{D}}) & \xrightarrow{\sigma \circ \Delta^L} & \mathcal{S}_\omega(\Sigma^0(\Gamma, c) \sqcup \widehat{\mathbb{D}}) \otimes \mathcal{O}_q[\mathcal{G}] \\ \cong \downarrow \kappa & & \cong \downarrow \kappa \otimes \mathrm{id} \\ \mathcal{S}_\omega(\Sigma^0(\Gamma, c)) & \xrightarrow{\mathrm{id} \otimes \epsilon} & \mathcal{S}_\omega(\Sigma^0(\Gamma, c)) \otimes \mathcal{O}_q[\mathcal{G}] \end{array}$$

Therefore the exactness of (31) implies that we have the exact sequence

$$(32) \quad 0 \rightarrow \mathcal{S}_\omega(\Sigma(\Gamma)) \xrightarrow{\iota} \mathcal{S}_\omega(\Sigma^0(\Gamma, c)) \xrightarrow{\Delta^{\mathcal{G}} - \mathrm{id} \otimes \epsilon} \mathcal{S}_\omega(\Sigma^0(\Gamma, c)) \otimes \mathcal{O}_q[\mathcal{G}].$$

Said differently,  $\iota(\mathcal{S}_\omega(\Sigma(\Gamma)))$  is the subalgebra of  $\mathcal{S}_\omega(\Sigma^0(\Gamma, c))$  of coinvariant vectors for the quantum gauge group coaction.

**Notation 4.4** For  $x \in \mathcal{O}_q[\mathrm{SL}_2]$  and  $v_0 \in V(\Gamma)$  the element of the form  $\bigotimes_v y_v$ , where  $y_v = 1$  for  $v \neq v_0$  and  $y_{v_0} = x$ , is denoted by  $x^{(v_0)} \in \mathcal{O}_q[\mathcal{G}] = \mathcal{O}_q[\mathrm{SL}_2]^{\otimes V(\Gamma)}$ .

Let  $\alpha$  be an arc of type either  $a$  or  $d$  and write  $v_1$  and  $v_2$  for the elements of  $\mathbb{V}$  corresponding to the boundary arcs containing  $s(\alpha)$  and  $t(\alpha)$ , respectively. The quantum gauge group coaction is characterized by the following formula illustrated in Figure 17:

$$(33) \quad \Delta^{\mathcal{G}}(\alpha_{ij}) = \sum_{a,b=\pm} \alpha_{ab} \otimes x_{jb}^{(v_2)} x_{ia}^{(v_1)}.$$

In order to prepare the comparison between stated skein algebras at  $\omega = +1$  and relative character varieties in the next subsection, let us derive from Theorem 1.1 an alternative

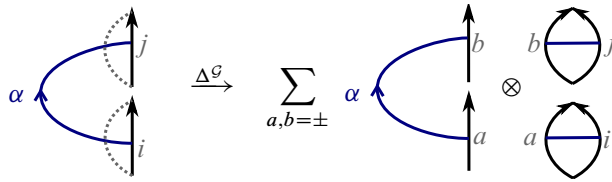


Figure 17: An illustration of (33).

presentation of  $\mathcal{S}_\omega(\Sigma)$ . During the rest of the section, we fix a finite presentation  $\mathbb{P} = (\mathbb{G}, \mathbb{RL})$  of  $\Pi_1(\Sigma_{\mathcal{P}}, \mathbb{V})$  such that every arc of  $\mathbb{G}$  is either of type  $a$  or  $d$ .

When comparing skein algebras with character varieties, there is a well-known sign issue which requires some attention. When  $\Sigma$  is closed, the skein algebra  $\mathcal{S}_{+1}(\Sigma)$  is generated by the classes of closed curves  $\gamma$  whereas the algebra  $\mathbb{C}[\mathcal{X}_{\text{SL}_2}(\Sigma)]$  of regular functions of the character variety is generated by curve functions  $\tau_\gamma$ , sending a class  $[\rho]$  of representation  $\rho: \pi_1(\Sigma_{\mathcal{P}}) \rightarrow \text{SL}_2(\mathbb{C})$  to  $\tau_\gamma([\rho]) := \text{tr}(\rho(\gamma))$ . However there is no isomorphism  $\mathcal{S}_{+1}(\Sigma) \cong \mathbb{C}[\mathcal{X}_{\text{SL}_2}(\Sigma)]$  sending  $\gamma$  to  $\tau_\gamma$ . Instead, we fix a spin structure on  $\Sigma_{\mathcal{P}}$  with associated Johnson quadratic form  $\Omega: H_1(\Sigma_{\mathcal{P}}; \mathbb{Z}/2\mathbb{Z}) \rightarrow \mathbb{Z}/2\mathbb{Z}$  and define  $w(\gamma) := 1 + \Omega([\gamma])$ . Then it follows from [7; 17; 44] that we have an isomorphism  $\mathcal{S}_{+1}(\Sigma) \cong \mathbb{C}[\mathcal{X}_{\text{SL}_2}(\Sigma)]$  sending  $\gamma$  to  $(-1)^{w(\gamma)}\tau_\gamma$ . A similar sign issue appears when dealing with stated skein algebras and relative character varieties; this was studied in [39] to which we refer for further details (see also [26; 48] for an elegant interpretation of this sign issue in term of *twisted character variety*).

In short, the authors defined in [39] the notion of *relative spin structure* to which one can associate a map  $w: \mathbb{G} \rightarrow \mathbb{Z}/2\mathbb{Z}$  having the property that for any simple relation  $R = \beta_k \star \dots \star \beta_1$ , one has  $\sum_{i=1}^k w(\beta_i) = 1$ . We will call a map  $w: \mathbb{G} \rightarrow \mathbb{Z}/2\mathbb{Z}$  satisfying this property a *spin function*.

**Notation 4.5** Let  $w$  be a spin function. For  $\alpha \in \mathbb{G}$ , we denote by  $U(\alpha)$  the  $2 \times 2$  matrix with coefficients in  $\mathcal{S}_\omega(\Sigma)$  defined by

$$(34) \quad U(\alpha) := \begin{cases} (-1)^{w(\alpha)} \omega C^{-1} M(\alpha) & \text{if } \alpha \text{ is of type } a, \\ (-1)^{w(\alpha)} C^{-1} M(\alpha) = (-1)^{w(\alpha)} N(\alpha) & \text{if } \alpha \text{ is of type } d. \end{cases}$$

**Proposition 4.6** (i) *The stated skein algebra  $\mathcal{S}_\omega(\Sigma)$  admits the alternative presentation with generators the elements  $U(\alpha)_i^j$  and with  $\alpha \in \mathbb{G}$  and  $i, j = \pm$ , together with the following relations:*

- *The  $q$ -determinant relations  $\det_q(U(\alpha)) = 1$  when  $\alpha$  is of type  $a$ , and  $\det_{q^2}(U(\alpha)) = 1$  when  $\alpha$  is of type  $d$ .*



- For  $R = \beta_k \star \cdots \star \beta_1 \in \mathbb{RL}$  a relation where  $l$  generators  $\beta_i$  are of type  $a$ , the trivial loop relation

$$(35) \quad U(\beta_k) \cdots U(\beta_1) = A^3 \omega^l.$$

- For each pair of generators in  $\mathbb{G}$ , the arc exchange relations obtained from the relations in Lemma 2.24 by replacing  $N(\alpha)$  by  $U(\alpha)$  if  $\alpha$  is of type  $d$  or by  $CN(\alpha)$  if  $\alpha$  is of type  $a$ .

(ii) The quantum gauge group coaction is characterized by the formula

$$(36) \quad \Delta^{\mathcal{G}}(U(\alpha)_i^j) = \sum_{a,b=\pm} U(\alpha)_a^b \otimes S(x_{bj})^{(v_2)} x_{ia}^{(v_1)},$$

where we use the same notation as in (33).

**Proof** It is clear from (34) that the matrix elements  $U(\alpha)_i^j$  generate the same algebra as the elements  $M(\alpha)_i^j = \alpha_{ij}$ , so they generate  $\mathcal{S}_\omega(\Sigma)$ . We need to check that the  $q$ -determinant, trivial loop and arc exchange relations for the elements  $\alpha_{ij}$  are equivalent to the relations of the proposition for the elements  $U(\alpha)_i^j$ . When  $\alpha \in \mathbb{G}$  is of type  $d$ , clearly the relation  $\det_{q^2}(N(\alpha)) = 1$  is equivalent to the relation  $\det_{q^2}(U(\alpha)) = 1$ . When  $\alpha \in \mathbb{G}$  is of type  $a$ , the equivalence

$$\det_q(M(\alpha)) = 1 \iff \det_q(U(\alpha)) = 1$$

follows from a straightforward computation (and is the reason for the  $\omega$  in the expression  $U(\alpha) = (-1)^{w(\alpha)} \omega C^{-1} M(\alpha)$ ). The equivalence between (10) and (35) is straightforward (and is responsible for the introduction of the spin function and for the  $(-1)^{w(\alpha)}$  factor in the definition of  $U(\alpha)$ ). The fact that the arc exchange relations are equivalent to the same relations with  $N(\alpha)$  replaced by  $U(\alpha)$  or  $CU(\alpha)$  depending whether  $\alpha$  is of type  $d$  or  $a$  follows from the definition of  $U(\alpha)$  and the fact that the arc exchange relations are homogeneous.

It remains to derive the formula (36) from (33). This is done by direct computation, left to the reader, using the fact that for the two  $2 \times 2$  matrices

$$X = \begin{pmatrix} x_{++} & x_{+-} \\ x_{-+} & x_{--} \end{pmatrix} \quad \text{and} \quad S(X) = \begin{pmatrix} S(x_{++}) & S(x_{+-}) \\ S(x_{-+}) & S(x_{--}) \end{pmatrix}$$

with coefficients in  $\mathcal{O}_q[\text{SL}_2]$ , one has  $S(X) = C^{-1t} X C$ . Figure 18 illustrates (36). In Figure 18, we use a special convention: we have drawn stated diagrams that go “outside” of  $\Sigma_{\mathcal{P}}$  in some small bigon neighborhoods of the boundary arcs. It must be

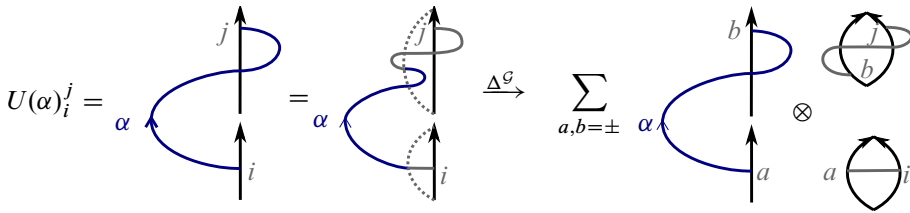


Figure 18: An illustration of (36).

understood that we need to apply a boundary skein relation in those neighborhoods. This convention permits us to draw the matrix coefficients  $(C^{-1}M(\alpha))_i^j$ . Note also that in Figure 18 we drop the scalar factor  $(-1)^{w(\alpha)}$ . □

### 4.2 Relative character varieties

Since the quantum moduli algebras are deformation quantizations of the (relative) character varieties studied by Fock and Rosly in [29], we briefly recall their construction and refer to [6] for a detailed survey.

Let  $\Sigma$  be a punctured surface and  $\mathbb{V} \subset \Sigma_{\mathcal{P}}$  be a finite subset which intersects each boundary arc exactly once and each connected component of  $\Sigma$  at least once. Denote by  $\mathring{\mathbb{V}} := \mathbb{V} \cap \mathring{\Sigma}_{\mathcal{P}}$  its (possibly empty) subset of inner points and let  $\Pi_1(\Sigma_{\mathcal{P}}, \mathbb{V})$  be the full subcategory of  $\Pi_1(\Sigma_{\mathcal{P}})$  generated by  $\mathbb{V}$ . The representation space  $\mathcal{R}_{\text{SL}_2}(\Sigma, \mathbb{V})$  is the set of functors  $\rho: \Pi_1(\Sigma, \mathbb{V}) \rightarrow \text{SL}_2(\mathbb{C})$ . The discrete gauge group is  $\mathcal{G}_{\mathbb{V}} := \text{SL}_2(\mathbb{C})^{\mathring{\mathbb{V}}}$  and it acts on  $\mathcal{R}_{\text{SL}_2}(\Sigma, \mathbb{V})$  by

$$(\rho \cdot g)(\alpha) := g(t(\alpha))^{-1} \rho(\alpha) g(s(\alpha)) \quad \text{for } \rho \in \mathcal{R}_{\text{SL}_2}(\Sigma, \mathbb{V}), g \in \mathcal{G}_{\mathbb{V}}, \alpha \in \Pi_1(\Sigma_{\mathcal{P}}, \mathbb{V}).$$

We claim that  $\mathcal{R}_{\text{SL}_2}(\Sigma, \mathbb{V})$  can be given the structure of affine variety in such a way that the action of the reducible algebraic group  $\mathcal{G}_{\mathbb{V}}$  is algebraic, so we can define the GIT quotient

$$\mathcal{X}_{\text{SL}_2}(\Sigma) := \mathcal{R}_{\text{SL}_2}(\Sigma, \mathbb{V}) // \mathcal{G}_{\mathbb{V}},$$

which we call the *relative character variety*. To prove this, consider a finite presentation  $\mathbb{P} = (\mathbb{G}, \mathbb{RL})$  of  $\Pi_1(\Sigma_{\mathcal{P}}, \mathbb{V})$  and write  $\mathbb{G} = (\alpha_1, \dots, \alpha_n)$  and  $\mathbb{RL} = (R_1, \dots, R_m)$ . Consider the regular map  $\mathcal{R}: \text{SL}_2(\mathbb{C})^{\mathbb{G}} \rightarrow \text{SL}_2(\mathbb{C})^{\mathbb{RL}}$  written  $\mathcal{R} = (\mathcal{R}_1, \dots, \mathcal{R}_m)$ , where the coordinate  $\mathcal{R}_i$  associated to a relation  $R_i = \alpha_{i_1}^{\varepsilon_1} \star \dots \star \alpha_{i_k}^{\varepsilon_k}$  is the polynomial function

$$\mathcal{R}_i(g_1, \dots, g_n) = g_{i_1}^{\varepsilon_1} \cdots g_{i_k}^{\varepsilon_k}.$$

Clearly one has  $\mathcal{R}_{\text{SL}_2}(\Sigma, \mathbb{V}) = \mathcal{R}^{-1}(\mathbb{1}_2, \dots, \mathbb{1}_2)$ , where  $\mathbb{1}_2$  is the identity matrix, so  $\mathcal{R}_{\text{SL}_2}(\Sigma, \mathbb{V})$  is a subvariety of  $\text{SL}_2(\mathbb{C})^{\mathbb{G}}$ .

Note that the algebra  $\mathbb{C}[\mathcal{R}_{\text{SL}_2}(\Sigma, \mathbb{V})]$  of regular functions lies in the exact sequence

$$(37) \quad \mathbb{C}[\text{SL}_2(\mathbb{C})]^{\otimes \mathbb{RL}} \xrightarrow{\mathcal{R}^* - \eta^{\otimes \mathbb{G}} \circ \epsilon^{\otimes \mathbb{RL}}} \mathbb{C}[\text{SL}_2(\mathbb{C})]^{\otimes \mathbb{G}} \rightarrow \mathbb{C}[\mathcal{R}_{\text{SL}_2}(\Sigma, \mathbb{V})] \rightarrow 0,$$

where  $\eta$  and  $\epsilon$  are the unit and counit of  $\mathbb{C}[\text{SL}_2]$ . So we have turned  $\mathcal{R}_{\text{SL}_2}(\Sigma, \mathbb{V})$  into an affine variety. Now the discrete gauge group action is induced by the Hopf comodule map  $\Delta^{\mathbb{G}}: \mathbb{C}[\mathcal{R}_{\text{SL}_2}(\Sigma, \mathbb{V})] \rightarrow \mathbb{C}[\mathcal{R}_{\text{SL}_2}(\Sigma, \mathbb{V})] \otimes \mathbb{C}[\mathcal{G}_{\mathbb{V}}]$ , which is the restriction of the right comodule map  $\tilde{\Delta}_{\mathbb{G}}: \mathbb{C}[\text{SL}_2(\mathbb{C})]^{\otimes \mathbb{G}} \rightarrow \mathbb{C}[\text{SL}_2(\mathbb{C})]^{\otimes \mathbb{G}} \otimes \mathbb{C}[\text{SL}_2(\mathbb{C})]^{\otimes \mathbb{V}}$  defined by

$$\tilde{\Delta}^{\mathbb{G}}(x^{(\alpha)}) = \sum x''^{(\alpha)} \otimes S(x''')^{(v_2)} x'^{(v_1)} \quad \text{for } x \in \mathcal{O}_q[\text{SL}_2] \text{ and } \alpha: v_1 \rightarrow v_2 \in \mathbb{G},$$

using Sweedler’s notation  $\Delta^{(2)}(x) = \sum x' \otimes x'' \otimes x'''$ . In particular, when  $x = x_{ij}$  with  $i, j \in \{-, +\}$ , the formula gives

$$(38) \quad \Delta^{\mathbb{G}}(x_{ij}^{(\alpha)}) = \sum_{a,b=\pm} x_{ab}^{(\alpha)} \otimes S(x_{bj})^{(v_2)} x_{ia}^{(v_1)}.$$

Note the analogy between (38) and (36).

Finally, the algebra of regular functions of the relative character variety is defined as the set of coinvariant vectors for this coaction, that is by the exact sequence

$$(39) \quad 0 \rightarrow \mathbb{C}[\mathcal{X}_{\text{SL}_2}(\Sigma)] \rightarrow \mathbb{C}[\mathcal{R}_{\text{SL}_2}(\Sigma)] \xrightarrow{\Delta^{\mathbb{G}} - \text{id} \otimes \epsilon} \mathbb{C}[\mathcal{R}_{\text{SL}_2}(\Sigma)] \otimes \mathbb{C}[\mathcal{G}_{\mathbb{V}}].$$

The relative character variety  $\mathcal{X}_{\text{SL}_2}(\Sigma)$  does not depend (up to unique isomorphism) on the choice of the triple  $(\mathbb{V}, \mathbb{G}, \mathbb{RL})$  used to define it, but only on  $\Sigma$ ; we refer to [36] for a proof. Note that in the particular case where  $\mathbb{V} \subset \partial \Sigma_{\mathcal{P}}$ , the gauge group is trivial so  $\mathcal{X}_{\text{SL}_2}(\Sigma) = \mathcal{R}_{\text{SL}_2}(\Sigma)$ . Moreover, if the presentation  $\mathbb{P}$  does not have any relations, then  $\mathcal{R}_{\text{SL}_2}(\Sigma) = \text{SL}_2(\mathbb{C})^{\mathbb{G}}$ . As we saw in Example 2.13, such a presentation  $\mathbb{P}$  always exists when  $\Sigma$  is a connected punctured surface with nontrivial boundary, therefore in that case one has

$$\mathcal{X}_{\text{SL}_2}(\Sigma) = \text{SL}_2(\mathbb{C})^{\mathbb{G}}.$$

Now consider an oriented ciliated graph  $(\Gamma, c)$  and consider the associated finite presentation  $(\mathbb{V}, \mathbb{G}, \mathbb{RL})$  of the groupoid  $\Pi_1(\Sigma_{\mathcal{P}}^0(\Gamma, c), \mathbb{V})$  associated to the open punctured surface defined in the previous subsection. The same triple  $(\mathbb{V}, \mathbb{G}, \mathbb{RL})$  also gives a finite presentation of  $\Pi_1(\Sigma_{\mathcal{P}}(\Gamma), \mathbb{V})$  associated to the closed punctured surface, where this time all elements of  $\mathbb{V}$  are inner vertices of  $\Sigma_{\mathcal{P}}(\Gamma)$ . Therefore one has

$$\mathcal{X}_{\text{SL}_2}(\Sigma^0(\Gamma, c)) = \mathcal{R}_{\text{SL}_2}(\Sigma(\Gamma)) = \text{SL}_2(\mathbb{C})^{\mathcal{E}(\Gamma)},$$

where as before  $\mathcal{E}(\Gamma)$  denotes the set of edges of  $\Gamma$ . So the exact sequence (39) can be rewritten as

$$0 \rightarrow \mathbb{C}[\mathcal{X}_{\mathrm{SL}_2}(\Sigma(\Gamma))] \rightarrow \mathbb{C}[\mathcal{X}_{\mathrm{SL}_2}(\Sigma^0(\Gamma, c))] \xrightarrow{\Delta^{\mathcal{G}} - \mathrm{id} \otimes \epsilon} \mathbb{C}[\mathcal{X}_{\mathrm{SL}_2}(\Sigma^0(\Gamma, c))] \otimes \mathbb{C}[\mathcal{G}_{\mathbb{V}}].$$

Note the analogy with the exact sequence (32). The main achievement of Fock and Rosly in [29] is the construction of Poisson structures on  $\mathbb{C}[\mathcal{X}_{\mathrm{SL}_2}(\Sigma^0(\Gamma, c))] = \mathbb{C}[\mathrm{SL}_2]^{\otimes \mathcal{E}(\Gamma)}$  and  $\mathbb{C}[\mathcal{G}_{\mathbb{V}}] = \mathbb{C}[\mathrm{SL}_2]^{\otimes \mathring{V}(\Gamma)}$  such that the coaction  $\Delta^{\mathcal{G}}$  is a Poisson morphism. Therefore, using the above exact sequence, the affine variety  $\mathcal{X}_{\mathrm{SL}_2}(\Sigma(\Gamma))$  receives a (quotient) Poisson structure. A good plan then is to show that this Poisson structure only depends on the surface  $\Sigma_{\mathcal{P}}(\Gamma)$  and not on  $(\Gamma, c)$ . This strategy permitted the authors of [29] to extend the Atiyah–Bott–Goldman Poisson structure from unpunctured closed surfaces to closed general punctured surfaces (see also [36] for a general treatment in the language of punctured surfaces rather than ciliated graphs and using groupoid cohomology, and for a Goldman type formula for the Poisson bracket).

Let us conclude this subsection with the following observation. It is well known that the (stated) skein algebra  $\mathcal{S}_{+1}(\Sigma)$  is isomorphic (though noncanonically) to the algebra  $\mathbb{C}[\mathcal{X}_{\mathrm{SL}_2}(\Sigma)]$  of regular functions of the (relative) character variety. For closed punctured surfaces this was shown by Bullock [17] under the assumption that  $\mathcal{S}_{+1}(\Sigma)$  is reduced; this assumption was proved in [44] (see also [23] for an alternative proof). For open punctured surfaces this was proved independently in [39, Theorem 1.3] and [26, Theorem 8.12] using triangulations of surfaces. Let us note that Theorem 1.1 gives a straightforward alternative proof of this result with the additional assumption that  $\mathcal{P} \neq \emptyset$ .

**Theorem 4.7** [17; 26; 39; 44] *The algebras  $\mathcal{S}_{+1}(\Sigma)$  (where  $\mathbb{k} = \mathbb{C}$ ) and  $\mathbb{C}[\mathcal{X}_{\mathrm{SL}_2}(\Sigma)]$  are isomorphic.*

**Proof** First suppose that  $\Sigma$  is an open connected punctured surface, let  $\mathbb{V}$  be such that each of its vertices are on the boundary (so the representation and relative character varieties are the same), let  $\mathbb{P} = (\mathbb{G}, \mathbb{RL})$  be a finite presentation of  $\Pi_1(\Sigma_{\mathcal{P}}, \mathbb{V})$  whose generators are either of type  $a$  or  $d$  and fix a spin function  $w$ . By (37), the algebra  $\mathbb{C}[\mathcal{X}_{\mathrm{SL}_2}(\Sigma)]$  is presented by the generators  $x_{ij}^{(\alpha)}$  for  $\alpha \in \mathbb{G}$  and  $i, j \in \{-, +\}$ , with

- the exchange relations  $x_{ij}^{(\alpha)} x_{kl}^{(\beta)} = x_{kl}^{(\beta)} x_{ij}^{(\alpha)}$  for all  $\alpha, \beta \in \mathbb{G}$  and  $i, j \in \{-, +\}$ ,
- the determinant relations  $\det(X(\alpha)) = 1$  for all  $\alpha \in \mathbb{G}$ ,
- the trivial loop relations  $X(\beta_k) \cdots X(\beta_1) = \mathbb{1}_2$  for  $R = \beta_k \star \cdots \star \beta_1 \in \mathbb{RL}$ ,

where we set

$$X(\alpha) := \begin{pmatrix} x_{++}^{(\alpha)} & x_{+-}^{(\alpha)} \\ x_{-+}^{(\alpha)} & x_{--}^{(\alpha)} \end{pmatrix}.$$

By comparing this presentation of  $\mathbb{C}[\mathcal{X}_{\text{SL}_2}(\Sigma)]$  with the presentation of  $\mathcal{S}_\omega(\Sigma)$  obtained in Proposition 4.6 by setting  $\omega = +1$ , we see that one has an isomorphism of algebras  $\Theta: \mathcal{S}_{+1}(\Sigma) \xrightarrow{\cong} \mathbb{C}[\mathcal{X}_{\text{SL}_2}(\Sigma)]$  sending  $U(\alpha)$  to  $X(\alpha)$ ; note that  $\mathcal{R} = \tau$  when  $\omega = +1$ , so all arc exchange relations become  $U(\alpha) \odot U(\beta) = \tau U(\alpha) \odot U(\beta) \tau$  giving relations  $\alpha_{ij} \beta_{kl} = \beta_{kl} \alpha_{ij}$ . Moreover, by comparing (38) and (36), we see that  $\Theta$  is equivariant for the gauge group coactions.

Now suppose that  $\Sigma$  is closed and connected with  $\mathcal{P} \neq \emptyset$ , and let  $(\Gamma, c)$  be a ciliated fat graph such that  $\Sigma(\Gamma) = \Sigma$ . By the preceding case, one has an equivariant isomorphism  $\Theta: \mathcal{S}_{+1}(\Sigma^0(\Gamma, c)) \xrightarrow{\cong} \mathbb{C}[\mathcal{X}_{\text{SL}_2}(\Sigma^0(\Gamma, c))]$ , so one has a commutative diagram

$$\begin{array}{ccccc} 0 \longrightarrow \mathcal{S}_{+1}(\Sigma(\Gamma)) & \longrightarrow & \mathcal{S}_{+1}(\Sigma^0(\Gamma, c)) & \xrightarrow{\Delta^{\mathcal{G}} - \text{id} \otimes \epsilon} & \mathcal{S}_{+1}(\Sigma^0(\Gamma, c)) \otimes \mathbb{C}[\mathcal{G}_{\mathbb{V}}] \\ & \cong \downarrow \exists! & \cong \downarrow \Theta & & \cong \downarrow \Theta \otimes \text{id} \\ 0 \longrightarrow \mathbb{C}[\mathcal{X}_{\text{SL}_2}(\Sigma(\Gamma))] & \longrightarrow & \mathbb{C}[\mathcal{X}_{\text{SL}_2}(\Sigma^0(\Gamma, c))] & \xrightarrow{\Delta^{\mathcal{G}} - \text{id} \otimes \epsilon} & \mathbb{C}[\mathcal{X}_{\text{SL}_2}(\Sigma^0(\Gamma, c))] \otimes \mathbb{C}[\mathcal{G}_{\mathbb{V}}] \end{array}$$

Since both lines are exact there exists an isomorphism  $\mathcal{S}_{+1}(\Sigma(\Gamma)) \xrightarrow{\cong} \mathbb{C}[\mathcal{X}_{\text{SL}_2}(\Sigma(\Gamma))]$  obtained by restriction of  $\Theta$ . □

### 4.3 Combinatorial quantizations of (relative) character varieties

The work of Fock and Rosly suggests a natural way of quantizing character varieties. The following problem was raised and solved independently by Alekseev, Grosse and Schomerus [2; 3] and Buffenoir and Roche [15] (see also [20] for a survey):

**Problem 4.8** Associate to each oriented ciliated graph  $(\Gamma, c)$  an (associative unital) algebra  $\mathcal{L}_\omega(\Gamma, c)$  over the ring  $\mathbb{k} := \mathbb{C}[\omega^{\pm 1}]$  satisfying:

(A1) As a  $\mathbb{k}$ -module,  $\mathcal{L}_\omega(\Gamma, c)$  is just the (free) module

$$\mathbb{C}[\mathcal{R}_{\text{SL}_2}(\Sigma^0(\Gamma, c))] \otimes_{\mathbb{C}} \mathbb{k} \cong \mathbb{C}[\text{SL}_2]^{\otimes \mathcal{E}(\Gamma)} \otimes_{\mathbb{C}} \mathbb{k}.$$

(A2) As before, write  $\mathcal{O}_q[\mathcal{G}] := \mathcal{O}_q[\text{SL}_2]^{\otimes \mathcal{V}(\Gamma)}$ . The linear map

$$\Delta^{\mathcal{G}}: \mathcal{L}_\omega(\Gamma, c) \rightarrow \mathcal{L}_\omega(\Gamma, c) \otimes \mathcal{O}_q[\mathcal{G}]$$

defined by the formulas

$$\Delta^{\mathcal{G}}(x_{ij}^{(\alpha)}) = \sum_{a,b=\pm} x_{ab}^{(\alpha)} \otimes S(x_{bj})^{(v_2)} x_{ia}^{(v_1)}$$

is a Hopf-comodule map. In particular, it is a morphism of algebras.

(Inv) The subalgebra  $\mathcal{L}_\omega^{\text{inv}}(\Gamma) \subset \mathcal{L}_\omega(\Gamma, c)$  defined by the exact sequence

$$0 \rightarrow \mathcal{L}_\omega^{\text{inv}}(\Gamma) \rightarrow \mathcal{L}_\omega(\Gamma, c) \xrightarrow{\Delta^G - \text{id} \otimes \epsilon} \mathcal{L}_\omega(\Gamma, c) \otimes \mathcal{O}_q[\mathcal{G}]$$

only depends (up to canonical isomorphism) on the (homeomorphism class of) surface  $S(\Gamma)$ .

(Q) Let  $\mathbb{k}_\hbar := \mathbb{C}[[\hbar]]$  and write  $\omega_\hbar := \exp(-i\pi)/(2\hbar) \in \mathbb{k}_\hbar$  so that  $\mu: \mathbb{k} \rightarrow \mathbb{k}_\hbar$  defined by  $\mu(\omega) := \omega_\hbar$  is a ring morphism. Then the  $\mathbb{k}_\hbar$  algebra  $\mathcal{L}_\omega^{\text{inv}}(\Gamma) \otimes_\mu \mathbb{k}_\hbar$  is a deformation quantization of the Poisson algebra  $\mathbb{C}[\mathcal{X}_{\text{SL}_2}(\Sigma(\Gamma))]$  equipped with its Fock–Rosly Poisson structure.

**Theorem 4.9** (Alekseev, Grosse and Schomerus [2; 3; 5], Buffenoir and Roche [15; 16]) *Problem 4.8 admits the solution  $\mathcal{L}_\omega(\Gamma, c) := \mathcal{L}_\omega(\Sigma^0(\Gamma, c))$ , where the  $\mathbb{k}$ -module isomorphism  $\mathcal{L}_\omega(\Gamma, c) \cong \mathbb{C}[\mathcal{R}_{\text{SL}_2}(\Sigma^0(\Gamma, c))] \otimes_{\mathbb{C}} \mathbb{k}$  is given by sending  $U(\alpha)$  to  $X(\alpha)$ .*

The algebras  $\mathcal{L}_\omega(\Gamma, c)$  are the so-called *quantum moduli algebras* and Theorem 1.3 is an obvious consequence of Theorem 1.1.

More precisely, the ciliated graphs considered in [15; 16] are those whose underlying graph is the 1-skeleton of some combinatorial triangulation of a Riemann surface. By combinatorial we mean that each edge has two distinct endpoints, so every arc is of type  $a$  and the only arc exchange relations among distinct arcs are in configurations (i) or (ii) (in the notation of Lemma 2.24). In [2; 3; 5] general ciliated graphs are considered, though in [3; 5] special attention is given to the quantum moduli algebras of the daisy graphs defined in Example 2.13 (they are called *standard graphs* in [3; 5]) and are further studied and related to stated skein algebras in [27]. In those daisy graphs, the arcs are of type  $d$  and the more complicated arc exchange relations in configurations (viii), (ix) and (x) appear under the name braid relations; see [3, Definition 12].

Note that, except for the study of the Poisson structure (which could have been easily done), we reproved Theorem 4.9. In [43], Meusburger and Wise proved that the solution of Problem 4.8 is unique, provided that we add some natural axioms for the operation of gluing graphs together. Actually the authors of [43] consider quantum moduli algebras associated to finite-dimensional ribbon algebras, whereas here we consider the infinite-dimensional one  $U_q\mathfrak{sl}_2$ , but their proof extends word-for-word to our context.

#### 4.4 Comparison with previous works

Let  $\Sigma^0$  be a connected punctured surface with one boundary component, one puncture on its boundary and possibly some inner punctures. Let  $(\Gamma, c)$  be its daisy graph

and  $\mathbb{P} = (\mathbb{G}, \emptyset)$  be the associated finite presentation as defined in Example 2.13 (so  $\Sigma^0 = \Sigma^0(\Gamma, c)$ ). In this case, since the presentation has no relations, one can consider the spin function  $w$  sending every generator to  $0 \in \mathbb{Z}/2\mathbb{Z}$ . Since every generator  $\alpha \in \mathbb{G}$  is of type  $d$ , the isomorphism  $\Psi: \mathcal{S}_\omega(\Sigma^0) \xrightarrow{\cong} \mathcal{L}_\omega(\Gamma, c)$  sends  $U(\alpha) = C^{-1}M(\alpha)$  to  $X(\alpha)$ . By precomposing with the reflection anti-involution  $\theta$ , one obtains an isomorphism

$$\Psi': \mathcal{S}_{\omega^{-1}}(\Sigma^0)^{op} \xrightarrow{\cong} \mathcal{L}_\omega(\Gamma, c),$$

which corresponds to Faitg’s isomorphism in [27]. Let us stress that our notation is quite different from that in [27]; in particular:

- The letter  $q$  in [27] is what we denoted by  $A$  (so our  $q$  corresponds to  $q^2$  in [27]).
- The letter  $\mathcal{R}$  in [27] is related to our  $\mathcal{R}$  by  $\mathcal{R} = \tau \circ \mathcal{R}$ .
- Faitg actually considered  $\mathcal{S}_{\omega^{-1}}(\Sigma^0)^{op}$ , the opposite of the stated skein algebra.

As Faitg, Jordan and Safronov kindly explained to the author, the existence of an isomorphism  $\Psi: \mathcal{S}_\omega(\Sigma^0) \xrightarrow{\cong} \mathcal{L}_\omega(\Gamma, c)$  could have been derived from [9; 31] as we now briefly explain using the notation in [31] to which we refer for further details. Set  $\mathbb{k} = \mathbb{C}[\omega^{\pm 1}]$  and fix a structure of a Riemann surface  $\Sigma$ . To any  $\mathbb{k}$ -ribbon category  $\mathcal{A}$ , one can associate a skein category  $\text{SkCat}_{\mathcal{A}}(\Sigma)$  whose objects are oriented embeddings of finitely many disjoint discs  $\mathbb{D} \rightarrow \Sigma$  colored by objects in  $\mathcal{A}$  and whose morphisms are framed  $\mathcal{A}$ -colored ribbon graphs in  $\Sigma \times [0, 1]$  considered up to skein relations; see [24, Section 4.2] for a precise definition. We denote by  $\mathbb{1} \in \text{SkCat}_{\mathcal{A}}(\Sigma^0)$  the empty set. Let  $\Sigma^0$  be obtained from a connected closed oriented surface  $\Sigma$  by removing an open disc. Fixing an arbitrary disc embedding  $\mathbb{D} \rightarrow \Sigma^0$  gives a functor  $\mathcal{P}: \mathcal{A} \rightarrow \text{SkCat}_{\mathcal{A}}(\Sigma^0)$  in an obvious way. Let  $\hat{\mathcal{A}} := \text{Fun}(\mathcal{A}^{op}, \text{Vect})$  be the free cocompletion of  $\mathcal{A}$  (which inherits a monoidal structure from  $\mathcal{A}$ ). The *internal skein algebra* is defined as the coend

$$\text{SkCat}_{\mathcal{A}}^{\text{int}}(\Sigma^0) := \int^{x \in \mathcal{A}} \text{Hom}_{\text{SkCat}_{\mathcal{A}}(\Sigma^0)}(\mathcal{P}(x), \mathbb{1}) \otimes x \in \hat{\mathcal{A}}.$$

The functor  $\text{Hom}_{\text{SkCat}_{\mathcal{A}}(\Sigma^0)}(\mathcal{P}(\cdot), \mathbb{1}): \mathcal{A}^{op} \rightarrow \text{Vect}$  has a natural lax monoidal structure, given by stacking ribbon graphs on top of each other, which endows  $\text{SkCat}_{\mathcal{A}}^{\text{int}}(\Sigma^0)$  with the structure of an algebra object in  $\hat{\mathcal{A}}$ . If  $\mathcal{A}$  is Tannakian, that is if it is equipped with a fully faithful monoidal functor  $\text{for}: \mathcal{A} \rightarrow \text{Vect}$ , then

$$\mathcal{S}_{\mathcal{A}}(\Sigma^0) := \text{for}(\text{SkCat}_{\mathcal{A}}^{\text{int}}(\Sigma^0)) = \int^{x \in \mathcal{A}} \text{Hom}_{\text{SkCat}_{\mathcal{A}}(\Sigma^0)}(\mathcal{P}(x), \mathbb{1}) \otimes \text{for}(x) \in \text{Vect}$$

is a unital associative algebra that we might call the *stated skein algebra* associated to  $\mathcal{A}$  and  $\Sigma^0$ . Let us consider two Tannakian ribbon categories: the (Cauchy closure

of the Temperley–Lieb category  $\text{TL}$  and the category of finite-dimensional  $U_q \mathfrak{sl}_2$  left modules  $\text{Rep}_q^{\text{fd}}(\text{SL}_2)$  (recall that  $q$  is generic here). The Tannakian structure forget:  $\text{Rep}_q^{\text{fd}}(\text{SL}_2) \rightarrow \text{Vect}$  is just the forgetful functor. It is well known that one has a monoidal braided equivalence of categories (which does not preserve the pivotal structure)  $G: \text{TL} \rightarrow \text{Rep}_q^{\text{fd}}(\text{SL}_2)$  sending the one strand ribbon  $[1] \in \text{TL}$  to the fundamental representations  $V$  of Section 2.1 with basis  $\{v_+, v_-\}$ , thus we get a Tannakian structure forget  $\circ G: \text{TL} \rightarrow \text{Vect}$ .

On the one hand, there is a natural algebra morphism

$$\Psi_1: \mathcal{S}_\omega(\Sigma^0) \rightarrow \mathcal{S}_{\text{TL}}(\Sigma^0)$$

sending the class  $[T, s]$  of a stated tangle, where  $\partial T$  has  $n$  elements, to the class of  $T \otimes v_s \in \text{Hom}_{\text{SkCat}_{\text{TL}}(\Sigma^0)}(\mathcal{P}([1]^{\otimes n}), \mathbb{1}) \otimes V^{\otimes n}$ , where  $v_s \in V^{\otimes n}$  is obtained from the state  $s$  by identifying the signs  $+$  and  $-$  with the basis vectors  $v_+$  and  $v_-$  of  $V$ . As noted in [31, Remark 2.21] and fully explored in [32], a detailed comparison of the definitions shows that  $\Psi_1$  is an isomorphism.

On the other hand, thanks to Cooke’s excision theorem in [24] and as proved in [31, Proposition 2.19], the internal skein algebra  $\text{SkCat}_{\mathcal{A}}^{\text{int}}(\Sigma^0)$  is isomorphic to the so-called moduli algebra  $\mathcal{A}_{\Sigma^0} = \underline{\text{End}}(\mathbb{1}) \in \widehat{A}$  introduced in [9, Definition 5.3]. The authors of [9, Theorem 5.14] defined an explicit isomorphism  $[\text{Rep}_q^{\text{fd}}(\text{SL}_2)]_{\Sigma^0} \cong \mathcal{L}_\omega(\Gamma)$ , so by composing the two isomorphisms, one get an isomorphism

$$\Psi_2: \mathcal{S}_{\text{TL}}(\Sigma^0) \xrightarrow{\cong} \mathcal{L}_\omega(\Gamma).$$

Putting  $\Psi_1$  and  $\Psi_2$  together, we get an alternative construction of Faitg’s isomorphism.

**Remark 4.10** The above construction generalizes the notion of a stated skein algebra  $\mathcal{S}_\mathcal{C}(\Sigma^0)$  to an arbitrary Tannakian ribbon category  $\mathcal{C}$  (how to replace  $\Sigma^0$  with an arbitrary punctured surface is obvious), and [9, Theorem 5.14] seems to permit us to give explicit finite presentations for  $\mathcal{S}_\mathcal{C}(\Sigma^0)$ . A detailed study of these generalized stated skein algebras will appear in a separate publication [25].

## 5 Concluding remarks

We conclude the paper by making some remarks concerning the usefulness of relating stated skein algebras and quantum moduli spaces (Theorem 1.3). We can see the stated skein algebras as defined by a huge set of generators (all stated tangles) and a huge set



of relations (isotopy and skein relations) whereas the quantum moduli algebra is defined by a finite subset of generators and by a finite subset of relations. Both presentations have their own advantages.

(i) The fact that the quantum moduli algebra  $\mathcal{L}_\omega^{\text{inv}}(\Gamma)$  only depends, up to canonical isomorphism, on the thickened surface  $S(\Gamma)$  (or equivalently  $\Sigma(\Gamma)$ ) is usually proved by defining elementary moves on graphs that preserve the thickened surface and showing that those elementary moves induce isomorphisms on the algebras. This strategy was pioneered by Fock and Rosly in the classical case of relative character varieties [29] and later carried on in [3; 16] for quantum moduli algebras (see also [43] for very detailed study). Thanks to the isomorphism  $\mathcal{L}_\omega^{\text{inv}}(\Gamma) \cong \mathcal{S}_\omega(\Sigma(\Gamma))$  (and the fact that stated skein algebras depend on surfaces rather than graphs), this fact is also an immediate consequence of Theorem 1.3. Also, the image of a closed curve  $\gamma$  through the reverse isomorphism  $\Psi^{-1}: \Sigma(\Gamma) \rightarrow \mathcal{L}_\omega^{\text{inv}}(\Gamma)$  is usually called its *holonomy*  $\text{Hol}(\gamma)$  or *Wilson loop operators*, and the expression of this holonomy in terms of generators as well as the proof of some composition properties is the subject of long and technical computations in [2; 3; 15; 16; 28; 43], whereas they become easy in the skein algebra setting.

(ii) Since the quantum moduli algebra  $\mathcal{L}_\omega(\Gamma, c)$  is quadratic homogeneous, we might have tried to prove that it is Koszul (proving that  $\underline{\mathcal{B}}^{\mathbb{G}}$  is free) without the help of the stated skein algebra. The standard technique to prove that the family  $\mathcal{B}$  of (26) is a PBW basis consists in examining the set of critical monomials of the form  $v_i v_j v_k$  (we use the notation of Section 3.2) where both  $v_i v_j$  and  $v_j v_k$  are leading terms. To such a critical monomial we associate a finite graph (which might have the shape of a diamond) and the diamond lemma implies that if each of these graphs is confluent (has a terminal object) then  $\mathcal{B}$  is a basis, so the quadratic algebra is Koszul; see [42, Section 4] for details. In our case, due to the huge amount of different kinds of relations in our presentation, this strategy would require us to verify the confluence of 6578 different graphs! This is way too much to be handled by hand. It is thanks to the fact that stated skein algebras have a lot of relations and generators that Lê was able to successfully use the diamond lemma in [40] to prove that  $\mathcal{B}$  is basis, and our proof that  $\underline{\mathcal{B}}^{\mathbb{G}}$  is a basis is directly derived from this fact. So proving the Koszulness of  $\mathcal{L}_\omega(\Gamma, c)$  without the help of stated skein algebras could have been a very difficult problem.

(iii) Even if we could find PBW bases for the algebras  $\mathcal{L}_\omega(\Gamma, c)$  without the help of skein algebras, finding bases for  $\mathcal{L}_\omega^{\text{inv}}(\Gamma)$  would be extremely difficult, since it is only defined as a kernel and no presentation is known. However, skein algebras  $\mathcal{S}_\omega(\Sigma(\Gamma)) \cong \mathcal{L}_\omega^{\text{inv}}(\Gamma)$  have well-known bases (of multicurves).

(iv) As we saw in Section 4.2, the fact that  $\mathcal{L}_{+1}(\Sigma, \mathbb{P})$  is isomorphic to the algebra of regular functions of the (relative) character variety  $\mathcal{X}_{\mathrm{SL}_2}(\Sigma)$  is very easy to prove, whereas relating the (stated) skein algebra  $\mathcal{S}_{+1}(\Sigma)$  to  $\mathbb{C}[\mathcal{X}_{\mathrm{SL}_2}(\Sigma)]$  is not so obvious (see [17; 44] for closed surfaces and [26; 39] for open ones).

(v) In [13] Bonahon and Wong proved that the Kauffman-bracket skein algebra  $\mathcal{S}_{+1}(\Sigma)$ , with deforming parameter  $+1$ , embeds into the center of the skein algebra  $\mathcal{S}_\zeta(\Sigma)$  with deforming parameter  $\zeta$  a root of unity of odd order (see also [41] for an alternative proof). This result was generalized in [39] to stated skein algebras as well (see also [11] for generalizations). In [8], Baseilhac and Roche showed that the construction of this so-called Chebyshev–Frobenius morphism is much easier in the context of quantum moduli algebras (that is, using the finite presentations of Theorem 1.1). Even though their study only concerns genus 0 surfaces, their proofs seem to generalize easily to general surfaces, providing simpler proofs for the results in [13; 39].

(vi) Bullock, Frohman and Kania-Bartoszyńska already proved in [19, Theorem 10] that  $\mathcal{L}_\omega^{\mathrm{inv}}(\Gamma)$  and  $\mathcal{S}_\omega(\Sigma(\Gamma))$  are isomorphic when  $\mathbb{k} = \mathbb{C}[[\hbar]]$  and  $\omega = -\exp(-\frac{1}{4}\hbar)$ . Their proof consists of defining an algebra morphism  $\Psi: \mathcal{L}_\omega^{\mathrm{inv}}(\Gamma) \rightarrow \mathcal{S}_\omega(\Sigma(\Gamma))$  (by techniques similar to what we did in Section 2.2), and noting that under the (mod  $\hbar$ ) identifications  $\mathcal{L}_\omega^{\mathrm{inv}}(\Gamma)/(\hbar) \cong \mathbb{C}[\mathcal{X}_{\mathrm{SL}_2}(\Sigma)]$  and  $\mathcal{S}_\omega(\Sigma(\Gamma))/(\hbar) \cong \mathcal{S}_{-1}(\Sigma(\Gamma))$ , the morphism  $\Psi$  reduces modulo  $\hbar$  to Bullock’s isomorphism  $\mathbb{C}[\mathcal{X}_{\mathrm{SL}_2}(\Sigma)] \cong \mathcal{S}_{-1}(\Sigma(\Gamma))$ . So the fact that the reduction of  $\Psi$  modulo  $\hbar$  is an isomorphism implies that  $\Psi$  is an isomorphism. This proof does not seem (at least to the author) to generalize to prove the identification  $\mathcal{L}_\omega^{\mathrm{inv}}(\Gamma) \cong \mathcal{S}_\omega(\Sigma(\Gamma))$  for more general rings (such as  $\mathbb{k} = \mathbb{C}$  and  $\omega$  a root of unity), whereas our Theorem 1.3 works in full generality. A second reason why the approach in [19] does not work at roots of unity is described in (vii).

(vii) The following important remark was kindly explained to us by the anonymous referee, whom the author warmly thanks. In traditional papers in lattice gauge field theory (like [3; 8]) the algebras  $\mathcal{L}_\omega(\Gamma, c)$  are seen as  $U_q\mathrm{sl}_2^{\otimes n}$ -modules instead of  $\mathcal{O}_q[\mathrm{SL}_2]^{\otimes n}$ -comodules (here  $n$  is the number of external vertices of  $\Gamma$ , ie the number of boundary arcs of  $\Sigma(\Gamma, c)$ ) and  $\mathcal{L}_\omega^{\mathrm{inv}}(\Gamma)$  is then defined as the algebra of  $U_q\mathrm{sl}_2^{\otimes n}$ -invariant vectors instead of  $\mathcal{O}_q[\mathrm{SL}_2]^{\otimes n}$ -coinvariant vectors. When  $q$  is generic, there is a perfect pairing between the two Hopf algebras  $U_q\mathrm{sl}_2$  and  $\mathcal{O}_q[\mathrm{SL}_2]$  so that both definitions coincide. However, at roots of unity, the induced morphism  $\mathcal{O}_q[\mathrm{SL}_2] \rightarrow U_q\mathrm{sl}_2^\circ$  is no longer injective nor surjective. As a consequence, the two definitions of  $\mathcal{L}_\omega^{\mathrm{inv}}(\Gamma)$  do not coincide anymore and Theorem 1.3 only holds for the definition used in the present

paper. For instance, consider the case where  $(\Gamma, c) = \text{⋈}$  so that  $\mathbf{m}_1 := \Sigma(\Gamma, c)$  is a once-punctured monogon, that is, a disc with one inner puncture and one boundary puncture. In this case  $\mathcal{S}_\omega(\mathbf{m}_1) \cong \mathcal{L}_\omega(\Gamma, c)$  is Majid's braided quantum group; see [26; 8]. On the one hand, when  $q := \omega^{-4}$  is a root of unity of odd order, Baseilhac and Roche have proved [8, page 41] that the subalgebra of  $U_q\mathfrak{sl}_2$ -invariant vectors coincides with the center of  $\mathcal{S}_\omega(\mathbf{m}_1)$  (denoted by  $\mathcal{L}_{0,1}^\varepsilon$  in [8]). This center is generated by the peripheral curve  $\gamma_p$  encircling the inner puncture  $p$  together with the image of the Chebyshev–Frobenius morphism. On the other hand, the  $\mathcal{O}_q[\text{SL}_2]$ -coinvariant vectors form the algebra  $\mathbb{C}[\gamma_p]$  generated by the peripheral curve, isomorphic to the skein algebra of a punctured disc  $\Sigma^0(\Gamma, c)$  as expected. Therefore the subalgebra of  $U_q\mathfrak{sl}_2$ -invariant vectors is bigger than the algebra of  $\mathcal{O}_q[\text{SL}_2]$ -coinvariant vectors and Theorem 1.3 would fail with the original definition of  $\mathcal{L}_\omega^{\text{inv}}(\Gamma)$  at roots of unity.

## References

- [1] **N Abdel, C Frohman**, *The localized skein algebra is Frobenius*, *Algebr. Geom. Topol.* 17 (2017) 3341–3373 MR Zbl
- [2] **A Y Alekseev, H Grosse, V Schomerus**, *Combinatorial quantization of the Hamiltonian Chern–Simons theory, I*, *Comm. Math. Phys.* 172 (1995) 317–358 MR Zbl
- [3] **A Y Alekseev, H Grosse, V Schomerus**, *Combinatorial quantization of the Hamiltonian Chern–Simons theory, II*, *Comm. Math. Phys.* 174 (1996) 561–604 MR Zbl
- [4] **A Y Alekseev, A Z Malkin**, *Symplectic structure of the moduli space of flat connection on a Riemann surface*, *Comm. Math. Phys.* 169 (1995) 99–119 MR Zbl
- [5] **A Y Alekseev, V Schomerus**, *Representation theory of Chern–Simons observables*, *Duke Math. J.* 85 (1996) 447–510 MR Zbl
- [6] **M Audin**, *Lectures on gauge theory and integrable systems*, from “Gauge theory and symplectic geometry” (J Hurtubise, F Lalonde, G Sabidussi, editors), *NATO Adv. Sci. Inst. Ser. C: Math. Phys. Sci.* 488, Kluwer, Dordrecht (1997) 1–48 MR Zbl
- [7] **J W Barrett**, *Skein spaces and spin structures*, *Math. Proc. Cambridge Philos. Soc.* 126 (1999) 267–275 MR Zbl
- [8] **S Baseilhac, P Roche**, *Unrestricted quantum moduli algebras, I: The case of punctured spheres*, *SIGMA Symmetry Integrability Geom. Methods Appl.* 18 (2022) art. id. 025 MR Zbl
- [9] **D Ben-Zvi, A Brochier, D Jordan**, *Integrating quantum groups over surfaces*, *J. Topol.* 11 (2018) 874–917 MR Zbl
- [10] **G M Bergman**, *The diamond lemma for ring theory*, *Adv. in Math.* 29 (1978) 178–218 MR Zbl

- [11] **W Bloomquist, T T Q Lê**, *The Chebyshev–Frobenius homomorphism for stated skein modules of 3–manifolds*, *Math. Z.* 301 (2022) 1063–1105 MR Zbl
- [12] **F Bonahon, H Wong**, *Quantum traces for representations of surface groups in  $SL_2(\mathbb{C})$* , *Geom. Topol.* 15 (2011) 1569–1615 MR Zbl
- [13] **F Bonahon, H Wong**, *Representations of the Kauffman bracket skein algebra, I: Invariants and miraculous cancellations*, *Invent. Math.* 204 (2016) 195–243 MR Zbl
- [14] **K A Brown, K R Goodearl**, *Lectures on algebraic quantum groups*, Birkhäuser, Basel (2002) MR Zbl
- [15] **E Buffenoir, P Roche**, *Two-dimensional lattice gauge theory based on a quantum group*, *Comm. Math. Phys.* 170 (1995) 669–698 MR Zbl
- [16] **E Buffenoir, P Roche**, *Link invariants and combinatorial quantization of Hamiltonian Chern–Simons theory*, *Comm. Math. Phys.* 181 (1996) 331–365 MR Zbl
- [17] **D Bullock**, *Rings of  $SL_2(\mathbb{C})$ –characters and the Kauffman bracket skein module*, *Comment. Math. Helv.* 72 (1997) 521–542 MR Zbl
- [18] **D Bullock**, *A finite set of generators for the Kauffman bracket skein algebra*, *Math. Z.* 231 (1999) 91–101 MR Zbl
- [19] **D Bullock, C Frohman, J Kania-Bartoszyńska**, *Topological interpretations of lattice gauge field theory*, *Comm. Math. Phys.* 198 (1998) 47–81 MR Zbl
- [20] **D Bullock, J Kania-Bartoszyńska, C Frohman**, *Skein quantization and lattice gauge field theory*, from “Knot theory and its applications” (C Adams, editor), 4–5, Elsevier, Oxford (1998) 811–824 MR Zbl
- [21] **D Bullock, J H Przytycki**, *Multiplicative structure of Kauffman bracket skein module quantizations*, *Proc. Amer. Math. Soc.* 128 (2000) 923–931 MR Zbl
- [22] **V Chari, A Pressley**, *A guide to quantum groups*, Cambridge Univ. Press (1995) MR Zbl
- [23] **L Charles, J Marché**, *Multicurves and regular functions on the representation variety of a surface in  $SU(2)$* , *Comment. Math. Helv.* 87 (2012) 409–431 MR Zbl
- [24] **J Cooke**, *Excision of skein categories and factorisation homology*, *Adv. Math.* 414 (2023) art. id. 108848 MR Zbl
- [25] **F Costantino, J Korinman, T T Q Lê**, *Stated skein algebras for Tannakian ribbon categories*, in preparation
- [26] **F Costantino, T T Q Lê**, *Stated skein algebras of surfaces*, *J. Eur. Math. Soc.* 24 (2022) 4063–4142 MR Zbl
- [27] **M Faitg**, *Holonomy and (stated) skein algebras in combinatorial quantization*, preprint (2020) arXiv 2003.08992
- [28] **M Faitg**, *Projective representations of mapping class groups in combinatorial quantization*, *Comm. Math. Phys.* 377 (2020) 161–198 MR Zbl

- [29] **V V Fock, A A Rosly**, *Poisson structure on moduli of flat connections on Riemann surfaces and the  $r$ -matrix*, from “Moscow Seminar in Mathematical Physics” (A Y Morozov, M A Olshanetsky, editors), Amer. Math. Soc. Transl. Ser. 2 191, Amer. Math. Soc., Providence, RI (1999) 67–86 MR Zbl
- [30] **C Frohman, J Kania-Bartoszynska**, *The structure of the Kauffman bracket skein algebra at roots of unity*, Math. Z. 289 (2018) 889–920 MR Zbl
- [31] **S Gunningham, D Jordan, P Safronov**, *The finiteness conjecture for skein modules*, Invent. Math. 232 (2023) 301–363 MR Zbl
- [32] **B Haïoun**, *Relating stated skein algebras and internal skein algebras*, SIGMA Symmetry Integrability Geom. Methods Appl. 18 (2022) art. id. 042 MR Zbl
- [33] **J Kamnitzer, P Tingley**, *The crystal commutator and Drinfeld’s unitarized  $R$ -matrix*, J. Algebraic Combin. 29 (2009) 315–335 MR Zbl
- [34] **A N Kirillov, N Reshetikhin**,  *$q$ -Weyl group and a multiplicative formula for universal  $R$ -matrices*, Comm. Math. Phys. 134 (1990) 421–431 MR Zbl
- [35] **J Korinman**, *Quantum groups and braiding operators in quantum Teichmüller theory*, preprint (2019) arXiv 1907.01732
- [36] **J Korinman**, *Triangular decomposition of character varieties*, preprint (2019) arXiv 1904.09022
- [37] **J Korinman**, *Stated skein algebras and their representations*, preprint (2021) arXiv 2105.09563
- [38] **J Korinman**, *Unicity for representations of reduced stated skein algebras*, Topology Appl. 293 (2021) art. id. 107570 MR Zbl
- [39] **J Korinman, A Quesney**, *Classical shadows of stated skein representations at roots of unity*, preprint (2019) arXiv 1905.03441
- [40] **T T Q Lê**, *Triangular decomposition of skein algebras*, Quantum Topol. 9 (2018) 591–632 MR Zbl
- [41] **T T Q Lê**, *Quantum Teichmüller spaces and quantum trace map*, J. Inst. Math. Jussieu 18 (2019) 249–291 MR Zbl
- [42] **J-L Loday, B Vallette**, *Algebraic operads*, Grundleh. Math. Wissen. 346, Springer (2012) MR Zbl
- [43] **C Meusburger, D K Wise**, *Hopf algebra gauge theory on a ribbon graph*, Rev. Math. Phys. 33 (2021) art. id. 2150016 MR Zbl
- [44] **J H Przytycki, A S Sikora**, *On skein algebras and  $Sl_2(\mathbb{C})$ -character varieties*, Topology 39 (2000) 115–148 MR Zbl
- [45] **N Reshetikhin, V G Turaev**, *Invariants of 3-manifolds via link polynomials and quantum groups*, Invent. Math. 103 (1991) 547–597 MR Zbl

- [46] **R Santharoubane**, *Algebraic generators of the skein algebra of a surface*, preprint (2018) arXiv 1803.09804
- [47] **N Snyder, P Tingley**, *The half-twist for  $U_q(\mathfrak{g})$  representations*, Algebra Number Theory 3 (2009) 809–834 MR Zbl
- [48] **DP Thurston**, *Positive basis for surface skein algebras*, Proc. Natl. Acad. Sci. USA 111 (2014) 9725–9732 MR Zbl
- [49] **P Tingley**, *A minus sign that used to annoy me but now I know why it is there (two constructions of the Jones polynomial)*, from “Proceedings of the 2014 Maui and 2015 Qinhuangdao conferences in honour of Vaughan F R Jones’ 60th birthday” (S Morrison, D Penneys, editors), Proc. Centre Math. Appl. Austral. Nat. Univ. 46, Austral. Nat. Univ., Canberra (2017) 415–427 MR Zbl
- [50] **V G Turaev**, *Skein quantization of Poisson algebras of loops on surfaces*, Ann. Sci. École Norm. Sup. 24 (1991) 635–704 MR Zbl
- [51] **E Witten**, *Quantum field theory and the Jones polynomial*, Comm. Math. Phys. 121 (1989) 351–399 MR Zbl

*Departement of Mathematics, Waseda University  
Tokyo, Japan*

julien.korinman@gmail.com

<https://sites.google.com/site/homepagejulienkorinman/>

Received: 4 January 2021      Revised: 8 September 2021

# On the functoriality of $\mathfrak{sl}_2$ tangle homology

ANNA BELIAKOVA

MATTHEW HOGANCAMP

KRZYSZTOF K PUTYRA

STEPHAN M WEHRLI

We construct an explicit equivalence between the (bi)category of  $\mathfrak{gl}_2$  webs and foams and the Bar-Natan (bi)category of Temperley–Lieb diagrams and cobordisms. With this equivalence we can fix functoriality of every link homology theory that factors through the Bar-Natan category. To achieve this, we define web versions of arc algebras and their quasihereditary covers, which provide strictly functorial tangle homologies. Furthermore, we construct explicit isomorphisms between these algebras and the original ones based on Temperley–Lieb cup diagrams. The immediate application is a strictly functorial version of the Beliakova–Putyra–Wehrli quantization of the annular link homology.

57K18; 18N25

1. Introduction	1303
2. Main players	1313
3. Shadings and a basis of foams	1324
4. Equivalences of foam and cobordism categories	1331
5. A diagrammatic TQFT on <b>Foam</b> ( $\emptyset$ )	1338
6. The Blanchet–Khovanov invariant	1346
7. Subquotient algebras and an invariant for all tangles	1352
References	1359

## 1 Introduction

In 1999 Khovanov [18] defined for any link in the 3–sphere a chain complex, whose homotopy type — hence, homology — is a link invariant and whose Euler characteristic is the Jones polynomial. It was later extended by Khovanov to tangles between even

collections of points [19] and then to all tangles by Brundan and Stroppel [6] and Chen and Khovanov [10]. The main advantage of the Khovanov homology with respect to the Jones polynomial is that link cobordisms induce chain maps between Khovanov's complexes; see Bar-Natan [2], Jacobsson [17], and Khovanov [22]. Even though the original construction is not strictly functorial — the sign of the chain map associated with a link cobordism depends on the decomposition of the cobordism into elementary pieces [17] — it was used by Rasmussen to provide a lower bound for the slice genus of a knot and a combinatorial proof of the Milnor conjecture [32].

In the last 15 years there were many attempts to fix the functoriality of Khovanov homology. In Caprau [8] and Clark, Morrison and Walker [11] this was done by modifying the Bar-Natan category [2], in which the construction of the complex can be naturally described, by taking into account orientations and by enlarging the ground ring; see also Vogel [35]. In 2014 Blanchet [5] proposed a more conceptual solution, which does not change the ring of scalars, but replaces circles and surfaces in the Bar-Natan category with *webs* and *foams*: certain planar trivalent graphs and singular cobordisms between them respectively. This construction, commonly referred to as  $\mathfrak{gl}_2$  homology, has been widely accepted as the most natural way to fix functoriality of Khovanov homology. The resulting, a priori potentially different,  $\mathfrak{gl}_2$  homology is known to coincide with Khovanov homology in case of links [5].

To obtain a computable invariant of tangles, Chen and Khovanov constructed a functor from the Bar-Natan bicategory to the bicategory of bimodules over extended Khovanov's arc algebras [10]. Because of its diagrammatic definition, it is straightforward to generalize this functor to the case of webs. However, it is no longer clear whether the new algebra or tangle invariant is isomorphic to those constructed by Chen and Khovanov. A partial solution to this problem, that considers only a special class of webs, was presented by Ehrig, Stroppel and Tubbenhauer [13; 14], but not much was known beyond this case. The special webs from [13; 14] are discussed in Example 6.6 below.

The Hochschild homology of the Chen–Khovanov invariant of an  $(n, n)$ -tangle  $T$  has been identified by Beliakova, Putyra and Wehrli in [4] with the annular Khovanov homology of the annular closure  $\hat{T}$  of the tangle. In the same paper the annular invariant has been quantized by deforming the Hochschild homology. The original goal of this paper was to make this quantized annular homology functorial, in order to construct its colored versions following Khovanov [21] and Cooper and Krushkal [12]. These quantized colored homologies are constructed in the follow up paper by the authors [3],



where we also show that both complexes coincide when the deformation parameter is generic; see also Putyra [28].

In order to obtain a strictly functorial quantized annular homology, we wanted first to understand the Ehrig–Stroppel–Tubbenhauer isomorphism between Khovanov’s arc algebras and their web algebras, and then reconstruct the Chen–Khovanov functor in the framework of webs and foams. However, after a chain of simplifications of their arguments, especially replacing the foam basis used in [13] with another one, more natural from the topological perspective, we understood the real reason why all the isomorphisms popped out: *foams and cobordisms constitute equivalent bicategories*. Despite sounding as a natural thing, the result is by no means obvious — compare eg Lauda, Queffelec and Rose [23, Section 3.1.2] or [14, page 186].

In this paper we construct a functor between foams and cobordisms by taking into account orientability of foams and another beautiful topological tool called *shadings*. They allow to think of a web (resp. a foam) as two transversely intersecting flat tangles (resp. surfaces), so that each of the two tangles or surfaces can be isotoped separately (see Lemma 3.6, the bicolored isotopy lemma). This approach makes many results on foams straightforward, but also leads to a simple diagrammatic representation of a basis of the space of foams bounded by a fixed web, on which the action of foams is very easy to compute. We use this basis to construct explicitly equivalences between the two bicategories in both directions and to obtain full web versions of the TQFT functors from [6; 10; 19].

To summarize, the above mentioned equivalence between foams and cobordisms gives an ultimate solution to all functoriality issues related with Khovanov homology. Our recipe is simple: precompose any link or tangle homology theory that factors through the Bar-Natan category with our equivalence and obtain a strictly functorial theory. In the following sections we discuss the above approach in more details.

## 1.1 The equivalence of foams and Bar-Natan cobordisms

In order to compute the Khovanov homology of a link  $L$ , one first picks its diagram  $D$  and constructs the *cube of resolutions* of  $D$ : a commutative diagram in the shape of the  $c$ -dimensional cube, where  $c$  equals the number of crossings in  $D$ , with vertices decorated by Kauffman resolutions of  $D$  and edges by saddle cobordisms between them [18]. Applying a 2-dimensional TQFT to this cube, changing signs of some maps, and collapsing the cube along diagonals results in an actual chain complex, which —

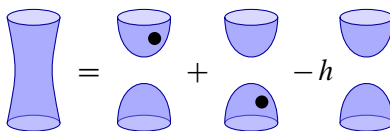
depending on the choice of the TQFT functor—computes the Khovanov homology of  $L$  or its deformation.

It was observed by Bar-Natan that most of the construction can be performed formally *before* applying a TQFT functor to get an invariant of a tangle  $T$  in the form of a formal complex  $[[T]]$  called the *Khovanov bracket* of  $T$  [2]. This complex is constructed in the *Bar-Natan bicategory*  $\mathbf{BN}$ , the locally additive graded bicategory with objects collections of points on a line, 1–morphisms generated by flat tangles, and 2–morphisms generated by isotopy classes of surfaces embedded in a 3–space and decorated by dots, modulo the following local relations:

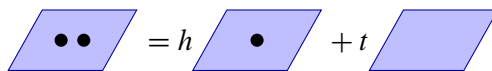
- *sphere evaluations:*

(1-1) 

- *neck-cutting relation:*

(1-2) 

- *dot-reduction:*

(1-3) 

Here  $h$  and  $t$  are fixed elements of the ring of scalars  $\mathbb{k}$ . When  $h = 0$ , then the neck-cutting relation evaluates a handle attached to a plane as a dot scaled by 2. Because of that it is common to think of a dot as “half” of a handle, even when 2 is not an invertible scalar. However, this interpretation is not correct if  $h \neq 0$ , in particular in the universal case  $\mathbb{k} = \mathbb{Z}[h, t]$ .

The formal bracket is projectively functorial [2]. Indeed, there is a way to associate a formal chain map with each Reidemeister move as well as any cobordism with a unique critical point. One constructs a formal chain map for any tangle cobordism by decomposing the cobordism into a sequence of the above elementary pieces and composing the associated maps; choosing a different decomposition may at most change the global sign of the map.

In Blanchet’s construction [5] the role of flat tangles is played by  $\mathfrak{gl}_2$ -webs, trivalent graphs with each edge colored blue or red,<sup>1</sup> and dotted surfaces are replaced with

<sup>1</sup>When compared to [5], blue edges are those with label 1 and red edges are those with label 2.

foams, which are singular cobordisms with each facet also colored blue or red. They constitute a bicategory **Foam**, where certain local relations between foams, including (1-1)–(1-3), are imposed; see Definition 2.6. Following [2] we can construct a formal complex  $[[T]]_{\mathbb{F}}$  in **Foam**, which we refer to as the *Blanchet–Khovanov bracket*.

The collection of blue edges of a web  $\omega$  is a flat tangle  $\omega_b$ , which we call the *underlying tangle* of  $\omega$ . Likewise, there is an *underlying surface*  $S_b$  associated with any foam  $S$ . It is tempting to consider a 2–functor **Foam**  $\rightarrow$  **BN** that forgets red edges in webs and red facets in foams. However, this operation is not compatible with relations between foams, and it is not clear at first how to solve this problem. For instance, it was observed in [23] that if such a functor exists, then it cannot be the identity on all foams with no red facets.

We resolved the above problem by taking into account the orientation of blue edges and facets. Shortly speaking, we fix an orientation for each flat tangle and surface in a canonical way, reinterpreting them as webs and foams respectively (recall that tangles and surfaces from **BN**, though orientable, come with no particular orientation). This results in a 2–functor, which, however, does not reach every object of **Foam**. In order to fix this we replace **BN** with the product  $\mathbf{wBN} := \mathbf{BN} \times \mathbb{Z}$ , where  $\mathbb{Z}$  is seen as a discrete bicategory. We use the extra integer to determine how many red points, edges, or facets have to be added to the right of the oriented blue points, tangle, or surface respectively.<sup>2</sup> This way we end up with a 2–functor  $\mathcal{E} : \mathbf{wBN} \rightarrow \mathbf{Foam}$ , such that every object of **Foam** is equivalent to one from the image of  $\mathcal{E}$ .

**Theorem A** *The 2–functor  $\mathcal{E} : \mathbf{wBN} \rightarrow \mathbf{Foam}$  is an equivalence of bicategories.*

From the point of view of representation theory,  $\mathcal{E}$  and its inverse can be understood as the categorification of the induction–restriction pair between representations of  $\mathfrak{sl}_2$  and  $\mathfrak{gl}_2$ .

There is also a local version of Theorem A. Having fixed a collection  $\Sigma$  of oriented blue and red points on  $\partial\mathbb{D}^2$ , write **Foam**( $\Sigma$ ) for the category of webs in  $\mathbb{D}^2$  bounded by  $\Sigma$  and foams in  $\mathbb{D}^2 \times [0, 1]$  between such webs. Likewise we consider the category **BN**( $\Sigma_b$ ) of flat tangles bounded by  $\Sigma_b$  and dotted surfaces between them, where  $\Sigma_b$  is the collection of blue points from  $\Sigma$ . We construct a functor  $\mathcal{E}_{\Sigma} : \mathbf{BN}(\Sigma_b) \rightarrow \mathbf{Foam}(\Sigma)$  in Section 4.1 by extending coherently all flat tangles to webs bounded by  $\Sigma$  and surfaces to foams.

<sup>2</sup>Compare this with the relation between the weight lattices of  $\mathfrak{sl}_2$  and  $\mathfrak{gl}_2$  — the latter is isomorphic to the product of the former with  $\mathbb{Z}$ .

**Theorem B** *The functor  $\mathcal{E}_\Sigma: \mathbf{BN}(\Sigma_b) \rightarrow \mathbf{Foam}(\Sigma)$  is an equivalence of categories.*

We construct the functor  $\mathcal{E}_\Sigma$  explicitly as well as its inverse  $\mathcal{E}_\Sigma^\vee$ . The latter not only forgets red facets of foams, but also scales them by a sign when necessary; we provide an explicit way to compute these signs in terms of the Blanchet evaluation of foams. When combined with a homological argument presented in Ozsváth, Rasmussen and Szabó [26] and Putyra [27], Theorem B implies that for every tangle  $T$  the image of the Khovanov bracket  $\llbracket T \rrbracket$  under  $\mathcal{E}_\Sigma$  is isomorphic to the Blanchet–Khovanov bracket  $\llbracket T \rrbracket_F$ . Hence, any TQFT functor on  $\mathbf{BN}(\Sigma_b)$  that leads to a tangle or link homology can be precomposed with  $\mathcal{E}_\Sigma^\vee$  to obtain a functor on  $\mathbf{Foam}(\Sigma)$  that computes the same homology groups, but which is strictly functorial with respect to tangle cobordisms.

**1.1.1 Main tools: shadings and bicolored isotopies** The key step in the proofs of Theorems A and B is to understand how foams with the same underlying surface are related. We achieve this by constructing foams from *shadings*. A shading is a union of two possibly intersecting surfaces: a nonoriented blue and an oriented red one, that are in general position in  $\mathbb{R}^3$ , together with a checkerboard black and white coloring of the connected components of their complement, called regions. Forgetting those red facets of a shading, the orientations of which disagree with the one induced from the white regions, results in a foam, and all foams can be constructed this way. The same applies to webs.

A particularly nice feature of representing foams by shadings is the flexibility of this construction, which we call the *bicolored isotopy argument*: deforming any of the two surfaces by an isotopy results in a foam that differs from the original one only up to a sign or replacing some dots with their duals; see Proposition 2.10 in Section 2.2 for a precise statement. This has a number of important consequences:

- closed foams can be evaluated (Theorem 2.14) using the bicolored isotopy argument by moving the blue and red facets away from each other;
- more generally, foams with the same boundary and underlying surfaces coincide up to a sign and types of dots (Proposition 2.10);
- a foam, the underlying surface of which is a product  $\omega \times [0, 1]$ , is invertible.

We then use the above to construct a basis of the space of foams bounded by a closed web  $\omega$ . It is given in terms of shadings of a plane that extend  $\omega$ , the blue loops of which may carry dots. The foam associated with such a picture  $\omega^+$  is given by attaching blue and red cups to the loops of  $\omega^+$  — red cups above all blue ones — and placing

a dot at the minimum of every blue cup attached to a loop that is marked by a dot. This leads to an explicit description of the tautological TQFT functor on  $\mathbf{Foam}(\emptyset)$  that associates the space  $\text{Hom}_{\mathbf{Foam}(\emptyset)}(\emptyset, \omega)$  with a closed web  $\omega$ , presented in Section 5. When compared with [14], our basis is not only easier to visualize, but also the formula for the action of foams involves fewer signs.

### 1.2 Functorial tangle homology

Khovanov extended his construction first to tangles with an even number of boundary points at each side [19]. For this he constructed a 2–functor  $\mathcal{F}_{Kh}^\circ : \mathbf{BN}^\circ \rightarrow \mathbf{Bimod}$ , where  $\mathbf{BN}^\circ$  is the subcategory of  $\mathbf{BN}$  with only even collections of points as objects. The 2–functor  $\mathcal{F}_{Kh}^\circ$  associates with a collection of  $2n$  points the *arc algebra*

$$(1-4) \quad H^n := \bigoplus_{a,b} \text{Hom}_{\mathbf{BN}}(a, b),$$

where  $a$  and  $b$  run through the set of Temperley–Lieb cup diagrams in  $\mathbb{R} \times (-\infty, 0]$  with  $2n$  boundary points at the top boundary line.<sup>3</sup> This algebra  $H^n$  is known to categorify the invariant subspace  $\text{Inv}(V^{\otimes n})$  of  $V^{\otimes n}$ , where  $V$  is the fundamental representation of  $\mathcal{U}_q(\mathfrak{sl}_2)$ . Cup diagrams parametrize indecomposable projective  $H^n$ –modules, which in turn correspond to elements of the canonical basis of  $V^{\otimes n}$ . Let  $CKh(T)$  be the chain complex associated with an  $(2n, 2n')$ –tangle  $T$ , ie the result of applying  $\mathcal{F}_{Kh}^\circ$  to  $\llbracket T \rrbracket$ . The functors  $CKh(T) \otimes (-)$  lift the action of tangles on  $\text{Inv}(V^{\otimes n})$  to the derived categories of the arc algebras [19].

In order to categorify the whole tensor power  $V^{\otimes n}$ , Chen and Khovanov considered a family of algebras  $A^{k,n-k}$ , where  $0 \leq k \leq n$ , each constructed as a subquotient of  $H^n$ . These algebras were discovered independently by Stroppel [34], who proved with Brundan [6; 7] that they are quasihereditary covers of arc algebras and Koszul. Furthermore, projective modules over  $A^{k,n-k}$  categorify the weight space  $V^{\otimes n}(\lambda)$  with  $\lambda = n - 2k$  [7; 10]. As in the case of arc algebras, there is a family of 2–functors  $\mathcal{F}_{Kh}^\lambda : \mathbf{BN} \rightarrow \mathbf{Bimod}$ , such that  $\mathcal{F}_{Kh}^\lambda$  assigns to a collection of  $n$  points the algebra  $A^{k,n-k}$  with  $\lambda = n - 2k$  [6; 10]. Write  $CKh(T; \lambda)$  for the result of applying  $\mathcal{F}_{Kh}^\lambda$  to  $\llbracket T \rrbracket$ . Then the functor  $CKh(T; \lambda) \otimes (-)$  lifts the action of  $T$  on the weight space  $V^{\otimes n}(\lambda)$ .

Using Theorem A we can construct a strictly functorial version of both Khovanov and Chen–Khovanov homologies by precomposing  $\mathcal{F}_{Kh}^\circ$  and  $\mathcal{F}_{Kh}^\lambda$  with  $\mathcal{E}^\vee$ . We provide a direct construction of both invariants.

<sup>3</sup>This presentation of  $H^n$  comes from Rozansky [33].

Following [14] we call the web version of  $H^n$  the *Blanchet–Khovanov algebra*. It is defined for any collection of oriented red and blue points  $\Sigma$  that is *balanced*, ie bounds a web, as the direct sum

$$\mathfrak{W}^{\mathcal{B}} := \bigoplus_{a,b \in \mathcal{B}} \text{Hom}_{\mathbf{Foam}(\Sigma)}(a, b),$$

where  $\mathcal{B}$  is a *cup basis* of webs bounded by  $\Sigma$ ; its elements play the role of cup diagrams for  $H^n$ . Although  $\mathfrak{W}^{\mathcal{B}}$  depends a priori on  $\mathcal{B}$ , we show that different choices of basis lead to isomorphic algebras. Moreover, there is a special basis of webs — the *red-over-blue* basis — such that forgetting red facets in cup foams is compatible with multiplication. In particular,  $\mathfrak{W}^{\mathcal{B}}$  admits a *positive basis*. This results immediately in an algebra isomorphism  $\mathfrak{W}^{\mathcal{B}} \cong H^n$ , where  $n$  is half of the blue points in  $\Sigma$ . We further extend this construction to a 2–functor  $\mathcal{F}_w^\circ : \mathbf{Foam}^\circ \rightarrow \mathbf{Bimod}$  following the construction of  $\mathcal{F}_{Kh}^\circ$ .

Suppose that  $T$  is an oriented tangle, the input and output of which are balanced. Then all resolutions of  $T$  are in  $\mathbf{Foam}^\circ$  and  $\mathcal{F}_w^\circ$  can be applied to  $\llbracket T \rrbracket_{\mathbb{F}}$  to produce a chain complex of bimodules  $C_{\mathfrak{W}}(T)$ . We call it the *Blanchet–Khovanov complex*.

**Theorem C** *The 2–functor  $\mathcal{F}_w^\circ$  is equivalent to  $\mathcal{F}_{Kh}^\circ \circ \mathcal{E}^\vee$ . In particular, the complexes  $C_{\mathfrak{W}}(T)$  and  $CKh(T)$  are isomorphic for any tangle  $T$  with balanced input and output.*

The construction of a web version of Chen–Khovanov algebras is more challenging. We first describe two extensions of a sequence  $\Sigma$  to a balanced one  $\Sigma^\circ$  by inserting extra blue points to the left and to the right of  $\Sigma$ . Then we pick a basis  $\mathcal{B}$  of webs bounded by  $\Sigma^\circ$  and the corresponding Blanchet–Khovanov algebra  $\mathfrak{W}^{\mathcal{B}}$ . The *extended Blanchet–Khovanov algebra*  $\mathfrak{A}^{\Sigma, \lambda}$ , where  $\lambda \in \mathbb{Z}$  has the same parity as the number of blue points in  $\Sigma$ , is a certain subquotient of  $\mathfrak{W}^{\mathcal{B}}$ . Following the same procedure we associate a bimodule with a web and a bimodule map with a foam for every  $\lambda \in \mathbb{Z}$ , obtaining a family of 2–functors  $\mathcal{F}_w^\lambda : \mathbf{Foam} \rightarrow \mathbf{Bimod}$ , each defined on the entire foam bicategory. As in the previous construction,  $\mathcal{F}_w^\lambda$  is compatible with relations between foams, so that applying it to  $\llbracket T \rrbracket_{\mathbb{F}}$  results in an invariant chain complex of bimodules  $C_{\mathfrak{W}}(T; \lambda)$ . We call it the *extended Blanchet–Khovanov complex* of  $T$ .

We construct an explicit isomorphism  $\mathfrak{A}^{\mathcal{B}, \lambda} \cong A^{k, n-k}$ , where  $n$  counts blue points in  $\Sigma$  and  $\lambda = n - 2k$ . Contrary to the previous case, it is not enough to forget red facets in cup foams to get the isomorphism, because the basic webs from  $\mathcal{B}$  may have too many blue arcs. This issue is resolved by *stabilization* — adding beneath webs and foams

extra blue arcs and disks respectively. We then extend this isomorphism to bimodules and prove the following fact:

**Theorem D** *The 2–functor  $\mathcal{F}_w^\lambda$  is equivalent to  $\mathcal{F}_{Kh}^\lambda \circ \mathcal{E}^\vee$ . In particular, the complexes  $C_{\mathfrak{M}}(T; \lambda)$  and  $CKh(T; \lambda)$  are isomorphic for any tangle  $T$ .*

All the isomorphisms are constructed explicitly and — in case nice bases are used — given by very simple formulas. Furthermore, by the discussion following Theorem B, the tangle homology computed with  $\mathcal{F}_w^\circ$  and  $\mathcal{F}_w^\lambda$  are isomorphic to the Khovanov and Chen–Khovanov invariants respectively.

### 1.3 Functoriality of quantized annular Khovanov homology

The above results allow us to construct a strictly functorial version of the quantized annular Khovanov homology, which was the motivation for this paper. Combining Theorem D with [4, Proposition 6.6] we get:

**Corollary E** *Suppose  $\mathbb{k}$  is flat over  $\mathbb{Z}[q^{\pm 1}]$ . Then the quantum Hochschild homology groups  $qHH_i(\mathfrak{A}^{\mathcal{B}, \lambda})$  with coefficients in  $\mathbb{k}$  vanish for  $i > 0$ , whereas the Chern character map*

$$h: K_0(\mathfrak{A}^{\mathcal{B}, \lambda}) \otimes_{\mathbb{Z}[q^{\pm 1}]} \mathbb{k} \rightarrow qHH_0(\mathfrak{A}^{\mathcal{B}, \lambda})$$

*is an isomorphism.*

Choose now an oriented tangle  $T$  that is bounded at both top and bottom by the same collection of oriented points  $\Sigma$ . We define for its annular closure  $\hat{T}$  the *quantum annular  $\mathfrak{gl}_2$  complex* as

$$Kh_{\mathbb{A}_q}(\hat{T}) := \bigoplus_{\lambda} qHH_{\bullet}(\mathfrak{A}^{\mathcal{B}, \lambda}, C_{\mathfrak{A}}(T; \lambda))$$

where  $\mathcal{B}$  is a cup basis of webs bounded by  $\Sigma$  and  $C_{\mathfrak{A}}(T; \lambda)$  — the chain complex of bimodules obtained by applying  $\mathcal{F}_w^\lambda$  to  $[[T]]_{\mathbb{F}}$ . Corollary E together with [4, Theorem B] imply the following:

**Corollary F** *The quantum annular  $\mathfrak{gl}_2$  homology  $Kh_{\mathbb{A}_q}(L)$  is a triply graded invariant of annular links that is strictly functorial with respect to annular link cobordisms. Moreover, it admits an action of  $\mathcal{U}_q(\mathfrak{gl}_2)$  that commutes with the differential and the maps induced by annular link cobordisms.*

It follows now from Theorem D and the following discussion that  $Kh_{\mathbb{A}_q}(L)$  is isomorphic with the quantized annular complex as constructed in [4].

## 1.4 Further generalizations

The Khovanov homology has been extended by Asaeda, Przytycki, and Sikora to links in thickened surfaces [1], but the functoriality has not been addressed until the recent paper of Queffelec and Wedrich [31]. There they have defined  $\mathfrak{gl}_2$  foams in thickened *oriented* surfaces, and the natural question is whether the results of this paper can be extended to show equivalence of the two constructions. This is addressed in a follow-up paper, where we also discuss foams in arbitrary 3-manifolds, including nonorientable ones.

Another natural question is about  $\mathfrak{gl}_N$  foams for  $N > 2$ ; see Khovanov [20], Mackaay, Stošić and Vaz [25], and Queffelec and Rose [30]. Again there are two (bi)categories involved: of *enhanced* and *not enhanced* foams, the latter allowing only facets of labels up to  $N - 1$ . We expect that a proper generalization of this paper would prove equivalence of both (bi)categories, hence also of the associated link homologies. Notice that functoriality of  $\mathfrak{gl}_N$  homology has been shown by Ehrig, Tubbenhauer and Wedrich in [15] using enhanced foams.

## 1.5 Organization of the paper

Section 2 provides a brief exposition of webs and foams. All the results presented there are well known, except perhaps the choice of defining relations. Section 3 discusses shadings, their connection to webs and foams, and bicolored isotopies. It ends with a construction of a basis of the space of foams bounded by a given web. The equivalence of bicategories **BN** and **Foam** together with the local versions are constructed in Section 4, in which we also compare the two versions of the Khovanov bracket. Finally, Sections 5–7 provide detailed constructions of TQFT functors: a description of the tautological functor on **Foam**( $\emptyset$ ) in terms of planar pictures, the constructions of the Blanchet–Khovanov algebras, their subquotients, and the 2-functors  $\mathcal{F}_w^\circ$  and  $\mathcal{F}_w^\lambda$ .

## 1.6 Conventions and notation

Throughout the paper we fix a commutative unital ring  $\mathbb{k}$  and linearity means  $\mathbb{k}$ -linearity. We denote by  $\{d\}$  the upward degree shift by  $d$ , ie  $M\{d\}_i = M_{i-d}$  for a graded module  $M$ . Hence, a homogeneous  $m \in M$  has degree  $\deg(m) + d$  when seen as an element of  $M\{d\}$ . We write  $\text{Com}_h(\mathbf{C})$  for the homotopy category of a linear category  $\mathbf{C}$ , the objects of which are formal complexes in  $\mathbf{C}$  and morphisms — homotopy classes of chain maps.



Manifolds are assumed to be smooth (or at least piecewise smooth when necessary) and submanifolds are neat—that is  $N \subset M$  is transverse to  $\partial M$  and  $\partial N = N \cap \partial M$ , see Hirsch [16]. Orientation of a surface  $S \subset \mathbb{R}^3$  is often identified with the *canonical normal vector field*  $\nu$ , defined by the property that for each  $p \in S$  the triple  $(e_1, e_2, \nu_p)$ , where  $(e_1, e_2)$  is an oriented basis of  $T_p S$ , is an oriented basis of  $T_p \mathbb{R}^3$ . Such a vector field is unique up to an isotopy and can be found by the right-hand rule.

### Acknowledgements

The authors are grateful to the organizers of the program *Homology theories in low-dimensional topology* in spring 2017 at the Isaac Newton Institute for Mathematical Sciences in Cambridge, where they have started to work on this project. Beliakova and Putyra are supported by the NCCR SwissMAP founded by Swiss National Science Foundation. Wehrli is partially supported by the Simons Foundation (grant 632059 Stephan Wehrli).

## 2 Main players

This section provides basic definitions and facts about webs and foams. Most of the material is well known [5; 8; 13; 23]; the exception is the choice of defining relations. The main purpose of this part is to fix notation and introduce terms used throughout the paper.

### 2.1 Webs

A *web* is an oriented trivalent graph with edges colored blue or red<sup>4</sup> in such a way, that at each vertex either two blue edges *merge* to a red one, or a red edge *splits* into two blue edges:



In this paper webs will be always embedded in a disk  $\mathbb{D}^2$  or a sphere  $\mathbb{S}^2$  with a fixed basepoint  $*$  that lies on  $\partial \mathbb{D}^2$  in the case of a disk. Edges of a web in a disk can be attached transversely to the boundary circle away from  $*$ ; each boundary point inherits

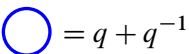
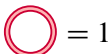
---

<sup>4</sup>Red edges are drawn as double thick lines to make the difference visible when the paper is printed black and white.

then both the color and orientation from the attached edge: outwards (resp. inwards) oriented edges terminate with positive (resp. negative) points. A web is *closed* if its boundary is empty.

**Remark 2.1** By moving the basepoint  $*$  to infinity, we can consider webs in  $\mathbb{D}^2$  or  $\mathbb{S}^2$  as embedded in a half plane  $\mathbb{R} \times (-\infty, 0]$  or a full plane  $\mathbb{R}^2$  respectively.

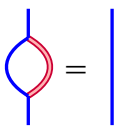
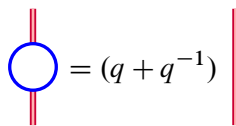
**Definition 2.2** We write *Web* for the module generated by isotopy<sup>5</sup> classes of webs in a disk, modulo the local<sup>6</sup> relations

(2-2)  

(2-3)  

where the webs above can carry any coherent orientation unless indicated. For each collection of oriented red and blue points  $\Sigma \subset \partial\mathbb{D}^2$  there is a submodule  $Web(\Sigma)$  generated by webs bounded by  $\Sigma$  and *Web* is the direct sum of all of them.

**Exercise 2.3** Show that webs satisfy the local relations

(2-4)  

Hint: Start with the left relation in (2-3).

Blue edges of a web  $\omega$  form a crossingless tangle  $\omega_b$ , which we call the *underlying tangle of  $\omega$* . In particular, it is a collection of disjoint circles when  $\omega$  is closed. Write  $\ell(\omega)$  for the number of blue loops in  $\omega_b$ . Let  $r(\omega)$  be a web, the underlying tangle of which is  $\omega_b$  with closed loops removed. We call it a *reduction of  $\omega$* . We construct reductions later in Section 3.2 using the bicolored isotopy argument and show the following fact, which implies in particular that  $r(\omega)$  does not depend on the placement of red edges.

**Proposition 2.4** Webs with same boundary and isotopic underlying tangles coincide in *Web*. In particular,  $\omega = (q + q^{-1})^{\ell(\omega)} r(\omega)$  for any web  $\omega$ .

<sup>5</sup>Isotopies are assumed to fix points on the boundary circle.

<sup>6</sup>The word *local* means that two webs are identified if there is a disk outside of which the webs coincide and inside they look like in the pictures.

The second statement in Proposition 2.4 is equivalent to [31, Lemma 2.1]. Let  $-\omega$  be the result of reversing orientation of all edges in a web  $\omega$ . This operation preserves the relations (2-2) and (2-3), hence it induces an involution on  $Web$ . It does not preserve the submodules  $Web(\Sigma)$ , but there is a pairing

$$(2-5) \quad (\omega, \omega') := (q + q^{-1})^{\ell(-\omega \cup \omega')},$$

which can be visualized by placing  $-\omega$  and  $\omega'$  on the lower and upper hemisphere of a sphere and applying Proposition 2.4 to the resulting web (entirely red webs evaluate to 1).

**Lemma 2.5** *The pairing (2-5) is nondegenerate.*

**Proof** Choose a nonzero  $w \in Web(\Sigma)$  and write it as a linear combination

$$c_1\omega_1 + \dots + c_r\omega_r$$

of pairwise nonisotopic webs  $\omega_1, \dots, \omega_r$ , the underlying tangles of which contain no loops. We may further assume that the polynomial  $c_1$  contains a term  $q^d$  with the maximal value of  $|d|$  among all  $c_i$ . Because  $\ell(-\omega_1 \cup \omega_i) < \ell(-\omega_1 \cup \omega_1)$  for any  $i \neq 1$ , the term  $q^d(\omega_1, \omega_1)$  is not canceled in the expansion of  $(w, \omega_1)$ . Hence,  $(w, \omega_1) \neq 0$ . □

## 2.2 Foams

A *foam* is a collection of *facets*, oriented blue and red<sup>7</sup> surfaces, embedded in a 3-ball  $B^3$  with boundary components attached transversely to  $\partial B^3$  or glued together along singular curves called *bindings* in a way, such that locally two blue facets merge into a red one in an orientation-preserving way as shown in Figure 1. Furthermore, blue facets may carry *dots*, but not the red ones, and bindings inherit orientation from blue facets. We say that a foam is *closed* if its boundary is empty. Otherwise it is bounded by a web in  $\partial B^3$ . Notice that blue facets alone, forgetting orientation, form a surface  $S_b$  with dots, the *underlying surface of S*. As in the case of webs, we fix a basepoint  $* \in \partial B^3$  away from  $\partial S$ . By moving it to infinity we can reinterpret foams as embedded in a half 3-space  $\mathbb{R}^2 \times (-\infty, 0]$ .

There is a canonical cyclic order of facets attached to a binding that follows the right-hand rule: point the thumb of your right hand along the binding curve and slightly bend the other fingers — they indicate the orientation of a small circle around the binding, hence a cyclic order of facets. We call a blue facet *positive* or *negative* depending on

<sup>7</sup>As in the case of webs, red facets of a foam are doubled in pictures.

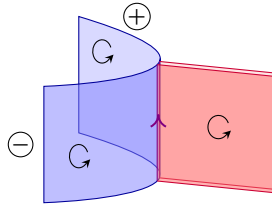


Figure 1: The local model for a foam. The orientation of the binding is coherent with the orientation of the blue facets, but opposite to the one induced from the red facet. The cyclic order is counterclockwise, when seen from above, so that the front blue facet is the negative one.

whether it succeeds or precedes the red facet respectively. For nonembedded foams this cyclic order is usually provided explicitly by drawing small arrows around the binding; see [5].

**Definition 2.6** We write *Foam* for the module generated by isotopy classes of foams in  $B^3$  with the following local relations imposed:

- *sphere evaluations:*

(2-6) = 0      = 1      = -1

- *neck-cutting relations:*

(2-7) = + - h -      = - -

- *dot-reduction and dot-moving relations:*

(2-8) = h + t      = h - -

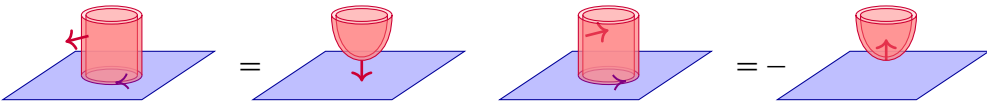
- *red facet detachments:*

(2-9) =      = -

(2-10) =      = -

Foams bounded by a web  $\omega \subset \mathbb{S}^2$  (with  $*$   $\notin \omega$ ) generate a submodule  $Foam(\omega)$ . As in the case of webs,  $Foam$  is the direct sum of all these submodules.

**Remark 2.7** The sign in (2-9) and (2-10) can be read easily from the direction of the canonical normal vector at the critical point on the red surface: it is positive exactly when the normal vector is directed towards the blue plane. For example, (2-9) can be written as



**Remark 2.8** Definition 2.6 does not follow [5], where foams were defined using the universal construction, from which a sufficient set of relations has been derived. Instead, it looks more like the one in [23]. However, our set of relations is smaller and more natural from the topological point of view. The parameters  $h$  and  $t$  have already appeared in [8; 13].

When  $\mathbb{k}$  is graded with  $h$  and  $t$  homogeneous in degree 2 and 4 respectively, then  $Foam$  is a graded module with a foam  $S$  being a homogeneous element in degree

$$(2-11) \quad \deg(S) := -\chi(S_b) + 2 \text{dots}(S).$$

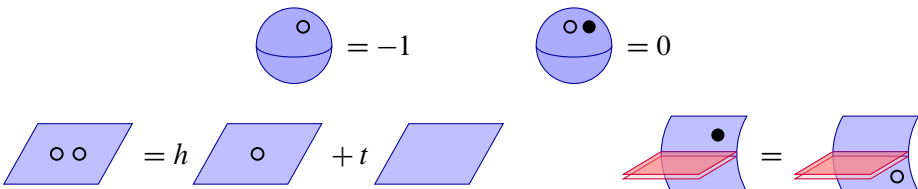
Here  $\chi(S_b)$  stands for the Euler characteristic of the underlying surface and  $\text{dots}(S)$  counts dots carried by the foam.

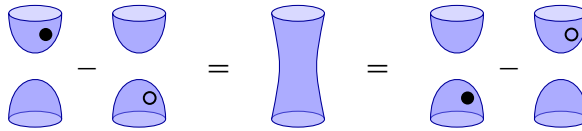
The dot-moving relation — the right one in (2-8) — takes a particularly simple form for  $h = 0$ : it allows to move a dot on the underlying surface at a cost of a sign. To have a similar interpretation in the general case, we introduce the *dual dot* as the difference:

$$(2-12) \quad \text{[parallelogram with white dot]} := h \text{ [parallelogram]} - \text{[parallelogram with black dot]}$$

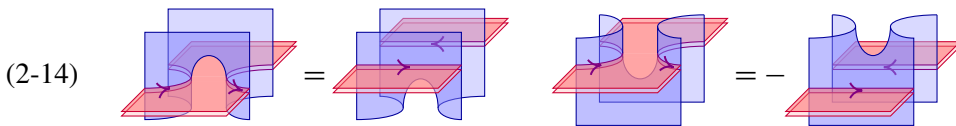
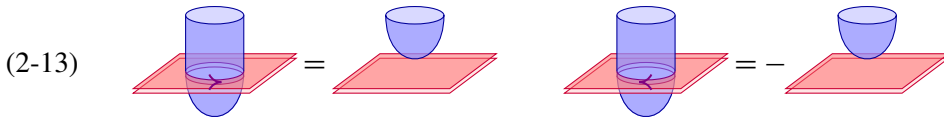
The following exercise lists several relations satisfied by dual dots.

**Exercise 2.9** Show the following equalities between foams:

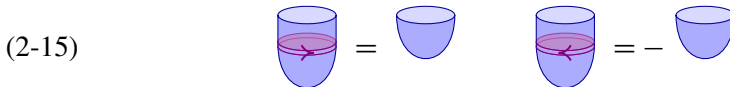




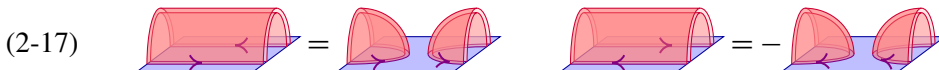
The detaching relations (2-9) and (2-10) can take many other forms. For instance, redrawing them to make red facets horizontal results in



Likewise, (2-9) together with (2-6) allow us to remove a red membrane attached to a blue cup



and other well-known relations arise by redrawing (2-10) and (2-15) in a way, such that blue facets form a horizontal plane and the boundary of red facets is vertical:



Notice that in each case the sign can be read from the direction of the normal vector as explained in Remark 2.7.

We interpret the above relations later as isotopies between two surfaces, a blue and a red one. This will be a key ingredient in the proofs of the two facts listed below. In what follows we write  $S \doteq S'$  if foams  $S$  and  $S'$  differ only by a sign and dualizing dots. For instance,  $S \doteq S'$  when  $S'$  is the result of moving a dot on the underlying surface of  $S$ .

**Proposition 2.10** *Let  $S$  and  $S'$  be foams with isotopic underlying surfaces and same boundary. Then  $S \doteq S'$  in Foam.*

We prove the above proposition at the end of Section 3.2. An important consequence of it is the uniqueness (up to a sign) of a foam  $\text{cup}(\omega)$ , the underlying surface of which is a collection of disjoint disks bounded by  $\omega_b$ . We call it the *cup foam* associated to  $\omega$ . Then for any family  $X$  of blue loops in  $\omega_b$  we denote by  $\text{cup}(\omega, X)$  the cup foam with a dot placed on every blue disk that bounds a curve from  $X$ . These foams constitute a linear basis of  $\text{Foam}(\omega)$  as shown in Section 3.3.

**Theorem 2.11** *Choose a closed web  $\omega$ . The set  $\{\text{cup}(\omega, X) \mid X \subset BL(\omega)\}$  is a linear basis of  $\text{Foam}(\omega)$ . In particular,  $\text{Foam}(\omega)$  is a free graded module of rank  $(q + q^{-1})^{\ell(\omega)}$ .*

### 2.3 Decategorification

Fix a collection of red and blue oriented points  $\Sigma \subset \mathbb{D}^2$ . A *foam with corners in  $\Sigma$*  is a foam  $S$  in  $\mathbb{D}^2 \times [0, 1]$  with  $S \cap \partial\mathbb{D}^2 = \Sigma \times [0, 1]$ . We gather them into a category  $\mathbf{Foam}(\Sigma)$ , in which

- objects are webs bounded by  $\Sigma$  with no relation imposed,
- morphisms from  $\omega_0$  to  $\omega_1$  are generated by foams with corners in  $\Sigma$ , with  $\omega_1$  at the top and  $-\omega_0$  at the bottom disk of  $\mathbb{D}^2 \times [0, 1]$ , modulo the relations (2-6)–(2-10), and
- the composition is given by stacking foams, one on top of the other.

We further enhance it to a graded additive category by introducing formal direct sums and formal degree shifts, so that objects are of the form  $\omega_1\{d_1\} \oplus \cdots \oplus \omega_r\{d_r\}$ , and redefining the degree of a foam  $S: \omega_0\{a\} \rightarrow \omega_1\{b\}$  as

$$(2-18) \quad \text{deg}(S) := (b - a) - \chi(S_b) + 2\text{dots}(S) + \frac{1}{2}\#\Sigma_b,$$

where, as before,  $\chi(S_b)$  is the Euler characteristic of the underlying surface of  $S$  and  $\text{dots}(S)$  counts dots on  $S$ , whereas  $\#\Sigma_b$  is the number of blue points in  $\Sigma$ . The reason for the last term is to make the identity foam a morphism of degree zero; it also makes the degree additive under the composition of foams. Furthermore, reinterpreting foams with corners as foams in  $B^3$  leads to an isomorphism of graded  $\mathbb{k}$ -modules

$$(2-19) \quad \text{Hom}_{\mathbf{Foam}(\Sigma)}(\omega, \omega') \cong \text{Foam}(-\omega \cup \omega')\{\frac{1}{2}\#\Sigma_b\}$$

for any webs  $\omega$  and  $\omega'$  bounded by  $\Sigma$ .

The orientation-reversing diffeomorphism of the thickened disk  $(p, t) \mapsto (p, 1 - t)$  induces a contravariant involutive functor

$$(2-20) \quad \text{Hom}_{\mathbf{Foam}(\Sigma)}(\omega, \omega') \ni S \mapsto S^! \in \text{Hom}_{\mathbf{Foam}(\Sigma)}(\omega', \omega)$$

that reflects a foam vertically and reverses orientation of its facets. We check directly that all the defining relations (2-6)–(2-10) are preserved.

Foams with corners categorify webs. Indeed, web relations are lifted to isomorphisms

$$(2-21) \quad \text{blue circle} \xleftrightarrow{\begin{matrix} \left[ \begin{array}{c} \text{blue dome} \\ \text{blue dome} \cdot -h \text{ blue dome} \\ \text{blue cup} \end{array} \right] \\ \left[ \begin{array}{c} \text{blue cup} \end{array} \right] \end{matrix}} \emptyset\{-1\} \oplus \emptyset\{+1\} \quad \text{red circle} \xleftrightarrow{\begin{matrix} - \text{red dome} \\ \text{red cup} \end{matrix}} \emptyset$$

$$(2-22) \quad \text{blue web} \xleftrightarrow{\pm \text{foam}} \text{blue web} \quad \text{red web} \xleftrightarrow{\begin{matrix} - \text{foam} \\ \text{foam} \end{matrix}} \text{red web}$$

where the sign in the bottom left corner depends on the orientation of the edges. Therefore, there is a well-defined epimorphism  $\gamma : \text{Web}(\Sigma) \rightarrow K_0(\mathbf{Foam}(\Sigma)) \otimes_{\mathbb{Z}[q^{\pm 1}]} \mathbb{k}$  that takes a web  $\omega$  to its class  $[\omega]$  in the Grothendieck group.

**Theorem 2.12** *The linear map  $\gamma : \text{Web}(\Sigma) \rightarrow K_0(\mathbf{Foam}(\Sigma)) \otimes_{\mathbb{Z}[q^{\pm 1}]} \mathbb{k}$  is an isomorphism.*

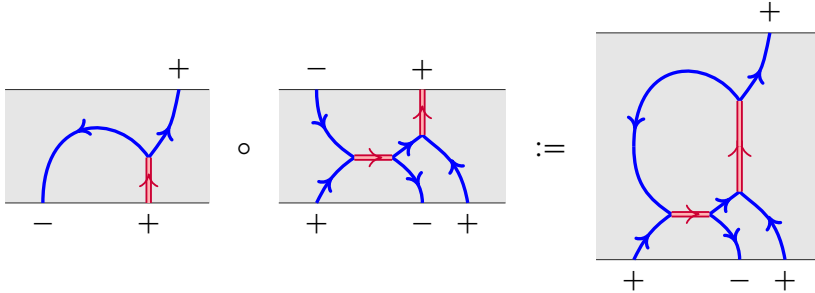
**Proof** We have to show that  $\gamma$  is injective. Consider a bilinear form  $\langle -, - \rangle$  on  $K_0(\mathbf{Foam}(\Sigma))$  defined for webs  $\omega$  and  $\omega'$  as  $\langle [\omega], [\omega'] \rangle := \text{rk}_q \text{Hom}_{\mathbf{Foam}(\Sigma)}(\omega, \omega')$ . It is well defined, because the rank of the morphism space depends only on the images of webs in the Grothendieck group. Theorem 2.11 and the isomorphism (2-19) imply together that  $\langle [\omega], [\omega'] \rangle = (\omega, \omega')$ , where the latter is the nondegenerate pairing from (2-5). Hence,  $\gamma(w) = 0$  forces  $(w, -) = 0$ , so  $w$  must be zero.  $\square$

### 2.4 Higher structures

It is common to consider webs embedded in a horizontal stripe  $\mathbb{R} \times [0, 1]$  instead of a disk. This is equivalent to picking two basepoints on  $\partial\mathbb{D}^2$ ,  $*$  and  $*$ ', and placing them at the left and right infinities respectively. Such webs are morphisms of a linear category **Web**, the objects of which are finite collections of oriented red and blue points



on a line, whereas the composition is defined by stacking stripes vertically:



Formally,  $\text{Hom}_{\mathbf{Web}}(\Sigma, \Sigma') = \text{Web}(-\Sigma \cup \Sigma')$ . This category is closely related to representations of  $\mathcal{U}_q(\mathfrak{gl}_2)$  [9]: there is a monoidal functor  $\mathcal{V}: \mathbf{Web} \rightarrow \mathbf{Rep}(\mathcal{U}_q(\mathfrak{gl}_2))$  such that

- a blue positive (resp. negative) point is assigned the fundamental representation  $V$  (resp. its dual  $V^*$ ) and a red positive (resp. negative) point — the determinant representation  $\wedge^2 V$  (resp.  $\wedge^2 V^*$ ), whereas a sequence of such points is assigned the tensor product of the corresponding representations,
- the merge and split webs (2-1) are assigned the canonical inclusion and quotient maps between representations, and
- cups and caps represent coevaluation and evaluation maps.

The relations between webs make the above functor faithful.

Define the *weight* of a point from  $\Sigma$  according to:

point	●+	●-	○+	○-
weight	+1	-1	+2	-2

The *total weight*  $w(\Sigma)$  of  $\Sigma$  is the sum of weights of its points. A quick analysis of the local model for webs (2-1) reveals that webs exist only between objects of the same weight. Hence, the category of webs decomposes into *weight blocks*  $\mathbf{Web}^k$ , each spanned by objects of weight  $k \in \mathbb{Z}$ . In particular,  $\text{Hom}_{\mathbf{Web}}(\emptyset, \Sigma) \neq 0$  only when  $w(\Sigma) = 0$ ; such collections are called *balanced*.

In a similar manner one collects the foam categories  $\mathbf{Foam}(\Sigma)$  into a bicategory  $\mathbf{Foam}$ , which also decomposes into blocks  $\mathbf{Foam}^k$  parametrized with  $k \in \mathbb{Z}$ . Theorem 2.12 can be then rephrased to say that  $\mathbf{Foam}^k$  categorifies  $\mathbf{Web}^k$ , ie the category of webs is obtained by replacing morphism categories of  $\mathbf{Foam}$  with their Grothendieck groups.

### 2.5 Blanchet evaluation formula

We end this section recalling the evaluation formula for closed foams in a 3–ball  $B^3$  following [5]. It requires two 2–dimensional TQFTs, one for blue and one for red facets. Each is uniquely determined by the (associative) commutative Frobenius algebra assigned to a circle. We choose the algebras

$$A_b := \mathbb{k}[X]/(X^2 - hX - t) \quad \text{and} \quad A_r := \mathbb{k}$$

for blue and red circles respectively, where  $h, t \in \mathbb{k}$  are fixed parameters (the standard choice is  $h = t = 0$ ). The comultiplications and counits are defined by the formulas

$$\begin{aligned} \Delta_b(1) &= 1 \otimes X + X \otimes 1 - h1 \otimes 1, & \Delta_b(X) &= X \otimes X + t1 \otimes 1, & \Delta_r(1) &= -1 \otimes 1, \\ \epsilon_b(1) &= 0, & \epsilon_b(X) &= 1, & \epsilon_r(1) &= -1. \end{aligned}$$

A dot on a blue surface is interpreted as the multiplication with  $X$ . Notice that  $h - X$ , which represents a dual dot, satisfies the polynomial relation defining  $A_b$ , so that  $\bar{X} := h - X$  extends to a conjugation compatible with multiplication. One checks directly that  $\overline{\Delta_b(a)} = -\Delta_b(\bar{a})$  and  $\epsilon_b(\bar{a}) = -\epsilon_b(a)$  for any  $a \in A_b$ .

When  $\mathbb{k}$  is graded with  $h$  and  $t$  homogeneous in degree 2 and 4 respectively, then we make  $A_b$  a graded algebra by setting  $\deg(X) = 2$ ; comultiplication and counit increase and decrease the degree by 2 respectively. Assigning now  $A_b\{-1\}$  to a blue circle produces a graded TQFT:  $\deg(1) = -1$  and  $\deg(X) = +1$ , in which case both multiplication and comultiplication are homogeneous in degree 1, matching the degree of a saddle. Likewise for the unit and counit. The other TQFT is upgraded by inheriting the grading on  $A_r$  from  $\mathbb{k}$ .

Assume that a closed foam  $S$  is obtained from a blue surface  $S_b$  and a red one  $S_r$  by identifying boundary circles  $C_i^+, C_i^- \subset \partial S_b$  with  $C_i^0 \subset \partial S_r$  for  $1 \leq i \leq m$ , such that  $C_i^+$  and  $C_i^-$  come from the positive and negative facet respectively. Let

$$Z_b(S_b) \in (A_b \otimes A_b)^{\otimes m} \quad \text{and} \quad Z_r(S_r) \in (A_r)^{\otimes m}$$

be the elements assigned by the two TQFTs to the blue and red surface, where the first factor in  $A_b \otimes A_b$  corresponds to  $C_i^+$  and the second to  $C_i^-$ . The evaluation assigns to  $S$  the value

$$(2-23) \quad Z(S) = \text{tr}^{\otimes m}(\pi^{\otimes m}(Z_b(S_b)) \otimes \eta^{\otimes m}(Z_r(S_r))) \in \mathbb{k},$$

where  $\pi: A_b \otimes A_b \rightarrow A_b$  sends  $x \otimes y$  to  $x\bar{y}$  and  $\eta: A_r \rightarrow A_b$  is the inclusion of algebras; the trace map  $\text{tr}: A_b \otimes A_b \rightarrow \mathbb{k}$  is the composition of the multiplication with the counit of  $A_b$ .

**Example 2.13** Let  $S$  be a blue sphere with a red disk inside and one dot, as shown below. It decomposes into three cups, two blue and a red one, where one of the blue cups carries a dot:



The orientation of the binding determines that the dotted cup is attached to the negative boundary. Hence,

$$Z_b(S_b) = 1 \otimes X, \quad Z_r(S_r) = 1,$$

resulting in  $Z(S) = \text{tr}(1 \otimes \bar{X}) = \epsilon_b(h - X) = -1$ .

The relations (2-15) and (2-6) evaluate the foam  $S$  from the example above to  $-1$  as well. This is not a coincidence: the defining relations were looked up in the kernel of  $Z$ . In fact, Proposition 2.10 implies a stronger statement. It was first proven in [5].

**Theorem 2.14** (cf [5]) *The evaluation (2-23) descends to an isomorphism*

$$Z : \text{Foam}(\emptyset) \rightarrow \mathbb{k}.$$

**Proof** We first check that  $Z$  is well defined, ie it preserves the relations (2-6)–(2-10). Those involving facets of one color can be checked directly, whereas moving a dot through an  $i^{\text{th}}$  binding corresponds to taking it from a facet attached to  $C_i^+$  (multiplication by  $X$ ) and placing it on the facet attached to  $C_i^-$  (multiplication by  $\bar{X} = h - X$ ). Hence, (2-8) is satisfied. We follow now Example 2.13 to compute

$$(2-24) \quad Z\left(\text{Sphere with red disk and dot on top}\right) = 0, \quad Z\left(\text{Sphere with red disk and dot on bottom}\right) = 1, \quad Z\left(\text{Sphere with red disk and dot on top, rotated}\right) = -1,$$

which immediately implies (2-9): using (2-7) cut both the red cylinder and the plane around the binding to obtain a sum of three foams, each consisting of a red cup, a blue plane, and a blue sphere with a red membrane inside. Two of these foams have an additional dot, one on the plane and the other the sphere; only the latter term survives and the sign comes from (2-24). We leave (2-10) as an exercise.

Assume now that  $Z(S) = 0$ . By Proposition 2.10,  $S$  coincides up to a sign with an entirely blue foam  $S'$ , which is the blue surface  $S_b$ , perhaps with some dots replaced with dual dots. However, applying the blue neck-cutting relation (2-7) to any component of positive genus reduces  $S'$  further to a sum of collections of dotted spheres. These in turn can be completely evaluated with (2-8) and (2-6). Hence,  $S' = Z(S') = 0$ , which shows that  $Z$  is invertible.  $\square$

### 3 Shadings and a basis of foams

This part is the backbone of the paper. We introduce here shadings of manifolds, use them to construct webs and foams, and prove the bicolored isotopy lemma: isotopic shadings encode equal webs and foams (the latter up to a sign and type of dots). Using this language we introduce then a basis of foams that is especially easy to visualize.

#### 3.1 Shadings and trivalent manifolds

A *shading* of a manifold  $M^n$  consists of two codimension 1 submanifolds, an oriented  $U_r$  and a nonoriented  $U_b$ , that are transverse to each other and to  $\partial M^n$ , together with a *checkerboard coloring* of  $M^n$ : a choice of color, white or black, for each connected component of the complement of  $U_r \cup U_b$ , such that any two components with a common facet have different colors. We refer to  $U_r$  and  $U_b$  as *red* and *blue* respectively. The components of the intersection  $U_r \cap U_b$  are called *bindings*; they decompose both  $U_r$  and  $U_b$  into *facets*. Finally, we refer to the components of the complement of  $U_r \cup U_b$  in  $M^n$  as *regions*.

**Lemma 3.1** *Assume  $M^n$  is simply connected and fix a point  $*$   $\in M^n$ . Then a pair of codimension 1 submanifolds of  $M^n$ , that are transverse to each other and away from  $*$ , determines a unique shading of  $M^n$  with the region containing  $*$  painted white.*

**Proof** Given a pair of transverse codimension 1 submanifolds  $(U_r, U_b)$  we construct a desired shading as follows. Given  $p \in M^n \setminus (U_r \cup U_b)$  choose a path  $\gamma$  from  $*$  to  $p$ , transverse to both  $U_r$  and  $U_b$ , and let  $d(\gamma) := \#(\gamma \cap U_r) + \#(\gamma \cap U_b)$  count the intersection points of  $\gamma$  with both submanifolds. Color  $p$  white or black depending on whether  $d(\gamma)$  is even or odd. Because  $M^n$  is simply connected, the parity of  $d(\gamma)$  does not depend on the choice of  $\gamma$  and the color of  $p$  is well defined.  $\square$

**Remark 3.2** It follows from Lemma 3.1 that every codimension 1 submanifold  $U$  of a simply connected manifold  $M^n$  admits a *standard orientation*: the one induced from white regions, when  $U$  is considered as a shading with  $U_r = \emptyset$ . In particular, every codimension 1 submanifold of  $M^n$  is orientable. When  $M^n$  is a line and  $U$  a collection of blue points, then the standard orientation on  $U$  is the alternating one. Likewise for the case  $M^n = \mathbb{S}^1$ , assuming the cardinality of  $U$  is even (otherwise it does not extend to a shading).

A *trivalent manifold* embedded in  $M^n$  is a generalization of webs and foams. It is a collection of *facets*, oriented codimension 1 submanifolds colored blue or red, with

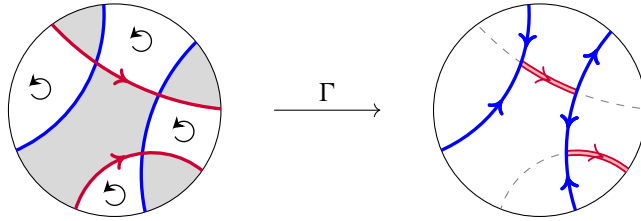


Figure 2: The construction of a web from a shading of a disk. The annihilated red edges are drawn as dashed lines on the right diagram.

boundary components attached transversely to  $\partial M^n$  or glued together along *bindings* in such a way that locally two blue facets merge into a red one. In other words, each point of a trivalent manifold has a neighborhood diffeomorphic to either  $\mathbb{R}^{n-1}$  or  $Y \times \mathbb{R}^{n-2}$ , where  $Y$  is an oriented *merge* or a *split* from (2-1).

Given a shading  $(U_r, U_b)$  of  $M^n$  we construct a trivalent manifold  $\Gamma(U_r, U_b)$  by examining the orientation on facets induced from white regions:

- Blue facets inherit the orientation.
- Red facets are preserved (“amplified”) if the induced orientation agrees with the given one or annihilated otherwise.

An example is presented in Figure 2. In particular,  $\Gamma(\emptyset, U_b)$  is  $U_b$  with its standard orientation as defined in Remark 3.2. It appears that every trivalent manifold arises this way, the proof of which is presented below and visualized in Figure 3. Hence, shadings can be considered as *completions* of trivalent manifolds, because of which we shall refer to shadings of  $\mathbb{D}^2$  and  $B^3$  as *completed webs* and *completed foams* respectively.

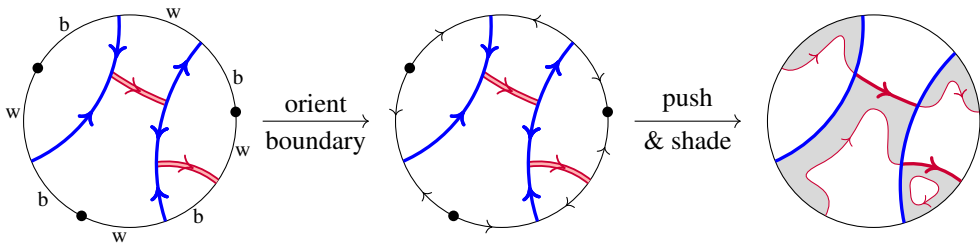


Figure 3: The construction of a shading from a planar web that extends a given shading of its boundary. The boundary of the disk is oriented in the middle picture, whereas the curves  $U_i^j$  are identified and pushed inwards the corresponding regions in the third picture.

**Lemma 3.3** Choose a trivalent manifold  $V \subset M^n$ , such that  $\partial V = \Gamma(\tilde{U}_r, \tilde{U}_b)$  for some shading  $(\tilde{U}_r, \tilde{U}_b)$  of  $\partial M^n$ . Then there exists a shading  $(U_r, U_b)$  of  $M^n$  that restricts to  $(\tilde{U}_r, \tilde{U}_b)$  on  $\partial M^n$  and satisfies  $\Gamma(U_r, U_b) = V$ .

**Proof** Consider the orientation of  $\partial M^n$  induced from  $M^n$  and reverse it at all black edges in the given shading. Then the boundary of any region  $R \subset M^n$  is a union of facets of  $V$  and regions in  $\partial M^n$ , such that oppositely oriented components meet only in two situations: when they are both contained in the boundary (so that they meet at a facet of  $\partial V$ ) or both are blue facets of  $V$  adjacent to a red facet outside of  $R$ . Consider the union of those components of  $\partial R$ , the orientation of which does not match the one induced from  $R$ . They constitute certain oriented  $(n-1)$ -dimensional submanifolds  $U_1, \dots, U_k$ . Taking a red colored copy  $U'_i$  of each  $U_i$ , push its interior inside  $R$  and paint the newly created region black. Repeating this for each region produces the desired shading.  $\square$

A useful consequence of Lemma 3.3 is that tangles and surfaces can be extended to webs and foams with given boundary. Recall that a collection  $\Sigma \subset \partial \mathbb{D}^2$  of oriented red and blue points is *balanced* if it bounds a web, which is equivalent to being of weight zero.

**Proposition 3.4** (1) Let  $\Sigma \subset \partial \mathbb{D}^2$  be a balanced collection of oriented red and blue points and  $\tau$  a tangle bounded by  $\Sigma_b$ . Then there exists a web  $\omega$  bounded by  $\Sigma$  with  $\omega_b = \tau$ .

(2) Let  $\omega \subset \partial B^3$  be a web and  $W$  a surface bounded by  $\omega_b$ . Then there is a foam  $S$  bounded by  $\omega$  with  $S_b = W$ .

**Proof** Extend  $\Sigma$  to a shading  $\tilde{\Sigma} = (\Sigma_r \cup \Sigma'_r, \Sigma_b)$ . Then  $\tilde{\Sigma}$  has an even number of points and the orientation of points from  $\Sigma$  matches the one induced from white regions. Let  $b$ ,  $r$ , and  $r'$  be the sums of orientations of blue points in  $\Sigma$ , red points in  $\Sigma$ , and red points added to  $\tilde{\Sigma}$  respectively. Then  $b + 2r = 0$ , because  $\Sigma$  is balanced, and  $b + r + r' = 0$ , because the orientation of points in  $\tilde{\Sigma}$  alternate. Subtracting the two equalities reveals that  $r - r' = 0$ . It follows that there is an oriented collection of disjoint intervals  $\tau_r \subset \mathbb{D}^2$  bounded by  $\tilde{\Sigma}_r$ , the orientation of which agree with the points from  $\Sigma_r$  and disagree with those from  $\Sigma'_r$ . Hence,  $\omega := \Gamma(\tau_r, \tau)$  is the desired web.

The second statement is even easier to show. Extend the web  $\omega$  to a shading  $\alpha$ . Then  $\alpha_r$  is a collection of disjoint loops and each such collection bounds a family  $W_r$  of disjoint disks in  $B^3$ . Therefore,  $S := \Gamma(W_r, W)$  is the desired foam.  $\square$

### 3.2 Bicolored isotopies

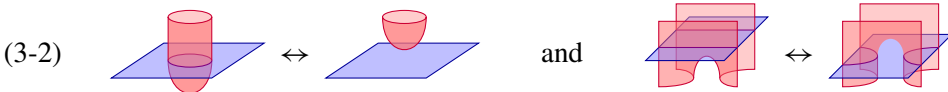
Choose an isotopy  $\Phi$  of  $M^n$  and a subset  $A$ . The set

$$\text{Tr}_\Phi(A) = \{(\Phi_t(a), t) \mid a \in A, t \in [0, 1]\}$$

is called the *trace of  $A \subset M^n$  under  $\Phi$*  [16]. We say that a pair of isotopies  $(\Phi, \Psi)$  of  $M^n$  is an *isotopy of a shading  $(U_r, U_b)$*  if  $(\text{Tr}_\Phi(U_r), \text{Tr}_\Phi(U_b))$  is a shading of  $M^n \times [0, 1]$  that coincides with  $(U_r, U_b)$  at the level  $t = 0$ . When  $M^n$  is a 2-disk, then a generic pair of isotopies can be encoded by a sequence of *bigon moves*



whereas in case of a 3-ball two moves are necessary:



In each move a shading of one side determines a shading of the other. Hence, we obtain the following characterization of isotopies of shadings in these cases.<sup>8</sup>

**Lemma 3.5** *Shadings  $(U_r, U_b)$  and  $(U'_r, U'_b)$  of  $\mathbb{D}^2$  or  $B^3$  are isotopic if and only if  $U_r$  is isotopic to  $U'_r$  and  $U_b$  is isotopic to  $U'_b$ .*

When a basepoint  $* \in M^n$  is present, then one must be careful how it behaves under the isotopy. There is no problem when  $\Psi$  and  $\Phi$  coincide at  $*$  (and in this paper we always assume that both  $\Psi$  and  $\Phi$  fix  $*$ ). Otherwise, the basepoint should stay at the same region if possible. However, when the region disappears, then the basepoint has to reappear in a white region. For instance, when  $*$  lies in the small bigon on the left-hand side of (3-1), then it reappears between the two strands on the right-hand side. Likewise, if  $*$  lies in the ball bounded in the left figure in (3-2), then in the right figure it must reappear between the blue plane and the red cup.

Recall from Section 2.2 that we write  $S \doteq S'$  for foams  $S$  and  $S'$  if they agree up to a sign and replacing some dots with their duals.

**Lemma 3.6** (bicolored isotopy) (1)  $\Gamma(\alpha_r, \alpha_b) = \Gamma(\alpha'_r, \alpha'_b)$  in *Web* if  $(\alpha_r, \alpha_b)$  and  $(\alpha'_r, \alpha'_b)$  are isotopic shadings of  $\mathbb{D}^2$ .  
 (2)  $\Gamma(W_r, W_b) \doteq \Gamma(W'_r, W'_b)$  in *Foam* if  $(W_r, W_b)$  and  $(W'_r, W'_b)$  are isotopic shadings of  $B^3$ .

<sup>8</sup>This can be extended to all manifolds by a detailed analysis of singular levels of a pair of isotopies.

**Proof** It is enough to consider the case of elementary isotopies. When applied to each side of the bigon move (3-1),  $\Gamma$  removes red edges in both pictures from the same side of the blue line. Hence,  $\Gamma(\alpha_r, \alpha_b)$  and  $\Gamma(\alpha'_r, \alpha'_b)$  are related by the left relation in either (2-3) or (2-4). Likewise, the moves (3-2) correspond to the detaching relations (2-9) and (2-10).  $\square$

The above result has far reaching consequences when paired with Lemma 3.3. The statements about comparing webs and foams with isotopic blue pieces follows, which in turn were used in the proof of Theorem 2.14 to show bijectivity of the Blanchet evaluation map  $Z$ .

**Proof of Proposition 2.4** Let  $\omega$  and  $\omega'$  have isotopic underlying tangles and take the trace of  $\omega_b$  under this isotopy as  $S_b$ ; it is the underlying surface of a foam  $S: \omega \rightarrow \omega'$  due to Proposition 3.4. Extend the foam to a shading  $(\tilde{S}_r, S_b)$  of  $\mathbb{D}^2 \times [0, 1]$ . When in generic position, it can be represented by a finite sequence of level sets, such that in between any two consecutive levels  $\tilde{S}_r$  has either no critical points (so that the level sets are related by the bicolored isotopy lemma) or a unique Morse type critical point — a cap, a cup, or a saddle — in which case the corresponding webs coincide (if the affected red edges are erased) or are identified by the right relations in (2-2) and (2-3) (if the red edges survive). Notice that  $S_b$  has no critical points.

For the second part, extend  $\omega$  to a shading  $(\tilde{\omega}_r, \omega_b)$  of  $\mathbb{D}^2$  and isotope closed blue loops, so that they do not intersect  $\omega_r$ . Applying  $\Gamma$  results in a new web  $\omega'$  that coincides with  $\omega$  as shown above. Removing blue circles from  $\omega'$  results in  $r(\omega)$  and the desired equality follows from (2-2).  $\square$

**Proof of Proposition 2.10** Let foams  $S_1$  and  $S_2$  have isotopic blue parts. Extend them to shadings  $W_1$  and  $W_2$  respectively and pick a ball  $\mathcal{O}$  in the interior of  $B^3$ , outside of which the red facets of the shadings coincide. Using Lemma 3.6 isotope blue facets away from  $\mathcal{O}$  (this may dualize dots), reducing the problem to showing equality for foams with only red facets. In such case, use the neck-cutting relation (2-7) to reduce each red surface to a collection of disjoint disks and spheres (the existence of such a system of cuts follows from the theory of incompressible surfaces). This may change the sign of the foam. The thesis follows, because each red sphere evaluates to  $-1$  and the disks are uniquely determined up to an isotopy by the boundary circles.  $\square$

It follows immediately from Proposition 2.10 that the foam used in the proof of Proposition 2.4 is invertible. That would be enough to prove the latter if we knew that



**Foam** categorifies *Web*. However, the proof of the categorification result is based on Theorem 2.11, which is proven only in the next section.

### 3.3 Cup foams

We will now apply the above results to show that cup foams, as defined in Section 2.2, constitute a free basis of spaces of foams. In particular, the category of foams is nondegenerate.

Let  $\omega$  be a closed web, so that  $\omega_b$  is a collection of blue loops. Orient them in a standard way (see Remark 3.2) and pick a foam  $I_\omega \in \text{Hom}_{\mathbf{Foam}(\emptyset)}(\omega_b, \omega)$  with  $\omega_b \times [0, 1]$  as its underlying surface; the existence of such a foam follows from Proposition 3.4. According to Proposition 2.10, there is a sign  $\text{sgn}(\omega) = \pm 1$  satisfying

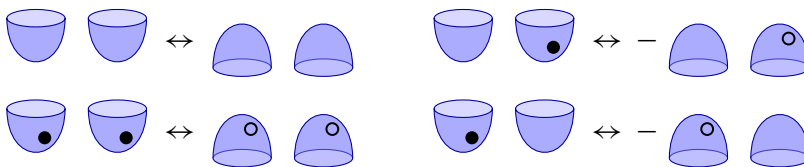
$$I_\omega^! I_\omega = \text{sgn}(\omega)(\omega_b \times [0, 1]),$$

where  $I_\omega^! \in \text{Hom}_{\mathbf{Foam}}(\omega, \omega_b)$  is the vertical flip of  $I_\omega$  as defined in (2-20). The sign can also be computed directly as  $\text{sgn}(\omega) = \text{sgn}(\omega)(C_\omega^\bullet C_\omega) = Z(C_\omega^\bullet I_\omega^! I_\omega C_\omega)$ , where  $C_\omega \in \text{Hom}_{\mathbf{Foam}}(\emptyset, \omega_b)$  is a collection of disks bounded by  $\omega_b$  and  $C_\omega^\bullet \in \text{Hom}_{\mathbf{Foam}}(\omega_b, \emptyset)$  is the same collection, except that each disk is decorated by a dot. Hence,  $\text{sgn}(\omega)$  is a well-defined integer, which we call the *sign of the web*  $\omega$ .

**Lemma 3.7** *The sign  $\text{sgn}(\omega)$  does not depend on the choice of  $I_\omega$ .*

**Proof** Let  $S \in \text{Hom}_{\mathbf{Foam}}(\omega_b, \omega)$  be another foam with  $S_b = \omega_b \times [0, 1]$ . Then  $S = \pm I_\omega$  by Proposition 2.10 and  $S^! S = I_\omega^! I_\omega$ , because the same sign relates  $S^!$  with  $I_\omega^!$ .  $\square$

Let  $BL(\omega)$  be the collection of blue loops in  $\omega$ . For each subset  $X \subset BL(\omega)$  we construct the cup foam  $\text{cup}(\omega, X)$  by attaching blue disks to the input of  $I_\omega$  and placing a dot on each disk bounded by a loop from  $X$ . Notice that red facets of  $\text{cup}(\omega, X)$  are above all dots and minima of blue facets. Therefore, we say that  $\text{cup}(\omega, X)$  is a *red-over-blue cup foam decorated by  $X$* . We construct likewise a *cap foam*  $\text{cap}(\omega, X) \in \text{Hom}_{\mathbf{Foam}(\emptyset)}(\omega, \emptyset)$  by reflecting  $\text{cup}(\omega, X)$  vertically and replacing each dot with the dual one scaled by  $-1$ . For instance, we have the following correspondence between cup and cap foams bounded by two blue loops:



Let us now represent a foam  $S \in \text{Hom}_{\text{Foam}(\varnothing)}(\omega, \omega')$  by a vertical cylinder labeled  $S$ , with  $\omega$  and  $\omega'$  at the bottom and top disk respectively. When no label is present, it is understood that  $\omega = \omega'$  and the cylinder represents the identity foam  $\omega \times [0, 1]$ . We emphasize the cases  $\omega = \varnothing$  and  $\omega' = \varnothing$  by drawing a cup or a cap instead and, to simplify notation, we decorate it directly with  $X \subset BL(\omega)$  when  $S$  is a cup or a cap foam:

$$\text{cup}(\omega, X) = \text{diagram of a cup with } \omega \text{ at the top and } X \text{ at the bottom} \quad \text{and} \quad \text{cap}(\omega, X) = \text{diagram of a cap with } X \text{ at the top and } \omega \text{ at the bottom}$$

Moreover,  $X^c := BL(\omega) \setminus X$  stands for the complement of a subset  $X \subset BL(\omega)$ .

**Lemma 3.8** *Foams satisfy the relations*

$$(3-3) \quad \text{diagram of a sphere with } Y \text{ at the top, } \omega \text{ at the middle, and } X \text{ at the bottom} = \begin{cases} \text{sgn}(\omega) & \text{if } Y = X^c, \\ 0 & \text{otherwise,} \end{cases}$$

$$(3-4) \quad \text{diagram of a cylinder with } \omega \text{ at the top and } \omega \text{ at the bottom} = \text{sgn}(\omega) \sum_{X \subset BL(\omega)} \left( \text{diagram of a cup } \omega, X + \text{diagram of a cap } \omega, X^c \right)$$

**Proof** From the construction of cup and cap foams,

$$\text{diagram of a sphere with } Y \text{ at the top, } \omega \text{ at the middle, and } X \text{ at the bottom} = \text{diagram of a cylinder with } Y \text{ at the top, } \omega_b \text{ at the top, } I_\omega^! \text{ in the middle, } \omega \text{ at the middle, } I_\omega \text{ in the middle, } \omega_b \text{ at the bottom, and } X \text{ at the bottom} = \text{sgn}(\omega) \text{diagram of a sphere with } Y \text{ at the top, } \omega_b \text{ at the middle, and } X \text{ at the bottom}$$

and the right-hand side is a collection of spheres, each carrying at most one regular and one dual dot, scaled by  $(-1)^{|Y|}$ . Such a sphere evaluates to 1 or  $-1$  when it carries either one regular or one dual dot respectively and vanishes otherwise (see Exercise 2.9). Hence, (3-3) follows.

The second relation follows from the equality  $\omega \times [0, 1] = \text{sgn}(\omega) I_\omega I_\omega^!$  and the neck-cutting relation from Exercise 2.9. □

We are ready to prove that cup foams form a linear basis of foams.

**Proof of Theorem 2.11** The first relation of Lemma 3.8 implies that cup foams are linearly independent. To show that they generate  $Foam(\omega) \cong \text{Hom}_{\mathbf{Foam}(\emptyset)}(\emptyset, \omega)$ , use the second relation to write a foam  $S$  bounded by  $\omega$  as a sum

$$\begin{array}{c} \omega \\ \text{---} \\ | \\ S \\ | \\ \text{---} \end{array} = \text{sgn}(\omega) \sum_{X \subset BL(\omega)} \begin{array}{c} \omega \\ \text{---} \\ | \\ X \\ | \\ \text{---} \end{array} + \begin{array}{c} X^c \\ \text{---} \\ | \\ \omega \\ | \\ \text{---} \\ S \end{array}$$

which is a linear combination of cup foams, because closed foams evaluate to scalars. Finally,

$$\text{deg}(\text{cup}(\omega, X)) = 2|X| - \ell,$$

as the underlying surface of the cup foam consists of  $\ell$  disks decorated by  $|X|$  dots, so that

$$\text{rk}_q \text{Foam}(\omega) = \sum_X q^{2|X| - \ell} = \sum_{s=0}^{\ell} \binom{\ell}{s} q^{2s - \ell} = (q + q^{-1})^{\ell}$$

as desired. □

### 4 Equivalences of foam and cobordism categories

In this section we prove Theorems A and B, which state that foams and Bar-Natan cobordisms constitute equivalent (bi)categories. Then we relate the complexes  $[[T]]$  and  $[[T]]_F$  associated with a tangle  $T$ .

#### 4.1 Embedding cobordisms into foams

Fix a balanced collection  $\Sigma \subset \partial\mathbb{D}^2$  away from a fixed basepoint  $* \in \partial\mathbb{D}^2$  and write  $\Sigma_b$  for the subset consisting of all blue points from  $\Sigma$ . Consider first the case when  $\Sigma = \Sigma_b$  and the points are oriented in a standard way as explained in Remark 3.2. This means that, when following the orientation of the boundary circle, the first point after the basepoint is negative and then the orientation alternates. Theorem B is in this case a direct consequence of Proposition 2.4 and Theorem 2.11: each web is isomorphic to an entirely blue one (and each such web is a flat tangle equipped with the standard orientation) and for such webs  $\omega$  and  $\omega'$  the cup basis of  $\text{Hom}_{\mathbf{Foam}(\Sigma)}(\omega, \omega')$  consists of foams with no

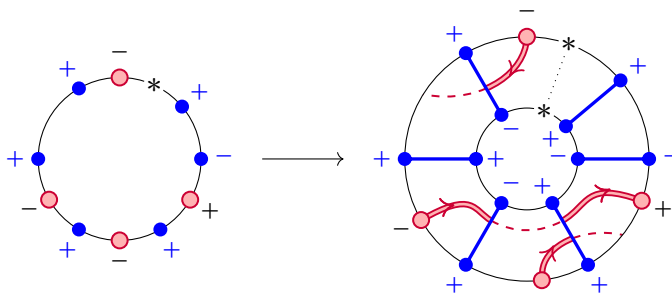


Figure 4: A collection of points  $\Sigma$  with four red and six blue points and an annular web  $E_\Sigma$  as in Remark 4.2. Dashed red lines are not part of the web, but they represent the additional red edges in the associated shading.

red facets. Hence, the naive map  $\text{Hom}_{\mathbf{BN}(\Sigma)}(\omega, \omega') \rightarrow \text{Hom}_{\mathbf{Foam}(\Sigma)}(\omega, \omega')$  that orients a cobordism in a standard way does the job. A little more work has to be done to cover the general case.

**Lemma 4.1** *There is a web  $E_\Sigma \subset \mathbb{S}^1 \times [0, 1]$  bounded by  $\Sigma$  at  $\mathbb{S}^1 \times \{1\}$  and standardly oriented  $\Sigma_b$  at  $\mathbb{S}^1 \times \{0\}$ , which is disjoint from  $\{*\} \times [0, 1]$  and with  $\Sigma_b \times [0, 1]$  as the underlying tangle.*

**Proof** Let  $\tau$  be a collection of radial blue intervals connecting blue points at  $\mathbb{S}^1 \times \{0\}$  with those at  $\mathbb{S}^1 \times \{1\}$ . Cut the annulus to a disk along  $\{*\} \times [0, 1]$  and apply Proposition 3.4 to get a desired web. □

**Remark 4.2** The extension of a tangle to a web is constructed in Lemma 3.3 from a shading of the disk, which is by no means unique. In case of an annulus, however, the situation is different: there is a unique up to an isotopy family of counterclockwise oriented arcs that bounds a given collection of oriented points at the outer boundary circle. Some of the arc may intersect the interval  $\{*\} \times [0, 1]$ ; moving them through the hole results in a preferred shading and a preferred web  $E_\Sigma$ .

Inserting a tangle inside the web  $E_\Sigma$  and a surface inside the foam  $E_\Sigma \times [0, 1]$  results in a functor  $\mathcal{E}_\Sigma: \mathbf{BN}(\Sigma_b) \rightarrow \mathbf{Foam}(\Sigma)$ , as it preserves units and composition.

**Theorem B** *The functor  $\mathcal{E}_\Sigma: \mathbf{BN}(\Sigma_b) \rightarrow \mathbf{Foam}(\Sigma)$  is an equivalence of categories.*

**Proof** It follows from Propositions 2.4 and 2.10 that  $\mathcal{E}_\Sigma$  is essentially surjective and full. Faithfulness follows from Theorem 2.11: both  $\text{Hom}_{\mathbf{BN}(\Sigma_b)}(\omega_b, \omega'_b)$  and

$\text{Hom}_{\mathbf{Foam}(\Sigma)}(\omega, \omega')$  are free graded modules of graded rank  $(q + q^{-1})^\ell$ , where  $\ell$  counts blue loops in  $-\omega \cup \omega'$ . □

### 4.2 A coherent way to forget red facets

The inverse functor to  $\mathcal{E}_\Sigma$  forgets red facets of a foam, but it may also change its sign and dualize some dots. To construct the functor explicitly, fix for each web  $\omega$  an invertible foam  $I_\omega \in \text{Hom}_{\mathbf{Foam}(\Sigma)}(\mathcal{E}_\Sigma(\omega_b), \omega)$ . According to Proposition 2.10, for every foam  $S: \omega \rightarrow \omega'$  there is a sign  $\text{sgn}(S) = \pm 1$  and a cobordism  $S_{\text{cob}}: \omega_b \rightarrow \omega'_b$ , which agrees with  $S_b$  up to dualizing some dots, that fit into a commuting square

$$(4-1) \quad \begin{array}{ccc} \omega & \xrightarrow{S} & \omega' \\ I_\omega \uparrow & & \uparrow I_{\omega'} \\ \mathcal{E}_\Sigma(\omega_b) & \xrightarrow{\text{sgn}(S)\mathcal{E}_\Sigma(S_{\text{cob}})} & \mathcal{E}_\Sigma(\omega'_b) \end{array}$$

Let us explain how both  $\text{sgn}(S)$  and  $S'_b$  can be obtained from the given data.

To construct  $S_{\text{cob}}$  take  $S_b$  and dualize all dots that are carried by blue facets with orientation opposite to the standard orientation on  $S_b$ . Indeed, the isotopy used in the proof of Proposition 2.10 to take red facets of  $I_{\omega'}^{-1} S I_\omega$  away from  $S_b$  involves an odd number of the dot migration moves (the right relation in (2-8), see also Exercise 2.9) for such dots, because this is the only move that reverses the local orientation around a dot.

The sign  $\text{sgn}(S)$  is uniquely defined by (4-1) in case  $S$  does not vanish and it can be computed then from the formula

$$\text{sgn}(S) = \frac{Z(C \cup S I_\omega)}{Z(C \cup I_{\omega'} \mathcal{E}_\Sigma(S_{\text{cob}}))},$$

where  $Z$  is the Blanchet evaluation map from Section 2.5 and  $C$  is a cup foam bounded by the web  $-\mathcal{E}_\Sigma(\omega_b) \cup \omega'$ , for which the two quantities being divided do not vanish.<sup>9</sup>

**Proposition 4.3** *The assignment*

$$\omega \mapsto \omega_b, \quad S \mapsto \text{sgn}(S)S_{\text{cob}}$$

*defines a functor  $\mathcal{E}_\Sigma^\vee: \mathbf{Foam}(\Sigma) \rightarrow \mathbf{BN}(\Sigma_b)$  inverse to  $\mathcal{E}_\Sigma$ .*

**Proof** Clearly,  $(S'' S')_{\text{cob}} = S''_{\text{cob}} S'_{\text{cob}}$  for composable foams  $S': \omega \rightarrow \omega'$  and  $S'': \omega' \rightarrow \omega''$ . Furthermore, the equality

$$S'' S' I_\omega = \text{sgn}(S'') \text{sgn}(S') (I_{\omega''} \mathcal{E}_\Sigma(S''_b) \mathcal{E}_\Sigma(S'_b)) = \text{sgn}(S'') \text{sgn}(S') (I_{\omega''} \mathcal{E}_\Sigma(S''_b S'_b))$$

<sup>9</sup>Explicitly,  $C = \text{cup}(-\mathcal{E}_\Sigma(\omega_b) \cup \omega', X)$  where  $X$  contains exactly one boundary circle of each genus 0 component of  $S_b$  that does not carry a dot.

forces  $\text{sgn}(S''S') = \text{sgn}(S'') \text{sgn}(S')$  if the composition  $S''S'$  does not vanish. Hence,  $\mathcal{E}_\Sigma^\vee$  is a functor. To end the proof, we check directly that  $\mathcal{E}_\Sigma^\vee \circ \mathcal{E}_\Sigma$  is the identity functor on  $\mathbf{BN}(\Sigma_b)$ , whereas the collection of the invertible foams  $I_\omega$  constitute a natural isomorphism between  $\mathcal{E}_\Sigma \circ \mathcal{E}_\Sigma^\vee$  and the identity functor on  $\mathbf{Foam}(\Sigma)$ .  $\square$

**Example 4.4** Let  $\omega$  be a blue circle oriented clockwise. This is the orientation induced from the unbounded region, hence standard, so that  $\omega = \omega_b$  and  $\mathcal{E}_\emptyset^\vee$  simply forgets orientation:

$$\mathcal{E}_\emptyset^\vee \left( \text{cup with blue arrow} \right) = \text{cup} \quad \text{and} \quad \mathcal{E}_\emptyset^\vee \left( \text{cup with blue arrow and dot} \right) = \text{cup with dot}$$

However, when  $\omega$  is oriented counterclockwise, then the invertible foam  $I_\omega$  is a cylinder with a red membrane, the canonical normal vector of which is oriented upwards. The membrane can be removed with (2-15) and (2-8). This gives no difference for the cup with no dot

$$\mathcal{E}_\emptyset^\vee \left( \text{cup with blue arrow} \right) = \mathcal{E}_\emptyset^\vee \left( \text{cup with blue arrow and red membrane} \right) = \mathcal{E}_\emptyset^\vee \left( \text{cup with blue arrow} \right) = \text{cup}$$

but in the presence of the dot, the dot is replaced by its dual:

$$\mathcal{E}_\emptyset^\vee \left( \text{cup with blue arrow and dot} \right) = \mathcal{E}_\emptyset^\vee \left( \text{cup with blue arrow and red membrane and dot} \right) = \mathcal{E}_\emptyset^\vee \left( - \text{cup with blue arrow and dot} + h \text{ cup} \right) = \text{cup with dual dot}$$

Recall that if  $h = 0$ , the dual dot equals *minus* the normal dot.

**Remark 4.5** Although the construction of  $\mathcal{E}_\Sigma^\vee$  depends on the choice of foams  $I_\omega$ , the functor is unique up to a unique natural isomorphism. To see this directly, suppose that  $\tilde{\mathcal{E}}_\Sigma^\vee$  is constructed using a different family of foams  $\tilde{I}_\omega$ . Then  $\tilde{I}_\omega = s(\omega)I_\omega$  for a well-defined sign  $s(\omega) = \pm 1$  and it follows from a direct computation that the collection of morphisms  $\iota_\omega := s(\omega) \cdot \omega_b \times [0, 1]$  is a natural isomorphism from  $\mathcal{E}_\Sigma^\vee$  to  $\tilde{\mathcal{E}}_\Sigma^\vee$ .

### 4.3 An equivalence of bicategories

Recall that a 1–morphism  $f : x \rightarrow y$  in a bicategory  $\mathbf{C}$  is an *equivalence* if there exists  $g : y \rightarrow x$  such that the compositions  $f \circ g$  and  $g \circ f$  are isomorphic to identity 1–morphisms. A 2–functor  $\mathcal{F} : \mathbf{C} \rightarrow \mathbf{D}$  is an *equivalence of bicategories* when it is

- a *local equivalence*, that is, the functor  $\mathcal{F}_{x,y} : \mathbf{C}(x, y) \rightarrow \mathbf{D}(\mathcal{F}(x), \mathcal{F}(y))$  is an equivalence of categories for all objects  $x, y$  of  $\mathbf{C}$ , and
- *essentially surjective*: each object of  $\mathbf{D}$  is equivalent to an object of the form  $\mathcal{F}(x)$ .

Indeed, the above conditions imply the existence of an inverse of  $\mathcal{F}$  [24].

There is a 2–functor

$$(4-2) \quad \mathcal{E}^0 : \mathbf{BN} \rightarrow \mathbf{Foam}$$

that equips points, tangles, and cobordisms with the standard orientation.<sup>10</sup> It is a local equivalence due to Theorem B, but not essentially surjective: objects from the image of  $\mathcal{E}^0$  have weight 0 or 1, so that the whole image is contained in  $\mathbf{Foam}^0 \sqcup \mathbf{Foam}^1$ . We fix this by enlarging the source bicategory to  $\mathbf{wBN} := \mathbf{BN} \times \mathbb{Z}$ , the product of  $\mathbf{BN}$  with  $\mathbb{Z}$  seen as a discrete bicategory. In other words, objects of  $\mathbf{wBN}$  are pairs  $(\Sigma, k)$  consisting of an object  $\Sigma$  from  $\mathbf{BN}$  and a number  $k \in \mathbb{Z}$ , whereas morphism categories are zero or copied from  $\mathbf{BN}$ ,

$$(4-3) \quad \mathbf{wBN}((\Sigma, k), (\Sigma', k')) := \begin{cases} \mathbf{BN}(\Sigma, \Sigma') & \text{if } k = k', \\ 0 & \text{otherwise.} \end{cases}$$

We then extend (4-2) to a 2–functor

$$(4-4) \quad \mathcal{E} : \mathbf{wBN} \rightarrow \mathbf{Foam}$$

in such a way that  $(\Sigma, k)$  is taken to the collection  $\mathcal{E}^0(\Sigma)$  with  $|k|$  red points added to the right, all positive when  $k > 0$  and negative otherwise. Likewise for 1– and 2–morphisms:  $\mathcal{E}$  takes a tangle  $\tau$  (resp. a cobordism  $W$ ), orients it in a standard way, and adds to the right  $|k|$  vertical red lines (resp. vertical red squares) with the appropriate orientation.

**Theorem A** *The 2–functor  $\mathcal{E} : \mathbf{wBN} \rightarrow \mathbf{Foam}$  is an equivalence of bicategories.*

**Proof** By Theorem B,  $\mathcal{E}$  is a local equivalence. Hence, it is enough to show that it is essentially surjective. For that choose an object  $\Sigma$  from  $\mathbf{Foam}$  and let  $k = \lfloor w(\Sigma)/2 \rfloor$ , where  $w(\Sigma)$  is the weight of  $\Sigma$ . Then  $\Sigma^0 := \mathcal{E}(\Sigma_b, k)$  has the same weight. Considering  $\mathbb{R} \times [0, 1]$  as a disk with two boundary points removed, we can apply Lemma 3.3 to the collection of points  $-\Sigma^0 \cup \Sigma$  to obtain a web  $E_\Sigma : \Sigma^0 \rightarrow \Sigma$  with vertical lines as the underlying tangle. Another application of Lemma 3.3 combined with Proposition 2.10 shows that it is an equivalence, with its mirror image the inverse 1–morphism.  $\square$

We write  $\mathcal{E}^\vee$  for the 2–functor inverse to  $\mathcal{E}$ . It can be constructed explicitly like the functors  $\mathcal{E}_\Sigma^\vee$ , except that the computation of signs requires not only a choice of

<sup>10</sup>Recall the convention that the basepoint  $*$  is placed at the left infinity, so that the left unbounded region is painted white. This implies in particular that the left most point of an object of  $\mathbf{BN}$  receives the positive orientation and the left most vertical strand of a 1–morphism is oriented upwards.

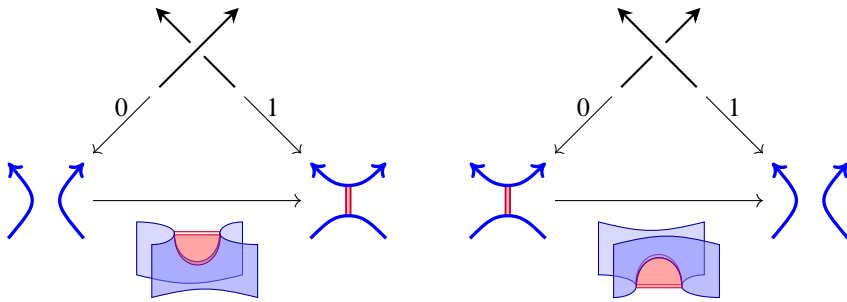
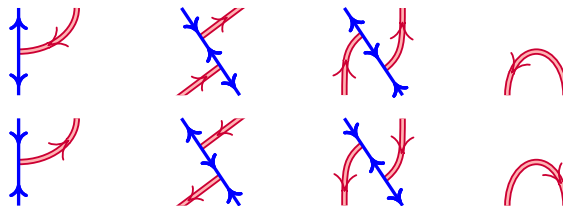


Figure 5: Web resolutions of a positive (to the left) and negative (to the right) crossing, together with the minimal foams between them.

isomorphisms between webs, but also a choice of equivalences between collections of points. For the latter one can use the webs



which are equivalences by Proposition 2.10. They can be used to construct an explicit equivalence from a collection  $\Sigma$  to  $\Sigma^0 = \mathcal{E}(\Sigma_b, \lfloor w(\Sigma)/2 \rfloor)$  by examining the points of  $\Sigma$  from left to right. The details are left to the reader.

#### 4.4 Comparison of Khovanov brackets

We finish this section by comparing two invariant complexes for a tangle  $T$ : the Khovanov bracket  $[[T]]$  from [2], which is a formal complex of objects from  $\mathbf{BN}(\partial T)$ , and the Blanchet–Khovanov bracket  $[[T]]_F$  constructed using webs and foams instead. In what follows we recall the construction of the latter — forgetting red edges in webs and red facets in foams recovers the former.

Let  $c$  be the number of crossings in  $T$ , out of which  $c_+$  are positive and  $c_-$  are negative. The first step to construct  $[[T]]_F$  is to compute the *cube of resolutions* of  $\mathcal{I}_F(T)$ : a commutative diagram with resolutions of  $T$  at vertices of the  $c$ –dimensional cube  $[0, 1]^c$ . Namely, a vertex  $\xi = (\xi_1, \dots, \xi_c) \in \{0, 1\}^c$  is decorated with the web  $T_\xi$  obtained from  $T$  by replacing each  $i^{\text{th}}$  crossing of the tangle with its resolution of type  $\xi_i$ , as shown in Figure 5. Let  $\xi'$  be another vertex, obtained from  $\xi$  by changing



one coordinate from 0 to 1. The directed edge  $\zeta: \xi \rightarrow \xi'$  is decorated with the minimal foam  $T_\zeta: T_\xi \rightarrow T_{\xi'}$ , which is a collection of vertical facets except over the region where the two resolutions do not match; here  $T_\zeta$  is a zip or an unzip as shown in Figure 5. It is evident that  $\mathcal{I}_F(T)$  commutes: directed paths between same vertices represent isotopic foams.

Pick a *sign assignment*  $\epsilon$ , that is a collection of signs  $\epsilon(\zeta) = \pm 1$ , one sign per edge in the cube, such that the product of signs around any square in the cube is equal to  $-1$ . The standard choice is  $\epsilon(\zeta) = (-1)^{s(\xi, \xi')}$ , where  $s(\xi, \xi')$  counts 1's left to the place at which  $\xi$  and  $\xi'$  disagree. Scaling each edge  $\zeta$  by  $\epsilon(\zeta)$  makes the cube anticommute and it can be shown that the isomorphism type of the cube is independent of the sign assignment—compare with [26, Lemma 2.2] or [27, Lemma 5.7]. The formal complex  $\llbracket T \rrbracket_F$  is obtained by flattening the cube along diagonals and shifting degrees accordingly. Explicitly,

$$\llbracket T \rrbracket_F^i := \bigoplus_{|\xi|=i+c_-} T_\xi \{c_- - c_+ - i\}$$

where  $|\xi| := \xi_1 + \dots + \xi_c$ , with the differential

$$d|_{T_\xi} = \sum_{\zeta: \xi \rightarrow \xi'} \epsilon(\zeta) T_\zeta.$$

The Khovanov bracket  $\llbracket T \rrbracket$  is constructed following the same steps, except that webs and foams are replaced with flat tangles and cobordisms. In particular, one has to erase in Figure 5 the red edges in resolutions and red facets in foams.

**Theorem 4.6** *The homotopy type of  $\llbracket T \rrbracket_F$  is an invariant of the tangle  $T$ , strictly functorial with respect to tangle cobordism. Its image under  $\mathcal{E}_{\partial T}^\vee$  is isomorphic to  $\llbracket T \rrbracket$ .*

**Proof** Following [2] one can show that  $\llbracket T \rrbracket$  is functorial up to a sign and strict functoriality is shown in [5] in the case of links, ie when  $T$  has no endpoints. From these two facts strict functoriality follows, because every tangle can be closed to a link.

To compare  $\llbracket T \rrbracket_F$  with  $\llbracket T \rrbracket$  consider the cube of resolutions  $\mathcal{I}_F(T)$  constructed in **Foam**( $\partial T$ ) and let  $\mathcal{I}(T)'$  be its image in **BN**( $\partial T$ ) under the equivalence of categories  $\mathcal{E}_{\partial T}^\vee$ . It differs from  $\mathcal{I}(T)$ , the cube of resolutions in **BN**( $\partial T$ ) that computes  $\llbracket T \rrbracket$ , only in signs at edges. Hence, the two cubes are isomorphic and the thesis follows.  $\square$

**Remark 4.7** The construction of  $\llbracket T \rrbracket_F$  can be easily extended to an invariant of knotted webs [29] and it is conjectured to be strictly functorial with respect to foams embedded in a four-dimensional space.

## 5 A diagrammatic TQFT on Foam( $\emptyset$ )

The assignment of the module  $Foam(\omega)$  to a closed web  $\omega$  extends to a functor

$$\text{Hom}_{\mathbf{Foam}(\emptyset)}(\emptyset, -) : \mathbf{Foam}(\emptyset) \rightarrow \mathbb{k}\text{-Mod}.$$

In what follows we provide a diagrammatic description of this functor by representing red-over-blue cup foams from  $Foam(\omega)$  using certain planar diagrams and examine how the diagrams change under the action of the linear maps associated with foams. In this section we assume that webs and foams are embedded in the plane  $\mathbb{R}^2 \times \{0\}$  and the half-space  $\mathbb{R}^2 \times (-\infty, 0]$  respectively.

### 5.1 A planar representation of cup foams

Let  $\omega$  be a bounded planar web and  $\omega^+$  its completion, which is a shading  $(\omega_r^+, \omega_b^+)$  satisfying  $\Gamma(\omega_r^+, \omega_b^+) = \omega$ . It is assumed that the basepoint  $*$  marks the unbounded region, so that the region is painted white. To simplify the picture and make the web  $\omega$  better visible, we do not color regions and we draw red edges as double or dashed lines depending on whether they survive or disappear after  $\Gamma$  is applied; see Figure 6. Furthermore, we allow to mark blue loops of  $\omega^+$  with (any number of) dots. We assign to such a planar diagram a completed foam  $\text{cup}(\omega^+) = \text{cup}_r(\omega^+) \cup \text{cup}_b(\omega^+)$  bounded by  $\omega^+$  that satisfies the following conditions:

- (CF1)  $\text{cup}_r(\omega^+) \subset \mathbb{R}^2 \times [-1, 0]$  and consists of disks that project injectively onto  $\mathbb{R}^2 \times \{0\}$ .
- (CF2)  $\text{cup}_b(\omega^+)$  is a collection of disks such that
 
$$\text{cup}_b(\omega^+) \cap (\mathbb{R}^2 \times [-1, 0]) = \omega_b^+ \times [-1, 0].$$
- (CF3) Each blue disk is decorated with as many dots as its boundary loop in  $\omega^+$ , all placed at heights smaller than  $-1$  (hence, below all red facets).

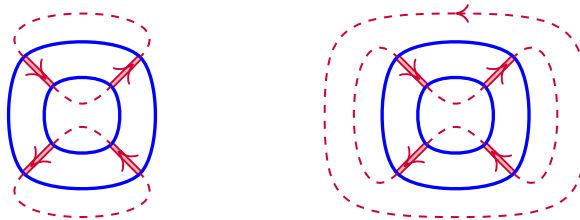


Figure 6: Two completions of the same web. The surrounding dashed circle in the right picture is required by the condition that the unbounded region is painted white.

The intersection of red and blue disks in  $\text{cup}(\omega^+)$  consists of intervals only; hence it is minimal among all completed cup foams bounded by  $\omega^+$ . Painting the unbounded region white extends to a unique shading supported by  $\text{cup}(\omega^+)$ . The resulting foam  $\Gamma(\text{cup}(\omega^+)) \in \text{Foam}(\omega)$  is a red-over-blue cup foam, ie its red facets are above all dots and minima of blue discs. We call it the *cup foam associated to  $\omega^+$* . The following observation is an immediate consequence of Theorem 2.11.

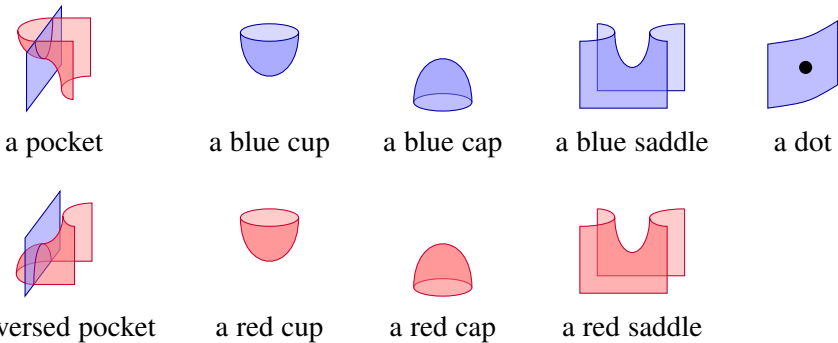
**Lemma 5.1** *Choose a completion  $\omega^+$  of  $\omega$  and consider the family of all dotted completed webs obtained from  $\omega^+$  by placing at most one dot on each blue loop. Then the corresponding cup foams form a linear basis of  $\text{Foam}(\omega)$ .*

Notice that dots in these pictures only mark loops. In particular, moving a dot along a loop—even passing through a crossing with a red strand—does not affect the cup foam represented by the diagram.

**Example 5.2** Let  $\omega$  be a blue circle. Then  $\text{Foam}(\omega)$  is generated by two blue cups: one with and the other without a dot. These are the cup foams associated to  $\omega^+$  when  $\omega$  is oriented clockwise, because this orientation is oriented from the unbounded region, hence  $\omega^+ = \omega$ . Otherwise,  $\omega^+$  is  $\omega$  surrounded by a dashed red circle, which results in the change of the sign of the cup with a dot; see Table 1. This is consistent with the computation from Example 4.4.

### 5.2 Action of foams

We now provide a description of the linear maps associated to foams in terms of the dotted completed webs. In fact, it is enough to analyze the *elementary completed foams*



because every foam can be decomposed into these.

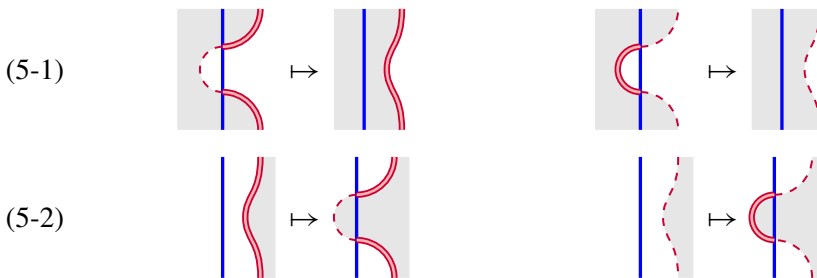
**Pockets and bicolored isotopies** A bicolored isotopy is a sequence of several bigon moves (3-1), which are realized by the pocket foams. When applied to a (completed) cup

web	completion	basis	
			$\leftrightarrow$
			$\leftrightarrow$
			$\leftrightarrow$
			$\leftrightarrow$
			$=$
			$= -$

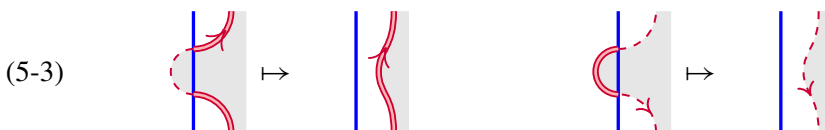
Table 1: A basis for foams bounded by a circle, represented as planar diagrams and as foams.

foam, it results in a collection of disks that may or may not satisfy (CF1); see Figure 7. In order to describe the map by local pictures, we shade the projection of the disk bounded by the red loop involved in the bigon move, ie the region bounded by the loop.

If the projection of the red disk is pushed through a blue arc, then the resulting cup foam is minimal — no double points in the projection are created. Hence, the associated map takes a dotted web to the result of applying the bigon move:



However, pulling the projection of a red disk off a blue arc creates double points, like in the right column of Figure 7. Indeed, the new red disk intersects the blue surface in a circle, so that either of (3-2) has to be applied. This may cost a sign, depending on the orientation of the edges:



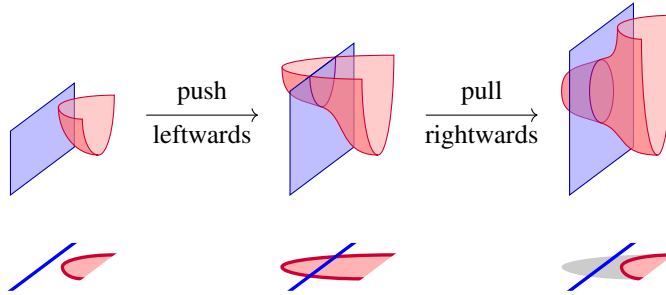
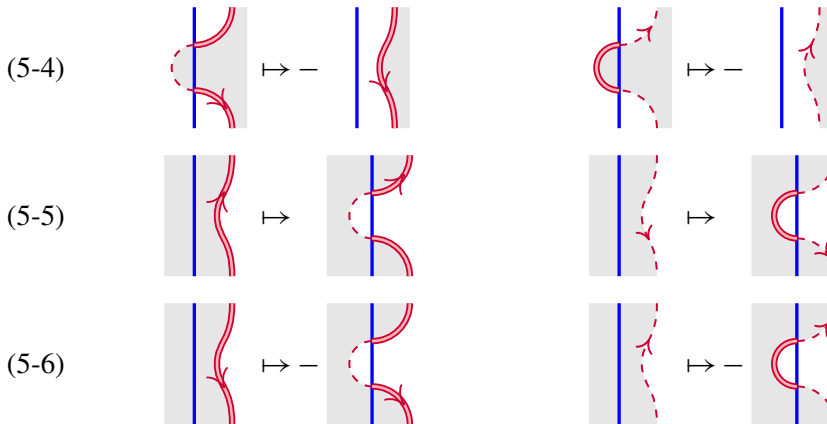


Figure 7: A cup foam with its projection on the horizontal plane (the left column) and the results of applying the bigon move twice (the middle and right columns). The middle foam is again a cup foam, but not the right one: the projection has double points — the shaded region outside of the cup — coming from the disk to the side of the vertical plane.



Indeed, the left moves in (5-3) and (5-4) are realized by detaching red cylinders with (2-9), whereas the right ones are by eliminating red caps with (2-16). Likewise, the relations (2-17) and (2-10) give the signs for the left and right sides respectively of both (5-5) and (5-5).

**Placing a dot** Placing a dot on  $\text{cup}(\omega^+)$  near the boundary violates (CF3). To obtain a minimal cup foam, the dot has to be moved down.

Let  $p$  be the projection of the dot onto the horizontal plane and assume that it does not lie on a red loop. We define the *nestedness*  $n(p)$  as the number of red loops encircling  $p$ . It counts red facets below  $p$  in the cup foam, hence, the number of times the dot-moving relation (2-8) has to be applied to move the dot from top to the bottom of a blue disk.

Therefore, placing a dot on a blue loop results in the map

$$(5-7) \quad / \mapsto \begin{cases} \bullet / & \text{if } n(p) \text{ is even,} \\ h / - \bullet & \text{if } n(p) \text{ is odd.} \end{cases}$$

**Blue cups, caps, and saddles** Suppose now that  $W$  is a completed foam with  $\omega^+$  at its bottom and a unique critical point that lies on the blue surface. In this case  $W \cup \text{cup}(\omega^+)$  is no longer a cup foam associated with the output of  $W$ : to have one, the critical point of  $W$  has to be slid downwards, below all red facets, and this may cost a sign. Moreover, a cap creates a sphere that has to be evaluated, whereas a saddle splitting one loop into two<sup>11</sup> creates a neck that has to be cut.

Let  $p$  be the projection of the critical point onto the horizontal plane and assume that  $p \notin \omega^+$ . We say that a red loop  $\gamma$  encircling  $p$  is *evenly distanced* if any generic path connecting  $p$  to a point  $q$  from a solid (resp. dashed) red arc of  $\gamma$  intersects blue circles in an even (resp. odd) number of points. Otherwise,  $\gamma$  is *oddly distanced*. Let  $s(p)$  count *oddly distanced* counterclockwise and *evenly distanced* clockwise red loops surrounding  $p$ . This corresponds to two types of red facets below  $p$ :

- (1) the facets that survive in the cup foam  $\Gamma(\omega^+)$  and with the canonical normal vector (see Remark 2.7) oriented downwards, and
- (2) those removed from  $\Gamma(\omega^+)$  and with the canonical normal vector oriented upwards.

These are exactly the situations, in which there is a sign in the relations (2-9) and (2-10). Hence  $s(p)$  determines the result of isotoping the blue critical point below all red facets. Therefore, the maps induced by critical blue points are the usual ones scaled by  $(-1)^{s(p)}$ :

- a cup:

$$(5-8) \quad \emptyset \mapsto (-1)^{s(p)} \bigcirc$$

- a cap:

$$(5-9) \quad \bigcirc \mapsto (-1)^{s(p)}$$

<sup>11</sup>Such a saddle is called a *split*. The other one, which joins two loops into one, is called a *merge*.

- a merge:

$$(5-10) \quad \text{[Two blue circles]} \mapsto (-1)^{s(p)} \text{[Blue saddle]}$$

- a split:

$$(5-11) \quad \text{[Blue saddle]} \mapsto (-1)^{s(p)} \left( \text{[Blue circle with dot]} + \text{[Blue circle with dot]} \right)$$

**Red cups, caps, and saddles** Placing a red cup at the top of  $\text{cup}(\omega^+)$  results in a cup foam. Hence, no sign appears. Conversely, capping off an isolated red circle creates a red sphere, which can be removed by (2-6) at a cost of sign. Hence, we obtain the maps:

- a cup:

$$(5-12) \quad \emptyset \mapsto \text{[Red cup]} \quad \emptyset \mapsto \text{[Red dashed cap]}$$

- a cap:

$$(5-13) \quad \text{[Red cup]} \mapsto -1 \quad \text{[Red dashed cap]} \mapsto 1$$

The behavior of merging and splitting saddles depends on whether the two red circles (those being merged or the result of a split) are nested or not. In the latter case, merging two loops takes a minimal cup foam to a minimal cup foam, whereas splitting a red circle creates a neck that has to be cut with (2-7) if it survives in the foam. Therefore, the corresponding maps satisfy the rules:

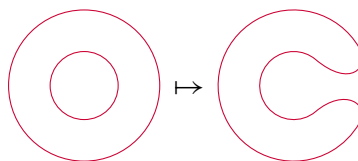
- a merge:

$$(5-14) \quad \text{[Red cup]} \text{ [Red dashed cap]} \mapsto \text{[Red saddle]} \quad \text{[Red dashed cap]} \text{ [Red dashed cap]} \mapsto \text{[Red dashed saddle]}$$

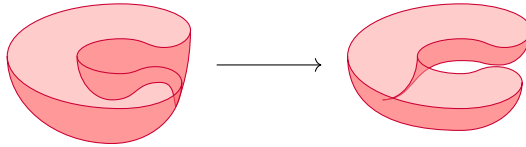
- a split:

$$(5-15) \quad \text{[Red saddle]} \mapsto - \text{[Red cup]} \text{ [Red dashed cap]} \quad \text{[Red dashed saddle]} \mapsto \text{[Red dashed cap]} \text{ [Red dashed cap]}$$

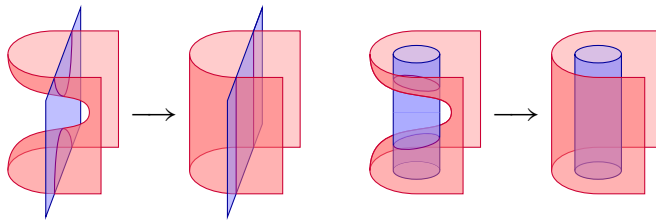
Merging nested red loops of a cup foam



results in a red disk, which does not project injectively onto the horizontal plane, but which can be isotoped to a croissant:



This isotopy can be described as a *finger move*: place your finger vertically near the saddle and move it inwards, pushing the red disk. The disk is then isotoped through every blue facet attached to a blue arc that cuts the inner red circle and through every cylinder attached to a blue circle surrounded by the inner red circle:



Depending on which red facets survive, each move represents two relations between foams. We leave it to the reader to check that the foams involved in the right move are always equal, whereas the left move costs a sign only in the two configurations



where the position of a saddle is marked with a cross. Let  $c$  be the number of such configurations. Then we end up with the following formula for merging nested red loops:

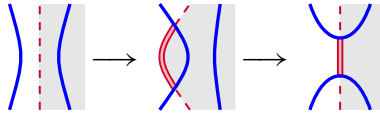
$$(5-16) \quad \begin{array}{c} \text{outer loop} \\ \text{inner loop} \end{array} \mapsto (-1)^c \begin{array}{c} \text{outer loop} \\ \text{inner loop} \end{array} \quad \begin{array}{c} \text{outer loop} \\ \text{inner loop} \end{array} \mapsto (-1)^c \begin{array}{c} \text{outer loop} \\ \text{inner loop} \end{array}$$

Dually, splitting a red loop into two nested loops is realized by the inverse of the finger move followed by stacking a red saddle over the little saddle connecting the two boundary loops. This creates a neck that has to be cut if it survives in the cup foam, so that in this case the sign is opposite to the one of the nested merge:

$$(5-17) \quad \begin{array}{c} \text{outer loop} \\ \text{inner loop} \end{array} \mapsto -(-1)^c \begin{array}{c} \text{outer loop} \\ \text{inner loop} \end{array} \quad \begin{array}{c} \text{outer loop} \\ \text{inner loop} \end{array} \mapsto (-1)^c \begin{array}{c} \text{outer loop} \\ \text{inner loop} \end{array}$$



**Other common foams** There are other moves of interest, such as blue saddles with vertical red facets (in particular, zips and unzips) or red cups and caps that intersect vertical blue facets. All can be represented as compositions of those described above. For instance, a zip is isotopic to a pocket move followed by a saddle:



As before, the shaded regions represent a projection of a red disk, and it is clear that the first move takes a basic cup foam to a basic cup foam, so that signs are governed by the second move. Therefore, the map induced by a zip is one of

- a merging zip:

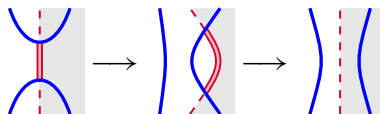
$$(5-18) \quad \text{Diagram} \mapsto (-1)^{s^+(p)} \text{Diagram}$$

- a splitting zip:

$$(5-19) \quad \text{Diagram} \mapsto (-1)^{s^+(p)} \left( \text{Diagram} + \text{Diagram} \right)$$

where  $s^+(p)$  is computed like  $s(p)$ , except that we take into account the loop passing through the created red edge if it is oriented counterclockwise.

In the unzip the saddle precedes the pocket and to ensure that the latter does not affect the sign, we perform the saddle to the side of the red disk attached to the red edge:



Therefore, the induced map is one of

- a merging unzip:

$$(5-20) \quad \text{Diagram} \mapsto (-1)^{s^-(p)} \text{Diagram}$$

- a splitting unzip:

$$(5-21) \quad \text{Diagram} \mapsto (-1)^{s^-(p)} \left( \text{Diagram} + \text{Diagram} \right)$$

where again  $s^-(p)$  is computed like  $s(p)$  without counting the loop passing through the removed red edge.

## 6 The Blanchet–Khovanov invariant of tangles with balanced boundaries

Let  $\mathbf{Foam}^\circ$  be the subcategory of  $\mathbf{Foam}$  generated by balanced sequences. In what follows we construct a TQFT functor  $\mathcal{F}_w^\circ: \mathbf{Foam}^\circ \rightarrow \mathbf{Bimod}$ . If  $T$  is an oriented tangle with balanced input and output collections of points, then its resolutions are in  $\mathbf{Foam}^\circ$ , so that applying  $\mathcal{F}_w^\circ$  to  $[[T]]_F$  results in a chain complex of bimodules. We then show that this chain complex is isomorphic to the Khovanov’s tangle invariant [19], but it admits a strictly functorial action of tangle cobordisms.

### 6.1 A linear basis of webs

A web  $\omega \subset \mathbb{R} \times (-\infty, 0]$  is called a *cup web* if its underlying tangle is a cup diagram, ie a collection of disjoint arcs. All cup webs with the same boundary and extending the same cup diagram coincide in  $\mathbf{Web}$  due to Proposition 2.4 (and are isomorphic as objects of  $\mathbf{Foam}$ ). Moreover, choosing a cup web for each cup diagram results in a basis of the space of webs with given boundary, which we call a *cup basis*.

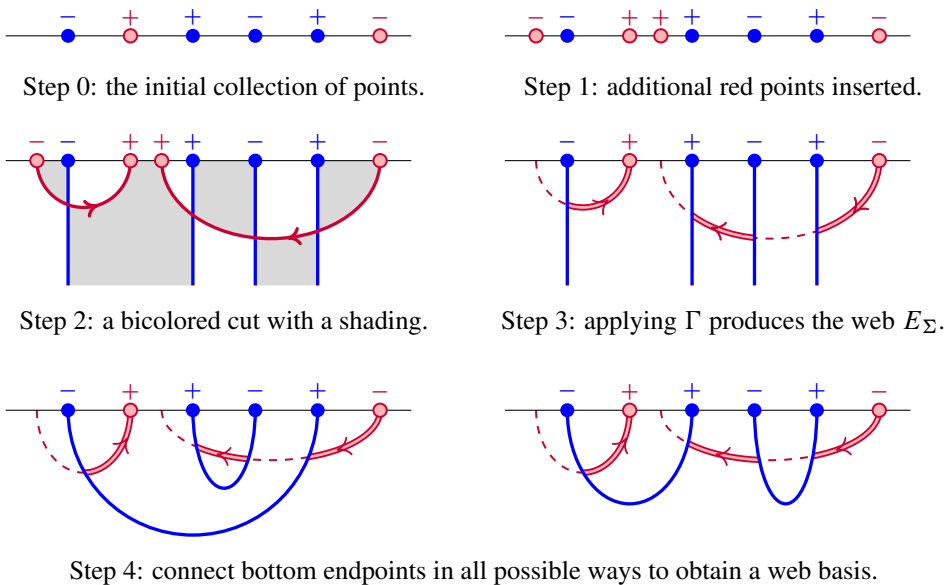


Figure 8: The construction of a cup basis for  $\Sigma = \bullet \circ \bullet \bullet \bullet \circ$ , together with completions of the cup webs (the edges erased in the third step are drawn as dashed arcs). The first three steps follow the proof of Proposition 3.4.

We describe now a particular nice cup basis of webs bounded by a balanced  $\Sigma$  (see also Figure 8). Let  $n$  be half of the number of blue points in  $\Sigma$  (being balanced,  $\Sigma$  has an even number of blue points). Proposition 3.4 provides an invertible web  $E_\Sigma: \Sigma_b \rightarrow \Sigma$  with  $2n$  vertical lines as blue edges. To obtain a cup basis, attach cup diagrams to the bottom of  $E_\Sigma$ . In other words, the basis is the image of cup diagrams under the equivalence  $\mathcal{E}_\Sigma$  from Section 4. We call it the *red-over-blue basis of type  $E_\Sigma$* , because all red edges in the webs appear above minima of blue cups.

### 6.2 Blanchet–Khovanov algebras

Fix a balanced collection of points  $\Sigma$  and let  $\mathcal{B}$  be a cup basis of  $Web(\Sigma)$ .

**Definition 6.1** The *Blanchet–Khovanov algebra*  $\mathfrak{M}^\mathcal{B}$  associated with  $\mathcal{B}$  is the direct sum of spaces of foams with corners

$$(6-1) \quad \mathfrak{M}^\mathcal{B} := \bigoplus_{a,b \in \mathcal{B}} \text{Hom}_{\mathbf{Foam}(\Sigma)}(a, b)$$

with multiplication given by the composition (and zero if foams cannot be composed).

**Remark 6.2** The above algebra appeared first in [14] for  $\Sigma$  a collection of positively oriented blue points followed by negatively oriented red points, the latter drawn in [14] at the bottom.

Choose a completion  $a^+$  for any cup web  $a$  and write  $a^!$  (resp.  $(a^+)^!$ ) for the result of reflecting  $a$  (resp.  $a^+$ ) along the horizontal line and reversing orientation of edges. Using the natural isomorphisms  $\text{Hom}_{\mathbf{Foam}(\Sigma)}(a, b) \cong \text{Foam}(b^!a)\{n\}$  we can represent elements of the algebra by dotted completed webs, in which case the multiplication is induced from the family of *generalized saddles*

$$(6-2) \quad (c^+)^!b^+ \sqcup (b^+)^!a^+ \xrightarrow{S_{c.b.a}} (c^+)^!a^+$$

each consisting of the identity foams  $(c^+)^! \times [0, 1]$  and  $a^+ \times [0, 1]$  glued to the half-rotation of  $b^+$  around the boundary line. These foams take a particularly nice form when  $\mathcal{B}$  is a red-over-blue basis, as they involve then only three types of moves:

- merging (5-10) and splitting (5-11) blue loops at points outside of all red circles,
- merging unnested red loops (5-14), and
- removing bigons *external* to the projection of red disks (5-1).

Hence, the product of two dotted diagrams is a positive linear combination of other diagrams.

**Corollary 6.3** *The algebra  $\mathfrak{W}^{\mathcal{B}}$  admits a positive basis.*

When  $\Sigma$  is a collection of blue points oriented in the alternating way and  $\mathcal{B}$  consists of oriented cup diagrams (ie webs with no red edges), then  $\mathfrak{W}^{\mathcal{B}}$  coincides with the arc algebra  $H^n$  from [19]. Indeed,  $\mathfrak{W}^{\mathcal{B}}$  is the image of  $H^n$  under the embedding of bicategories  $\mathcal{E}: \mathbf{BN} \rightarrow \mathbf{Foam}$ . However, when  $\mathcal{B}$  is not a red-over-blue basis, then the generalized saddles (6-2) may involve moves on red arcs that cost a sign, such as splits (5-15) or nested saddles (5-16) and (5-17). Hence, cup foams do not constitute a positive basis of the algebra in such case. Yet, it is still isomorphic to the arc algebra.

**Theorem 6.4** *Let  $\Sigma$  be a balanced collection of points with  $2n$  blue points. Then there is an algebra isomorphism  $\mathfrak{W}^{\mathcal{B}} \cong H^n$  for any cup basis  $\mathcal{B}$  of webs bounded by  $\Sigma$ . When  $\mathcal{B}$  is a red-over-blue basis, then the isomorphism simply forgets red facets of basic cup foams.*

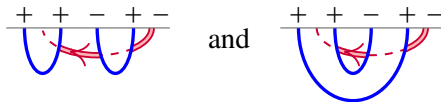
**Proof** Assume first that  $\mathcal{B}$  is a red-over-blue basis of type  $E_{\Sigma}$ . Then  $\mathfrak{W}^{\mathcal{B}}$  is the image of  $H^n$  under the equivalence of categories  $\mathcal{E}_{\Sigma}$ , which equips a collection of dotted cups with its standard orientation. The inverse of  $E_{\Sigma}$  simply forgets red edges in webs and red facets in foams. Hence, the thesis follows.

Let now  $\mathcal{B}'$  be any cup basis and pick for each cup web  $a' \in \mathcal{B}'$  the isomorphic cup web  $a \in \mathcal{B}$ , an invertible foam  $I_a \in \mathbf{Foam}(a, a')$ , and  $s(a) = \pm 1$  such that  $I_a^! I_a = s(a) \omega_a \times [0, 1]$ . Then the collection of linear isomorphisms

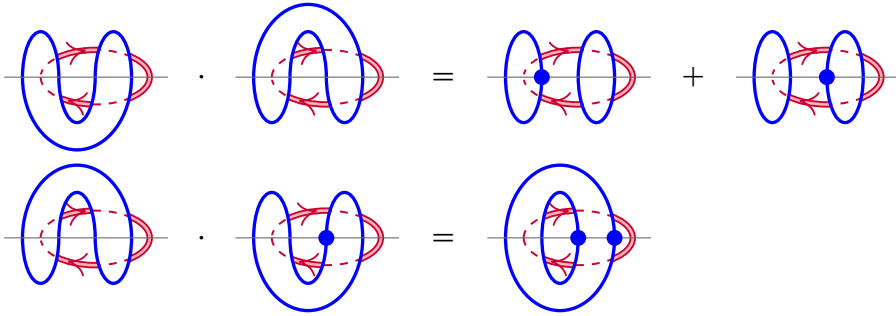
$$\varphi_{ba}: \mathbf{Foam}(a', b') \xrightarrow{\cong} \mathbf{Foam}(a, b), \quad S \mapsto s(b) I_b^! S I_a,$$

constitutes an isomorphism of algebras  $\mathfrak{W}^{\mathcal{B}'} \cong \mathfrak{W}^{\mathcal{B}}$ , where the latter is isomorphic to  $H^n$ . □

**Example 6.5** Let  $\Sigma = \begin{matrix} + & + & - & + & - \\ \bullet & \bullet & \bullet & \bullet & \circ \end{matrix}$ . Then the cup basis  $\mathcal{B}$  consists of two elements



that form four pairs: two of them have two blue loops and each of the other two has one blue loop. Hence,  $\dim \mathfrak{W}^{\mathcal{B}} = 12$ . The multiplication of any two diagrams involves only merges and splits of blue loops (5-10)–(5-11) with  $s(p) = 0$ , an unnested merge of red loops (5-14) and bigon moves (5-1). None of them introduces signs, so that the product is always a positive sum of diagrams, like in  $H^2$ . For instance:



Erasing red edges recovers the usual diagrammatic calculus of  $H^2$ .

**Example 6.6** Recall the Blanchet–Khovanov algebra from [14]: it is defined using webs that have only vertical red edges,  $2n$  positive blue endpoints at the top and positive  $n$  red endpoints at the bottom. We call them here *EST webs*. To fit this construction into our framework, we move the bottom endpoints rightwards and to the top by appending a collection of nested red cups; see Figure 9. Contrary to the case of red-over-blue bases, minima of red cups in EST webs appear below blue cups. This is the reason why the minus sign appears in the formula for multiplication quite often, already in the case of four points. Yet Theorem 6.4 provides a direct isomorphism between this algebra and Khovanov’s arc algebra. Such an isomorphism was explicitly constructed in [13] by providing a sign for each generator, then checking directly that these signs result in a homomorphism of algebras.

### 6.3 Blanchet–Khovanov bimodules

Pick now two balanced collections  $\Sigma$  and  $\Sigma'$  with cup bases  $\mathcal{B}$  and  $\mathcal{B}'$  respectively. We assign to a web  $\omega : \Sigma \rightarrow \Sigma'$  its *Blanchet–Khovanov bimodule*  $\mathcal{F}_\omega^\circ$ , which is the

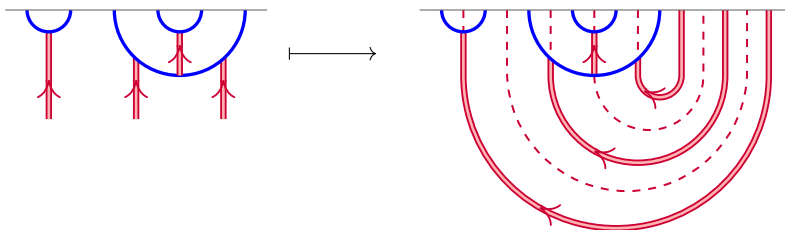


Figure 9: Turning red edges rightwards and to the top produces a cup web from an EST web. There is a natural completion, visualized by dashed arcs, with minima of red cups below all blue ones.

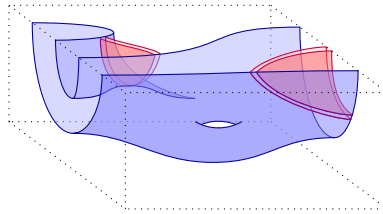


Figure 10: An element of  $\mathbf{Foam}(a, \omega, b)$  is a foam in a cube, bounded by the webs  $\omega, a$ , and  $b$  at the top and opposite vertical facets of the cube respectively.

$(\mathfrak{W}^{\mathcal{B}}, \mathfrak{W}^{\mathcal{B}'})$ –bimodule

$$(6-3) \quad \mathcal{F}_w^\circ(\omega) := \bigoplus_{a \in \mathcal{B}, b \in \mathcal{B}'} \mathbf{Foam}(a, \omega, b),$$

where  $\mathbf{Foam}(a, \omega, b)$  is the space of foams bounded by  $b^1 \cup \omega \cup a$  and seen as foams in a cube with  $\omega$  at the top facet, whereas  $a$  and  $b$  lie on opposite vertical facets; see Figure 10. The algebras  $\mathfrak{W}^{\mathcal{B}}$  and  $\mathfrak{W}^{\mathcal{B}'}$  act on the left and on the right respectively, and there is a diagrammatic presentation of this bimodule as explained in Section 5. Moreover, placing a foam  $S \in \mathbf{Foam}(\omega, \omega')$  on top results in a bimodule map  $\mathcal{F}_w^\circ(S): \mathcal{F}_w^\circ(\omega) \rightarrow \mathcal{F}_w^\circ(\omega')$ . The assignment  $S \mapsto \mathcal{F}_w^\circ(S)$  is clearly functorial in  $S$  and it preserves the foam relations (2-6)–(2-10). In particular,  $\mathcal{F}_w^\circ(\omega) \cong \mathcal{F}_w^\circ(\omega')$  if  $S$  is invertible. Finally, horizontal composition of foams induces a canonical homomorphism of graded bimodules

$$(6-4) \quad \mathcal{F}_w^\circ(\omega') \otimes_{\mathfrak{W}^{\Sigma'}} \mathcal{F}_w^\circ(\omega) \rightarrow \mathcal{F}_w^\circ(\omega' \omega)$$

for any pair of composable webs  $\omega: \Sigma \rightarrow \Sigma'$  and  $\omega': \Sigma' \rightarrow \Sigma''$ . The proof of [19, Theorem 1] can be adapted to our framework to show that (6-4) is an isomorphism.

**Remark 6.7** Contrary to the case of Blanchet–Khovanov algebras, the isomorphism (6-4) may not take a pair of cup foams into a positive combination of cup foams. When using the diagrammatics of completed webs, (6-4) is induced by a collection of generalized saddles, the description of which — contrary to the case of algebras — may involve moves on red loops that cost a sign, such as (5-15)–(5-17). However, this is not the case when both webs have only blue endpoints, oriented in an alternating way — in this case all red loops lie inside the webs and are not affected when webs are composed.

As in the case of algebras,  $\mathcal{F}_w^\circ(\omega)$  coincides with the arc bimodule  $\mathcal{F}_{Kh}^\circ(\omega)$  defined in [19] when  $\omega$  is a standardly oriented flat tangle and both  $\mathcal{B}$  and  $\mathcal{B}'$  are collections of

cup diagrams. Although in general  $\mathcal{F}_w^\circ(\omega)$  is not a priori a bimodule over arc algebras, it can be made such through the algebra isomorphisms  $H^n \cong \mathfrak{W}^{\mathcal{B}}$  and  $H^{n'} \cong \mathfrak{W}^{\mathcal{B}'}$  provided by Theorem 6.4. Hence, it makes sense to compare  $\mathcal{F}_w^\circ(\omega)$  with  $\mathcal{F}_{Kh}^\circ(\omega_b)$ .

**Theorem 6.8** *Let  $\omega: \Sigma \rightarrow \Sigma'$  be a web between balanced collections of points with  $2n$  and  $2n'$  blue points respectively. Then there is an isomorphism of  $(H^n, H^{n'})$ -bimodules  $\mathcal{F}_w^\circ(\omega) \cong \mathcal{F}_{Kh}^\circ(\omega_b)$ . The isomorphism simply forgets red facets of basic cup foams when  $\mathcal{B}$  and  $\mathcal{B}'$  are red-over-blue web bases.*

**Proof** Assume first that  $\mathcal{B}$  and  $\mathcal{B}'$  are red-over-blue cup bases of types  $E_\Sigma$  and  $E_{\Sigma'}$  respectively. Fix a foam  $I_\omega$  in a cube with vertical rectangles as blue facets, bounded by  $\omega_b$  and  $\omega$  at the bottom and top facets, and with  $E_\Sigma$  and  $E_{\Sigma'}$  at appropriate vertical facets. Placing it on top of an element from  $\mathbf{Foam}(a, \omega_b, b)$  results in a  $\mathbb{k}$ -linear isomorphism  $\mathbf{Foam}(a, \omega_b, b) \cong \mathbf{Foam}(E_\Sigma \cup a, \omega, E_{\Sigma'} \cup b)$ . It is straightforward to check that these isomorphisms are compatible with the action of the arc algebras, so that they constitute an isomorphism of bimodules  $\mathcal{F}_{Kh}^\circ(\omega_b) \cong \mathcal{F}_w^\circ(\omega)$ ; it takes a collection of dotted cups to a basic cup foam. Forgetting red facets is the inverse map.

The general case is reduced to the above as in the proof of Theorem 6.4: choose a collection of invertible foams, one per  $a \in \mathcal{B}$  and one per  $b \in \mathcal{B}'$ , and glue them to the sides of foams generating  $\mathcal{F}_{Kh}^\circ(\omega)$ . □

**Remark 6.9** When  $\mathcal{B}$  is a red-over-blue basis of type  $E_\Sigma$ , then the action of  $H^n$  can be understood pictorially as follows: a dotted surface  $S \in H^n$  is standardly oriented and combined with  $E_\Sigma \times [0, 1]$  before acting on  $\mathcal{F}_{Kh}^\circ(\omega)$ . The same applies to  $H^{n'}$  if  $\mathcal{B}'$  is a red-over-blue basis.

We say that a linear basis  $\{x_1, \dots, x_d\}$  of an  $(A, B)$ -bimodule is *positive with respect to bases*  $\{a_i\}$  of  $A$  and  $\{b_j\}$  of  $B$ , when each  $a_i x_k$  and  $x_k b_j$  has positive coefficients in this basis. Because dotted cups constitute a positive basis of arc bimodules, Theorem 6.8 implies the existence of a positive basis for Blanchet–Khovanov bimodules.

**Corollary 6.10** *Suppose that both  $\mathcal{B}$  and  $\mathcal{B}'$  are red-over-blue cup bases of webs. Then basic cup foams constitute a positive basis for  $\mathcal{F}_w^\circ(\omega)$ .*

Although the formulas for the actions of the algebras on a Blanchet–Khovanov bimodule involve no signs when red-over-blue bases as used, this is not the case for action of

foams: the square

$$(6-5) \quad \begin{array}{ccc} \mathcal{F}_w^\circ(\omega) & \xrightarrow{\mathcal{F}_w^\circ(S)} & \mathcal{F}_w^\circ(\omega') \\ I_\omega \uparrow \cong & & I_{\omega'} \uparrow \cong \\ \mathcal{F}_{Kh}^\circ(\omega_b) & \xrightarrow{\mathcal{F}_{Kh}^\circ(S_b)} & \mathcal{F}_{Kh}^\circ(\omega'_b) \end{array}$$

commutes only up to a sign, where we abuse the notation and denote the isomorphism from the proof of Theorem 6.8 by the same symbol  $I_\omega$  as the foam used to construct it. However, the sign does not depend on the direct summand of the bimodule: it is determined by the configuration of red loops (see Section 5) and the configuration is the same for all closures  $b^! \omega a$ .

### 6.4 A functorial homology for tangles with balanced boundaries

The previous sections describe a morphism of bicategories  $\mathcal{F}_w^\circ: \mathbf{Foam}^\circ \rightarrow \mathbf{Bimod}$ , which we extend naturally to  $Com_{/h}(\mathbf{Foam}^\circ)$ . As mentioned in the introduction, we can apply it to the formal bracket  $\llbracket T \rrbracket_F$  of a tangle  $T$  with balanced input and output, producing a chain complex  $C_{\mathfrak{W}}(T)$ . Invariance and functoriality of the bracket implies that the homotopy type of  $C_{\mathfrak{W}}(T)$  is an invariant of the tangle  $T$  that is functorial with respect to tangle cobordisms.

**Theorem C** *The 2–functor  $\mathcal{F}_w^\circ$  is equivalent to  $\mathcal{F}_{Kh}^\circ \circ \mathcal{E}^\vee$ . In particular, the complexes  $C_{\mathfrak{W}}(T)$  and  $CKh(T)$  are isomorphic for any tangle  $T$  with balanced input and output.*

**Proof** The two functors coincide on objects by Theorem 6.4 and on 1–morphisms by Theorem 6.8. Furthermore, the collection of isomorphisms  $I_\omega$  is natural in  $\omega$ , because the square (6-5) commutes when  $S_b$  is replaced with  $\mathcal{E}^\vee(S)$ . Indeed, the sign relating  $\mathcal{F}_w^\circ(S) \circ I_\omega$  with  $I_{\omega'} \circ \mathcal{F}_{Kh}^\circ(S_b)$  is exactly the one provided by  $\mathcal{E}^\vee$ . The last statement is a direct consequence of Theorem 4.6. □

## 7 Subquotient algebras and an invariant for all tangles

Inspired by [10] we use the TQFT from the previous section to define a family of 2–functors  $\mathcal{F}_w^\lambda: \mathbf{Foam} \rightarrow \mathbf{Bimod}$  parametrized by  $\lambda \in \mathbb{Z}$ , which are defined on the whole bicategory of foams. As before, these 2–functors lead to invariant chain complexes for tangles that are strictly functorial versions of the Chen–Khovanov tangle invariants. Contrary to the previous sections, we assume here that  $h = t = 0$ . In particular, a foam vanishes when it has a blue facet with two dots.



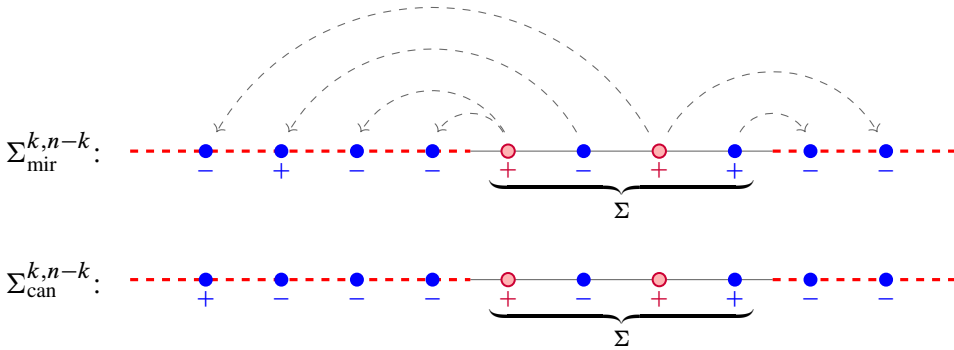


Figure 11: A visualization of two ways to balance a sequence for  $k = 2$ . For the mirror balancing (above) replace each red point with two blue points first, then reflect the left  $k + m = 4$  points to the left and the remaining ones to the right platform, changing their orientations. In the canonical balancing (below) the points on each platform are ordered with respect to their orientation.

### 7.1 Balancing

Suppose that  $\Sigma$  has  $m$  red and  $n$  blue points and choose  $0 \leq k \leq n$ . We say that a sequence  $\Sigma^\circ$  on a line with platforms is a *balancing of  $\Sigma$  of type  $n - 2k$*  if it is balanced and obtained from  $\Sigma$  by placing  $\ell$  and  $r$  blue points to the left and right of  $\Sigma$  respectively, where  $r - \ell = n - 2k$ . We say that the extra points lie on *platforms*, which are drawn as dashed lines. In what follows we describe two methods to balance a given sequence; see Figure 11.

The *mirror balancing*  $\Sigma_{\text{mir}}^{k,n-k}$  of  $\Sigma$  of type  $\lambda = n - 2k$  is constructed as follows. First, replace each red point by two blue points oriented the same way and call the new sequence  $\Sigma'$ . Then  $\Sigma_{\text{mir}}^{k,n-k}$  is obtained from  $\Sigma$  by reflecting the first  $k + m$  points of  $\Sigma'$  on the left and the remaining ones on the right platform, so that both sequences appear on the platforms in a reversed order; we also change the orientation of these points (compare with Figure 11). It is a balanced sequence, which is an alternating sequence of blue points if  $\Sigma$  is. However, it depends heavily on the orientations of points of  $\Sigma$ . The next construction does not share this drawback.

The *canonical balancing*  $\Sigma_{\text{can}}^{k,n-k}$  of type  $\lambda = n - 2k$  is constructed again by placing  $k + m$  points on the left and  $n - k + m$  points on the right platform, except that now we order the points in such a way that, when read from left to right, positive points on each platform appear first. Moreover, we want the minimal number of negative

(resp. positive) points on the left (resp. right) platform. This leads to one of the following distributions, depending on the total weight  $w = w(\Sigma)$  of the sequence  $\Sigma$ .

If  $|w| \geq |\lambda|$ :

Left platform: place  $\frac{1}{2}(|w| - \lambda)$  points of type  $-\text{sgn}(w)$ , then fill with  $+$ 's.

Right platform: place  $\frac{1}{2}(|w| + \lambda)$  points of type  $-\text{sgn}(w)$ , then fill with  $-$ 's.

If  $|w| < \lambda$ :

Left platform: fill with  $+$ 's.

Right platform: place  $\frac{1}{2}(\lambda - w)$  points of type  $+$ , then fill with  $-$ 's.

If  $|w| < -\lambda$ :

Left platform: place  $\frac{1}{2}(w - \lambda)$  points of type  $-$ , then fill with  $+$ 's.

Right platform: fill with  $-$ 's.

We check directly that in each case we obtain a balanced sequence with at most  $k + m$  negative and at most  $n - k + m$  positive points on the left and right platform respectively.

**Remark 7.1** The distribution of points on platforms in  $\Sigma_{\text{can}}^{k,n-k}$  depends only on the total weight  $w$  of the sequence and the type  $\lambda$  of the balancing, but not directly on the number of points nor their orientation. This is why we call it *canonical*.

## 7.2 Webs and foams with platforms

We now allow foams to meet the side vertical facets of the ambient cube in collections of horizontal blue lines. More precisely, fix a web  $\omega : \Sigma_0 \rightarrow \Sigma_1$  together with balanced collections  $\Sigma_0^\circ$  and  $\Sigma_1^\circ$ , such that the first  $\ell$  and last  $r$  points of both  $\Sigma_1^\circ$  and  $\Sigma_0^\circ$  are blue, oriented the same way, and removing them recovers  $\Sigma_1$  and  $\Sigma_2$  respectively. Given cup webs  $a$  and  $b$  bounded by  $\Sigma_0^\circ$  and  $\Sigma_1^\circ$  respectively, we write  $\widetilde{\text{Foam}}^{(\ell,r)}(a, \omega, b)$  for the space of foams embedded in a cube with the following boundary:

- the web  $\omega$  at the top facet of the cube,
- $\ell$  and  $r$  horizontal blue lines at the vertical facets to the left and to the right of  $\omega$  respectively, and
- the cup webs  $a$  and  $b$  at the vertical facets attached to the input and output of  $\omega$ .

Figure 12 provides examples of such foams. We say that such a foam is *violating* if it has a connected component that either

- meets a platform in more than one line, or
- intersects a platform and carries a dot.

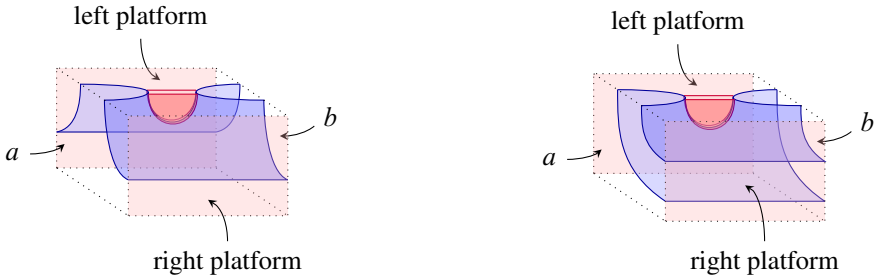


Figure 12: Examples of foams with platforms. The right foam is violating, because it is connected and intersects the right platform (the front facet in the picture) in two lines. The left foam is not violating unless it carries a dot.

It is straightforward to check that the property of being a violating foam is preserved by foam relations. Hence, violating foams generate a linear subspace of  $\widetilde{\mathbf{Foam}}^{(\ell,r)}(a, \omega, b)$ . We write  $\mathbf{Foam}^{(\ell,r)}(a, \omega, b)$  for the quotient space, or simply  $\mathbf{Foam}^{(\ell,r)}(a, b)$  when  $\omega$  is the identity web.

Gluing foams horizontally results in a linear map

$$\widetilde{\mathbf{Foam}}^{(\ell,r)}(a, \omega_0, b) \otimes \widetilde{\mathbf{Foam}}^{(\ell,r)}(b, \omega_1, c) \rightarrow \widetilde{\mathbf{Foam}}^{(\ell,r)}(a, \omega_0\omega_1, c)$$

and it is straightforward to notice that a foam  $S'S$  is violating when either  $S$  or  $S'$  is a violating foam. Hence, there is an induced linear map

$$\mathbf{Foam}^{(\ell,r)}(a, \omega_0, b) \otimes \mathbf{Foam}^{(\ell,r)}(b, \omega_1, c) \rightarrow \mathbf{Foam}^{(\ell,r)}(a, \omega_0\omega_1, c).$$

Consider now webs with platforms as discussed in Section 7.1. Their blue arcs fall into three families visualized in Figure 13:

- *inner arcs*, with at least one endpoint not on a platform,
- *outer arcs*, with each endpoint on a different platform, and
- *violating arcs*, with both endpoints on the same platform.

Webs with no violating arcs and no red endpoints on platforms are *admissible*. Outer arcs of an admissible web are nested one in another and the most nested one of them

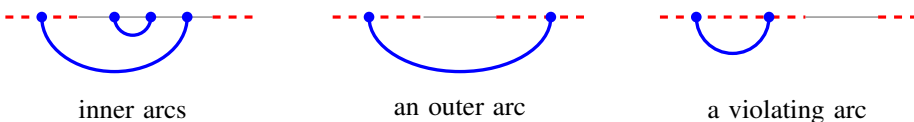


Figure 13: Three types of blue arcs in a cup diagram with platforms.

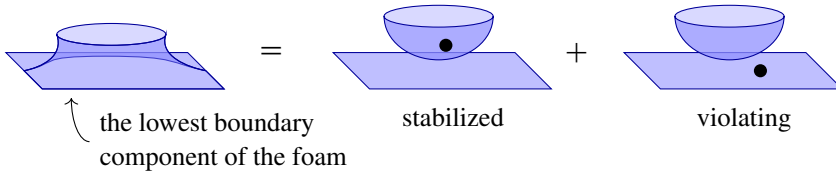


Figure 14: A way to destabilize a foam bounded by stabilized webs.

encloses all inner arcs. Notice that  $\mathbf{Foam}^{(\ell,r)}(a, \omega, b) = 0$  when either  $a$  or  $b$  has a violating arc.

**Lemma 7.2** *Let  $\Sigma^\circ$  be a balancing of  $\Sigma$  with  $\ell$  and  $r$  points on the left and on the right platform, and let  $n$  count the blue points of  $\Sigma$ . Then an admissible cup web bounded by  $\Sigma^\circ$  has at least  $(\ell + r - n)/2$  outer arcs.*

**Proof** An admissible web bounded by  $\Sigma^\circ$  has at most  $n$  inner arcs, so that at least  $(\ell + r) - n$  points from the platforms must be connected by outer arcs.  $\square$

Given a cup basis  $\mathcal{B}$  of webs bounded by  $\Sigma^\circ$ , we shall write  $\mathcal{B}^{\ell,r}$  for the subset of admissible webs with  $\ell$  and  $r$  points on the left and right platform respectively.

### 7.3 Stabilization

We say that a foam  $\widehat{S}$  is a *stabilization* of a foam  $S$ , when it is obtained by placing a blue horizontal rectangle below  $S$ . Likewise, *stabilizing a web* means adding an additional outer arc. It follows that  $\widehat{S}$  is a violating foam if and only if the foam  $S$  is violating. Hence, there is a well-defined injection

$$(7-1) \quad \mathbf{Foam}^{(\ell,r)}(a, \omega, b) \xrightarrow{(-)} \mathbf{Foam}^{(\ell+1,r+1)}(\widehat{a}, \omega, \widehat{b})$$

where  $\widehat{a}$  and  $\widehat{b}$  are appropriate stabilizations of the webs. It is also surjective: by applying the neck-cutting relation (2-7) we can write every foam  $S \in \mathbf{Foam}^{(\ell,r)}(\widehat{a}, \omega, \widehat{b})$  as a sum  $S_0 + S_1$ , such that the lowest blue boundary curve bounds a blue disk in each  $S_i$ , see Figure 14. Furthermore, stabilization is natural with respect to placing foams on top as well as to the horizontal composition of foams, ie

$$W \cup_\omega \widehat{S} = \widehat{W \cup_\omega S} \quad \text{and} \quad \widehat{S'} \widehat{S} = \widehat{S' S}$$

for any  $S \in \mathbf{Foam}^{(\ell,r)}(a, \omega, b)$ ,  $S' \in \mathbf{Foam}^{(\ell,r)}(b, \omega', c)$  and  $W : \omega \rightarrow \omega''$ .

Let  $\mathcal{B}$  be a cup basis of webs bounded by  $\Sigma^\circ$  and  $\mathcal{B}^{\ell,r}$  the subset of admissible webs. We write  $\widehat{\mathcal{B}}^{\ell,r}$  for the set of stabilized basic webs; they are bounded by a bigger

collection  $\widehat{\Sigma}^\circ$ . It is in general only a subset of a basis of admissible webs bounded by  $\widehat{\Sigma}^\circ$ . However, it is a basis when platforms carry sufficiently many points.

**Lemma 7.3** *Let  $\ell$  and  $r$  count points of  $\Sigma^\circ$  on the left and on the right platform, whereas  $n$  is the number of the remaining blue points. Then  $\widehat{\mathcal{B}}^{\ell,r}$  is a cup basis if  $\ell + r \geq n$ .*

**Proof** The collection  $\widehat{\Sigma}^\circ$  has  $\ell + 1$  points on the left and  $r + 1$  points on the right platform. Hence, by Lemma 7.2, every admissible web bounded by  $\widehat{\Sigma}^\circ$  has an outer arc.  $\square$

### 7.4 Subquotient algebras and bimodules

We are now ready to construct a foam version of the subquotient algebras and bimodules from [10]. Let  $\Sigma$  be a sequence of  $n$  blue and  $m$  red points, and pick  $\lambda = n - 2k$  with  $0 \leq k \leq n$ . Choose a balancing  $\Sigma^\circ$  with  $k + m$  and  $n - k + m$  points on the left and right platform respectively and a cup basis  $\mathcal{B}$  for webs bounded by it; the admissible webs form a subset  $\mathcal{B}^{k+m,n-k+m}$ .

**Definition 7.4** *The extended Blanchet–Khovanov algebra  $\mathfrak{A}^{\mathcal{B},\lambda}$  is the direct sum of spaces of foams with platforms*

$$(7-2) \quad \mathfrak{A}^{\mathcal{B},\lambda} := \bigoplus_{a,b \in \mathcal{B}^{k+m,n-k+m}} \mathbf{Foam}^{(k+m,n-k+m)}(a,b)$$

with multiplication given by the composition (and zero if foams cannot be composed).

It follows from the definition that  $\mathfrak{A}^{\mathcal{B},\lambda}$  is a quotient of a subalgebra of  $\mathfrak{W}^{\mathcal{B}}$ . In particular, it inherits the description in terms of dotted completed webs with the following modifications:

- the horizontal line, along which cup webs are glued, has platform sections on its sides;
- we allow only completions of admissible cup foams; and
- such a diagram vanishes when it contains a blue loop intersecting a platform at least twice or a blue loop intersecting a platform and carrying a dot at the same time.

In particular,  $\mathfrak{A}^{\mathcal{B},\lambda}$  is isomorphic to the Chen–Khovanov algebra  $A^{k,n-k}$  when  $\Sigma$  consists of  $n$  blue points that are oriented in an alternating way and  $\Sigma^\circ$  is the mirror balancing. With the help of the stabilization map (7-1) we can find such isomorphisms for all extended Blanchet–Khovanov algebras.

**Theorem 7.5** *Let  $\Sigma$  be a collection of  $m$  red and  $n$  blue points with a balancing  $\Sigma^\circ$  of type  $\lambda = n - 2k$ , where  $0 \leq k \leq n$ . Then there is an algebra isomorphism  $\mathfrak{A}^{\mathcal{B},\lambda} \cong A^{k,n-k}$  for any cup basis  $\mathcal{B}$  of webs bounded by  $\Sigma^\circ$ . When  $\mathcal{B}$  is a red-over-blue basis, the isomorphism simply forgets red facets of basic cup foams and drops  $m$  lowest blue rectangles.*

**Proof** We assume that  $\mathcal{B}$  is a red-over-blue basis — the general case is proven the same way as in Theorem 6.4. Let  $\mathcal{B}_0$  be the collection of admissible cup diagrams with  $2n$  blue endpoints,  $k$  of which are on the left and  $n - k$  on the right platform. It follows from Lemma 7.2 that each cup web from  $\mathcal{B}$  has  $m$  outer arcs, so that it is constructed by placing an invertible web  $E$  on top of cup diagrams from  $\mathcal{B}_0$  stabilized  $m$  times. Hence, as a  $\mathbb{k}$ -module,  $\mathfrak{A}^{\mathcal{B},\lambda}$  is isomorphic to  $\text{stab}^{(m)}(A^{k,n-k})$ , the algebra  $A^{k,n-k}$  stabilized  $m$  times. Because stabilization is compatible with horizontal composition of foams, we obtain an algebra isomorphism

$$A^{k,n-k} \xrightarrow{\widehat{(-)}^{(m)}} \text{stab}^{(m)}(A^{k,n-k}) \xrightarrow{(E \times [0,1]) \cup (-)} \mathfrak{A}^{\mathcal{B},\lambda}$$

which takes a dotted surface with platforms, adds  $m$  extra horizontal rectangles below, and then glues  $E \times [0, 1]$  to it along the top and platforms. The inverse of this map simply forgets red facets and drops the extra  $m$  blue rectangles as desired.  $\square$

We follow the same ideas to construct a collection of bimodules for a web  $\omega : \Sigma_0 \rightarrow \Sigma_1$ . Let  $n_i$  be the number of blue points in  $\Sigma_i$ . Choose  $0 \leq k_i \leq n_i$  for  $i = 0, 1$  such that  $n_0 - 2k_0 = n_1 - 2k_1 =: \lambda$ , and let  $\Sigma_0^\circ$  and  $\Sigma_1^\circ$  be the *canonical* balancings of type  $\lambda$ , except that we stabilize one of them, so that both sequences have the same numbers of points on platforms:  $\ell$  on the left and  $r$  on the right one. Notice that  $\Sigma_0$  and  $\Sigma_1$  have the same weight, so that their balancings agree on platforms. Choose cup bases  $\mathcal{B}_0$  and  $\mathcal{B}_1$  for  $\Sigma_0^\circ$  and  $\Sigma_1^\circ$  respectively. We assign to the web  $\omega$  the  $(\mathfrak{A}^{\mathcal{B}_0,\lambda}, \mathfrak{A}^{\mathcal{B}_1,\lambda})$ -bimodule

$$\mathcal{F}_\omega^\lambda(\omega) := \bigoplus_{a \in \mathcal{B}_0^{\ell,r}, b \in \mathcal{B}_1^{\ell,r}} \mathbf{Foam}^{(\ell,r)}(a, \omega, b),$$

which we call the *extended Blanchet–Khovanov bimodule of weight  $\lambda$* . The algebras  $\mathfrak{A}^{\mathcal{B}_0,\lambda}$  and  $\mathfrak{A}^{\mathcal{B}_1,\lambda}$  act on the left and right by composing foams horizontally, stabilized sufficiently many times when necessary. It should be also clear that placing a foam  $W : \omega \rightarrow \omega'$  on top induces a bimodule map  $\mathcal{F}_W^\lambda(W) : \mathcal{F}_\omega^\lambda(\omega) \rightarrow \mathcal{F}_\omega^\lambda(\omega')$ , and that taking tensor products over the algebras corresponds to composing webs

$$\mathcal{F}_\omega^\lambda(\omega) \otimes_{\mathfrak{A}^{\mathcal{B},\lambda}} \mathcal{F}_\omega^\lambda(\omega') \xrightarrow{\cong} \mathcal{F}_\omega^\lambda(\omega' \omega).$$

As before, Theorem 7.5 allows us to think of  $\mathcal{F}_w^\lambda(\omega)$  as a bimodule over the algebras  $A^{k_i, n_i - k_i}$ , so that we can compare it with the bimodule  $C_{CK}(\omega_b; \lambda)$  assigned to the flat tangle  $\omega_b$  in [10]. We leave the following as an easy exercise:

**Theorem 7.6** *Let  $\omega: \Sigma \rightarrow \Sigma'$  be a web between collections of points with  $n$  and  $n'$  blue points, respectively, and choose  $0 \leq k \leq n$  and  $0 \leq k' \leq n'$ , such that*

$$n - 2k = n' - 2k' = \lambda.$$

*Then there is an isomorphism of  $(A^{k, n-k}, A^{k', n'-k'})$ -bimodules  $\mathcal{F}_w^\lambda(\omega) \cong C_{CK}(\omega_b, \lambda)$ , which — up to stabilization — forgets red facets of cup foams when  $\mathcal{F}_w^\lambda(\omega)$  is constructed using red-over-blue web bases.*

### 7.5 A functorial homology for all tangles

The above sections describe a family of 2-functors  $\mathcal{F}_w^\lambda: \mathbf{Foam} \rightarrow \mathbf{Bimod}$  parametrized with  $\lambda \in \mathbb{Z}$ , which — as before — we extend to the bicategory of formal complexes  $Com_{/h}(\mathbf{Foam})$ . Applying  $\mathcal{F}_w^\lambda$  to the bracket  $[[T]]_F$  of a tangle  $T$  results in a chain complexes of bimodules, the homotopy type of which is an invariant of  $T$  and which is functorial with respect to tangle cobordisms.

**Theorem D** *The 2-functor  $\mathcal{F}_w^\lambda$  is equivalent to  $\mathcal{F}_{Kh}^\lambda \circ \mathcal{E}^\vee$ . In particular, the complexes  $C_{\mathfrak{M}}(T; \lambda)$  and  $CKh(T; \lambda)$  are isomorphic for any tangle  $T$ .*

**Proof** The equivalence of  $\mathcal{F}_w^\lambda$  and  $\mathcal{F}_{Kh}^\lambda \circ \mathcal{E}^\vee$  follows from Theorems 7.5 and 7.6 along the same lines as in the proof of Theorem C. The second statement is a direct consequence of Theorem 4.6. □

## References

- [1] **MM Asaeda, JH Przytycki, A S Sikora**, *Categorification of the Kauffman bracket skein module of I-bundles over surfaces*, *Algebr. Geom. Topol.* 4 (2004) 1177–1210 MR Zbl
- [2] **D Bar-Natan**, *Khovanov’s homology for tangles and cobordisms*, *Geom. Topol.* 9 (2005) 1443–1499 MR Zbl
- [3] **A Beliakova, M Hogancamp, K K Putyra, SM Wehrli**, *On unification of colored annular  $\mathfrak{sl}_2$  knot homology*, preprint (2023) arXiv 2305.02977
- [4] **A Beliakova, K K Putyra, SM Wehrli**, *Quantum link homology via trace functor, I*, *Invent. Math.* 215 (2019) 383–492 MR Zbl
- [5] **C Blanchet**, *An oriented model for Khovanov homology*, *J. Knot Theory Ramifications* 19 (2010) 291–312 MR Zbl

- [6] **J Brundan, C Stroppel**, *Highest weight categories arising from Khovanov's diagram algebra, II: Koszulity*, Transform. Groups 15 (2010) 1–45 MR Zbl
- [7] **J Brundan, C Stroppel**, *Highest weight categories arising from Khovanov's diagram algebra I: Cellularity*, Mosc. Math. J. 11 (2011) 685–722, 821–822 MR Zbl
- [8] **C Caprau**, *An  $\mathfrak{sl}(2)$  tangle homology and seamed cobordisms*, preprint (2007) arXiv 0707.3051
- [9] **S Cautis, J Kamnitzer, S Morrison**, *Webs and quantum skew Howe duality*, Math. Ann. 360 (2014) 351–390 MR Zbl
- [10] **Y Chen, M Khovanov**, *An invariant of tangle cobordisms via subquotients of arc rings*, Fund. Math. 225 (2014) 23–44 MR Zbl
- [11] **D Clark, S Morrison, K Walker**, *Fixing the functoriality of Khovanov homology*, Geom. Topol. 13 (2009) 1499–1582 MR Zbl
- [12] **B Cooper, V Krushkal**, *Categorification of the Jones–Wenzl projectors*, Quantum Topol. 3 (2012) 139–180 MR Zbl
- [13] **M Ehrig, C Stroppel, D Tubbenhauer**, *Generic  $\mathfrak{gl}_2$ -foams, web and arc algebras*, preprint (2016) arXiv 1601.08010
- [14] **M Ehrig, C Stroppel, D Tubbenhauer**, *The Blanchet–Khovanov algebras*, from “Categorification and higher representation theory” (A Beliakova, A D Lauda, editors), Contemp. Math. 683, Amer. Math. Soc., Providence, RI (2017) 183–226 MR Zbl
- [15] **M Ehrig, D Tubbenhauer, P Wedrich**, *Functoriality of colored link homologies*, Proc. Lond. Math. Soc. 117 (2018) 996–1040 MR Zbl
- [16] **M W Hirsch**, *Differential topology*, Graduate Texts in Math. 33, Springer (1976) MR Zbl
- [17] **M Jacobsson**, *An invariant of link cobordisms from Khovanov homology*, Algebr. Geom. Topol. 4 (2004) 1211–1251 MR Zbl
- [18] **M Khovanov**, *A categorification of the Jones polynomial*, Duke Math. J. 101 (2000) 359–426 MR Zbl
- [19] **M Khovanov**, *A functor-valued invariant of tangles*, Algebr. Geom. Topol. 2 (2002) 665–741 MR Zbl
- [20] **M Khovanov**,  *$\mathfrak{sl}(3)$  link homology*, Algebr. Geom. Topol. 4 (2004) 1045–1081 MR Zbl
- [21] **M Khovanov**, *Categorifications of the colored Jones polynomial*, J. Knot Theory Ramifications 14 (2005) 111–130 MR Zbl
- [22] **M Khovanov**, *An invariant of tangle cobordisms*, Trans. Amer. Math. Soc. 358 (2006) 315–327 MR Zbl
- [23] **A D Lauda, H Queffelec, D E V Rose**, *Khovanov homology is a skew Howe 2-representation of categorified quantum  $\mathfrak{sl}_m$* , Algebr. Geom. Topol. 15 (2015) 2517–2608 MR Zbl



- [24] **T Leinster**, *Basic bicategories*, preprint (1998) arXiv 9810017
- [25] **M Mackaay, M Stošić, P Vaz**,  $\mathfrak{sl}(N)$ -link homology ( $N \geq 4$ ) using foams and the Kapustin–Li formula, *Geom. Topol.* 13 (2009) 1075–1128 MR Zbl
- [26] **P S Ozsváth, J Rasmussen, Z Szabó**, *Odd Khovanov homology*, *Algebr. Geom. Topol.* 13 (2013) 1465–1488 MR Zbl
- [27] **K K Putyra**, *A 2-category of chronological cobordisms and odd Khovanov homology*, from “Knots in Poland, III: Part III” (J H Przytycki, P Traczyk, editors), Banach Center Publ. 103, Polish Acad. Sci. Inst. Math., Warsaw (2014) 291–355 MR Zbl
- [28] **K K Putyra**, *A quantum colored  $\mathfrak{sl}(2)$  knot homology: three approaches, same invariant*, conference recording, Erwin Schrödinger International Institute for Mathematics and Physics (2019) Available at <http://www.categorification.net/esi19>
- [29] **H Queffelec**, *Skein modules from skew Howe duality and affine extensions*, *Symmetry Integrability Geom. Methods Appl.* 11 (2015) art.id.030 MR Zbl
- [30] **H Queffelec, D E V Rose**, *The  $\mathfrak{sl}_n$  foam 2-category: A combinatorial formulation of Khovanov–Rozansky homology via categorical skew Howe duality*, *Adv. Math.* 302 (2016) 1251–1339 MR Zbl
- [31] **H Queffelec, P Wedrich**, *Khovanov homology and categorification of skein modules*, preprint (2018) arXiv 1806.03416
- [32] **J Rasmussen**, *Khovanov homology and the slice genus*, *Invent. Math.* 182 (2010) 419–447 MR Zbl
- [33] **L Rozansky**, *A categorification of the stable  $SU(2)$  Witten–Reshetikhin–Turaev invariant of links in  $S^2 \times S^1$* , preprint (2010) arXiv 1011.1958
- [34] **C Stroppel**, *Parabolic category  $\mathbb{C}$ , perverse sheaves on Grassmannians, Springer fibres and Khovanov homology*, *Compos. Math.* 145 (2009) 954–992 MR Zbl
- [35] **P Vogel**, *Functoriality of Khovanov homology*, preprint (2015) arXiv 1505.04545

*Institut für Mathematik, Universität Zürich  
Zürich, Switzerland*

*Department of Mathematics, Northeastern University  
Boston, MA, United States*

*Institute for Theoretical Studies, ETH Zürich  
Zürich, Switzerland*

*Mathematics Department, Syracuse University  
Syracuse, NY, United States*

anna@math.uzh.ch, m.hogancamp@northeastern.edu,  
krzyszttof.putyra@math.uzh.ch, smwehrli@syr.edu

Received: 15 February 2021      Revised: 4 August 2021



# Asymptotic translation lengths and normal generation for pseudo-Anosov monodromies of fibered 3–manifolds

HYUNGRYUL BAIK

EIKO KIN

HYUNSHIK SHIN

CHENXI WU

Let  $M$  be a hyperbolic fibered 3–manifold. We study properties of sequences  $(S_{\alpha_n}, \psi_{\alpha_n})$  of fibers and monodromies for primitive integral classes in the fibered cone of  $M$ . The main object is the asymptotic translation length  $\ell_C(\psi_{\alpha_n})$  of the pseudo-Anosov monodromy  $\psi_{\alpha_n}$  on the curve complex. We first show that there exists a constant  $C > 0$  depending only on the fibered cone such that for any primitive integral class  $(S, \psi)$  in the fibered cone,  $\ell_C(\psi)$  is bounded from above by  $C/|\chi(S)|$ . We also obtain a moral connection between  $\ell_C(\psi)$  and the normal generating property of  $\psi$  in the mapping class group on  $S$ . We show that for all but finitely many primitive integral classes  $(S, \psi)$  in an arbitrary 2–dimensional slice of the fibered cone,  $\psi$  normally generates the mapping class group on  $S$ . In the second half of the paper, we study if it is possible to obtain a continuous extension of normalized asymptotic translation lengths on the curve complex as a function on the fibered face. An analogous question for normalized entropy has been answered affirmatively by Fried and the question for normalized asymptotic translation length on the arc complex in the fully punctured case has been answered negatively by Strenner. We show that such an extension in the case of the curve complex does not exist in general by explicit computation for sequences in the fibered cone of the magic manifold.

30F60, 37E30; 32G15, 37B40

## 1 Introduction

Let  $M$  be a hyperbolic fibered 3–manifold. Thurston introduced the so-called Thurston norm on the first cohomology group of  $M$ , and showed that the unit norm ball is a finite sided polyhedron. Let  $F$  be a top-dimensional face of this polyhedron and consider a primitive integral class contained in the open cone  $\mathcal{C} = \mathcal{C}_F$  over  $F$ . Thurston showed

that if this cohomology class corresponds to a fibration of  $M$  over the circle  $S^1$ , then all primitive integral classes in  $\mathcal{C}$  correspond to fibrations of  $M$  over  $S^1$ . In such a case, we call  $F$  a *fibred face* and the open cone  $\mathcal{C}$  a *fibred cone*.

For each primitive integral class  $\alpha \in \mathcal{C}$ , let  $(S_\alpha, \psi_\alpha)$  be the pair of corresponding fiber and its monodromy. Since  $M$  is hyperbolic, the monodromy  $\psi_\alpha$  is pseudo-Anosov by Thurston's hyperbolization theorem; see, for example Farb and Margalit [8, Theorem 13.4]. We will study the asymptotic translation length of  $\psi_\alpha$  on the curve complex of the surface  $S_\alpha$  and the normal generators of mapping class groups  $\text{Mod}(S_\alpha)$ .

Let  $G$  be a group acting isometrically on a metric space  $(X, d_X)$ . For  $h \in G$ , the *asymptotic translation length* (or *stable length*) of  $h$  is defined by

$$\ell_X(h) = \liminf_{n \rightarrow \infty} \frac{d_X(x, h^n x)}{n},$$

where  $x$  is a point in  $X$ . It is not hard to see that  $\ell_X(h)$  is independent of the choice of  $x$ .

For a surface  $S$ , let  $\mathcal{T}(S)$  be the Teichmüller space of  $S$  and let  $\mathcal{C}(S)$  be the curve complex of  $S$ . Since  $\psi_\alpha$  acts by an isometry on both  $\mathcal{T}(S_\alpha)$  and  $\mathcal{C}(S_\alpha)$ , one can consider the asymptotic translation lengths of  $\psi_\alpha$  on  $\mathcal{T}(S_\alpha)$  and on  $\mathcal{C}(S_\alpha)$ , denoted by  $\ell_{\mathcal{T}}(\psi_\alpha)$  and  $\ell_{\mathcal{C}}(\psi_\alpha)$  respectively.

There has been a lot of work on  $\ell_{\mathcal{T}}(\psi_\alpha)$  for primitive integral classes  $\alpha$  in the fibred cone; see Fathi, Laudenbach and Poénaru [9], Fried [10; 11], Long and Oertel [22], Matsumoto [26], and McMullen [27]. In the case of  $\ell_{\mathcal{C}}(\psi_\alpha)$ , there has also been some progress in the literature; see Aougab and Taylor [1], Bowditch [5], Farb, Leininger and Margalit [7], Gadre, Hironaka, Kent and Leininger [12], Gadre and Tsai [13], Masur and Minsky [24], Valdivia [34; 35], and the authors [2; 3; 16].

The following is a general upper bound of  $\ell_{\mathcal{C}}(\psi_\alpha)$  in the fibred cone in terms of the Euler characteristic  $\chi(S_\alpha)$  of  $S_\alpha$ .

**Theorem 1.1** [3] *Let  $F$  be a fibred face of a closed hyperbolic fibred 3-manifold  $M$  and  $K$  be a compact subset of  $\text{int}(F)$ , the interior of  $F$ . Then there exists a constant  $C$  depending on  $K$  such that for any sequence  $(S_{\alpha_n}, \psi_{\alpha_n})$  of primitive integral classes which is contained in the intersection between the cone over  $K$  and a  $(d+1)$ -dimensional rational subspace of  $H^1(M)$ ,*

$$\ell_{\mathcal{C}}(\psi_{\alpha_n}) \leq \frac{C}{|\chi(S_{\alpha_n})|^{1+1/d}}.$$

Here a  $(d + 1)$ -dimensional rational subspace of  $H^1(M)$  means a subspace of  $H^1(M)$  which admits a basis  $v_1, \dots, v_{d+1} \in H^1(M; \mathbb{Q})$ . We note that in [3] the above theorem was stated in the case of closed hyperbolic fibered 3-manifolds, but almost the same proof can be adapted to the case of compact hyperbolic fibered 3-manifolds possibly with boundary; see Remark 2.5.

Two additional questions naturally arise from Theorem 1.1. First, what can we say if the sequence is not contained in the cone over any compact subset of the fibered face  $F$ ? For instance, given a sequence that has a subsequence converging projectively to the boundary  $\partial F$ , can we determine the upper bound of the asymptotic translation length of the pseudo-Anosov monodromies? We answer the first question in the following theorem.

**Theorem 3.1** *Let  $F$  be a fibered face of a compact hyperbolic fibered 3-manifold, possibly with boundary. Then there exists a constant  $C$  depending on  $F$  such that for any primitive integral class  $(S, \psi) \in \mathcal{C}_F$ ,*

$$\ell_C(\psi) \leq \frac{C}{|\chi(S)|}.$$

We make a couple of remarks regarding Theorem 3.1. We first note that a version of Theorem 3.1 was obtained by Schleimer in [30]. Even though he used different language, [30, Theorem 4.4] can be reinterpreted to give a statement of the form of Theorem 3.1 when the manifold is closed. We give an alternative argument which works for the nonclosed case as well.

Secondly we remark that the upper bound in Theorem 3.1 is optimal. In Lemma 4.12, we give an explicit sequence  $(S_{\alpha_n}, \psi_{\alpha_n})$  converging projectively to a point in  $\partial F$  such that the asymptotic translation length of the corresponding pseudo-Anosov monodromy is comparable to  $1/|\chi(S_{\alpha_n})|$ . That is, there exists a constant  $C$  such that

$$\frac{1}{C} \frac{1}{|\chi(S_{\alpha_n})|} \leq \ell_C(\psi_{\alpha_n}) \leq \frac{C}{|\chi(S_{\alpha_n})|}.$$

In general, for real-valued functions  $A(x)$  and  $B(x)$ , we say that  $A(x)$  is *comparable* to  $B(x)$  if there exists a constant  $C$  independent of  $x$  such that  $1/C \leq A(x)/B(x) \leq C$ , and we denote it by  $A(x) \asymp B(x)$ .

The second question is whether the upper bound in Theorem 1.1 is sharp. It is noted in [3] that the bound is optimal for  $d = 1$ . We show that it is also optimal when  $d = 2$  by constructing an example coming from the magic manifold  $N$ , which is the exterior of some 3-component link in the 3-sphere  $S^3$ .

**Theorem 4.13** *Let  $F$  be a fibered face of the magic manifold. Then there exist two points,  $b_0 \in \partial F$  and  $c_0 \in \text{int}(F)$ , which satisfy the following:*

- (1) *For any  $r \in \mathbb{Q} \cap [1, 2)$ , there exists a sequence  $(S_{\alpha_n}, \psi_{\alpha_n})$  of primitive integral classes in  $\mathcal{C}_F$  converging projectively to  $b_0$  as  $n \rightarrow \infty$  such that*

$$\ell_C(\psi_{\alpha_n}) \asymp \frac{1}{|\chi(S_{\alpha_n})|^r}.$$

- (2) *For any  $r \in \mathbb{Q} \cap [\frac{3}{2}, 2]$ , there exists a sequence  $(S_{\alpha_n}, \psi_{\alpha_n})$  of primitive integral classes in  $\mathcal{C}_F$  converging projectively to  $c_0$  as  $n \rightarrow \infty$  such that*

$$\ell_C(\psi_{\alpha_n}) \asymp \frac{1}{|\chi(S_{\alpha_n})|^r}.$$

*In particular, the upper bound in Theorem 1.1 is optimal when  $d = 2$ .*

As an immediate corollary of Theorem 4.13, we conclude that there is no normalization of the asymptotic translation length function defined on the rational classes of the fibered face, which continuously extends to the whole fibered face. More precisely, we have the following.

**Corollary 4.15** *Let  $F$  be a fibered face of the magic manifold  $N$ . For  $\alpha \in F \cap H^1(N; \mathbb{Q})$ , let  $(S_{\tilde{\alpha}}, \psi_{\tilde{\alpha}})$  be the fiber and pseudo-Anosov monodromy corresponding to the primitive integral class  $\tilde{\alpha}$  lying on the ray of  $\alpha$  passing through the origin. Then there is no normalization of the asymptotic translation length function*

$$F \cap H^1(N; \mathbb{Q}) \rightarrow \mathbb{R}_{\geq 0}, \quad \alpha \mapsto \ell_C(\psi_{\tilde{\alpha}}),$$

*in terms of the Euler characteristic  $\chi(S_{\tilde{\alpha}})$  which admits a continuous extension on  $F$ .*

For the arc complex, Strenner defined in [31] the normalized asymptotic translation length function  $\mu_d$  for each integer  $d \geq 1$  on the rational classes of a fibered face with the fully punctured condition. Strenner proved in the same paper that the functions  $\mu_d$  for  $d \geq 2$  are typically nowhere continuous. His result and Corollary 4.15 stand in contrast to Fried's result [10]. See also Matsumoto [26] and McMullen [27]. They proved that the normalized entropy function of pseudo-Anosov monodromies has a continuous extension on the fibered face, which is strictly convex.

Now we turn our attention to normal generation of mapping class groups. Let  $S = S_{g,n}$  be an orientable surface of genus  $g$  with  $n$  punctures, possibly  $n = 0$ . We denote  $S_{g,0}$  by  $S_g$ . We say that an element  $h$  of a group  $G$  normally generates  $G$  if the normal

closure of  $h$  is equal to  $G$ . For a given primitive class  $(S_\alpha, \psi_\alpha)$  in the fibered cone  $\mathcal{C}$ , when does  $\psi_\alpha$  normally generate  $\text{Mod}(S_\alpha)$ ? Normal generation in the mapping class group has been studied by many authors. For instance, D Long [21] asked if there exist pseudo-Anosov normal generators. Later Ivanov asked in [14] what properties are satisfied by the pseudo-Anosov normal generators. A work of Lanier and Margalit [20] (partially) answered the questions of Long and Ivanov. In particular, they showed that for a pseudo-Anosov element  $f \in \text{Mod}(S_g)$ , if the stretch factor  $\lambda(f)$  is smaller than  $\sqrt{2}$ , then  $f$  normally generates  $\text{Mod}(S_g)$ . The normal closure of random elements was studied as well, for instance by Maher and Tiozzo [23]. They showed that with asymptotic probability 1, the normal closure of a random element is free. This in particular implies that random elements are not normal generators.

This connects to our brief discussion of asymptotic translation length, since the logarithm of the stretch factor,  $\log \lambda(f)$ , is equal to  $\ell_{\mathcal{T}}(f)$ . In other words, if a pseudo-Anosov element of  $\text{Mod}(S)$  is contained in some proper normal subgroup, then its asymptotic translation length on the Teichmüller space cannot be too small. It is natural to ask an analogous statement for the curve complexes, ie if a pseudo-Anosov element of  $\text{Mod}(S)$  is contained in some proper normal subgroup, then its asymptotic translation length on the curve complex cannot be too small in some sense. The following question was raised by Dan Margalit (via personal communication).

**Question 1.2** For a subgroup  $H$  of  $\text{Mod}(S_g)$ , set

$$L_{\mathcal{C}}(H) = \min\{\ell_{\mathcal{C}}(f) \mid f \text{ is pseudo-Anosov and } f \in H\}.$$

Is there a constant  $C > 0$  such that

$$L_{\mathcal{C}}(H) \geq \frac{C}{g}$$

for any  $g \geq 2$  and for any proper normal subgroup  $H$  of  $\text{Mod}(S_g)$ ?

As a partial evidence toward this question, it is shown by Baik and Shin [2] that

$$L_{\mathcal{C}}(\mathcal{I}_g) \asymp \frac{1}{g},$$

where  $\mathcal{I}_g$  is the Torelli group, ie the proper normal subgroup of  $\text{Mod}(S_g)$  whose action on the first homology is trivial. In fact, by [2, Theorem 3.2],

$$L_{\mathcal{C}}(\mathcal{I}_g) \geq \frac{1}{96(g-1)}$$

for all  $g \geq 2$ .

Combining with Theorem 3.1, we propose the following conjecture regarding the normal generators of mapping class groups contained in the fibered cone which was originally asked as a question by Dan Margalit (via personal communication).

**Conjecture 1.3** Let  $F$  be a fibered face of a closed hyperbolic fibered 3–manifold  $M$ . Then for all but finitely many primitive classes  $(S_\alpha, \psi_\alpha) \in \mathcal{C}_F$ ,  $\psi_\alpha$  normally generates  $\text{Mod}(S_\alpha)$ .

We give a partial answer when primitive integral classes are contained in a 2–dimensional rational subspace of  $H^1(M)$ . See also Remark 3.7.

**Theorem 3.4** Let  $F$  be a fibered face of a closed hyperbolic fibered 3–manifold  $M$ , and let  $L$  be a 2–dimensional rational subspace of  $H^1(M)$ . Then for all but finitely many primitive integral classes  $(S, \psi)$  in  $\mathcal{C}_F \cap L$ ,  $\psi$  normally generates  $\text{Mod}(S)$ . In particular, if the rank of  $H^1(M)$  equals 2, then Conjecture 1.3 is true.

## Acknowledgements

We thank Susumu Hirose, Michael Landry and Dan Margalit for helpful comments. We appreciate the referees for their valuable comments. Baik was partially supported by Samsung Science & Technology Foundation grant SSTF-BA1702-01. Kin was supported by Grant-in-Aid for Scientific Research (C) (18K03299, 21K03247), JSPS. Shin was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2017R1D1A1B03035017).

## 2 Arithmetic sequences in the fibered cone

For a hyperbolic 3–manifold  $M$ , possibly with boundary  $\partial M$ , Thurston [32] defined a norm  $\|\cdot\|$  on  $H_2(M, \partial M; \mathbb{R})$ . It turns out the unit norm ball  $B_M$  with respect to the Thurston norm is a finite-sided polyhedron. Let  $F$  be a top-dimensional face of  $B_M$ . We consider an open cone  $\mathcal{C} = \mathcal{C}_F$  over  $F$ . Thurston showed that if  $M$  is a fibered 3–manifold, then either all integral points in  $\mathcal{C}$  are fibered or none of them are fibered. (When an integral cohomology class corresponds to a fibration of  $M$  over  $S^1$ , we say the integral point is fibered.) In the former case, we call  $\mathcal{C}$  a *fibered cone*. We denote by  $\overline{\mathcal{C}}$  the closure of the fibered cone  $\mathcal{C}$ .



By abuse of notation, the first cohomology classes are treated as their dual second homology classes throughout this paper without explicitly mentioning it. Furthermore, we will write a primitive integral class  $\alpha \in H^1(M)$  as a pair  $(S, \psi)$  when  $S$  and  $\psi$  are the fiber and the monodromy for the fibration over  $S^1$  corresponding to  $\alpha$ .

In this section, we will show a key property of infinite arithmetic sequences in a fibered cone for the proof of Theorem 3.4. Here by an arithmetic sequence we mean a sequence  $(\alpha + n\beta)_{n \in \mathbb{Z}_{\geq 0}}$  where  $\alpha$  (resp.  $\beta$ ) is a primitive integral class in a fibered cone  $\mathcal{C}$  (resp. the closure  $\overline{\mathcal{C}}$  of the fibered cone  $\mathcal{C}$ ). We first need to find some criterion for a given element of the mapping class group to be a normal generator. In [20], the so-called *well-suited curve criterion* is introduced. Roughly speaking, this criterion says that if there is a simple closed curve  $c$  such that the configuration of  $c \cup f(c)$  is simple enough, then  $f$  is a normal generator for the mapping class group.

Here we state one special case that we need and show its proof for the sake of completeness. For more general statements, see [20, Sections 2, 7 and 9]. For a closed curve  $c$  in the surface  $S_g$  without specified orientation,  $[c]$  means the homology class in  $H_1(S_g)$  with arbitrary orientation.

**Lemma 2.1** [20, Lemma 2.3] *Let  $f \in \text{Mod}(S_g)$  for  $g \geq 3$ . Suppose that there is a nonseparating curve  $c$  in  $S_g$  such that  $c$  and  $f(c)$  are disjoint and*

$$\pm[c] \neq [f(c)] \in H_1(S_g).$$

*Then the normal closure of  $f$  is  $\text{Mod}(S_g)$ .*

**Proof** Let  $f$  and  $c$  be as in the statement of the lemma. Then one can find nonseparating curves  $a, b, d, x$  and  $y$  which satisfy the following conditions.

- $a, b, c$  and  $d$  bound a subsurface  $S$  of  $S_g$  which is homeomorphic to a 4-punctured sphere.
- Each of the triple of curves  $(a, b, x)$ ,  $(b, d, y)$  and  $(b, c, f(c))$  bounds a pair of pants contained in  $S$ .
- No two of the curves  $a, b, c, d, x, y$  and  $f(c)$  are homologous.

To see the existence of such curves, start with Figure 1, left, which is the surface of genus 0 with four boundary components (ie, a 4-punctured sphere) labeled  $A, B, C$  and  $D$ . Glue a pair of pants along the boundary components labeled  $A$  and  $B$ , and glue another pair of pants along the boundary components labeled  $C$  and  $D$ . Then we get a

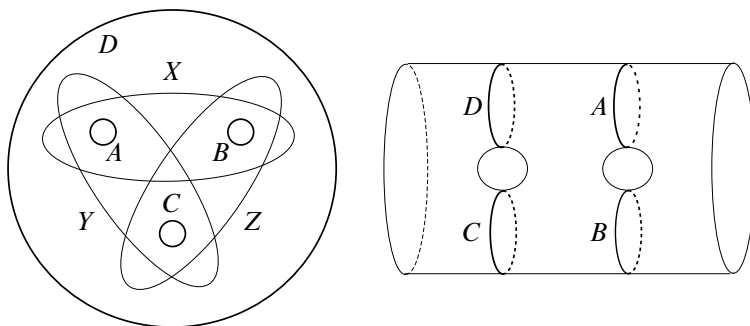


Figure 1: Left: a 4-punctured sphere. Right: a genus 2 surface with two boundary components.

surface of genus 2 with two boundary components (Figure 1, right). Along the two boundary components, we glue in another surface of genus  $k \geq 0$  with two boundary components. The resulting surface is a closed surface of genus  $3 + k$ . We take  $k$  so that  $3 + k = g$  which is the genus of our given surface  $S_g$ . This is our model surface, and we let  $\Sigma$  denote the model surface. If we set  $a = A, b = B, c = C, d = D, x = X, y = Y$  and  $f(c) = Z$ , then the above conditions are satisfied by construction.

By the classification of the compact orientable surfaces, for any two pairs of disjoint nonhomologous simple closed curves on the surface, there exists a homeomorphism which maps one pair to the other. (This is a special case of the so-called change of coordinates principle. See for instance [8].) Hence, there exists a homeomorphism  $\Phi$  from  $\Sigma$  to  $S_g$  such that  $\Phi(C) = c$  and  $\Phi(Z) = f(c)$ . Now set  $a = \Phi(A), b = \Phi(B), d = \Phi(D), x = \Phi(X)$  and  $y = \Phi(Y)$ . Then we get the desired set of curves  $a, b, d, x$  and  $y$  which satisfy all the conditions together with  $c$  and  $f(c)$ .

For any curve  $\gamma$  on  $S_g$ , let  $T_\gamma$  be the left-handed Dehn twist about  $\gamma$ . Then by the lantern relation, we have  $T_a T_b T_c T_d = T_{f(c)} T_x T_y$ . Using the commutativity of the Dehn twists about disjoint curves, one can rewrite the lantern relation as

$$T_d = T_c^{-1} T_{f(c)} T_a^{-1} T_x T_b^{-1} T_y.$$

Note that  $T_c^{-1} T_{f(c)} = T_c^{-1} (f T_c f^{-1}) = (T_c^{-1} f T_c) f^{-1}$  which is contained in the normal closure of  $f$ .

As before, by the change of coordinates principle, there exists an orientation-preserving homeomorphism  $h$  of  $S_g$  such that  $h(c) = a$  and  $h(f(c)) = x$ . Then

$$T_a^{-1} T_x = T_{h(c)}^{-1} T_{h(f(c))} = h^{-1} T_c^{-1} T_{f(c)} h,$$

ie it is just a conjugate of  $T_c^{-1}T_{f(c)}$ . Hence  $T_a^{-1}T_x$  is in the normal closure of  $f$ . Similarly,  $T_b^{-1}T_y$  is also contained in the normal closure of  $f$ .

This shows that  $T_d$  lies in the normal closure of  $f$ . From the fact that there exists only one mapping class group orbit of nonseparating simple closed curves and the Dehn twists about nonseparating simple closed curves generate the mapping class group, we can now conclude that the entire mapping class group  $\text{Mod}(S_g)$  is contained in the normal closure of  $f$ . □

Now we prove the key proposition on the sequences in the fibered cone.

**Proposition 2.2** *Let  $\mathcal{C}$  be a fibered cone for a closed hyperbolic fibered 3-manifold  $M$ . Let  $\alpha \in \mathcal{C}$  and  $\beta \in \overline{\mathcal{C}}$  be integral classes. Then there is some integer  $n_0 > 0$  depending on  $\alpha$  and  $\beta$  which satisfies the following. If  $(S, \psi) = \alpha + n\beta \in \mathcal{C}$  is a primitive integral class for  $n \geq n_0$ , then there is an essential simple closed curve  $c$  on  $S$  such that  $c, \psi(c), \dots, \psi^{n-1}(c)$  are disjoint, and  $\pm[c] \neq [\psi(c)]$  in  $H_1(S)$ .*

**Proof** Let  $n$  be a positive integer such that  $\alpha + n\beta$  is a primitive integral class. Let  $S_\alpha$  and  $S_\beta$  be embedded surfaces in  $M$  which represent  $\alpha$  and  $\beta$  respectively. Note that their orientations are assigned, and each connected component of those surfaces has genus at least 2, since  $M$  is a closed hyperbolic 3-manifold. In what follows, we explain how to choose these representatives more explicitly.

For any primitive integral class in  $\mathcal{C}$ , one obtains a suspension flow  $\mathcal{F}$  of the monodromy. Fried showed that when  $M$  is a closed hyperbolic fibered 3-manifold, the flow  $\mathcal{F}$  is an invariant of  $\mathcal{C}$  in the following sense: if one considers the suspension flows from two primitive integral classes in  $\mathcal{C}$ , then they are the same flow up to reparametrization and conjugation by homeomorphisms on  $M$ . Moreover Fried showed that if an embedded surface  $S$  in  $M$  is a fiber for a primitive integral class in  $\mathcal{C}$ , then  $S$  can be made transverse to  $\mathcal{F}$ , and the first return map along the flow  $\mathcal{F}$  represents the monodromy; see [11] and [9, Theorem 14.11 and Lemma 14.12].

Surely  $S_\alpha$  can be made transverse to  $\mathcal{F}$ , since  $\alpha \in \mathcal{C}$ . If  $\beta \in \mathcal{C}$ , then the same holds for  $S_\beta$ . However if  $\beta \in \partial\mathcal{C} = \overline{\mathcal{C}} \setminus \mathcal{C}$ , then this may or may not be possible for representatives of  $\beta$ . The transverse surface theorems by Mosher [28] and Landry [19] including the case of compact hyperbolic 3-manifolds tells us that, for any integral class  $\beta \in \overline{\mathcal{C}}$ , there exists a flow  $\widehat{\mathcal{F}}$ , which is semiconjugate to  $\mathcal{F}$ , such that a representative  $S_\beta$  of  $\beta$  is transverse to  $\widehat{\mathcal{F}}$ . Here  $\widehat{\mathcal{F}}$  is obtained from  $\mathcal{F}$  by using the dynamic blowup of some

(possibly empty) singular periodic orbits of  $\mathcal{F}$ . The flow  $\widehat{\mathcal{F}}$  is called a dynamic blowup of  $\mathcal{F}$  for  $\beta \in \overline{\mathcal{C}}$ . (The dynamic blowups of  $\mathcal{F}$  may not be unique.) For more details of the dynamic blowup of singular orbits, see [28, pages 8–9] and [19, Section 3.1].

We now explain some relevant properties of  $\widehat{\mathcal{F}}$  which are needed in the proof of Proposition 2.2. The new flow  $\widehat{\mathcal{F}}$  is obtained from  $\mathcal{F}$  by replacing the singular orbits of  $\mathcal{F}$  by a set of annuli such that flow lines in the interior of each annulus spiral toward boundary components of the annulus. Moreover  $S_\alpha \cap \mathcal{A}$  is a union of embedded trees in  $S_\alpha$ , where  $\mathcal{A}$  is the collection of annuli created during the finitely many blowups of singular orbits. When  $\beta \in \mathcal{C}$ , it is shown in the transverse surface theorem that  $\widehat{\mathcal{F}}$  is obtained by dynamically blowing up  $\mathcal{F}$  along an empty collection of periodic orbits, and hence  $\widehat{\mathcal{F}}$  is the same as  $\mathcal{F}$ . Now  $S_\beta$  is transverse to  $\widehat{\mathcal{F}}$ . From the construction of  $\widehat{\mathcal{F}}$ , we may suppose that  $S_\alpha$  is still transverse to  $\widehat{\mathcal{F}}$ .

For any positive integer  $n$ , we can consider  $n$  parallel copies of  $S_\beta$ , say  $S_1, \dots, S_n$  such that the  $S_i$  are very close to each other. Whenever we are in this situation, the  $n$  copies of  $S_i$  are labeled so that for  $1 \leq i < n$ ,  $S_i$  gets mapped to  $S_{i+1}$  by the flow  $\widehat{\mathcal{F}}$  before touching any other  $S_j$ . Note that  $n$  is not fixed.

We now describe the surgery, ie cut and paste, on  $S_\alpha, S_1, \dots, S_n$  along the intersection locus to get a surface  $S$  which represents  $\alpha + n\beta$ . Along each component of the intersection between  $S_\alpha$  and each copy of  $S_\beta$ , we cut those surfaces. Locally there are four sheets of surfaces, two from  $S_\alpha$  and two from the copy of  $S_\beta$ . Glue one sheet from  $S_\alpha$  to one sheet from  $S_\beta$  so that the orientations on those sheets match up. One can do the same for the other two remaining sheets. The resulting surface  $S$  represents  $\alpha + n\beta$ . Clearly  $S$  is transverse to  $\widehat{\mathcal{F}}$ . We note that this is a standard operation. For instance, it is the same as the oriented sum in [6].

The transversality of  $S$  to  $\widehat{\mathcal{F}}$  implies two things. First of all, this means  $S$  is transverse to  $\mathcal{A}$ . Since the original flow  $\mathcal{F}$  is obtained from  $\widehat{\mathcal{F}}$  by collapsing the annuli in  $\mathcal{A}$  to singular orbits of  $\mathcal{F}$ ,  $S$  is transverse to  $\mathcal{F}$  after the collapsing. Second, the intersection  $S \cap \mathcal{A}$  is a collection of trees on  $S$  by transversality together with the construction of  $\mathcal{A}$  in the dynamic blowup. Now let  $\widehat{\Psi}$  and  $\Psi$  be the first return maps on  $S$  for  $\widehat{\mathcal{F}}$  and  $\mathcal{F}$ , respectively. Since  $\widehat{\Psi}$  and  $\Psi$  differ only on the trees and each tree is contractible,  $\widehat{\Psi}$  and  $\Psi$  are clearly homotopic to each other. Therefore  $\widehat{\Psi}$  represents the monodromy  $\psi = [\Psi]$  for  $\alpha + n\beta$ .

Note that because all  $S_i$  are parallel copies of  $S_\beta$ , any curve or region on  $S_\beta$  gives rise to a curve or region on each of the  $S_i$  that are parallel to it. Hence, in what follows,

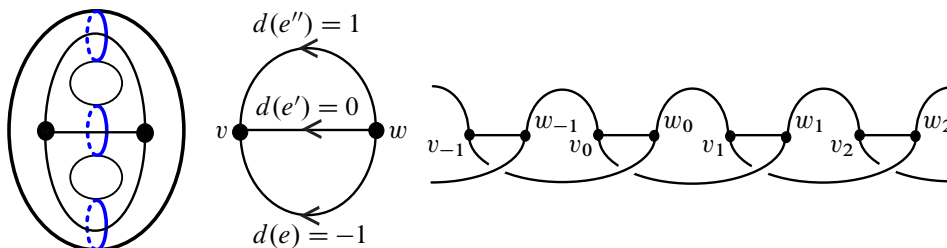


Figure 2: Left: a multicurve  $C$  together with its 3-regular graph  $G$  on  $S_\beta \simeq$  closed surface of genus 2. Middle: an example of a cochain  $d$  on  $G$ : for three edges from  $w$  to  $v$ , their values are  $-1$ ,  $0$  and  $1$  respectively. Right: a  $\mathbb{Z}$ -fold cover  $G'$  corresponding to  $d$  of the middle diagram.

whenever we specify any multicurve on  $S_\beta$  we implicitly specify multicurves on all of the  $S_i$  which are parallel to each other.

Let  $C$  be a multicurve on  $S_\beta$  such that all the connected components of  $S_\beta \setminus C$  have genus 0 with three ends (Figure 2, left). Furthermore, we assume that every intersecting curve between  $S_\alpha$  and  $S_\beta$  is parallel to one of the curves in  $C$ . Such a multicurve  $C$  always exists. To construct one, group the intersecting curves between  $S_\alpha$  and  $S_\beta$  into parallel families, choose one in each parallel family and use them to form a multicurve  $C'$ . Now, if some connected component of  $S_\beta \setminus C'$  has genus greater than 0, or has more than three ends, then we can add an extra curve to  $C'$  to break it into components of lower complexity, and repeat this process until all the connected components of  $S_\beta \setminus C'$  have genus 0 with three ends.

Now we make use of the graph theoretic lemma below.

**Lemma 2.3** *Let  $G$  be a 3-regular finite graph. Let  $d$  be an integer valued cellular cochain on  $G$  whose value on each edge is bounded above by  $k \geq 0$ , and let  $G'$  be the  $\mathbb{Z}$ -fold cover constructed from  $d$  (ie the vertices of  $G'$  are  $\mathbb{Z}$ -copies of the vertices of  $G$  and each edge  $e$  in  $G$  from  $w$  to  $v$  is lifted to edges from the  $j^{\text{th}}$  lift of  $w$  to the  $(j+d(e))^{\text{th}}$  lift of  $v$ ; see Figure 2, middle and right). Then there is some  $R$  depending only on  $k$  and the number of edges  $|E(G)|$  of  $G$  such that  $G'$  has a simple loop  $\gamma'$  of length no more than  $2R$ .*

**Proof** Suppose there are no such loops of length less than  $2R$  in  $G'$  for any  $R$ . Then the  $R$ -neighborhood (ie neighborhood with radius  $R$  assigning each edge length 1) of any vertex  $v_0$  in  $G'$  must be a tree whose vertices have valence 1 or 3. Hence it contains  $3 \times (2^R - 1)$  edges. However, such a neighborhood must contain at most

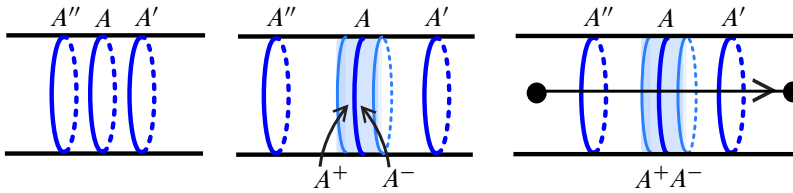


Figure 3: Left: parallel curves on  $S_\beta$  which are some components of the intersection between  $S_\alpha$  and  $S_\beta$ . Middle: an annular neighborhood of  $A$  and the side of  $A^\pm$ . Right: for an edge  $e$  starting from the side of  $A^+$ ,  $A$  contributes to  $d(e)$  by  $+1$ .

$(2Rk + 1)|E(G)|$  edges. (This is because in  $R$  steps, one can travel up at most  $Rk$  levels, ie  $Rk$  copies of the fundamental domain, or travel down at most  $Rk$  levels. Together with the original level, there are  $(2Rk + 1)$  levels in total that one might be able to pass through, and hence there are at most  $(2Rk + 1)|E(G)|$  edges in them.)

Since the exponential function grows faster than the linear one, one can set  $R$  sufficiently large to reach a contradiction. □

We continue the proof of Proposition 2.2. Note that the multicurve  $C$  above gives a pants decomposition of  $S_\beta$ . Let  $G$  be the 3-regular graph where each vertex corresponds to a pair of pants in the pants decomposition of  $S_\beta$ , and each edge corresponds to the component of the multicurve between two pairs of pants; see Figure 2, left. Now we define the cochain  $d$  on  $G$  which only depends on  $S_\alpha$  and  $S_\beta$  as follows; see Figure 3.

Consider the surface  $S$  obtained from the cut and paste construction of  $S_\alpha$  and  $n$  copies of  $S_\beta$ . If a curve  $A$  is one component of the intersection between  $S_\alpha$  and  $S_\beta$ , we cut  $S_\beta$  along  $A$  (hence we cut each copy of  $S_\beta$  along a curve corresponding to  $A$ ) which results in two boundary curves for each copy of  $S_\beta$ , say  $A^+$  and  $A^-$ . The labeling  $A^+$  and  $A^-$  are determined as follows: in the surface obtained from  $S_\alpha$  and the copies of  $S_\beta$  via the cut and paste construction, an annular piece of  $S_\alpha$  connecting the  $i^{\text{th}}$  copy of  $S_\beta$  to the  $(i + 1)^{\text{th}}$  copy of  $S_\beta$  is attached to the  $i^{\text{th}}$  copy of  $S_\beta$  along  $A^+$  (the index of each copy of  $S_\beta$  is understood as an integer modulo  $n$ ). We label the other boundary component  $A^-$ .

Now the labeling on each copy of  $S_\beta$  is well defined, and if one considers an annular neighborhood of  $A$ , then one can make sense of the statement that one side is the side of  $A^+$  and the other side is the side of  $A^-$ .

Let us consider an edge  $e$  on  $G$  which intersects the curve  $A$ . If  $e$  is with the orientation so that it passes from the side of  $A^+$  to that of  $A^-$ , then  $A$  contributes to  $d(e)$  by  $+1$ ,

and  $A$  contributes to  $d(e^{-1})$  by  $-1$ , where  $e^{-1}$  is the same edge as  $e$  with the opposite orientation. The number  $d(e)$  is obtained by summing up all the contributions of curves in  $S_\alpha \cap S_\beta$  that the edge  $e$  passes through. Note that the cochain  $d$  does not depend on  $n$  but only on  $S_\alpha$  and  $S_\beta$ , since we consider copies of  $S_\beta$  very close to each other, the intersection with  $S_\alpha$  looks exactly the same in any copy of  $S_\beta$ .

Let  $k$  be the maximum of the values of  $d$  on all edges on  $G$ , and let  $R$  be the constant from Lemma 2.3. Now let  $n$  be any integer so that  $n \geq 2Rk + 2$ , and consider the surface  $S$  obtained from  $S_\alpha$  and  $n$  copies of  $S_\beta$  by a cut and paste construction. (In other words, here we will argue that the integer  $n_0$  in Proposition 2.2 can be chosen as  $2Rk + 2$ .) Let  $\gamma'$  be a simple loop in  $G'$  in Lemma 2.3. The fact that  $|d(e)| \leq k$  implies that  $\gamma'$  passes through at most  $2Rk + 1$  consecutive fundamental domains of the deck group action on  $G'$ . The embedding of these  $2Rk + 1$  fundamental domains, together with one more, to  $2Rk + 2$  copies of  $S_\beta$  after the surgery, sends  $\gamma'$  to some simple loop  $\gamma$  on the surface  $S$ . (To do that, pick a point in each pant in  $S_i$ . Now pick a starting vertex  $v_0$  on  $\gamma'$ , and let  $\gamma$  start at the point associated to the corresponding pant in  $S_{Rk+2}$ . Now, we travel along  $\gamma'$ , and for each edge, connect the points in the two pants associated with the two end points of the edge. The construction of  $G'$  and Lemma 2.3 imply that the resulting path  $\gamma$  will also be closed.)

Let  $c \in C$  be a component of the multicurve on  $S_\beta$  and let  $c_i$  be the corresponding copies of  $c$  on the  $i^{\text{th}}$  copy  $S_i$  of  $S_\beta$ . Suppose that  $c$  is chosen such that  $c_l$  is crossed by  $\gamma$  once for some  $l$ , and that  $\gamma$  does not cross the lowest copy  $S_1$ ; see Figure 4. One can choose such a  $c$  since the length of  $\gamma'$  is no more than  $2R$ . Note that all  $c_i$  survive under surgery because they do not cross the intersections between  $S_i$  and  $S_\alpha$ . Furthermore, except for the top  $c_n$ , their images under the first return map are  $\psi(c_i) = c_{i+1}$ . By the construction of  $S$ , it follows that  $c_1, \psi(c_1) = c_2, \dots, \psi^{n-1}(c_1) = c_n$  are disjoint. For the proof of Proposition 2.2, we only need to show that  $[c_2] \pm [c_1]$  is not homologous to 0. (This also implies that  $c_1$  on  $S$  is essential.) To do so, one only needs to show that

$$(\psi_*^{l-2} + \psi_*^{l-3} + \dots + \text{id}_*)([c_2] - [c_1]) = [c_l] - [c_1]$$

and

$$(\psi_*^{l-2} - \psi_*^{l-3} + \dots + (-1)^{l-2} \text{id}_*)([c_2] + [c_1]) = [c_l] + (-1)^{l-2}[c_1]$$

are not 0. Since  $\gamma$  passes through  $c_l$  and it does not pass through  $c_1$ , the simple closed curves  $c_l$  and  $c_1$  do not bound a subsurface. Therefore  $[c_l] \neq \pm [c_1]$ . This completes the proof of Proposition 2.2. □

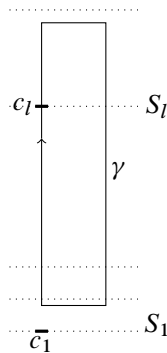


Figure 4: The horizontal line segments (with dots) represent the copies  $S_1, S_2, \dots$  of  $S_\beta$ , and the curve with arrow represents the loop  $\gamma$  which passes through  $S_l$  but not the lowest copy  $S_1$  of  $S_\beta$ .

We now consider a compact hyperbolic fibered 3–manifold  $M$ . In order to obtain an estimate for the asymptotic translation length of monodromies from the arithmetic sequences in the fibered cone for  $M$ , we show the following variant of Proposition 2.2.

**Proposition 2.4** *Let  $\mathcal{C}$  be a fibered cone for a compact hyperbolic fibered 3–manifold  $M$  possibly with boundary, let  $\alpha \in \mathcal{C}$  and  $\beta \in \overline{\mathcal{C}}$  be integral classes, and suppose  $(S, \psi) = \alpha + n\beta \in \mathcal{C}$  is a primitive integral class for an integer  $n \geq 2$ . Then there is an essential simple closed curve  $c$  on  $S$  or essential arc on  $S$  such that  $c, \psi(c), \dots, \psi^{n-1}(c)$  are disjoint. In particular,*

$$\ell_c(\psi) \leq \frac{2}{n-1}.$$

**Proof** Let  $\mathcal{F}$  be the suspension flow for the fibered cone  $\mathcal{C}$ . In [19, Appendix A], Landry generalized Fried’s theory on the fibered cone (for closed hyperbolic fibered 3–manifolds) to the case of compact hyperbolic fibered 3–manifolds  $M$  possibly with boundary. In particular  $\mathcal{F}$  is an invariant of  $\mathcal{C}$  as well. Then we use the transverse surface theorem [19; 28] for compact hyperbolic fibered 3–manifolds  $M$  again. Let  $\widehat{\mathcal{F}}$  be a dynamic blowup of  $\mathcal{F}$  for  $\beta \in \overline{\mathcal{C}}$ . We can take representatives  $S_\alpha$  and  $S_\beta$  of  $\alpha$  and  $\beta$  respectively so that  $S_\alpha$  and  $S_\beta$  are transverse and they intersect the new flow  $\widehat{\mathcal{F}}$  transversely. We may assume that  $S_\alpha$  and  $S_\beta$  intersect minimally, ie the number of components of the intersection between  $S_\alpha$  and  $S_\beta$  is minimal among all representatives of  $\alpha$  and  $\beta$ . The surface obtained from  $S_\alpha$  and  $S_\beta$  by a cut and paste construction is a fiber of the fibration associated with  $\alpha + \beta \in \mathcal{C}$ . This implies that  $S_\alpha$  and  $S_\beta$  are minimal representatives of  $\alpha$  and  $\beta$ .



Do surgery (as in the proof of Proposition 2.2) at the intersection locus of  $S_\alpha$  and  $n$  copies of  $S_\beta$  to obtain a surface  $S$  representing  $\alpha + n\beta$ . We now find the desired essential simple closed curve on  $S$  or an essential arc  $c$  on  $S$ . Let  $c$  be one of the intersection curves or arcs between  $S_\alpha$  and  $S_\beta$ , and let  $S_1$  be the lowest copy of  $S_\beta$ . The fact that  $c$  is essential on  $S_\alpha$  and on  $S_\beta$  follows from the fact that the intersection between  $S_\alpha$  and  $S_\beta$  is minimal; see [32] or [6, Lemma 5.8]. It is not hard to see from the cut and paste construction that  $c$  is also essential on  $S$ .

From the choice of  $c$ , it follows that  $c$  and  $\psi^{n-1}(c)$  are disjoint. They are distinct in the arc and curve complex  $\mathcal{AC}(S)$ , since  $\psi$  is pseudo-Anosov. Thus the distance between  $c$  and  $\psi^{n-1}(c)$  in  $\mathcal{AC}(S)$  equals 1. This implies that  $(n - 1)\ell_{\mathcal{AC}}(\psi) = \ell_{\mathcal{AC}}(\psi^{n-1}) \leq 1$  — cf [16, Lemma 2.1] — where  $\ell_{\mathcal{AC}}(\psi)$  is the asymptotic translation length of  $\psi$  on  $\mathcal{AC}(S)$ . It is known that the inclusion map  $\mathcal{C}(S) \rightarrow \mathcal{AC}(S)$  is 2-bilipschitz; see, for instance, [25, Lemma 2.2] or [18]. In particular, this tells us that

$$\ell_{\mathcal{C}}(\psi) \leq 2\ell_{\mathcal{AC}}(\psi).$$

Thus we have  $\ell_{\mathcal{C}}(\psi) \leq 2\ell_{\mathcal{AC}}(\psi) \leq 2/(n - 1)$ . □

**Remark 2.5** In [3], Theorem 1.1 was proved in the case of closed hyperbolic fibered 3-manifolds. We note that almost the same proof can be adapted to the case of compact hyperbolic fibered 3-manifold. In fact, one only needs to modify the last paragraph (after Lemma 8) in the proof of [3, Theorem 5] to allow  $\gamma$  and  $\gamma'$  to be either an essential simple closed curve or an essential simple arc. Then one obtains the same conclusion of Theorem 1.1 by the fact that inclusion map  $\mathcal{C}(S) \rightarrow \mathcal{AC}(S)$  is 2-bilipschitz as in the proof of Proposition 2.4 in this paper.

### 3 Applications of arithmetic sequences

#### 3.1 Asymptotic translation lengths in fibered cones

In this section, we show the following estimate for the asymptotic translation lengths in the curve complexes.

**Theorem 3.1** *Let  $F$  be a fibered face of a compact hyperbolic fibered 3-manifold possibly with boundary. Then there exists a constant  $C$  depending on  $F$  such that for any primitive integral class  $(S, \psi) \in \mathcal{C}_F$ ,*

$$\ell_{\mathcal{C}}(\psi) \leq \frac{C}{|\chi(S)|}.$$

To prove this theorem, we need the following lemma about rational cones. Here a rational cone in Euclidean space  $\mathbb{R}^m$  is the set of the points of the form

$$\{\mathbf{x} = (x_1, \dots, x_m) \in \mathbb{R}^m \mid A\mathbf{x}^t \geq \mathbf{0}\}$$

for some  $k \times m$  matrix  $A$  with integer entries (where  $\mathbf{x}^t$  is the transpose of  $\mathbf{x}$ .) We further assume that this set has nonempty interior.

**Lemma 3.2** *Let  $P$  be a rational cone in  $\mathbb{R}^m$ , and let  $\text{int}(P)$  be its interior. Then there exist two nonempty finite sets  $\Omega_0 \subset \text{int}(P) \cap \mathbb{Z}^m$  and  $\Omega \subset P \cap \mathbb{Z}^m$  such that*

$$\text{int}(P) \cap \mathbb{Z}^m = \left\{ a + \sum_{b \in \Omega} k_b b \mid a \in \Omega_0, k_b \in \mathbb{Z}, k_b \geq 0 \right\}.$$

**Proof** It is a classical result—see [33, Proposition 3.4]—that  $P \cap \mathbb{Z}^m$  is a finitely generated monoid. Let  $\Omega$  be a finite set of generators of  $P \cap \mathbb{Z}^m$ , and let

$$\Omega_0 = \left\{ \sum_{b \in W} b \mid W \subset \Omega, W \not\subset F \text{ for all faces } F \text{ of } \partial P \right\}.$$

Here a face of  $\partial P$  is a polytope of dimension  $m - 1$  which is the intersection of  $\partial P$  with a  $(m - 1)$ -dimensional subspace of  $\mathbb{R}^m$ . Note that  $W$  can possibly contain only a single point in  $\text{int}(P)$ . Clearly  $\Omega_0$  is a finite set with at most  $2^{|\Omega|}$  elements.

Note that a linear combination of elements in  $\Omega$  with nonnegative coefficients lie on a face of  $\partial P$  if and only if all the coefficients for those generators that are not on this face are 0. In other words, if  $\sum_{b \in \Omega} k_b b$  is in  $\text{int}(P)$  and  $k_b$  are all nonnegative, then the set  $\{b \in \Omega \mid k_b \geq 1\}$  must not be contained in any face of  $\partial P$ . Hence

$$\text{int}(P) \cap \mathbb{Z}^m = \left\{ a + \sum_{b \in \Omega} k_b b \mid a \in \Omega_0, k_b \in \mathbb{Z}, k_b \geq 0 \right\}$$

and in particular  $\Omega_0 \subset \text{int}(P) \cap \mathbb{Z}^m$  as we desire. □

Here is an example of the two finite sets  $\Omega_0$  and  $\Omega$  for a rational cone in  $\mathbb{R}^2$ .

**Example 3.3** Let us consider the following rational cone in  $\mathbb{R}^2$ .

$$P = \left\{ \mathbf{x} = (x_1, x_2) \in \mathbb{R}^2 \mid \begin{pmatrix} 0 & 1 \\ 3 & -2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \geq \begin{pmatrix} 0 \\ 0 \end{pmatrix} \right\}.$$

One can take  $\Omega = \{b_1 = (1, 0), b_2 = (1, 1), b_3 = (2, 3)\}$  as a set of generators of  $P \cap \mathbb{Z}^2$ . There are two faces of  $\partial P$ . One is  $\{(x, 0) \mid x \geq 0\}$  which contains  $\{b_1\}$  as a

subset, and the other is  $\{(x, \frac{3}{2}x) \mid x \geq 0\}$  which contains  $\{b_3\}$  as a subset. One sees that  $\Omega_0$  consists of five elements,  $b_2, b_1 + b_2 = (2, 1), b_1 + b_3 = (3, 3), b_2 + b_3 = (3, 4)$  and  $b_1 + b_2 + b_3 = (4, 4)$ .

**Proof of Theorem 3.1** For a fibered cone  $\mathcal{C}$ , the closure  $\overline{\mathcal{C}}$  is a rational cone in  $H^1(M)$ , because the unit Thurston norm ball is a polytope whose vertices are rational points [32]. By Lemma 3.2, if an integral class  $\delta$  is in  $\mathcal{C}$ , then it can always be written of the form  $\delta = a + \sum_{b \in \Omega} k_b b$ , where  $a \in \Omega_0$  and  $k_b$  is a nonnegative integer. If  $S$  is a norm-minimizing surface of  $\delta$ , then we have  $\|\delta\| = |\chi(S)|$  and it is bounded above by

$$\max\left(1, \max_{b \in \Omega}(k_b)\right) \left(\|a\| + \sum_{b \in \Omega} \|b\|\right).$$

Hence, when  $|\chi(S)| > \max_{a \in \Omega_0} \|a\| + \sum_{b \in \Omega} \|b\|$ ,

$$|\chi(S)| \leq \max_{b \in \Omega}(k_b) \left(\|a\| + \sum_{b \in \Omega} \|b\|\right).$$

Therefore,

$$\max_{b \in \Omega}(k_b) \geq \frac{|\chi(S)|}{\|a\| + \sum_{b \in \Omega} \|b\|} \geq \frac{|\chi(S)|}{\max_{a \in \Omega_0} \|a\| + \sum_{b \in \Omega} \|b\|}.$$

Let  $b_m$  be a  $b$  in  $\Omega$  that maximizes  $k_b$ . We set  $\alpha = a + \sum_{b \in \Omega, b \neq b_m} k_b b, \beta = b_m$  and  $n = k_{b_m}$ . We have  $\alpha \in \mathcal{C}$  and  $\beta \in \overline{\mathcal{C}}$ . Then  $\delta$  is given by  $\delta = \alpha + n\beta$  with

$$n \geq \frac{|\chi(S)|}{\max_{a \in \Omega_0} \|a\| + \sum_{b \in \Omega} \|b\|}.$$

Note that the denominator in the right hand side only depends on the fibered cone. Now, when  $\|\delta\| = |\chi(S)| > \max_{a \in \Omega_0} \|a\| + \sum_{b \in \Omega} \|b\|$ , the conclusion of the theorem follows directly from Proposition 2.4. The remaining case  $\|\delta\| \leq \max_{a \in \Omega_0} \|a\| + \sum_{b \in \Omega} \|b\|$  consisting of finitely many primitive integral classes  $\delta$ ; hence the theorem is proved.  $\square$

### 3.2 Normal generation in the fibered cone

In this section we prove the following theorem as a partial answer to Conjecture 1.3.

**Theorem 3.4** *Let  $F$  be a fibered face of a closed hyperbolic fibered 3-manifold  $M$ , and let  $L$  be a 2-dimensional rational subspace of  $H^1(M)$ . Then for all but finitely many primitive integral classes  $(S, \psi)$  in  $\mathcal{C}_F \cap L$ ,  $\psi$  normally generates  $\text{Mod}(S)$ . In particular, if the rank of  $H^1(M)$  equals 2, then Conjecture 1.3 is true.*

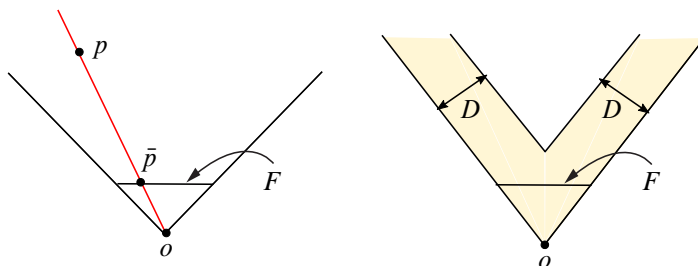


Figure 5: Left: the fibered face  $F$  in the fibered cone  $\mathcal{C}$  ( $p$  and  $\bar{p}$  lie on the same ray in  $\mathcal{C}_F$  passing through the origin). Right: the subset  $\mathcal{N}_D \subset \mathcal{C}$ .

For the proof of Theorem 3.4, we first prove the following result.

**Theorem 3.5** *Let  $\mathcal{C}$  be a fibered cone of a closed hyperbolic fibered 3–manifold  $M$ . Then there exists some  $x \in \mathcal{C}$  such that for each primitive integral class  $(S, \psi) \in x + \mathcal{C}$ ,  $\psi$  normally generates  $\text{Mod}(S)$ , where  $x + \mathcal{C} = \{x + v \mid v \in \mathcal{C}\}$ .*

**Proof** Let  $d$  be any Euclidean metric on  $H^1(M)$ . Let  $F$  be the fibered face corresponding to  $\mathcal{C}$ . For every point  $p \in \mathcal{C}$ , let  $\bar{p}$  be the intersection of  $F$  with the ray starting from the origin and passing  $p$  (Figure 5, left). By [27, Corollary 5.4], we have a real analytic, strictly concave and degree 1 homogeneous function  $y = 1/\log K(\cdot)$  defined on  $\mathcal{C}$ , such that the stretch factor  $\lambda(p)$  for  $p \in \mathcal{C}$  is equal to  $K(p)$  and  $y(p) = 1/\log K(p) \rightarrow 0$  as  $p \rightarrow \partial F$ . The concavity implies that there must be some  $k > 0$  (independent of the choice of  $\bar{p}$ ) such that

$$\frac{1}{\log(K(\bar{p}))} \geq k \cdot d(\bar{p}, \partial\mathcal{C}).$$

A way to see the existence of  $k$  is as follows. Concavity of  $y$  implies that there is some point  $p_0 \in F$  where  $y(p_0) > 0$ . Then, for any point  $\bar{p} \in F$ , consider the line segment from  $p_0$  to the boundary of  $F$  passing through  $\bar{p}$ . Then concavity of  $y$  means that on this line segment,  $y$  is bounded from below by the linear function  $L$  which takes value 0 at one end and  $y(p_0)$  at another end. Hence it has a slope  $s = s(\bar{p})$  that depends on  $\bar{p}$  and  $s = s(\bar{p})$  is continuous on  $\bar{p}$ . On the other hand, the function  $d(\cdot, \partial\mathcal{C})$ , restricted to this line segment, is piecewise linear, and hence it is also bounded from above by a linear function  $L'$  taking value 0 at the end on  $\partial F$ . We choose such linear function  $L'$  with the smallest slope  $s' = s'(\bar{p})$ . Then  $s' = s'(\bar{p})$  is continuous on  $\bar{p}$ . Now  $k$  can be chosen as any number below the ratio  $s/s'$  between these two slopes. As both slopes depend continuously on  $\bar{p}$ , and  $F$  has compact closure, we can choose a universal  $k$  that works on the whole face  $F$ .

Furthermore, the degree 1 homogeneity implies that

$$\frac{1}{\log(K(p))} = \frac{d(0, p)}{d(0, \bar{p})} \cdot \frac{1}{\log(K(\bar{p}))}.$$

For  $D > 0$ , we consider the set  $\mathcal{N}_D$  (Figure 5, right),

$$\mathcal{N}_D = \{p \in \mathcal{C} \mid d(p, \partial\mathcal{C}) \leq D\}.$$

From the above computation, the stretch factor for  $p \in \mathcal{C} \setminus \mathcal{N}_D$  satisfies

$$\begin{aligned} \lambda(p) &= e^{\log K(p)} = (e^{\log K(\bar{p})})^{d(0, \bar{p})/d(0, p)} \\ &\leq (e^{1/(kd(\bar{p}, \partial\mathcal{C}))})^{d(0, \bar{p})/d(0, p)} = e^{1/(kd(p, \partial\mathcal{C}))} \leq e^{1/(kD)}. \end{aligned}$$

Hence as long as  $D$  is sufficiently large,  $\lambda(p)$  can be made to be as close to 1 as needed. In particular it is smaller than  $\sqrt{2}$  when  $D$  is large enough. This together with [20, Theorem 1.2] shows that for some  $D$ , all primitive integral classes in  $\mathcal{C} \setminus \mathcal{N}_D$  are normal generators. The theorem now follows by picking an arbitrary  $x \in \mathcal{C} \setminus \mathcal{N}_D$ , due to the fact that the boundary of  $\mathcal{N}_D$  must be parallel to that of  $\partial\mathcal{C}$  itself; see Figure 5, right.  $\square$

The next result follows immediately from Lemma 2.1 and Proposition 2.2.

**Theorem 3.6** *Let  $\mathcal{C}$  be a fibered cone of a closed hyperbolic fibered 3-manifold. Suppose that  $(S_{\alpha_n}, \psi_{\alpha_n})$  is a sequence of primitive integral classes in  $\mathcal{C}$  such that  $\alpha_n = v + nw$ , where  $v \in \mathcal{C}$  and  $w \in \overline{\mathcal{C}}$  are fixed integral classes. Then  $\psi_{\alpha_n}$  normally generates  $\text{Mod}(S_{\alpha_n})$  for sufficiently large  $n$ .*

We are now ready to prove Theorem 3.4.

**Proof of Theorem 3.4** Let  $L$  be a 2-dimensional rational subspace of  $H^1(M)$  satisfying the assumption of Theorem 3.4. Theorem 3.5 says that there is some  $x \in \mathcal{C}$  such that all primitive integral classes  $(S, \psi)$  in  $x + \mathcal{C}$  normally generate  $\text{Mod}(S)$ . In particular this holds for all primitive integral classes in  $(x + \mathcal{C}) \cap L$ . Because  $L$  is of dimension 2, the integral classes in  $(\mathcal{C} \setminus (x + \mathcal{C})) \cap L$  are the union of finitely many sequences of the form  $(v + nw)_{n \in \mathbb{N}}$ , where  $v \in \mathcal{C}$  and  $w \in \overline{\mathcal{C}}$ . Thus by Theorem 3.6, for all but finitely many primitive integral classes  $(S, \psi)$  in  $(\mathcal{C} \setminus (x + \mathcal{C})) \cap L$ ,  $\psi$  normally generates  $\text{Mod}(S)$ .  $\square$

**Remark 3.7** Our approach to Theorem 3.4 does not work when the dimension of the rational subspace  $L$  of  $H^1(M)$  is more than 2. This is because in this case, the intersection  $(\mathcal{C} \setminus (x + \mathcal{C})) \cap L$  no longer consists of finitely many sequences of primitive integral classes of the form  $v + nw$ , where  $v \in \mathcal{C}$  and  $w \in \overline{\mathcal{C}}$ .

## 4 Sequences in the fibered cone of the magic manifold

Let  $\mathcal{C}_3$  be the 3-chain link in  $S^3$  as in Figure 6, left. The magic manifold  $N$  is the exterior of  $\mathcal{C}_3$  (hence  $\partial N$  consists of three boundary tori), and it is a hyperbolic and fibered 3-manifold. We give some background on invariant train tracks in Section 4.1 and we discuss the fibered cone of  $N$  in Section 4.2. We compute the upper and lower bounds of the asymptotic translation length of particular sequences in the fibered cone of  $N$  in Sections 4.3 and 4.4. Then we prove Theorem 4.13 in Section 4.5.

### 4.1 Invariant train tracks for pseudo-Anosov maps

For definitions and basic results on train tracks, see [4; 8; 29]. Let  $\psi : S \rightarrow S$  be a pseudo-Anosov homeomorphism defined on a surface  $S$  possibly with boundary/punctures. When  $S$  is a punctured surface, we say that  $\psi$  is *fully punctured* if the set of singularities of the unstable foliation for  $\psi$  is contained in the set of punctures of  $S$ .

Let  $\tau$  be an invariant train track for  $\psi$ . Then  $\psi : S \rightarrow S$  induces a map on  $\tau$  to itself which takes switches (vertices) to themselves. Such a map is called the *train track map*. By abuse of notation, we denote the train track map on  $\tau$  also by  $\psi : \tau \rightarrow \tau$ . Following [4, Section 3.3], we say that a branch  $e$  of  $\tau$  is *real* if there exists an integer  $m \geq 1$  such that  $\psi^m(e)$  passes through all branches of  $\tau$ . Otherwise we say that  $e$  is *infinitesimal*. The train track map  $\psi : \tau \rightarrow \tau$  induces a finite digraph  $\Gamma$  by taking a vertex for each real branch of  $\tau$ , and then adding  $m_{ij}$  directed edges from the  $j^{\text{th}}$  real branch  $e_j$  to the  $i^{\text{th}}$  real branch  $e_i$ , where  $m_{i,j}$  is the number of times that the image  $\psi(e_j)$  under the train track map  $\psi$  passes through  $e_i$  in either direction.

For the lower bound of  $\ell_{\mathcal{C}}(\psi)$ , we recall a result of Gadre and Tsai. The following statement is a consequence of [13, Lemma 5.2] together with the proof of [13, Theorem 5.1].

**Proposition 4.1** *Let  $\psi \in \text{Mod}(S_{g,n})$  be a pseudo-Anosov element and let  $\tau$  be an invariant train track for  $\psi$ . Suppose that  $r$  is a positive integer such that for any real branch  $e$  of  $\tau$ ,  $\psi^r(e)$  passes through every real branch. If we set  $h = r + 24|\chi(S_{g,n})| - 8n$ , then  $\psi^h(e)$  passes through every branch of  $\tau$  (including infinitesimal branches). Moreover, if we set*

$$w = h + 6|\chi(S_{g,n})| - 2n = r + 30|\chi(S_{g,n})| - 10n \leq r + 30|\chi(S_{g,n})|,$$

then we have

$$\ell_{\mathcal{C}}(\psi) \geq \frac{1}{w} \geq \frac{1}{r + 30|\chi(S_{g,n})|}.$$

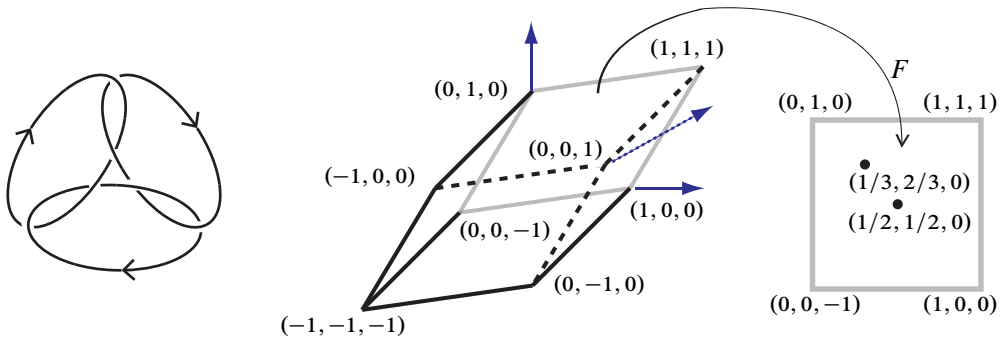


Figure 6: Left: the 3-chain link  $C_3$ . Right: the Thurston norm ball of  $N$  and fibered face  $F$ .

### 4.2 Fibered cones of the magic manifold

We consider coordinates of integral classes in fibered cones of  $N$ . We assign orientations of the three components of  $C_3$  as in Figure 6, left. Let  $S_\alpha$ ,  $S_\beta$  and  $S_\gamma$  be the oriented 2-punctured disks bounded by these components of  $C_3$ . We set  $\alpha = [S_\alpha]$ ,  $\beta = [S_\beta]$  and  $\gamma = [S_\gamma]$  in  $H_2(N, \partial N; \mathbb{Z}) \simeq H^1(N; \mathbb{Z})$ . Then  $\alpha$ ,  $\beta$  and  $\gamma$  form a basis of  $H_2(N, \partial N; \mathbb{Z})$ . We denote by  $(x, y, z)$  the class  $x\alpha + y\beta + z\gamma$ . The Thurston norm ball  $B_N$  is the parallelepiped with vertices  $\pm\alpha = \pm(1, 0, 0)$ ,  $\pm\beta = \pm(0, 1, 0)$ ,  $\pm\gamma = \pm(0, 0, 1)$  and  $\pm(\alpha + \beta + \gamma) = \pm(1, 1, 1)$ ; see Figure 6, right.

A symmetry of  $C_3$  tells us that every top-dimensional face of  $B_N$  is a fibered face. Moreover all fibered faces of  $N$  are permuted transitively by homeomorphisms of  $N$ . Hence they have the same topological types in their fibers and the same dynamics of their monodromies. To study monodromies of fibrations on  $N$ , it suffices to pick a particular fibered face, say  $F$  with vertices  $(1, 0, 0)$ ,  $(1, 1, 1)$ ,  $(0, 1, 0)$  and  $(0, 0, -1)$ ; see Figure 6, right. For a primitive integral class  $(S, \psi) \in \mathcal{C}_F$ , the monodromy  $\psi$  is pseudo-Anosov defined on  $S$  with boundary components, since  $\partial N \neq \emptyset$ . Each connected component of  $\partial S$  is a simple closed curve which lies on one of the boundary tori of  $N$ . By abusing notation, we often regard boundary components of  $S$  as punctures of  $S$  by crushing each boundary component to a puncture. Hence we think of  $\psi$  as a pseudo-Anosov map defined on the punctured surface  $S$ . Such ambiguity does not matter for our purpose since the computation of the asymptotic translation lengths of the pseudo-Anosov monodromies on the curve complex will not be affected. Under this convention, one sees that for any primitive integral class  $(S, \psi) \in \mathcal{C}_F$ , the pseudo-Anosov monodromy  $\psi$  is fully punctured; see for example [15].

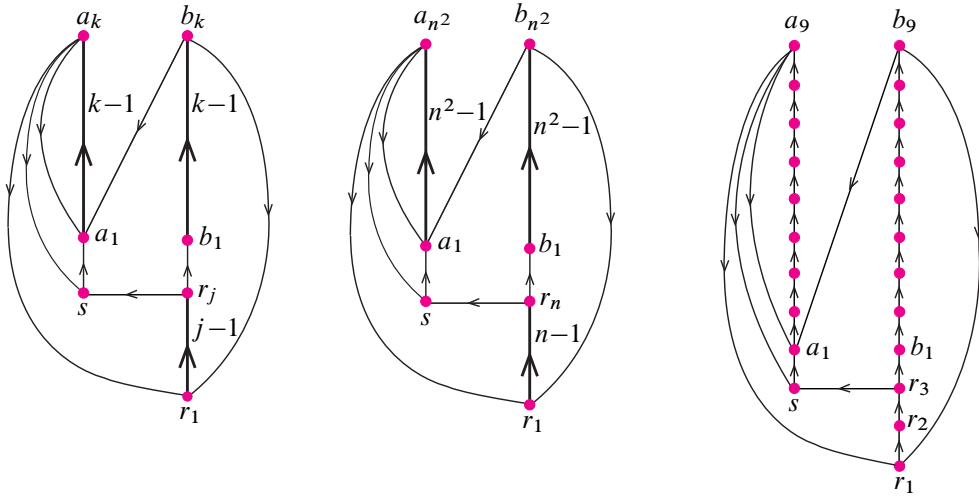


Figure 7: Digraphs  $\Gamma_{(1,j,k)+}$ , left,  $\Gamma_{(1,n,n^2)+}$ , middle, and  $\Gamma_{(1,3,9)+}$ , right.

The open face  $\text{int}(F)$  is written by

$$\text{int}(F) = \{(x, y, z) \mid x + y - z = 1, x > 0, y > 0, x > z, y > z\}.$$

This implies that  $(x, y, z) \in \mathcal{C}_F$  if and only if  $x > 0, y > 0, x > z$  and  $y > z$ . The next lemma tells us the topological type of the corresponding fiber  $S_{(x,y,z)}$ .

**Lemma 4.2** [17] *For a primitive integral class  $(x, y, z) \in \mathcal{C}_F$ , let  $|\partial S_{(x,y,z)}|$  denote the number of the boundary components of  $S_{(x,y,z)}$ . The Thurston norm*

$$\|(x, y, z)\| = |\chi(S_{(x,y,z)})|$$

*equals  $x + y - z$ , and  $|\partial S_{(x,y,z)}|$  is given by*

$$|\partial S_{(x,y,z)}| = \gcd(x, y + z) + \gcd(y, z + x) + \gcd(z, x + y).$$

*More precisely, each term in the right-hand side expresses the number of boundary components of  $S_{(x,y,z)}$  which lie on one of the boundary tori of  $N$ .*

We introduce another coordinate,  $(i, j, k)_+$ . For  $i, j, k \geq 0$ , define

$$(i, j, k)_+ = i(1, 1, 1) + j(0, 1, 0) + k(1, 1, 0) = (i + k, i + j + k, i).$$

Note that  $(1, 1, 0) \in \mathcal{C}_F$ , but  $(0, 1, 0) \notin \mathcal{C}_F$  and  $(1, 1, 1) \notin \mathcal{C}_F$  (in fact the two classes lie on  $\partial F$ ); see Figure 6, right. We denote by  $\overline{(i, j, k)_+}$ , the class with the Thurston norm 1 which is projectively equal to  $(i, j, k)_+$ .



If  $i, j$  and  $k$  are integers with  $i \geq 0, j \geq 0$  and  $k > 0$ , then  $(i, j, k)_+ \in \mathcal{C}_F$ . If  $(i, j, k)_+$  is a primitive integral class in  $\mathcal{C}_F$ , then we let  $(S_{(i,j,k)_+}, \psi_{(i,j,k)_+})$  be the pair of the fiber and its monodromy. In [15, Section 3], the second author constructs an invariant train track  $\tau = \tau_{(i,j,k)_+}$  and the digraph  $\Gamma = \Gamma_{(i,j,k)_+}$  of the train track map  $\psi = \psi_{(i,j,k)_+} : \tau \rightarrow \tau$  for each primitive integral class  $(i, j, k)_+ \in \mathcal{C}_F$ . Figure 7, left, illustrates  $\Gamma = \Gamma_{(1,j,k)_+}$  when  $i = 1, j > 0$  and  $k > 0$ ; see also [15, Figure 22(4)]. The vertices in the left column of  $\Gamma$  are denoted by  $s, a_1, \dots, a_k$  from bottom to top; vertices in the right column of  $\Gamma$  are denoted by  $r_1, \dots, r_j, b_1, \dots, b_k$  from bottom to top. (Recall that each vertex of  $\Gamma$  corresponds to a real branch of  $\tau$ .) The numbers  $j - 1$  and  $k - 1$  near the “thick” edges of  $\Gamma$  indicate their lengths of paths. For instance, the edge  $r_1 \xrightarrow{j-1} r_j$  from  $r_1$  to  $r_j$  indicates the edge path  $r_1 \rightarrow \dots \rightarrow r_{j-1} \rightarrow r_j$ . See Figure 7, right, for the concrete example. When  $j = 1$  or  $k = 1$ , the corresponding “thick” edges collapse; see Figure 11.

### 4.3 Computing the lower bounds

For fixed positive integers  $p$  and  $q$ , we consider the sequence

$$(1, n^p, n^q)_+ = (1 + n^q, 1 + n^p + n^q, 1) \in \mathcal{C}_F$$

for a varying positive integer  $n$ . The integral class  $(1, n^p, n^q)_+$  is primitive, since  $\gcd(1, n^p, n^q) = 1$ . From the formula of the Thurston norm in Lemma 4.2, it is immediate to see the following lemma. See also Figure 6, right.

**Lemma 4.3** *Let  $(\overline{1, n^p, n^q})_+$  be the projective class of  $(1, n^p, n^q)_+$ .*

- (1) *If  $p = q$ , then  $(\overline{1, n^p, n^q})_+ \rightarrow (\frac{1}{3}, \frac{2}{3}, 0) \in \text{int}(F)$  as  $n \rightarrow \infty$ .*
- (2) *If  $p < q$ , then  $(\overline{1, n^p, n^q})_+ \rightarrow (\frac{1}{2}, \frac{1}{2}, 0) \in \text{int}(F)$  as  $n \rightarrow \infty$ .*
- (3) *If  $p > q$ , then  $(\overline{1, n^p, n^q})_+ \rightarrow (0, 1, 0) \in \partial F$  as  $n \rightarrow \infty$ .*

Here we consider the following three cases:  $q < p < 2q, p < q \leq 2p$  and  $2p \leq q$ . We define

$$k = k_{p,q} = \begin{cases} n^q(2n^q + 1) & \text{if } q < p < 2q, \\ n^q(2n^p + 1) & \text{if } p < q \leq 2p, \\ n^q(2n^{q-p} + 1) & \text{if } 2p \leq q. \end{cases}$$

**Proposition 4.4** *For any two vertices  $v$  and  $w$  of  $\Gamma = \Gamma_{(1,n^p,n^q)_+}$ , there exists an edge path from  $v$  to  $w$  of length  $k + 2n^p + 3n^q$ .*

In other words, if we set  $k' = k_{p,q} + 2n^p + 3n^q$ , then for any real branch  $v$  of  $\tau$ ,  $\psi^{k'}(v)$  passes through every real branch. For the proof of Proposition 4.4, we need some lemmas. Recall that  $s$  is the bottom vertex in the left column of  $\Gamma$ . Let  $v_0$  be the top vertex  $a_{n^q}$  in the left column of  $\Gamma$ ; see Figure 9.

**Lemma 4.5** *For any vertex  $v$  in the left column of  $\Gamma$ , there exists an edge path from  $s$  to  $v$  of length  $k$ .*

**Proof** We have an edge path  $s \rightarrow a_1 \xrightarrow{n^q-1} a_{n^q} = v_0$  from  $s$  to  $v_0$  of length  $n^q$ . For the proof of the lemma, it suffices to show that for any vertex  $v$  in the left column of  $\Gamma$ , there exists an edge path from  $v_0$  to  $v$  of length  $k - n^q$ . Then the desired path can be obtained from the concatenation of the two paths, the path from  $s$  to  $v_0$  and the path from  $v_0$  to  $v$ . Equivalently, we show that for any  $i = 0, \dots, n^q$ , there exists a cycle based at  $v_0$  of length  $k - n^q + i$ .

It is easy to find two cycles based at  $v_0$  in  $\Gamma$  of lengths  $n^q$  and  $n^q + 1$ ; see Figure 7, left. We have another cycle based at  $v_0$  in  $\Gamma$  of length  $n^p + n^q + 1$ ,

$$v_0 = a_{n^q} \rightarrow r_1 \xrightarrow{n^p-1} r_{n^p} \rightarrow s \rightarrow a_1 \xrightarrow{n^q-1} a_{n^q} = v_0$$

We show that repeated use of these three cycles is enough to produce the cycles we desire. Suppose  $q < p < 2q$ . Then  $k - n^q = 2n^{2q}$ . We now show that for any  $i = 0, \dots, n^q$ , there exist nonnegative integers  $a, b$  and  $c$  such that

$$an^q + b(n^q + 1) + c(n^q + n^p + 1) = 2n^{2q} + i.$$

This is done by setting  $c = 0, b = i$  and  $a = 2n^q - i$ .

Suppose  $p < q \leq 2p$ . Then  $k - n^q = 2n^{p+q}$ . We claim that for any  $i = 0, \dots, n^q$ , there exist nonnegative integers  $a, b$ , and  $c$  such that

$$an^q + b(n^q + 1) + c(n^q + n^p + 1) = 2n^{p+q} + i.$$

This can be done by setting

$$c = \left\lfloor \frac{i}{n^p + 1} \right\rfloor, \quad b = i - (n^p + 1) \left\lfloor \frac{i}{n^p + 1} \right\rfloor, \quad a = 2n^p - b - c,$$

where  $\lfloor \cdot \rfloor$  is the floor function. Here  $b$  and  $c$  are nonnegative integers by definition, and  $b$  is the remainder of  $i$  divided by  $n^p + 1$ . Hence  $b$  must be no larger than  $n^p$ . On the other hand  $c \leq n^{q-p}$ , because  $i \leq n^q < n^{q-p}(n^p + 1)$ . Thus  $b + c \leq n^p + n^{q-p} \leq 2n^p$ , which implies that  $a$  is nonnegative.

Lastly, suppose  $2p \leq q$ . Then  $k - n^q = 2n^{2q-p}$ . We claim that for any  $i = 0, \dots, n^q$ , there exist nonnegative integers  $a, b$ , and  $c$  such that

$$an^q + b(n^q + 1) + c(n^q + n^p + 1) = 2n^{2q-p} + i.$$

This can be done by setting

$$c = \left\lfloor \frac{i}{n^p + 1} \right\rfloor, \quad b = i - (n^p + 1) \left\lfloor \frac{i}{n^p + 1} \right\rfloor, \quad a = 2n^{q-p} - b - c.$$

Here  $b$  and  $c$  are nonnegative integers by definition, and  $b$  is the remainder of  $i$  divided by  $n^p + 1$ . Hence  $b$  must be no larger than  $n^p$ . On the other hand  $c \leq n^{q-p}$ , because  $i \leq n^q < n^{q-p}(n^p + 1)$ . Thus  $b + c \leq n^p + n^{q-p} \leq 2n^{q-p}$ , which says that  $a$  is nonnegative. This finishes the proof.  $\square$

**Lemma 4.6** *For any vertex  $v$  in the left column of  $\Gamma$  and for any  $m \geq 0$ , there exists an edge path from  $s$  to  $v$  of length  $k + m$ .*

**Proof** Let  $v$  be any vertex in the left column of  $\Gamma$ . For any  $m \geq 0$ , one can find a vertex  $v'$  in the left column of  $\Gamma$  such that there is an edge path from  $v'$  to  $v$  of length  $m$ . (To see this, use the above cycles based at  $v_0$  of lengths  $n^q$  and  $n^q + 1$ .) Lemma 4.5 tells us that there exists an edge path from  $s$  to  $v'$  of length  $k$ . The concatenation of these edge paths is a desired edge path of length  $k + m$ .  $\square$

**Lemma 4.7** *For any vertex  $v$  in the right column of  $\Gamma$  and for any  $m \geq 0$ , there exists an edge path from  $s$  to  $v$  of length  $k + n^p + n^q + m$ .*

**Proof** Let  $v$  be an arbitrary vertex in the right column of  $\Gamma$ . Then there exists an edge path from  $v_0$  to  $v$  of length  $\ell$  with  $1 \leq \ell \leq n^p + n^q$ . To see this, use the path

$$v_0 = a_{n^q} \rightarrow r_1 \xrightarrow{n^p-1} r_{n^q} \rightarrow b_1 \xrightarrow{n^q-1} b_{n^q}$$

from  $v_0$  to  $b_{n^q}$ . On the other hand, Lemma 4.6 tells us that there exists an edge path from  $s$  to  $v_0$  of length  $k + (n^p + n^q - \ell) + m$ . Here  $(n^p + n^q - \ell) + m$  plays the role of  $m$  in Lemma 4.6. Concatenating these two paths, one obtains an edge path from  $s$  to  $v$  of length  $k + n^p + n^q + m$ .  $\square$

By Lemmas 4.6 and 4.7, we immediately have the following lemma.

**Lemma 4.8** *For any vertex  $v$  of  $\Gamma$  and for any  $m \geq 0$ , there exists an edge path from  $s$  to  $v$  of length  $k + n^p + n^q + m$ .*

We are now ready to prove Proposition 4.4.

**Proof of Proposition 4.4** Note that for any vertex  $v$ , there exists an edge path from  $v$  to  $s$  of length  $0 \leq \ell \leq n^p + 2n^q$ . To see this, one can use the edge path of length  $n^p + 2n^q$  passing through all vertices of  $\Gamma$ ,

$$r_1 \xrightarrow{n^p-1} r_{n^q} \rightarrow b_1 \xrightarrow{n^q-1} b_{n^q} \rightarrow a_1 \xrightarrow{n^q-1} a_{n^q} \rightarrow s.$$

By Lemma 4.8 there exists an edge path from  $s$  to any vertex  $w$  of length exactly  $k + (2n^p + 3n^q - \ell)$ , since  $2n^p + 3n^q - \ell \geq n^p + n^q$ . The concatenation of the two paths has length  $k + 2n^p + 3n^q$ . □

Now we are ready to compute the lower bounds. For real-valued functions  $A(x)$  and  $B(x)$ , we write  $A(x) \gtrsim B(x)$  if there is a constant  $C > 0$  independent of  $x$  such that  $A(x) \geq C \cdot B(x)$ .

**Theorem 4.9** *The sequence  $(1, n^p, n^q)_+$  in  $\mathcal{C}_F$  satisfies*

$$\ell_C(\psi_{(1, n^p, n^q)_+}) \gtrsim \begin{cases} 1/n^{2q} & \text{if } q < p < 2q, \\ 1/n^{p+q} & \text{if } p < q \leq 2p, \\ 1/n^{2q-p} & \text{if } 2p \leq q. \end{cases}$$

**Proof** By Lemma 4.2, it is not hard to see that

$$(k_{p,q} + 2n^p + 3n^q) + 30|\chi(S_{(1, n^p, n^q)_+})| \asymp \begin{cases} n^{2q} & \text{if } q < p < 2q, \\ n^{p+q} & \text{if } p < q \leq 2p, \\ n^{2q-p} & \text{if } 2p \leq q. \end{cases}$$

Then the desired claim follows from Propositions 4.1 and 4.4. □

### 4.4 Computing the upper bounds

To prove Theorem 4.13, we will also compute the upper bound of the asymptotic translation length of  $\psi_{(1, n^p, n^q)_+}$ .

**Theorem 4.10** *For any fixed positive integers  $p$  and  $q$  with  $q < p < 2q$ , the sequence  $(1, n^p, n^q)_+$  of primitive integral classes in  $\mathcal{C}_F$  converges projectively to  $(0, 1, 0) \in \partial F$  as  $n \rightarrow \infty$ , and*

$$\ell_C(\psi_{(1, n^p, n^q)_+}) \leq \frac{4}{n^{2q}}.$$

The first half of Theorem 4.10 follows from Lemma 4.3(3). For the rest of the proof, we first introduce the dual arcs of real branches of train tracks. Consider an invariant train track  $\tau$  for the monodromy  $\psi$  defined on the fiber  $S$  of a fibration on  $N$ . If we

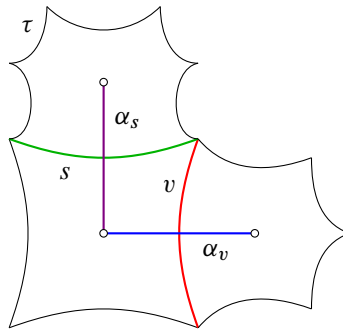


Figure 8: Cell decomposition, branches, and dual arcs.

think of the surface  $S$  with boundary as the punctured surface which is again denoted by  $S$  abusing notation, each component of the complement  $S \setminus \tau$  of the train track is a once-punctured ideal polygon, because  $\psi$  is fully punctured. Consider the cell decomposition of  $S$  corresponding to  $\tau$ . That is, 0-cells are switches of  $\tau$ , 1-cells are branches of  $\tau$ , and 2-cells are ideal polygons of  $S \setminus \tau$ .

Given a real branch  $v$ , the *dual arc*  $\alpha_v$  of  $v$  is defined to be the edge of the dual cell complex that connects the punctures in two polygons (possibly the same polygon) sharing the real branch  $v$ ; see Figure 8.

Notice that the dual arc  $\alpha_v$  is an essential arc. In order to see this, consider a rectangle associated with the real branch  $v$ , contained in a Markov partition for a pseudo-Anosov homeomorphism which represents  $\psi$ . Then  $v$  corresponds to leaves of the unstable foliation and the dual arc  $\alpha_v$  corresponds to leaves of the stable foliation in this rectangle. If the dual arc is not essential, then this implies that the real branch  $v$  cannot support a positive transverse measure, which is a contradiction to a property of pseudo-Anosov homeomorphisms.

Readers may notice that the dual arc associated to a real branch is a general notion for fully punctured pseudo-Anosov homeomorphisms. More precisely, if  $\tau$  is an invariant train track for a fully punctured pseudo-Anosov  $\psi$ , then for a real branch  $v$  of  $\tau$ , one can define the dual arc  $\alpha_v$  which is essential.

**Proof of Theorem 4.10** Let  $(S, \psi) = (S_{(1, n^p, n^q)_+}, \psi_{(1, n^p, n^q)_+})$  be the pair of the fiber and its monodromy for  $(1, n^p, n^q)_+$ . Let  $\Gamma$  be the digraph of the train track  $\tau$  for  $(1, n^p, n^q)_+$ , and let  $\psi_* : V(\Gamma) \rightarrow V(\Gamma)$  be the induced map, where  $V(\Gamma)$  is the set of vertices of  $\Gamma$ . The map  $\psi_*$  can be read off Figure 9.

Here is the outline of the proof. We will compute the upper bound of the asymptotic translation length  $\ell_{\mathcal{AC}}(\psi)$  of  $\psi$  on the arc and curve complex  $\mathcal{AC}(S)$ . Since  $\mathcal{C}(S)$  and  $\mathcal{AC}(S)$  are quasi-isometric, this gives an upper bound on  $\mathcal{C}(S)$ . We show that there are distinct vertices  $t$  and  $v$  in  $\Gamma$ , ie distinct real branches  $t$  and  $v$  of  $\tau$ , such that  $\psi_*^{n^{2q}}(t)$  does not contain  $v$ . Using this fact, we also show that there are disjoint arcs  $\beta_t$  and  $\alpha_v$  in  $\mathcal{AC}(S)$  such that  $\psi^{n^{2q}}(\beta_t)$  and  $\alpha_v$  are disjoint. This implies that the distance in  $\mathcal{AC}(S)$  satisfies  $d_{\mathcal{AC}}(\beta_t, \psi^{n^{2q}}(\beta_t)) \leq 2$ , and we deduce that  $\ell_{\mathcal{AC}}(\psi) \leq 2/n^{2q}$ .

**Step 1**  $\mathcal{C}(S)$  and  $\mathcal{AC}(S)$  are quasi-isometric.

**Proof** Just recall that the inclusion map  $\mathcal{C}(S) \rightarrow \mathcal{AC}(S)$  is 2-bilipschitz. □

Hence for the proof of Theorem 4.10, it is enough to show that the asymptotic translation length  $\psi$  on  $\mathcal{AC}(S)$  satisfies

$$\ell_{\mathcal{AC}}(\psi) \leq \frac{2}{n^{2q}}.$$

**Step 2** Let  $t$  be the vertex  $b_{n^q}$  of  $\Gamma$ . Then  $\psi_*^{n^{2q}}(t)$  doesn't contain all vertices in  $\Gamma$ .

**Proof** We will show that there is a vertex  $v$  that is not contained in  $\psi_*^{n^{2q}}(t)$ . Consider the partition  $\{A, B, R_1, R_2, \dots, R_{n^{p-q}}\}$  of vertices  $a_i, b_i,$  and  $r_i$  of  $\Gamma$ , where each partition element consists of  $n^q$  vertices as in Figure 9. Under the iteration of the  $(n^q)^{\text{th}}$  power  $\psi_*^{n^q}$  of  $\psi_*$ , one can see that

$$\begin{aligned} \psi_*^{n^q}(t) &= \{a_{n^q}, r_{n^q}\}, \\ \psi_*^{2n^q}(t) &= \{a_{n^q}, a_{n^q-1}, r_{n^q}, r_{2n^q}\}, \\ \psi_*^{3n^q}(t) &= \{a_{n^q}, a_{n^q-1}, a_{n^q-2}, r_{n^q}, r_{n^q-1}, r_{2n^q}, r_{3n^q}\}, \\ &\vdots \end{aligned}$$

and that the number of vertices in each partition element contained in  $\psi_*^{j \cdot n^q}(t)$  is increasing by at most one as  $j$  increases. Hence one can see that there are vertices in each  $R_k$  ( $k = 1, \dots, n^{p-q}$ ) that are not contained in  $\psi_*^{n^{2q}}(t)$ . More precisely, consider  $R_1 = \{r_1, r_2, \dots, r_{n^q}\}$ . One can check that for vertices in  $R_1$ , the image  $\psi_*^{j \cdot n^q}(t)$  contains only

$$\{r_{n^q}, r_{n^q-1}, \dots, r_{n^q-j+2}\} \subset R_1$$

for  $2 \leq j \leq n^q$ . Therefore  $\psi_*^{n^{2q}}(t)$  does not contain  $r_1$ , and we may choose  $v$  to be  $r_1$ . □

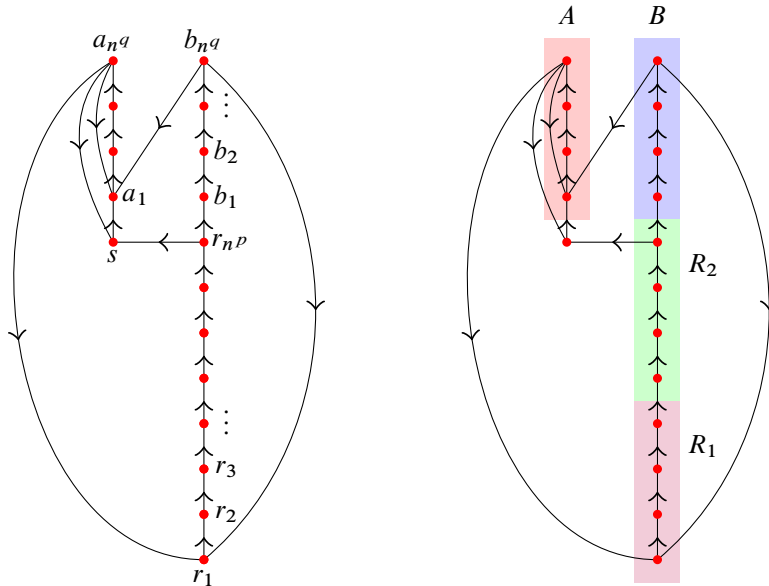


Figure 9: Left: digraph  $\Gamma_{(1,n^p,n^q)_+}$ . Right: digraph  $\Gamma_{(1,2^3,2^2)_+}$  with partition  $\{A, B, R_1, R_2\}$ .

**Step 3** *There are distinct arcs  $\alpha_v$  and  $\beta_t$  in  $AC(S)$  such that  $\psi^{n^{2q}}(\beta_t)$  and  $\alpha_v$  are disjoint.*

Before proving Step 3, we first discuss some properties of the primitive integral class  $(1, j, k)_+$  with  $j > 0$  and  $k > 0$ . Recall that  $r_1, \dots, r_j, b_1, \dots, b_k$  are vertices of  $\Gamma = \Gamma_{(1,j,k)_+}$  which lie on the right column of  $\Gamma$  (Figure 7, left). There is a single ideal polygon  $P = P_{(1,j,k)_+}$  containing a single puncture  $c_P$  of the fiber  $S = S_{(1,j,k)_+}$  such that the two endpoints of each real branch  $b_i$  ( $i = 1, \dots, k$ ) are switches (of  $\tau$ ) in the boundary  $\partial P$  of  $P$ ; see Figure 10. From the construction of  $\tau$  in [15], it follows that  $\partial P$  consists of periodic branches, ie infinitesimal branches, and  $\psi = \psi_{(1,j,k)_+}$  maps  $c_P$  to itself (and hence the ideal polygon  $P$  is preserved by  $\psi$ ). To see  $\psi(c_P) = c_P$ , we consider the fiber  $S = S_{(i,j,k)_+}$  with boundary. (So we now think of the above  $c_P$  as a boundary component of  $S$ .) By using Lemma 4.2 for the primitive integral class  $(1, j, k)_+$ , we see that there is a boundary torus  $T$  of  $N$  such that  $c_P$  is the only boundary component of  $S$  which lies on  $T$ . This implies  $c_P$  is preserved by  $\psi$ .

For the real branch  $r_i$  ( $i = 1, \dots, j$ ), consider its dual arc  $\alpha_{r_i}$ . Let  $c_{r_i}$  and  $c'_{r_i}$  be boundary components in  $\partial S$  which are connected by  $\alpha_{r_i}$ . (Possibly  $c_{r_i} = c'_{r_i}$ .) Then there is another boundary torus  $T'$  of  $N$  on which the both  $c_{r_i}$  and  $c'_{r_i}$  lie.

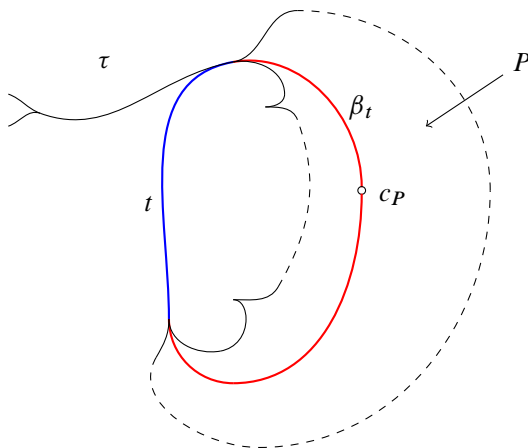


Figure 10: Part of the train track  $\tau$ . The ideal polygon  $P$ , real branch  $t = b_{n^q}$ , and arc  $\beta_t$  based at  $c_P$ .

**Proof of Step 3** Consider the primitive integral class  $(1, n^p, n^q)_+$  in question. The two endpoints of the real branch  $t = b_{n^q}$  are switches (of  $\tau$ ) in  $\partial P$ . Join  $c_P$  and each endpoint of the real branch  $t$  by an arc and then we obtain an arc  $\beta_t$  in  $S$ ; see Figure 10. Since  $t$  is a real branch, one sees that the arc  $\beta_t$  is essential. Since  $\psi$  maps  $c_P$  to itself,  $\psi^\ell(\beta_t)$  is an essential arc based at the same  $c_P$  for each  $\ell > 0$ . Moreover  $\psi^\ell(\beta_t)$  is not homotopic to  $\beta_t$  for each  $\ell > 0$ , since  $\psi$  is pseudo-Anosov. Let us consider the dual arc  $\alpha_v$  of  $v = r_1$ . Recall that  $c_v$  and  $c'_v$  which are connected by  $\alpha_v$  lie on a boundary torus  $T'$  of  $N$ , yet  $c_P$  lies on the different boundary torus  $T$  of  $N$ . The arc  $\beta_t$  has endpoints at  $c_P$ , and hence  $\beta_t$  is not homotopic to  $\alpha_v$ .

Now we prove that  $\psi^{2q}(\beta_t)$  and  $\alpha_v$  are disjoint. The ideal polygon  $P$  is preserved by  $\psi$ , and  $\psi^{n^{2q}}(t)$  is carried by  $\tau$  since  $\tau$  is invariant under  $\psi$ . Moreover, since  $\psi^{n^{2q}}(t)$  does not pass through  $v$  by the proof of Step 2, it follows that  $\psi^{n^{2q}}(\beta_t)$  is disjoint from  $v$ , and hence also disjoint from its dual arc  $\alpha_v$ .  $\square$

**Step 4**

$$\ell_{\mathcal{AC}}(\psi) \leq \frac{2}{n^{2q}}.$$

**Proof** Clearly  $\beta_t$  and  $\alpha_v$  are disjoint. Since  $\psi^{n^{2q}}(\beta_t)$  is an essential arc based at  $c_P$ , we have  $\psi^{n^{2q}}(\beta_t) \neq \alpha_v$  in  $\mathcal{AC}(S)$  by the same argument as in the proof of Step 3. This together with the fact that the geometric intersection number  $i(\psi^{n^{2q}}(\beta_t), \alpha_v) = 0$



implies that  $\beta_t$  and  $\psi^{n^{2q}}(\beta_t)$  are at most distance 2 in  $\mathcal{AC}(S)$ , ie  $d_{\mathcal{AC}}(\beta_t, \psi^{n^{2q}}(\beta_t)) \leq 2$ . By the definition of the asymptotic translation length, it follows that

$$\ell_{\mathcal{AC}}(\psi) \leq \frac{2}{n^{2q}}. \quad \square$$

Thus we have finished the proof of Theorem 4.10. □

**Theorem 4.11** *For any fixed positive integers  $p$  and  $q$  with  $2p \leq q$ , the sequence  $(1, n^p, n^q)_+$  of primitive integral classes in  $\mathcal{C}_F$  converges projectively to  $(\frac{1}{2}, \frac{1}{2}, 0)$  in  $\text{int}(F)$  as  $n \rightarrow \infty$ , and*

$$\ell_{\mathcal{C}}(\psi_{(1, n^p, n^q)_+}) \leq \frac{C}{n^{2q-p}},$$

where  $C$  is a constant independent of  $n$ .

**Proof** The first half of the claim follows from Lemma 4.3(2). For the rest of the proof, let  $\psi = \psi_{(1, n^p, n^q)_+}$ . Consider the digraph  $\Gamma = \Gamma_{(1, n^p, n^q)_+}$  and the induced map  $\psi_* : V(\Gamma) \rightarrow V(\Gamma)$ . Let  $t$  be the vertex  $b_{n^q}$  of  $\Gamma$ . By using a similar argument as in Step 2 of the proof of Theorem 4.10, one can show that the set of vertices  $\psi_*^{j \cdot n^q}(t)$  is contained in  $V(\Gamma) \setminus R$  for  $j = 1, \dots, \lfloor (n^q - 1)/(n^p + 1) \rfloor$ , where  $R = \{r_1, r_2, \dots, r_{n^p}\}$ . In other words, each vertex in  $R$  is not contained in  $\psi_*^{j \cdot n^q}(t)$  for such  $j$ . In particular, if we set  $D = D(n) = \lfloor (n^q - 1)/(n^p + 1) \rfloor$ , then  $r_1$  is not contained in  $\psi_*^{Dn^q}(t)$ . Then we consider the two arcs  $\beta_t$  and  $\alpha_v$  as in Step 3 of the proof of Theorem 4.10. By the same argument, it follows that  $\beta_t, \alpha_t$  and  $\psi^{Dn^q}(\beta_t)$  are distinct elements in  $\mathcal{AC}(S)$ . Moreover we have  $i(\psi^{Dn^q}(\beta_t), \alpha_v) = 0$  and  $i(\beta_t, \alpha_v) = 0$ . Therefore  $\beta_t$  and  $\psi^{Dn^q}(\beta_t)$  are at most distance 2 in  $\mathcal{AC}(S)$ , and we have  $\ell_{\mathcal{AC}}(\psi) \leq 2/(Dn^q)$ , which implies that  $\ell_{\mathcal{C}}(\psi) \leq 4/(Dn^q)$ . Since  $Dn^q \asymp n^{2q-p}$ , we have finished the proof. □

### 4.5 The behaviors of asymptotic translation lengths

We prove the following lemma which implies that the upper bound of Theorem 3.1 is optimal.

**Lemma 4.12** *The sequence  $(1, n, 1)_+$  of primitive integral classes in  $\mathcal{C}_F$  converges projectively to a point in  $\partial F$  as  $n \rightarrow \infty$ , and*

$$\ell_{\mathcal{C}}(\psi_{(1, n, 1)_+}) \asymp \frac{1}{|\chi(S_{(1, n, 1)_+})|}.$$

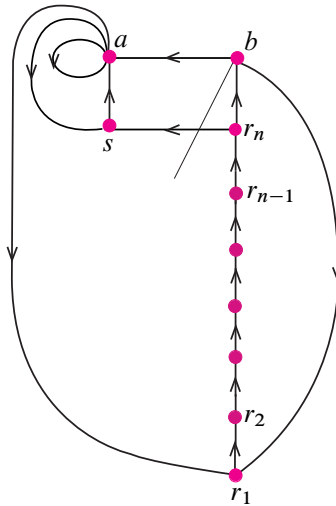


Figure 11: The digraph  $\Gamma_{(1,n,1)_+}$ .

**Proof** The first half of the claim follows from that fact that  $\overline{(1, n, 1)_+} \rightarrow (0, 1, 0) \in \partial F$  as  $n \rightarrow \infty$ . Since  $|\chi(S_{(1,n,1)_+})| = n + 3$ , it is enough to prove that  $\ell_C(\psi_{(1,n,1)_+}) \asymp 1/n$ . By the digraph  $\Gamma = \Gamma_{(1,n,1)_+}$  (see Figure 11) together with Proposition 4.1, it is not hard to see that  $\ell_C(\psi_{(1,n,1)_+}) \gtrsim 1/n$ .

Now we compute the upper bound. Let  $(S, \psi) = (S_{(1,n,1)_+}, \psi_{(1,n,1)_+})$  and let  $t$  be the vertex  $b$  of  $\Gamma$ . We have

$$\psi_*(t) = \{r_1\}, \psi_*^2(t) = \{r_2\}, \dots, \psi_*^n(t) = \{r_n\}.$$

In particular this implies that  $\psi^n(t)$  does not pass through the real branch  $r_1$  of  $\tau = \tau_{(1,n,1)_+}$ . We consider the essential arc  $\beta_t$  for  $t$  as in the proof of Theorem 4.10, and consider the dual arc  $\alpha_{r_1}$  of  $r_1$ . By the same argument as in the proof of Theorem 4.10, one sees that the three arcs  $\beta_t, \psi^n(\beta_t)$  and  $\alpha_{r_1}$  are distinct elements in  $\mathcal{AC}(S)$ . Furthermore for the geometric intersection numbers between arcs, we have  $i(\beta_t, \alpha_{r_1}) = 0$  and  $i(\psi^n(\beta_t), \alpha_{r_1}) = 0$ . Therefore  $\beta_t$  and  $\psi^n(\beta_t)$  are at most distance 2 in  $\mathcal{AC}(S)$ , and we have  $\ell_{\mathcal{AC}}(\psi) \leq 2/n$ , which gives the desired upper bound  $\ell_C(\psi) \leq 4/n$ .  $\square$

Now we are ready to prove the following theorem.

**Theorem 4.13** *Let  $F$  be a fibered face of the magic manifold. Then there exist two points,  $b_0 \in \partial F$  and  $c_0 \in \text{int}(F)$ , which satisfy the following:*

- (1) For any  $r \in \mathbb{Q} \cap [1, 2)$ , there exists a sequence  $(S_{\alpha_n}, \psi_{\alpha_n})$  of primitive integral classes in  $\mathcal{C}_F$  converging projectively to  $\mathfrak{b}_0$  as  $n \rightarrow \infty$  such that

$$\ell_C(\psi_{\alpha_n}) \asymp \frac{1}{|\chi(S_{\alpha_n})|^r}.$$

- (2) For any  $r \in \mathbb{Q} \cap [\frac{3}{2}, 2]$ , there exists a sequence  $(S_{\alpha_n}, \psi_{\alpha_n})$  of primitive integral classes in  $\mathcal{C}_F$  converging projectively to  $\mathfrak{c}_0$  as  $n \rightarrow \infty$  such that

$$\ell_C(\psi_{\alpha_n}) \asymp \frac{1}{|\chi(S_{\alpha_n})|^r}.$$

In particular, the upper bound in Theorem 1.1 is optimal when  $d = 2$ .

**Proof** Because of the symmetry of the Thurston norm ball  $B_N$ , it suffices to prove the theorem for the fibered face as we picked in Section 4.2. For (1), if  $1 < r < 2$ , let  $p$  and  $q$  be positive integers such that  $r = 2q/p$  with  $q < p < 2q$ . By Lemma 4.3, the sequence  $(1, n^p, n^q)_+$  converges projectively to  $(0, 1, 0) \in \partial F$ . By Theorems 4.9 and 4.10, we have  $\ell_C(\psi_{(1, n^p, n^q)_+}) \asymp 1/n^{2q}$ . Since we have  $\|(1, n^p, n^q)_+\| \asymp n^p$ , it follows that

$$\ell_C(\psi_{(1, n^p, n^q)_+}) \asymp \frac{1}{|\chi(S_{(1, n^p, n^q)_+})|^{2q/p}} = \frac{1}{|\chi(S_{(1, n^p, n^q)_+})|^r},$$

where  $r = 2q/p \in (1, 2)$ . If  $r = 1$ , it follows from Lemma 4.12.

For (2), if  $\frac{3}{2} \leq r < 2$ , let  $p$  and  $q$  be positive integers such that  $r = 2 - p/q$  with  $2p \leq q$ . By Lemma 4.3, the sequence  $(1, n^p, n^q)_+$  converges projectively to  $(\frac{1}{2}, \frac{1}{2}, 0) \in \text{int}(F)$  as  $n \rightarrow \infty$ . By Theorems 4.9 and 4.11, we have  $\ell_C(\psi_{\alpha_n}) \asymp 1/n^{2q-p}$ . Since we have  $\|(1, n^p, n^q)_+\| \asymp n^q$ , it follows that

$$\ell_C(\psi_{(1, n^p, n^q)_+}) \asymp \frac{1}{|\chi(S_{(1, n^p, n^q)_+})|^{2-p/q}} = \frac{1}{|\chi(S_{(1, n^p, n^q)_+})|^r},$$

where  $r = 2 - p/q \in [\frac{3}{2}, 2)$ . For  $r = 2$ , one can choose a sequence of primitive integral classes contained in the intersection between the cone over some compact set  $K \subset \text{int}(F)$  and some 2-dimensional rational subspace of  $H^1(M)$ , eg the sequence  $(1, n, n)_+$ . Then the sequence satisfies the desired property from [3, Corollary 1].

Finally we consider the upper bound in Theorem 1.1 when  $d = 2$ . If  $(p, q) = (1, 2)$ , then

$$\ell_C(\psi_{(1, n, n^2)_+}) \asymp \frac{1}{|\chi(S_{(1, n, n^2)_+})|^{1+1/2}}.$$

Then Theorem 1.1 implies that the sequence  $(1, n, n^2)_+$  of primitive integral classes can not be contained in any finite union of 2–dimensional rational subspaces of  $H^1(N)$ . The fibered cone  $\mathcal{C}_F$  is a  $(2+1)$ –dimensional rational subspace of  $H^1(N)$ . Thus Theorem 1.1 is optimal when  $d = 2$ .  $\square$

In light of Theorem 4.13(1), we ask the following question.

**Question 4.14** Let  $F$  be a fibered face of a compact hyperbolic fibered 3–manifold. Does there exist a sequence  $(S_{\alpha_n}, \psi_{\alpha_n})$  of primitive integral classes in  $\mathcal{C}_F$  converging projectively to  $\partial F$  as  $n \rightarrow \infty$  such that  $\ell_C(\psi_{\alpha_n}) \asymp 1/|\chi(S_{\alpha_n})|^2$ ?

By Theorem 4.13, we immediately have the following corollary.

**Corollary 4.15** Let  $F$  be a fibered face of the magic manifold  $N$ . For  $\alpha \in F \cap H^1(N; \mathbb{Q})$ , let  $(S_{\tilde{\alpha}}, \psi_{\tilde{\alpha}})$  be the fiber and pseudo-Anosov monodromy corresponding to the primitive integral class  $\tilde{\alpha}$  lying on the ray of  $\alpha$  passing through the origin. Then there is no normalization of the asymptotic translation length function

$$F \cap H^1(N; \mathbb{Q}) \rightarrow \mathbb{R}_{\geq 0}, \quad \alpha \mapsto \ell_C(\psi_{\tilde{\alpha}}),$$

in terms of the Euler characteristic  $\chi(S_{\tilde{\alpha}})$  which admits a continuous extension on  $F$ .

## References

- [1] **T Aougab, S J Taylor**, *Pseudo-Anosovs optimizing the ratio of Teichmüller to curve graph translation length*, from “In the tradition of Ahlfors–Bers, VII” (A S Basmajian, Y N Minsky, A W Reid, editors), Contemp. Math. 696, Amer. Math. Soc., Providence, RI (2017) 17–28 MR Zbl
- [2] **H Baik, H Shin**, *Minimal asymptotic translation lengths of Torelli groups and pure braid groups on the curve graph*, Int. Math. Res. Not. 2020 (2020) 9974–9987 MR Zbl
- [3] **H Baik, H Shin, C Wu**, *An upper bound on the asymptotic translation lengths on the curve graph and fibered faces*, Indiana Univ. Math. J. 70 (2021) 1625–1637 MR Zbl
- [4] **M Bestvina, M Handel**, *Train-tracks for surface homeomorphisms*, Topology 34 (1995) 109–140 MR Zbl
- [5] **B H Bowditch**, *Tight geodesics in the curve complex*, Invent. Math. 171 (2008) 281–300 MR Zbl
- [6] **D Calegari**, *Foliations and the geometry of 3–manifolds*, Oxford Univ. Press (2007) MR Zbl

- [7] **B Farb, C J Leininger, D Margalit**, *The lower central series and pseudo-Anosov dilatations*, Amer. J. Math. 130 (2008) 799–827 MR Zbl
- [8] **B Farb, D Margalit**, *A primer on mapping class groups*, Princeton Mathematical Series 49, Princeton Univ. Press (2012) MR Zbl
- [9] **A Fathi, F Laudenbach, V Poénaru** (editors), *Travaux de Thurston sur les surfaces*, Astérisque 66–67, Soc. Math. France, Paris (1979) MR Zbl
- [10] **D Fried**, *Flow equivalence, hyperbolic systems and a new zeta function for flows*, Comment. Math. Helv. 57 (1982) 237–259 MR Zbl
- [11] **D Fried**, *The geometry of cross sections to flows*, Topology 21 (1982) 353–371 MR Zbl
- [12] **V Gadre, E Hironaka, R P Kent, IV, C J Leininger**, *Lipschitz constants to curve complexes*, Math. Res. Lett. 20 (2013) 647–656 MR Zbl
- [13] **V Gadre, C-Y Tsai**, *Minimal pseudo-Anosov translation lengths on the complex of curves*, Geom. Topol. 15 (2011) 1297–1312 MR Zbl
- [14] **N V Ivanov**, *Fifteen problems about the mapping class groups*, from “Problems on mapping class groups and related topics” (B Farb, editor), Proc. Sympos. Pure Math. 74, Amer. Math. Soc., Providence, RI (2006) 71–80 MR Zbl
- [15] **E Kin**, *Dynamics of the monodromies of the fibrations on the magic 3-manifold*, New York J. Math. 21 (2015) 547–599 MR Zbl
- [16] **E Kin, H Shin**, *Small asymptotic translation lengths of pseudo-Anosov maps on the curve complex*, Groups Geom. Dyn. 13 (2019) 883–907 MR Zbl
- [17] **E Kin, M Takasawa**, *Pseudo-Anosov braids with small entropy and the magic 3-manifold*, Comm. Anal. Geom. 19 (2011) 705–758 MR Zbl
- [18] **M Korkmaz, A Papadopoulos**, *On the arc and curve complex of a surface*, Math. Proc. Cambridge Philos. Soc. 148 (2010) 473–483 MR Zbl
- [19] **M Landry**, *Stable loops and almost transverse surfaces*, Groups Geom. Dyn. 17 (2023) 35–75 MR Zbl
- [20] **J Lanier, D Margalit**, *Normal generators for mapping class groups are abundant*, preprint (2018) Available at <http://people.math.gatech.edu/~dmargalit7/papers/normal.pdf>
- [21] **DD Long**, *A note on the normal subgroups of mapping class groups*, Math. Proc. Cambridge Philos. Soc. 99 (1986) 79–87 MR Zbl
- [22] **DD Long, U Oertel**, *Hyperbolic surface bundles over the circle*, from “Progress in knot theory and related topics” (M Boileau, M Domergue, Y Mathieu, K Millett, editors), Travaux en Cours 56, Hermann, Paris (1997) 121–142 MR Zbl
- [23] **J Maher, G Tiozzo**, *Random walks, WPD actions, and the Cremona group*, Proc. Lond. Math. Soc. 123 (2021) 153–202 MR Zbl

- [24] **H A Masur, Y N Minsky**, *Geometry of the complex of curves, I: Hyperbolicity*, Invent. Math. 138 (1999) 103–149 MR Zbl
- [25] **H A Masur, Y N Minsky**, *Geometry of the complex of curves, II: Hierarchical structure*, Geom. Funct. Anal. 10 (2000) 902–974 MR Zbl
- [26] **S Matsumoto**, *Topological entropy and Thurston’s norm of atoroidal surface bundles over the circle*, J. Fac. Sci. Univ. Tokyo Sect. IA Math. 34 (1987) 763–778 MR Zbl
- [27] **C T McMullen**, *Polynomial invariants for fibered 3–manifolds and Teichmüller geodesics for foliations*, Ann. Sci. École Norm. Sup. 33 (2000) 519–560 MR Zbl
- [28] **L Mosher**, *Surfaces and branched surfaces transverse to pseudo-Anosov flows on 3–manifolds*, J. Differential Geom. 34 (1991) 1–36 MR Zbl
- [29] **A Papadopoulos, R C Penner**, *A characterization of pseudo-Anosov foliations*, Pacific J. Math. 130 (1987) 359–377 MR Zbl
- [30] **S Schleimer**, *Strongly irreducible surface automorphisms*, from “Topology and geometry of manifolds” (G Matić, C McCrory, editors), Proc. Sympos. Pure Math. 71, Amer. Math. Soc., Providence, RI (2003) 287–296 MR Zbl
- [31] **B Strenner**, *Fibrations of 3–manifolds and asymptotic translation length in the arc complex*, preprint (2018) arXiv 1810.07236
- [32] **W P Thurston**, *A norm for the homology of 3–manifolds*, Mem. Amer. Math. Soc. 339, Amer. Math. Soc., Providence, RI (1986) MR Zbl
- [33] **W P Thurston**, *Entropy in dimension one*, from “Frontiers in complex dynamics” (A Bonifant, M Lyubich, S Sutherland, editors), Princeton Math. Ser. 51, Princeton Univ. Press (2014) 339–384 MR Zbl
- [34] **A D Valdivia**, *Asymptotic translation length in the curve complex*, New York J. Math. 20 (2014) 989–999 MR Zbl
- [35] **A D Valdivia**, *Lipschitz constants to curve complexes for punctured surfaces*, Topology Appl. 216 (2017) 137–145 MR Zbl

*Department of Mathematical Sciences, KAIST  
Daejeon, South Korea*

*Center for Education in Liberal Arts and Sciences, Osaka University  
Osaka, Japan*

*Reinsurance Group of America  
Chesterfield, MO, United States*

*Department of Mathematics, University of Wisconsin at Madison  
Madison, WI, United States*

hrbaik@kaist.ac.kr, kin.eiko.celas@osaka-u.ac.jp, hshin153@gmail.com,  
cwu367@math.wisc.edu

Received: 28 March 2021      Revised: 10 August 2021

# Geometric triangulations and highly twisted links

SOPHIE L HAM

JESSICA S PURCELL

It is conjectured that every cusped hyperbolic 3–manifold admits a geometric triangulation, that is, it can be decomposed into positive volume ideal hyperbolic tetrahedra. We show that sufficiently highly twisted knots admit a geometric triangulation. In addition, by extending work of Guéritaud and Schleimer, we also give quantified versions of this result for infinite families of examples.

57K10, 57K31, 57K32, 57R05

1. Introduction	1399
2. Fully augmented links	1404
3. Layered solid tori	1412
4. Angle structures	1414
5. Dehn filling	1430
6. Borromean rings and related links	1433
7. Doubling layered solid tori	1436
8. Dehn filling the Borromean rings	1446
9. Fully augmented 2–bridge links	1451
References	1461

## 1 Introduction

A *topological triangulation* of a 3–manifold  $M$  is a decomposition of  $M$  into tetrahedra or ideal tetrahedra such that the result of gluing yields a manifold homeomorphic to  $M$ . Every compact 3–manifold with boundary consisting of tori has interior that admits a topological ideal triangulation; see Bing [4] and Moise [22].



Figure 1: Constructing a fully augmented link.

A *geometric triangulation* is a much stronger notion. It is an ideal triangulation of a cusped hyperbolic 3–manifold  $M$  such that each tetrahedron is positively oriented and has a hyperbolic structure of strictly positive volume, and such that the result of gluing gives  $M$  a smooth manifold structure with its complete hyperbolic metric. It is still unknown whether every finite volume hyperbolic 3–manifold admits a geometric triangulation, and there are currently only a few families which provably admit one. These include 2–bridge knots and punctured torus bundles (due to Guéritaud and Futer [15]), and all the manifolds of the SnapPy census (see Culler, Dunfield, and Weeks [6]), as well as manifolds built from isometric Platonic solids; see Goerner [13; 14]. On the other hand, Choi has shown that there exists an orbifold with no geometric triangulation [5].

We prove that a large family of knots admit geometric triangulations. To state the main result, we recall the following definitions.

A *twist region* of a link diagram consists of a portion of the diagram where two strands twist around each other maximally. More carefully, let  $D(K)$  be a diagram of a link  $K \subset S^3$ . Two distinct crossings of the diagram are *twist equivalent* if there exists a simple closed curve on the diagram that runs transversely through the two crossings, and is disjoint from the diagram elsewhere. The collection of all twist equivalent crossings forms a *twist region*.

Note that one can perform flypes on a link diagram until all twist equivalent crossings line up in a row, forming bigons between them. Suppose every simple closed curve that meets the diagram transversely only in two crossings has the property that it bounds a region of the diagram consisting only of bigons, or possibly contains no crossings. If this holds for every simple closed curve, the diagram is called *twist-reduced*. Figure 1, left, shows an example of a twist-reduced diagram with exactly two twist regions.

**Theorem 5.5** *For every  $n \geq 2$ , there exists a constant  $A_n$  depending on  $n$ , such that if  $K$  is a link in  $S^3$  with a prime, twist-reduced diagram with  $n$  twist regions, and at least  $A_n$  crossings in each twist region, then  $S^3 - K$  admits a geometric triangulation.*



The proof uses links called *fully augmented links*. These are obtained by starting with a twist-reduced diagram of any knot or link, and for each twist-region adding a simple unknot called a *crossing circle* encircling the twist-region. We further remove all pairs of crossings in each twist region; see Figure 1.

The result has explicit geometric properties, and can be subdivided into geometric tetrahedra. The original link complement is obtained by Dehn filling the crossing circles. We complete the proof of Theorem 5.5 by arguing that Dehn filling can be performed in a way that gives a geometric triangulation.

In fact, we can prove a result that is more general than Theorem 5.5, allowing any Dehn fillings on crossing circles and indeed leaving some crossing circles unfilled:

**Theorem 5.4** *Let  $L$  be a hyperbolic fully augmented link with  $n \geq 2$  crossing circles. Then there exist constants  $A_1, \dots, A_n$  such that if  $M$  is a manifold obtained by Dehn filling the crossing circle cusps of  $S^3 - L$  along slopes  $s_1, \dots, s_n$  whose lengths satisfy  $\text{len}(s_i) \geq A_i$  for each  $i = 1, \dots, n$ , then  $M$  admits a geometric triangulation. Allowing some collection of  $s_i = \infty$ , meaning leaving some crossing circle cusps unfilled, also admits a geometric triangulation.*

Guéritaud and Schleimer considered geometric triangulations and Dehn filling [16]. They showed that if a cusped manifold satisfies certain “genericity” conditions, then Dehn filling can be performed via geometric triangulation.<sup>1</sup> Unfortunately, the usual geometric decomposition of a fully augmented link, as in Agol and Thurston [2] or Futer and Purcell [11] and Purcell [27], fails Guéritaud and Schleimer’s genericity conditions. Nevertheless, we may adjust the decomposition to give a triangulation satisfying the Guéritaud–Schleimer conditions. This is the idea of the proof of Theorem 5.5.

Highly twisted knots, as in Theorem 5.5, are known to have other useful geometric properties. For example, they can be shown to be hyperbolic when there are at least six crossings in each twist region [11]. When there are at least seven, there are bounds on the volumes of such knots and links; see Futer, Kalfagianni, and Purcell [10]. With at least 116 crossings per twist region, there are bounds on their cusp geometry; see Purcell [25]. The results of Theorem 5.5 are not as nice as these other results, because we do not have an effective universal bound on the number of crossings per twist region required to guarantee that a knot admits a geometric triangulation. Nevertheless, we conjecture such a bound holds.

<sup>1</sup>In fact, they showed something stronger: that Dehn filling gives a triangulation that is actually *canonical*, ie dual to the Ford–Voronoi domain, but we will not consider canonical decompositions here.

To obtain effective results, we need to generalise and sharpen results of Guéritaud and Schleimer, and we do this in the second half of the paper. This allows us to present two effective results, which guarantee geometric triangulations of new infinite families of cusped hyperbolic 3-manifolds.

**Theorem 8.6** *Let  $L$  be a fully augmented link with exactly two crossing circles. Let  $M$  be a manifold obtained by Dehn filling the crossing circles of  $S^3 - L$  along slopes  $m_1, m_2 \in (\mathbb{Q} \cup \{1/0\}) - \{0, 1/0, \pm 1, \pm 2\}$ . Then  $M$  admits a geometric triangulation.*

There are three fully augmented links with exactly two crossing circles. One is the Borromean rings and the others are closely related; these are shown in Figure 13. The Dehn fillings of these links include double twist knots, which were already known to admit geometric triangulations by Guéritaud and Futer [15]. They also include large families of cusped hyperbolic manifolds that do not embed in  $S^3$ . More generally:

**Theorem 9.12** *Let  $L$  be a result of taking the standard diagram of a 2-bridge link, and then fully augmenting the link, such that  $L$  has  $n > 2$  crossing circles (and no half-twists). Let  $s_1, s_2, \dots, s_n \in \mathbb{Q} \cup \{1/0\}$  be slopes, one for each crossing circle, that are all positive or all negative. Suppose finally that  $s_1$  and  $s_n$  are the slopes on the crossing circles on either end of the diagram, and the slopes satisfy*

$$s_1, s_n \notin \{0/1, 1/0, \pm 1/1, \pm 2/1\} \quad \text{and} \quad s_2, \dots, s_{n-1} \notin \{0/1, 1/0, \pm 1/1\}.$$

*Then the manifold obtained by Dehn filling  $S^3 - L$  along these slopes on its crossing circles admits a geometric triangulation.*

For ease of notation, we will refer to a link such as  $L$  in the above theorem as a *fully augmented 2-bridge link*. That is, a fully augmented 2-bridge link is obtained by fully augmenting the standard diagram of a 2-bridge link.

The Borromean rings and other links of Theorem 8.6 form examples of fully augmented 2-bridge links, and therefore instances of Theorem 8.6 also follow from Theorem 9.12. However, we prove the theorems separately to build up tools.

The manifolds included in Theorem 9.12 include many 2-bridge links, obtained by setting each  $s_j = 1/m_j$  where  $m_j$  is an integer with appropriate sign. Such a Dehn filling gives a 2-bridge link with a diagram with at least two crossings per twist region, an even number of crossings in each twist region, and conditions on signs of twisting. All 2-bridge links were already known to admit geometric triangulations [15]. However, again Theorem 9.12 also includes infinitely many additional manifolds obtained by different Dehn fillings.

## 1.1 More on geometric triangulations

It is known that every cusped hyperbolic 3–manifold has a decomposition into convex ideal polyhedra, due to work of Epstein and Penner [7]. The convex polyhedra may be further subdivided into tetrahedra, but the result may not give a geometric triangulation. The difficulty is that the subdivision involves triangulating the polygonal faces of the polyhedra, and these triangulations may not be consistent with each other under gluing. To solve this problem, flat tetrahedra are inserted between identified faces of the polyhedra; see Petronio and Porti for more discussion [24].

If we pass to finite covers, then geometric triangulations exist by work of Luo, Schleimer, and Tillmann [21]: every cusped hyperbolic 3–manifold admits a finite cover with a geometric triangulation.

If we relax the restriction that the tetrahedra glue to give a complete hyperbolic metric, and only require that the dihedral angles of each tetrahedron are strictly positive and sum to  $2\pi$  around each edge of the triangulation, then the result is called an *angle structure* (or sometimes a *strict angle structure*). Geometric triangulations admit angle structures. Moreover, Hodgson, Rubinstein, and Segerman show that many 3–manifolds admit an angle structure, including all hyperbolic link complements in  $S^3$  [18]. However, they note that the triangulations they find are not generally geometric.

There was some hope in the past that a class of triangulations introduced by Agol [1], called *veering triangulations*, give geometric triangulations. Indeed it was shown by Hodgson, Rubinstein, Segerman, and Tillmann [19] and by Futer and Guéritaud [9] that veering triangulations admit angle structures. However Hodgson, Issa, and Segerman found a 13–tetrahedron veering triangulation that is not geometric [17], and recently Futer, Taylor, and Worden showed that a random veering triangulation is not geometric [12]. Thus, tools to exhibit geometric triangulations must come from other directions.

Why geometric triangulations? Various results become easier with geometric triangulations. For example, Neumann and Zagier showed that certain useful bounds exist on the volume of a hyperbolic 3–manifold that admits a geometric triangulation [23], although this can be proven in general with more work; see Petronio and Porti [24]. Similarly, Benedetti and Petronio give a straightforward proof of Thurston’s hyperbolic Dehn surgery theorem using geometric triangulations [3]. Choi finds nice conditions on the deformation variety for manifolds admitting geometric triangulations [5]. In summary, such triangulations seem to lead to simpler proofs, and more manageable geometry.

## 1.2 Organisation

The paper is organised as follows. Sections 2 through 5 give the proof of Theorem 5.5 on more general highly twisted knots. We recall fully augmented links in Section 2, layered solid tori in Sections 3 and 4, and put this together with Dehn filling in Section 5.

Sections 6 through 8 give the proof of Theorem 8.6, on Dehn fillings of links with two crossing circles, and build up the new machinery required for both Theorem 8.6 and Theorem 9.12.

Finally, in Section 9, we complete the proof of Theorem 9.12, on Dehn fillings of fully augmented 2–bridge links.

**Acknowledgements** We thank B Nimershiem for helping us to improve the exposition in Section 4. We also thank the referees for their comments, which helped us improve the paper. Both authors were supported in part by the Australian Research Council.

## 2 Fully augmented links

The links of the main theorem, Theorem 5.5, are obtained by Dehn filling a parent link, called a fully augmented link. In this section, we review fully augmented links and their geometry, and show that they admit geometric triangulations.

Begin with any twist-reduced diagram of a link. As in Figure 1, middle, for each set of twist equivalent crossings, insert a single unknotted curve that encircles the bigons of the twist region. If a twist region consists of only a single crossing, there are two ways to insert this link component; either will do. These unknotted components are chosen to be disjoint, and to bound discs that are punctured by exactly two strands of the original link. We call them *crossing circles*. A *fully augmented link* is a link obtained by adding a single crossing circle to every twist region of a twist-reduced diagram, and then removing all crossings that bound a bigon. That is, crossings are removed in pairs. The resulting diagram consists of crossing circles that are perpendicular to the plane of projection and strands that lie on the plane of projection except possibly for single crossings in the neighbourhood of a crossing circle; see Figure 1, right.

Agol and Thurston studied the geometry of fully augmented links using a decomposition into ideal polyhedra [2]. In particular, they show that every fully augmented link admits a decomposition into two identical totally geodesic polyhedra that determine a circle packing on  $\mathbb{C}$ . The result we need is the following:

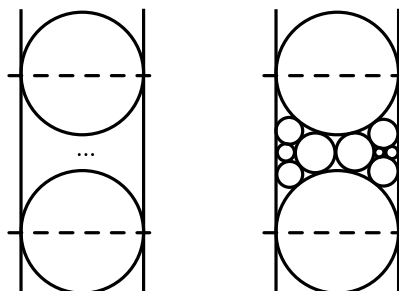


Figure 2: Left: the general form of a circle packing determining a polyhedron, with a vertex that projects to a crossing circle cusp at the point at infinity. The dashed lines show two parallel shaded faces meeting at infinity. The region between the circles and lines will be filled with a circle packing. Note, white and shaded faces through infinity cut out a rectangle. Right: a specific example.

**Proposition 2.1** *A fully augmented link decomposes into the union of two identical ideal polyhedra with the following properties.*

- (1) *Each polyhedron is convex, right-angled, with a checkerboard colouring of its faces, shaded and white. The shaded faces are all ideal triangles, each a subset of a 2-punctured disc bounded by a crossing circle.*
- (2) *Each polyhedron is determined by a circle packing on  $\mathbb{R}^2 \cup \{\infty\}$ , with white faces lifting to planes in  $\mathbb{H}^3$  whose boundaries are given by the circles. Shaded faces lift to planes with boundaries given by the dual circle packing.*
- (3) *Embed the ideal polyhedron in  $\mathbb{H}^3$  as a convex right-angled polyhedron. Each ideal vertex projects to a link component, or more precisely, the boundary of a sufficiently small horoball neighbourhood of an ideal vertex projects to a subset of a horospherical torus about a link component.*

*Apply an isometry so that an ideal vertex corresponding to a crossing circle lies at the point at infinity in the upper half space model of  $\mathbb{H}^3$ . Then two white faces form parallel vertical planes meeting the point at infinity, with two shaded faces forming perpendicular parallel vertical planes, cutting out a rectangle. Two other white faces, defining circles tangent to the white parallel vertical planes, meet the shaded parallel vertical planes at right angles. This forms a rectangle with two circles; see Figure 2.*

*Moreover, none of the four ideal vertices on  $\mathbb{R}^2$  corresponding to the corners of the rectangle project to crossing circles.*

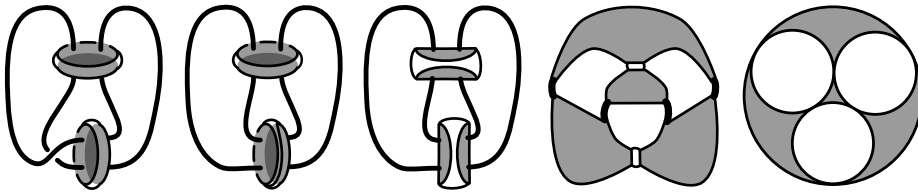


Figure 3: Left to right: Slice along shaded faces bounded by 2-punctured discs and unwind single crossings. Then slice along the white plane of projection. Shrinking remnants of the link to ideal vertices gives the ideal polyhedron. The circle packing is obtained by lifting the polyhedron to  $\mathbb{H}^3$ , and taking boundaries of white faces.

Proofs of Proposition 2.1 can be found in various places, including [11; 27]. We review the details briefly as they are important to our argument.

**Proof of Proposition 2.1** There are two types of totally geodesic surface in the complement of a fully augmented link, and these will form the white and shaded faces of the polyhedra. The first totally geodesic surface comes from embedded 2-punctured discs bounded by crossing circles; colour each of these shaded.

The second comes from a surface related to the plane of projection. If the fully augmented link has no crossings on the plane of projection, then it is preserved by a reflection in the plane of projection and the white surface is the plane of projection. This reflection is realised by an isometry fixing the white surface pointwise, so the white surface is totally geodesic. If the link admits single crossings adjacent to crossing circles, reflection in the plane of projection will change the direction of each crossing. However, a full twist about the adjacent crossing circle is a homeomorphism of the complement, and it returns the link to its original position. The combination of reflection followed by twists is an isometry fixing a surface pointwise; this is the white surface. Again it is totally geodesic.

To obtain the decomposition into ideal polyhedra, first cut along each shaded 2-punctured disc. Near single crossings, rotate one copy of the 2-punctured disc by  $180^\circ$  to remove the crossing from the diagram. The white face then lies on the projection plane. Slice along the projection plane. This process is shown in Figure 3.

The result is two ideal polyhedra. We now show that these satisfy the properties stated. First, the checkerboard colouring is as claimed, by construction. The involution described above is the reflection through white faces taking one polyhedron to another.

Also, note that under the involution, shaded 2-punctured discs are taken to their reflections in the projection plane, hence must still be geodesic. It follows that they are perpendicular to white faces, and so the polyhedron is right-angled.

The circle packing comes from the totally geodesic white faces. These faces are all disjoint, and correspond to regions of the plane of projection. They lift to a collection of geodesic planes in  $\mathbb{H}^3$ , whose boundaries form a collection of circles that are tangent exactly when two white faces are adjacent across a strand of the link, or meet a common crossing circle. The shaded faces lift to ideal triangles, dual to the white circles. Thus this corresponds to a circle packing by shaded circles dual to a circle packing of white circles. The intersections of the exteriors of planes in  $\mathbb{H}^3$  defined by the circles gives a convex region with all right-angled dihedral angles; this is the geometric structure on the ideal polyhedron.

The fact that the cusp is as claimed follows from the fact that each ideal vertex of the polyhedron is 4-valent, so moving one to infinity gives a rectangle, and each shaded face is an ideal triangle, so beneath a vertical shaded face lies a single white circle. The shaded ideal triangle is obtained by slicing a 2-punctured disc through the projection plane. One vertex corresponds to an arc of the crossing circle above (or below) the projection plane, and the other two vertices correspond to strands of the link running through the crossing circle. Thus, exactly one of the ideal vertices of the shaded triangle corresponds to a crossing circle. Because the ideal vertex corresponding to the crossing circle lies at infinity, the other two ideal vertices, lying at points of intersection of vertical shaded and white planes, must correspond to link strands on the plane of projection. These are other vertices of the rectangle on  $\mathbb{R}^2$ .  $\square$

We will show that fully augmented links admit a geometric triangulation coming from the decomposition into polyhedra of Proposition 2.1. To do so, we will show that appropriate neighbourhoods of crossing circles can be triangulated separately.

Consider a polyhedron from the decomposition of a fully augmented link, as in Proposition 2.1. Arrange the polyhedron in  $\mathbb{H}^3$  so that the point at infinity projects to a crossing circle cusp, with vertical planes cutting out a rectangle on  $\partial\mathbb{H}^3$ . Then there is a unique circle on  $\partial\mathbb{H}^3$  meeting each vertex of the rectangle. It intersects exactly four of the circles in the circle packing corresponding to white faces of the polyhedron; see Figure 4. The circle is the boundary of a geodesic plane in  $\mathbb{H}^3$ . The intersection of the plane with the polyhedron determines a totally geodesic rectangular surface.

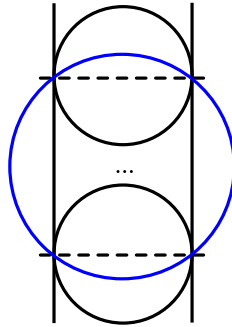


Figure 4: For each ideal vertex corresponding to a crossing circle, there exists a circle running through all four vertices of the rectangle of Proposition 2.1(3). Each of these defines a plane in  $\mathbb{H}^3$ , and their intersection with the polyhedron defines a rectangle  $R$ .

**Lemma 2.2** *Let  $R_1, \dots, R_n$  denote the totally geodesic rectangular surfaces arising as above, one for each crossing circle vertex. Then either the interiors of the rectangles are pairwise disjoint, or there exist exactly two crossing circles, each ideal polyhedron is a regular ideal octahedron, and the rectangle  $R_1 = R_2$  cuts each polyhedron into two square prisms.*

**Proof** Let  $C$  denote the circle running through the four vertices of the rectangle of a crossing circle cusp. Consider the intersection of this circle  $C$  and the circles corresponding to white faces of the polyhedron. The circle  $C$  intersects the two vertical planes that form two of the four edges of the rectangle; this gives two intersections. Additionally, the circle  $C$  intersects the two hemispheres bounded by circles meeting the vertical planes, as at the top and bottom of Figure 4.

The fact that  $C$  cannot meet any other white faces of the polyhedron now follows from the fact that it encloses the region on  $\partial\mathbb{H}^3$  bounded by the parallel vertical planes and by the two white circles tangent to them. All other white circles are completely contained in this region. Therefore, the hemispheres they determine cannot intersect  $C$ .

Now we consider the intersections of two rectangular surfaces  $R_1$  and  $R_2$  arising from circles  $C_1$  and  $C_2$  from different crossing circles. Arrange the polyhedron so that the crossing circle vertex corresponding to  $R_1$  lies at infinity, and  $R_1$  determines a circle  $C_1$  running through four vertices of a rectangle. The rectangular surface  $R_2$  lies on a geodesic hemisphere  $H_2$  whose boundary on  $\partial\mathbb{H}^3$  is a circle  $C_2$ . By the argument above,  $C_2$  meets exactly four of the white circles in the circle packing; these



intersections cut  $C_2$  into four circular arcs. Note that the endpoints of each arc are ideal vertices of the polyhedron.

If the surfaces  $R_1$  and  $R_2$  intersect, then the circles  $C_1$  and  $C_2$  intersect. Because the circular arcs of  $C_2$  lie inside white circles of the circle packing, this is possible only if a point of intersection of  $C_1$  and  $C_2$  occurs within one of the circles of the circle packing. Because  $C_1$  only meets the two parallel sides of the rectangle and two circles tangent to them, the circles where  $C_1$  and  $C_2$  intersect must be among these circles.

Next note that  $C_1$  and  $C_2$  must intersect twice within the same circle of the circle packing, else the ideal vertices met by circular arcs of  $C_1$  and  $C_2$  interleave on a circle. However,  $C_1$  runs through the outermost ideal vertices on each of the four circles under consideration: there are no additional ideal vertices outside the rectangle for  $C_2$  to meet for interleaving.

Finally, the intersection points either lie on the vertices of the rectangle defining  $R_1$ , or lie outside that rectangle, because  $C_1$  lies outside that rectangle. But only points inside the rectangle lie inside the polyhedron, so intersections outside the rectangle cannot give intersections of  $R_1$  and  $R_2$ , which both are embedded in the polyhedron.

The only remaining possibility is that  $C_1$  and  $C_2$  both run through at least two of the same ideal vertices on the rectangle defining  $R_1$ . If exactly two,  $R_1$  and  $R_2$  share an edge, but have disjoint interiors as claimed. If more than two, then they must share all four ideal vertices, and  $R_1$  and  $R_2$  agree. In this case, the polyhedron is determined: it must be a regular ideal octahedron with  $R_1 = R_2$  cutting off an ideal vertex corresponding to a crossing circle on either side. The fully augmented link can only have two crossings circles, corresponding to the ideal vertices used to define  $R_1$  and  $R_2$ . The rectangles  $R_1 = R_2$  cut the octahedron into two pyramids over a square base.  $\square$

**Proposition 2.3** *Every fully augmented link admits a geometric triangulation with the following properties:*

- (1) *Each crossing circle meets exactly four tetrahedra, two in each polyhedron.*
- (2) *The triangulation is symmetric across the white faces. That is, a reflection across white faces preserves the triangulation.*

**Proof** Begin with the ideal polyhedral decomposition of Proposition 2.1. For either one of the two symmetric ideal polyhedra, cut off each ideal vertex corresponding to a

crossing circle by cutting along the rectangles of Lemma 2.2. This splits the polyhedron into  $n$  pyramids over a rectangular base corresponding to crossing circles, where  $n$  is the number of crossing circles, and one remaining convex ideal polyhedron  $P$ . In the case that there are just two crossing circles, the remaining convex ideal polyhedron is degenerate: it is just the rectangle  $R_1 = R_2$ . In all other cases, it has 3-dimensional interior.

Split each rectangular pyramid into two geometric tetrahedra by choosing a diagonal of the rectangle and cutting along it.

When we reglue into the fully augmented link, the choices of diagonals on the rectangles  $R_1, \dots, R_n$  are mapped to ideal edges on the convex polyhedron  $P$ . When  $R_1 = R_2$ , and  $P$  is degenerate, choosing the same diagonal gives the desired triangulation.

When  $P$  is nondegenerate, we triangulate it by coning. Choose any ideal vertex  $w$  of  $P$ . For any face not containing  $w$  that is not already an ideal triangle, subdivide the face into ideal triangles in any way by adding ideal edges. Then take cones from  $w$  over all the triangles in faces disjoint from  $w$ . Because  $P$  is convex, the result is a division of  $P$  into geometric ideal tetrahedra.

Transfer the triangulation on the first polyhedron to the second by reflection in the white surface. This gives both polyhedra exactly the same subdivision, up to reflection.

Now note that the polyhedra glue by reflection in the white faces, so no new flat tetrahedra need to be introduced in these faces to obtain the gluing. All shaded faces are ideal triangles, which are glued by isometry and again no flat tetrahedra need to be introduced. Thus the decomposition of the polyhedra into geometric tetrahedra described above gives a decomposition of the fully augmented link complement into geometric tetrahedra.

Finally, the fact that the geometric triangulation satisfies the two properties of the theorem follows by construction: each crossing circle meets two rectangles, hence four tetrahedra, and the triangulation is preserved by the reflection in the white faces.  $\square$

The previous lemma gives a geometric triangulation of a fully augmented link, but four tetrahedra meet each crossing circle. We need a triangulation for which only two tetrahedra meet each crossing circle:

**Proposition 2.4** *Every fully augmented link admits a geometric triangulation with the property that each crossing circle meets exactly two tetrahedra.*

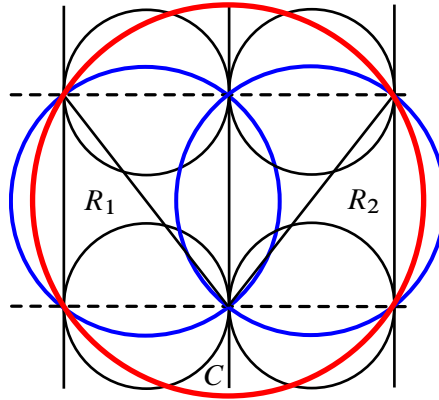


Figure 5: Glue both polyhedra in the decomposition of a fully augmented link along a white face corresponding to a vertical plane in  $\mathbb{H}^3$ . Two rectangles,  $R_1$  and  $R_2$ , are glued as shown. A larger circle  $C$  runs through vertices of both  $R_1$  and  $R_2$ .

**Proof** Begin with the geometric triangulation of Proposition 2.3 and consider the crossing circles. These are triangulated by exactly four tetrahedra, two in each polyhedron. For each of the two tetrahedra in one polyhedron, one face lies on a totally geodesic ideal rectangle coming from Lemma 2.2 embedded in  $\mathbb{H}^3$ . Call the two rectangles, one for each polyhedron,  $R_1$  and  $R_2$ .

Adjust the geometric triangulation as follows. At an ideal vertex corresponding to a crossing circle, two polyhedra are glued along a white face. The result of gluing both polyhedra together along such a face is shown in Figure 5. Note, rectangles  $R_1$  and  $R_2$  are glued along an edge. The boundary of the two glued polyhedra forms an even larger rectangle, with boundary the outermost parallel lines in Figure 5 and the dashed lines.

There is a circle  $C$  running through each vertex of that rectangle, shown in red in Figure 5. Note that the hemisphere defined by  $C$  in  $\mathbb{H}^3$  meets an edge of  $R_1$  and an edge of  $R_2$ . Cutting along the hemisphere cuts the two polyhedra along a rectangle  $R$ . The region bounded by  $R_1$ ,  $R_2$ ,  $R$  and the two polyhedra is a solid with three ideal quadrilateral faces and two ideal triangle faces; it forms a prism over an ideal triangle. The region between  $R$  and infinity forms a neighbourhood of the crossing circle vertex. Triangulate it by adding an edge along a diagonal of  $R$  and then coning to infinity. This gives two geometric ideal tetrahedra meeting the crossing circle vertex. These are the only tetrahedra meeting this vertex.

It remains to triangulate the prism over the ideal triangle bounded by  $R_1$ ,  $R_2$ , and  $R$ . The rectangles  $R_1$ ,  $R_2$ , and  $R$  all have been triangulated by a choice of diagonal; the one on  $R$  comes from the paragraph just above, and those on  $R_1$  and  $R_2$  come from Proposition 2.3. These three edges determine two ideal triangles whose interiors are disjoint in the interior of the triangular prism. They divide the prism into three geometric tetrahedra.  $\square$

### 3 Layered solid tori

To obtain highly twisted links, we will be performing Dehn filling on the crossing circle cusps of fully augmented links, using the triangulation of Proposition 2.4. We need a triangulation of the solid torus used in the Dehn filling. The triangulation that will work in this setting is a *layered solid torus*, first described by Jaco and Rubinstein [20]. In this section, we will review the construction of layered solid tori, and how they can be used to triangulate a Dehn filling of a triangulated manifold such as a fully augmented link.

The boundary of a layered solid torus consists of two ideal triangles whose union is a triangulation of a punctured torus. The space of all such triangulations of punctured tori is described by the Farey graph. Guéritaud and Schleimer present a description of the layered solid torus using the combinatorics of the Farey graph [16], and then glue this into the boundary of a manifold to be Dehn filled. We will follow their presentation.

#### 3.1 Review of layered solid tori

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

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

Here  $i(\cdot, \cdot)$  denotes geometric intersection number of slopes on the torus. We draw an ideal geodesic between each pair  $a/b$  and  $c/d$  with  $|ad - bc| = 1$ . This gives the *Farey triangulation*.

Any triangulation of a once-punctured torus consists of three slopes on the boundary of the torus, with each pair of slopes having geometric intersection number 1. Denote the

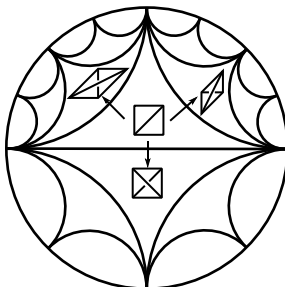


Figure 6: Each triangle in the Farey graph determines a triangulation of a punctured torus. Moving across an edge replaces one of the three slopes of the triangulation by a different slope.

slopes by  $p$ ,  $q$ , and  $r$ . Note that this triple determines a triangle in the Farey triangulation. Moving across an edge of the Farey triangulation changes the triangulation by replacing one slope with another, say  $r'$  replaces  $r$ ; see Figure 6.

In the case that we wish to perform a Dehn filling by attaching a solid torus to a triangulated once-punctured torus, there are four important slopes involved. Three of the slopes are the slopes of the initial triangulation of the once-punctured solid torus. In our setting, these will typically either be  $\{0/1, 1/0, 1/1\}$  or  $\{0/1, 1/0, -1/1\}$ . They form an *initial triangle* in the Farey graph. The last slope is  $m$ , the slope of the Dehn filling.

Now consider the geodesic in  $\mathbb{H}^2$  from the centre of the initial triangle to the slope  $m \subset \partial\mathbb{H}^2$ . This passes through a sequence of triangles in the Farey graph by crossing edges of the Farey triangulation. In particular, there will be a finite sequence of triangles, each determined by three slopes,

$$(T_0, T_1, \dots, T_N) = (pqr, pqr', \dots, stm),$$

with initial triangle  $T_0$  and final triangle  $T_N$  such that  $m$  is not a slope of any previous triangle in the sequence. For our purposes, we will require that  $N \geq 2$ . Thus we do not allow  $m$  to be a slope of the initial triangle  $T_0$  nor a slope of the three triangles adjacent to  $T_0$ .

We build a layered solid torus by stacking a tetrahedron onto a once-punctured torus, initially triangulated by the slopes of  $T_0$ , and replacing one slope with another at each step as we stack. That is, two consecutive once-punctured tori always have two slopes in common and two that differ by a diagonal exchange. The diagonal exchange is

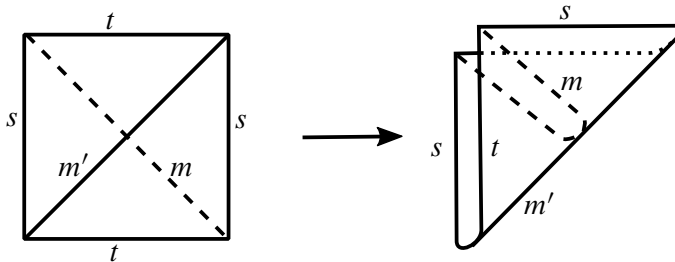


Figure 7: Folding  $m$  makes it homotopically trivial.

obtained in three dimensions by layering a tetrahedron onto a given punctured torus such that the diagonal on one side matches the diagonal to be replaced. In Figure 6, note that the diagonal exchanges have been drawn in such a way to indicate the tetrahedra.

For each triangle in the path from  $T_0$  to  $T_{N-1}$ , layer on a tetrahedron, obtaining a collection of tetrahedra homotopy equivalent to  $T^2 \times [0, 1]$ . At the  $k^{\text{th}}$  step, the boundary component  $T^2 \times \{0\}$  has the triangulation of  $T_0$  and  $T^2 \times \{1\}$  has the triangulation of  $T_k$ . Continue until  $k = N - 1$ , obtaining a triangulated complex with boundary consisting of two once-punctured tori, one triangulated by  $T_0$  and the other by  $T_{N-1}$ . Recall that  $m$  is a slope of  $T_N$  — notice that we are not adding on a tetrahedron corresponding to  $T_N$ .

If we stop at  $T_{N-1}$  (not  $T_N$ ), then one further diagonal exchange will give the slope  $m$ . That is,  $m$  is not one of the slopes of the triangulation of  $T_{N-1}$ , but a single diagonal exchange replaces the triangulation  $T_{N-1}$  with  $T_N$ , which is a triangulation consisting of two slopes  $s$  and  $t$  in common with  $T_{N-1}$  and the slope  $m$  cutting across a slope  $m'$  of  $T_{N-1}$ .

Recall, we are trying to obtain a triangulation of a solid torus for which the slope  $m$  is homotopically trivial. To homotopically kill the slope  $m$ , fold the two triangles of  $T_{N-1}$  across the diagonal slope  $m'$ . Gluing the two triangles on one boundary component of  $T^2 \times I$  in this manner gives a quotient homeomorphic to a solid torus, with boundary still triangulated by  $T_0$ . Inside, the slopes  $t$  and  $s$  are identified. The slope  $m$  has been folded onto itself, meaning it is now homotopically trivial; see Figure 7.

## 4 Angle structures

In order to prove that the triangulations we construct are geometric, we will use tools from the theory of angle structures on 3-manifolds. (These are also often called *strict*

angle structures in the literature.) We are following the lead of Guéritaud and Schleimer in [16], who use angle structures to show that layered solid tori admit geometric triangulations. The results we need are only slight generalisations of Guéritaud and Schleimer’s work, and the proofs follow almost immediately. However, we believe it is useful to step through the results and many of the proofs here as well. Not only does that make this paper more self-contained, but it also sets up a number of tools that we will need later in the paper when we further generalise to different triangulations of solid tori. Thus in this section we review angle structures, relevant results such as the Casson–Rivin theorem, and we work through the proof that layered solid tori admit geometric triangulations using angle structures.

**Definition 4.1** An *angle structure* on an ideal triangulation  $\tau$  of a 3–manifold  $M$  (possibly with boundary) is an assignment of dihedral angles on each tetrahedron such that opposite edges of the tetrahedron carry the same angle, and such that

- (1) all angles lie in the range  $(0, \pi)$ ,
- (2) around each ideal vertex of a tetrahedron, the dihedral angles sum to  $\pi$ ,
- (3) around each edge in the interior of  $M$ , the dihedral angles sum to  $2\pi$ .

The set of all angle structures for the triangulation  $\tau$  is denoted by  $\mathcal{A}(\tau)$ .

An angle structure on an ideal tetrahedron uniquely determines a hyperbolic structure on that tetrahedron. However, an angle structure on a triangulated 3–manifold is not as restrictive as a geometric triangulation. While one can assemble a space from hyperbolic triangles determined by the angles, under the gluing there may be shearing along edges. Thus the structure does not necessarily give a hyperbolic structure on  $M$ .

However, an angle structure determines a volume by summing the volumes of the hyperbolic ideal tetrahedra with the dihedral angles given by the angle structures. That is, recall that a hyperbolic ideal tetrahedron with dihedral angles  $\alpha$ ,  $\beta$ , and  $\gamma$  has volume  $\Lambda(\alpha) + \Lambda(\beta) + \Lambda(\gamma)$ , where  $\Lambda$  is the Lobachevsky function. Define the volume functional  $\mathcal{V}: \mathcal{A}(\tau) \rightarrow \mathbb{R}$  as follows. For  $p \in \mathcal{A}(\tau) \subset \mathbb{R}^{3n}$ , assign to the angle structure  $p = (p_1, p_2, p_3, \dots, p_{3n})$  the real number

$$\mathcal{V}(p) = \Lambda(p_1) + \Lambda(p_2) + \Lambda(p_3) + \dots + \Lambda(p_{3n}).$$

The volume functional is a convex function on  $\mathcal{A}(\tau)$ . That means it either takes its maximum on the interior of the space  $\mathcal{A}(\tau)$ , or there is no maximum in  $\mathcal{A}$ , and  $\mathcal{V}$  is maximised on the boundary of the closure  $\overline{\mathcal{A}(\tau)}$ ; see for example [8].

The following theorem, proved independently by Casson and Rivin, will allow us to use angle structures to obtain a geometric triangulation in the case that the maximum occurs in the interior of the space  $\mathcal{A}(\tau)$ .

**Theorem 4.2** (Casson and Rivin) *Let  $M$  be an orientable 3–manifold with boundary consisting of tori, and let  $\tau$  be an ideal triangulation of  $M$ . Then a point  $p \in \mathcal{A}(\tau)$  corresponds to a complete hyperbolic metric on the interior of  $M$  if and only if  $p$  maximises the volume functional  $\mathcal{V}: \mathcal{A}(\tau) \rightarrow \mathbb{R}$ .*

The proof of Theorem 4.2 follows from work in [28]. A different proof that includes a nice exposition is given by Futer and Guéritaud [8].

#### 4.1 Angle structures on layered solid tori

This subsection is devoted to the following proposition and its proof, which guarantees an angle structure on a layered solid torus. The result is essentially [16, Proposition 10], and the proof is very similar, but our statement is slightly more general. Additionally, parts of the proof will be needed in a later section, so we include the full argument.

**Proposition 4.3** *Let  $p, q$ , and  $r$  be slopes on the torus that bound a triangle in the Farey graph in  $\mathbb{H}^2$ . Let  $m$  be a slope separated from the triangle  $(p, q, r)$  by at least one triangle; that is, the geodesic  $\gamma$  in  $\mathbb{H}^2$  from the centre of triangle  $(p, q, r)$  to  $m$  intersects at least three triangles (one containing  $m$ , one containing  $(p, q, r)$ , and at least one more). Relabel  $p, q$ , and  $r$  if necessary so that the geodesic  $\gamma$  exits the triangle  $(p, q, r)$  by crossing the edge  $(p, q)$ , and exits the next triangle  $(p, q, r')$  by crossing the edge  $(q, r')$ . (Thus  $r$  is the first slope to disappear from the triangulation, and  $p$  is the second.) Assign to  $p, q$ , and  $r$  **exterior** dihedral angles  $\theta_p, \theta_q$ , and  $\theta_r$ , respectively, satisfying*

$$(4.4) \quad \theta_p + \theta_q + \theta_r = \pi, \quad -\pi < \theta_p, \theta_q < \pi, \quad \text{and} \quad 0 < \theta_r < \pi.$$

Finally, consider the layered solid torus  $T$  with boundary  $\partial T$  a punctured torus triangulated with slopes  $p, q$ , and  $r$ , and meridian the slope  $m$ . Set interior dihedral angles at edges of slope  $p, q$ , and  $r$  of  $\partial T$  equal to  $\pi - \theta_p, \pi - \theta_q$ , and  $\pi - \theta_r$ , respectively.

There exists an angle structure on  $T$  with the given interior dihedral angles if and only if

$$(4.5) \quad i(m, p)\theta_p + i(m, q)\theta_q + i(m, r)\theta_r > 2\pi,$$

where  $i(a, b)$  denotes geometric intersection number.



**Remark 4.6** Guéritaud and Schleimer actually require  $0 \leq \theta_p, \theta_q$  in the statement of their proposition. However, most of their argument applies equally well if one of  $\theta_p$  or  $\theta_q$ , say  $\theta_p$ , is negative, provided  $\pi - \theta_p < 2\pi$ . The other conditions will then imply that  $\theta_q$  is positive, and that  $\theta_p + \theta_r$  is positive, and these conditions suffice to prove the proposition.

We start by setting up notation. Let  $T$  be a layered solid torus constructed by following a geodesic  $\gamma$  in the Farey graph in  $\mathbb{H}^2$  from the centre of triangle  $(p, q, r)$  to the slope  $m$ , with  $\gamma$  intersecting at least three triangles. Let  $(T_0, T_1, \dots, T_{N-1}, T_N)$  denote the sequence of triangles. As in the statement of Proposition 4.3, we label so that  $r$  is the first slope to be replaced by diagonal exchange and  $p$  is the second.

Let  $\Delta_1, \dots, \Delta_{N-1}$  denote the tetrahedra in  $T$ , constructed as in Section 3. Thus  $\Delta_1$  meets the boundary of the layered solid torus in slopes  $(p, q, r)$ , and the tetrahedron  $\Delta_{N-1}$  is folded on itself to form the solid torus with  $m$  homotopically trivial.

**Lemma 4.7** *The solid torus  $T$  has a single ideal vertex. A horosphere about this ideal vertex intersects each tetrahedron of  $T$  in four triangles, arranged corner to corner so that their outer boundary forms a hexagon, with opposite angles agreeing. For tetrahedra  $\Delta_1, \dots, \Delta_{N-2}$ , an inner boundary is also a hexagon, with inner boundary of the triangles of  $\Delta_i$  identified to the outer boundary of the triangles of  $\Delta_{i+1}$ . For tetrahedron  $\Delta_{N-1}$ , the four triangles form a solid hexagon.*

**Proof** Consider the boundary of any layered solid torus. This is a 1-punctured torus triangulated by two triangles. A path that stays on the 1-punctured torus that runs once around the puncture will run over exactly six triangles; these form a hexagon in the cusp neighbourhood of the solid torus; see Figure 8. Stripping the  $k$  outermost tetrahedra off a layered solid torus yields a smaller layered solid torus for  $k < N - 1$ ; its boundary still forms a hexagon as in Figure 8.

The innermost tetrahedron has its two inside triangles folded together. This gives one of the hexagons shown in Figure 9. □

For the tetrahedron  $\Delta_i$ , label the (interior) dihedral angles by  $x_i, y_i$ , and  $z_i$ , with  $x_i + y_i + z_i = \pi$ . By adjusting these labels, we may ensure that  $z_i$  is the angle assigned to the slope that is covered by  $\Delta_i$ , and that  $x_i$  and  $y_i$  are chosen to be in alphabetical order when we run around one of the cusp triangles in anticlockwise order. These labels agree with the choices in Figures 8 and 9.

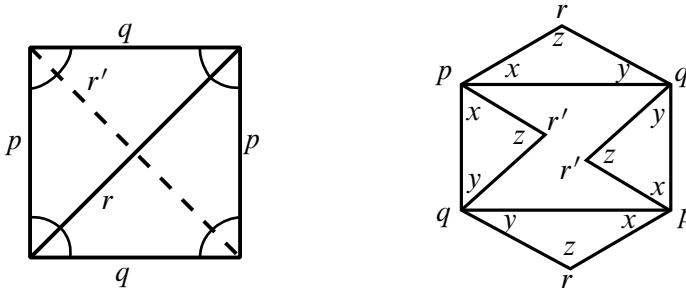


Figure 8: Left: a path encircling the puncture of the 1-punctured torus meets exactly six triangles, meeting slopes  $p, q, r, p, q,$  and  $r,$  in order. Right: These triangles lift to give a hexagon in the cusp neighbourhood of a layered solid torus. The tetrahedron that effects the diagonal exchange from  $r$  to  $r'$  is glued to the hexagon along two faces, forming a new hexagon in the interior.

Since opposite edges of a tetrahedron have the same angles, this choice of angles  $x_i, y_i,$  and  $z_i$  completely determines the angles on the hexagons. We summarise the result in the following lemma.

**Lemma 4.8** For  $i \in \{1, \dots, N - 2\},$  two opposite interior angles of the outer hexagon of  $\Delta_i$  are  $z_i,$  two opposite exterior angles of the inner hexagon are  $z_i,$  and at the four vertices shared by both hexagons, two angles  $x_i$  meet at two of the opposite vertices, and two angles  $y_i$  meet at two other opposite vertices.

For  $\Delta_{N-1},$  the interior angles of the solid hexagon are either  $z_{N-1}, 2x_{N-1} + z_{N-1},$  and  $2y_{N-1}$  (with opposite angles agreeing), or  $z_{N-1}, 2y_{N-1} + z_{N-1},$  and  $2x_{N-1}$  (with opposite angles agreeing).

Gluing tetrahedron  $\Delta_{i+1}$  to  $\Delta_i$  in the construction of the layered solid torus corresponds to performing a diagonal exchange in the triangulation of the boundary. One of the

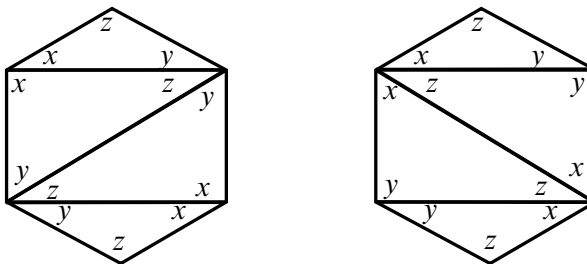


Figure 9: The last tetrahedron in the layered solid torus has two interior triangles folded together. The two possible cases are shown.

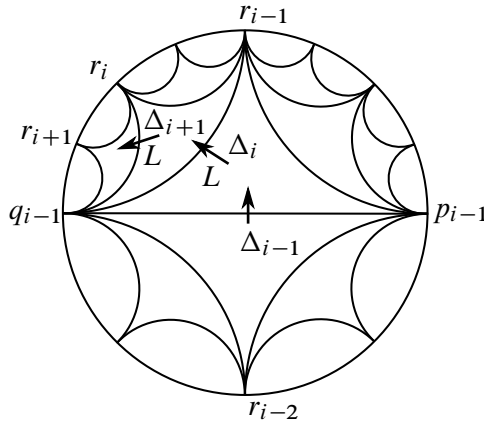


Figure 10: Turning left then left again in the Farey graph. Tetrahedron  $\Delta_{i-1}$  has inner boundary with slopes  $p_{i-1}$ ,  $q_{i-1}$ , and  $r_{i-1}$ , with angles  $x_{i-1}$ ,  $y_{i-1}$ , and  $z_{i-1}$ , respectively. Adding  $\Delta_i$  in the  $L$  direction removes the slope  $p_{i-1}$  from the inner boundary, replacing it with slope  $r_i$ , with angle  $z_i$ . Adding  $\Delta_{i+1}$  removes slope  $r_{i-1}$ , replacing it with slope  $r_{i+1}$ , with angle  $z_{i+1}$ .

three edges on the punctured torus boundary is covered by this move. In the cusp picture of the hexagons, gluing  $\Delta_{i+1}$  to  $\Delta_i$  glues four triangles to the inner hexagon. Two opposite vertices are covered by the triangles and two new vertices are added to the interior; see Figure 11 for an example.

The labelling of Lemma 4.8 implies that two vertices of the inner hexagon formed at the  $i^{\text{th}}$  step by  $\Delta_i$  have interior angle  $2\pi - z_i$ . These vertices were just added at the previous step by diagonal exchange. Since the path  $\gamma$  in the Farey graph is a geodesic, these vertices will not be covered in the next step. Thus there are two choices for vertices to cover. We call the choices  $L$  and  $R$ , referring to a choice of direction in the Farey graph, as follows. After crossing the first edge in the Farey graph,  $L$  and  $R$  are determined by the direction the geodesic  $\gamma$  takes in the Farey graph, left or right. Except in the last triangle of the Farey graph, this corresponds to attaching a tetrahedron and covering a diagonal. Label the corresponding tetrahedron  $\Delta_i$  with an  $L$  or  $R$ , for  $i = 2, \dots, N - 1$ ; see Figure 10 for an example.

We need to consider the interior angles of each hexagon. When values of the  $z_i$  are given, we will choose the  $x_i$  and  $y_i$  so that the interior angles form a Euclidean hexagon at each step. Consider the outermost hexagon. The slopes of the edges of the outermost hexagon are  $p$ ,  $q$ , and  $r$ , and their interior angles are  $\pi - \theta_p$ ,  $\pi - \theta_q$ , and  $\pi - \theta_r$ , respectively, as in Proposition 4.3. These are chosen so that the sum of all interior

angles is  $4\pi$ , as usual for a Euclidean hexagon. Since tetrahedron  $\Delta_1$  covers the edge of slope  $r$ , the angle  $z_1$  must agree with the interior angle along the slope  $r$ , or  $z_1 = \pi - \theta_r$ . Now we consider the next hexagon.

**Lemma 4.9** *Let  $\theta_r$ ,  $\theta_p$ , and  $\theta_q$  denote exterior dihedral angles as in Proposition 4.3. In particular, recall that  $r$  is the first slope covered, and  $p$  is the second.*

*For the first tetrahedron  $\Delta_1$ , set  $z_1 = \pi - \theta_r$ . Suppose  $z_2 \in (0, \pi)$ , and define a new variable  $z_0 = \pi + \theta_p$ .*

*The tetrahedron  $\Delta_2$  has either an  $L$  or an  $R$  label. Assign the same label to  $\Delta_1$ , so both are labelled  $L$  or both are labelled  $R$ .*

- *If both  $\Delta_1$  and  $\Delta_2$  are labelled  $L$ , set  $x_1 = \pi - \frac{1}{2}(z_0 + z_2)$ , and  $y_1 = \pi - z_1 - x_1$ .*
- *If both  $\Delta_1$  and  $\Delta_2$  are labelled  $R$ , set  $y_1 = \pi - \frac{1}{2}(z_0 + z_2)$ , and  $x_1 = \pi - z_1 - y_1$ .*

*Then the values of the interior angles of the hexagon between  $\Delta_1$  and  $\Delta_2$  are  $z_2$  (at the two edges of slope  $p$ ),  $2\pi - z_1$ , and  $z_1 - z_2$ .*

**Proof** One of the interior angles is immediate: the angle at the newly added edge of slope  $r'$  is  $2\pi - z_1$ .

If  $\Delta_2$  is labelled  $L$ , then the slope  $p$  is given angle  $x_1$  in  $\Delta_1$ , as in Figure 8, right. Otherwise it is given angle  $y_1$  in  $\Delta_1$ , based on our orientation conventions. Assume first that  $\Delta_2$  is labelled  $L$ .

Before adding  $\Delta_1$ , the interior angle at the edge of slope  $p$  was  $\pi - \theta_p$ . After adding  $\Delta_1$ , it decreases by  $2x_1$ . Thus the interior angle is

$$\pi - \theta_p - 2x_1 = \pi - \theta_p - 2\pi + (z_0 + z_2) = z_2.$$

Similarly, after adding  $\Delta_1$  the interior angle at the edge of slope  $q$  becomes

$$\begin{aligned} \pi - \theta_q - 2y_1 &= \pi - (\pi - \theta_p - \theta_r) - 2(\pi - z_1 - x_1) \\ &= \theta_p + \theta_r - 2\pi + 2z_1 + 2\pi - z_0 - z_2 = z_1 - z_2. \end{aligned}$$

Similar equations hold, only switching the roles of  $x_1$  and  $y_1$ , if  $\Delta_2$  is labelled  $R$ .  $\square$

We will deal with the last tetrahedron  $\Delta_{N-1}$  separately. For the others:

**Lemma 4.10** *Let  $\theta_p$ ,  $\theta_q$ , and  $\theta_r$  denote the exterior dihedral angles as they did in Proposition 4.3. Suppose  $z_0 = \pi + \theta_p$ ,  $z_1 = \pi - \theta_r$ , and  $z_2, \dots, z_{N-1}$  lie in  $(0, \pi)$ .*

For  $i = 1, \dots, N - 2$ , assign the angles  $x_i$  and  $y_i$  as below, with assignments depending on the labels ( $L$  or  $R$ ) of  $\Delta_i$  and  $\Delta_{i+1}$ :

- If  $\Delta_i$  and  $\Delta_{i+1}$  are labelled  $L$ , set  $x_i = \pi - \frac{1}{2}(z_{i-1} + z_{i+1})$ , and  $y_i = \pi - z_i - x_i$ .
- If  $\Delta_i$  and  $\Delta_{i+1}$  are labelled  $R$ , set  $y_i = \pi - \frac{1}{2}(z_{i-1} + z_{i+1})$ , and  $x_i = \pi - z_i - y_i$ .
- If  $\Delta_i$  is labelled  $L$  and  $\Delta_{i+1}$  labelled  $R$ , set  $y_i = \frac{1}{2}(z_{i-1} - z_i - z_{i+1})$ , and  $x_i = \pi - z_i - y_i$ .
- If  $\Delta_i$  is labelled  $R$  and  $\Delta_{i+1}$  labelled  $L$ , set  $x_i = \frac{1}{2}(z_{i-1} - z_i - z_{i+1})$ , and  $y_i = \pi - z_i - x_i$ .

Then for  $i = 1, \dots, N - 2$ , the hexagon between tetrahedra  $\Delta_i$  and  $\Delta_{i+1}$  has interior angles  $z_{i+1}$ ,  $2\pi - z_i$ , and  $z_i - z_{i+1}$ .

Moreover, for any interior edge obtained by layering tetrahedra  $\Delta_1, \dots, \Delta_{N-2}$ , the sum of the dihedral angles about that edge is  $2\pi$ .

**Proof** The proof is by induction. We will show that after layering tetrahedron  $\Delta_{i+1}$  onto tetrahedra  $\Delta_1, \dots, \Delta_i$ , the interior edges of hexagons are as claimed, and the sum of dihedral angles around all interior edges is  $2\pi$ .

By Lemma 4.9, the interior angles of the hexagon are as claimed when  $i = 1$ . When layering  $\Delta_1$  onto the tetrahedra outside of the layered solid torus, there are no interior edges created, so the statement on interior edges is vacuously true.

Now assume by induction that the interior angles of the hexagon between  $\Delta_{i-1}$  and  $\Delta_i$  are as claimed in the lemma, and that dihedral angles sum to  $2\pi$  around any interior edges in the layering of tetrahedra  $\Delta_1, \dots, \Delta_i$ . Consider  $\Delta_{i+1}$ .

The argument is mainly a matter of bookkeeping, particularly keeping track of labels on tetrahedra when turning left or right. We have illustrated the process carefully for the case that  $\Delta_i$  and  $\Delta_{i+1}$  are both labelled  $L$ . Figure 10 shows the path in the Farey graph. What is important at each step is which slope is covered by the diagonal exchange effected by adding the next tetrahedron. Thus  $\Delta_i$  covers a slope  $p_{i-1}$  and  $\Delta_{i+1}$  covers a slope  $r_{i-1}$ .

Figure 11, left, shows the effect on the cusp triangulation. In that figure, the outermost hexagon lies on the outside of  $\Delta_{i-1}$ , with the thick lines the hexagon between  $\Delta_{i-1}$  and  $\Delta_i$ . The edges of  $\Delta_{i-1}$  with slopes  $r_{i-1}$  are both assigned angle  $z_{i-1}$ . In the figure, slope  $r_{i-1}$  is marked by the red dot.

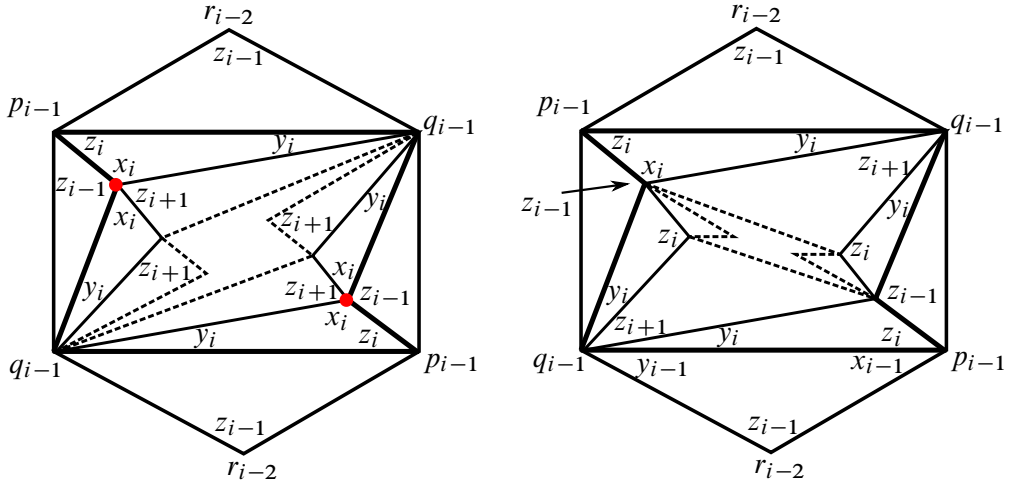


Figure 11: Left: the cusp diagram of the portion of a layered solid torus obtained by turning left, then left again. The red dot indicates an edge of the triangulation that is surrounded by the three tetrahedra  $\Delta_{i-1}$ ,  $\Delta_i$ , and  $\Delta_{i+1}$ . Right: turning left then right. The vertices of the outer hexagon for  $\Delta_{i-1}$  are adjacent to these three tetrahedra, and to no other interior tetrahedra.

Adding tetrahedron  $\Delta_i$  gives a new hexagon, indicated by the thinner line in Figure 11, left, between  $\Delta_i$  and  $\Delta_{i+1}$ . The edges of  $\Delta_i$  with slopes  $p_{i-1}$  and  $r_i$  are assigned angle  $z_i$ . Our orientation convention then ensures that the edge of slope  $r_{i-1}$  is assigned angle  $x_i$ .

Finally we add tetrahedron  $\Delta_{i+1}$ . This gives a new innermost hexagon, indicated by the dashed lines in Figure 11, left. The edge of slope  $r_{i-1}$  is assigned angle  $z_{i+1}$ .

First we consider the interior angles of the hexagon between  $\Delta_i$  and  $\Delta_{i+1}$ . One of these is  $2\pi - z_i$ , as desired. The other two are obtained by subtracting  $2x_i$  and  $2y_i$  from interior angles of the hexagon at the previous step. In particular, we have angles

$$2\pi - z_{i-1} - 2x_i = 2\pi - z_{i-1} - 2\pi + z_{i-1} + z_{i+1} = z_{i+1}$$

and

$$z_{i-1} - z_i - 2y_i = z_{i-1} - z_i - 2\pi + 2z_i + (2\pi - z_{i-1} - z_{i+1}) = z_i - z_{i+1},$$

as desired.

Notice that after adding tetrahedron  $\Delta_{i+1}$ , the edge of slope  $r_{i-1}$  is completely surrounded by tetrahedra  $\Delta_{i-1}$ ,  $\Delta_i$ , and  $\Delta_{i+1}$ , and thus it becomes an interior edge.

Notice also that this is the only new interior edge obtained by adding  $\Delta_{i+1}$ . Thus we only need to ensure the sum of dihedral angles about this edge is  $2\pi$ . We read the dihedral angles off of Figure 11, left:

$$z_{i-1} + 2x_i + z_{i+1} = 2\pi.$$

This will hold if and only if  $x_i$  satisfies the requirements of the lemma.

A very similar pair of pictures, Farey graph and cusp triangulation, gives the result in the case where  $\Delta_i$  and  $\Delta_{i+1}$  are both labelled  $R$ . In this case, however, the slope  $q_{i-1}$  will be covered by  $\Delta_i$ . Again,  $r_{i-1}$  will then be covered by  $\Delta_{i+1}$ , but by turning right, the angles adjacent to the slope  $r_{i-1}$  in this case will be  $z_{i-1}$ , two copies of  $y_i$ , and  $z_{i+1}$ . Thus this case differs from the previous only by switching the roles of  $x_i$  and  $y_i$ .

If we first turn left then turn right, the slope  $p_{i-1}$  is covered first by  $\Delta_i$ , then  $q_{i-1}$  by  $\Delta_{i+1}$ ; see Figure 11, right. The interior angles of the hexagon between  $\Delta_i$  and  $\Delta_{i+1}$  are  $2\pi - z_i$ ,  $2\pi - z_{i-1} - 2x_i$ , and  $z_{i-1} - z_i - 2y_i$ . The last two simplify as follows:

$$\begin{aligned} 2\pi - z_{i-1} - 2x_i &= 2\pi - z_{i-1} - 2\pi + 2z_i + 2y_i \\ &= 2\pi - z_{i-1} - 2\pi + 2z_i + (z_{i-1} - z_i - z_{i+1}) = z_i - z_{i+1}, \\ z_{i-1} - z_i - 2y_i &= z_{i-1} - z_i - z_{i-1} + z_i + z_{i+1} = z_{i+1}. \end{aligned}$$

Finally, in this case, none of the newly added edges are surrounded by the three tetrahedra  $\Delta_{i-1}$ ,  $\Delta_i$ , and  $\Delta_{i+1}$ . However, adding  $\Delta_{i+1}$  may have created an interior edge at  $q_{i-1}$  if  $q_{i-1}$  does not lie on the boundary of the layered solid torus. By induction, we know that the interior angle of the hexagon between  $\Delta_{i-1}$  and  $\Delta_i$  at the edge of slope  $q_{i-1}$  must be  $z_{i-1} - z_i$ . To this we add two angles  $y_i$  coming from  $\Delta_i$ , and one angle  $z_{i+1}$  from  $\Delta_{i+1}$ .

In particular, the angles will fit into the Euclidean hexagon, and therefore have the correct angle sum, if and only if

$$2y_i + z_{i+1} = z_{i-1} - z_i.$$

This holds if and only if  $y_i$  satisfies the requirement of the lemma.

The case of  $R$  followed by  $L$  is nearly identical, with the roles of  $x_i$  and  $y_i$  switched. Thus by induction, the result holds for  $i = 1, \dots, N - 2$ . □

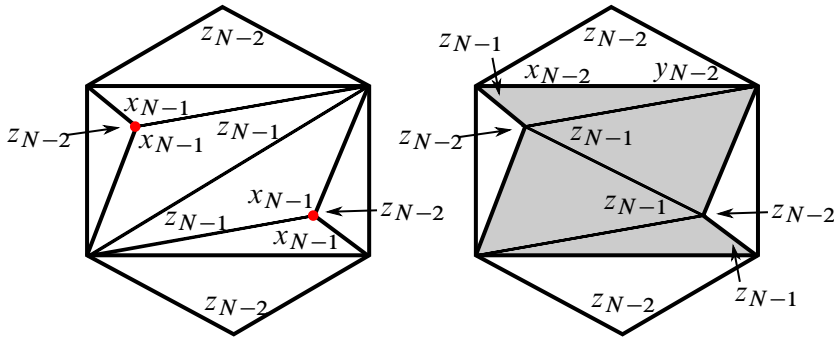


Figure 12: Shown are both cases when  $\Delta_{N-1}$  is labelled  $L$ . On the left, the empty tetrahedron  $\Delta_N$  is labelled  $L$ , and on the right, the empty tetrahedron  $\Delta_N$  is labelled  $R$ .

**Lemma 4.11** Consider the last tetrahedron  $\Delta_{N-1}$ . Assign a label  $L$  or  $R$  to an empty tetrahedron  $\Delta_N$  depending on whether  $\gamma$  turns left or right when running into the final triangle  $T_N$  of the Farey complex, and set  $z_N = 0$ . Define the angles  $x_{N-1}$  and  $y_{N-1}$  in terms of  $z_N, z_{N-1}$ , and  $z_{N-2}$  depending on the labels  $L$  or  $R$  on  $\Delta_{N-1}$  and  $\Delta_N$  exactly as in Lemma 4.10. Then the sum of dihedral angles is  $2\pi$  around the interior edges in the layered solid torus that are surrounded by  $\Delta_{N-2}$  and  $\Delta_{N-1}$ .

**Proof** The proof is very similar to that of Lemma 4.10. The cusp triangulations for cases  $LL$  and  $LR$  are shown in Figure 12.

In the case  $LL$ , exactly one edge in the interior of the solid torus is surrounded by  $\Delta_{N-2}$  and  $\Delta_{N-1}$ . The sum of the angles around this edge is

$$z_{N-2} + 2x_{N-1} = z_{N-2} + 2\pi - z_{N-2} - z_N = 2\pi,$$

since  $z_N = 0$ . Thus the sum is  $2\pi$  in the  $LL$  case when  $i = N - 1$ .

In the case  $LR$ , an interior edge is surrounded by  $\Delta_{N-1}$  and  $\Delta_{N-2}$ , and the sum of angles around the edge must be

$$\begin{aligned} z_{N-2} + 2x_{N-1} + z_{N-1} &= z_{N-2} + z_{N-1} + 2(\pi - z_{N-1} - y_{N-1}) \\ &= z_{N-2} + z_{N-1} + 2\pi - 2z_{N-1} - z_{N-2} + z_{N-1} + z_N = 2\pi. \end{aligned}$$

The cases  $RR$  and  $RL$  hold similarly. □

**Lemma 4.12** Let  $\theta_p, \theta_q$ , and  $\theta_r$  be as in Proposition 4.3. Let

$$(z_0 = \pi + \theta_p, z_1 = \pi - \theta_r, z_2, \dots, z_{N-1}, z_N = 0)$$



be a sequence of numbers with  $z_i \in (0, \pi)$  for  $i = 1, \dots, N - 1$ . Let  $x_i$  or  $y_i$  be defined in terms of the sequence of the  $z_j$  via the equations of Lemma 4.10. Then  $x_i, y_i$ , and  $z_i$  give an angle structure on the layered solid torus if and only if, for each  $i = 1, \dots, N - 1$ , the sequence satisfies

$$\begin{cases} z_{i-1} > z_i + z_{i+1} & \text{if } \Delta_i \text{ and } \Delta_{i+1} \text{ are labelled } RL \text{ or } LR \text{ (hinge condition),} \\ z_{i-1} + z_{i+1} > 2z_i & \text{if } \Delta_i \text{ and } \Delta_{i+1} \text{ are labelled } RR \text{ or } LL \text{ (convexity condition),} \end{cases}$$

and additionally  $z_2 < \pi - \theta_p$ .

Moreover, if they give an angle structure, then the sequence is strictly decreasing.

**Proof** Suppose first that we have an angle structure. Then  $x_i, y_i, z_i \in (0, \pi)$  for  $i = 1, \dots, N - 1$ , and  $x_i + y_i + z_i = \pi$ . We can use this equation along with the equations of Lemma 4.10 to write both  $x_i$  and  $y_i$  in terms of  $z_{i-1}, z_i$ , and  $z_{i+1}$ .

In the *LL* or *RR* case, each of the  $x_i$  and  $y_i$  are one of

$$(4.13) \quad \frac{1}{2}(z_{i-1} - 2z_i + z_{i+1}) \quad \text{and} \quad \pi - \frac{1}{2}(z_{i-1} + z_{i+1}).$$

Thus, because we are assuming we have an angle structure, we have:

$$0 < \frac{1}{2}(z_{i-1} - 2z_i + z_{i+1}) < \pi \quad \text{and} \quad 0 < \pi - \frac{1}{2}(z_{i-1} + z_{i+1}) < \pi.$$

The first inequality on the left implies the convexity equation. When  $i = 1$ , the first inequality on the right implies  $z_2 < \pi - \theta_p$ .

In the *RL* or *LR* case, each of the  $x_i$  and  $y_i$  are one of

$$(4.14) \quad \frac{1}{2}(z_{i-1} - z_i - z_{i+1}) \quad \text{and} \quad \pi - \frac{1}{2}(z_{i-1} + z_i + z_{i+1}).$$

Because we have an angle structure,

$$0 < \frac{1}{2}(z_{i-1} - z_i - z_{i+1}) < \pi \quad \text{and} \quad 0 < \pi - \frac{1}{2}(z_{i-1} + z_i + z_{i+1}) < \pi.$$

Again the first inequality on the left implies the hinge equation. This concludes one direction of the proof.

Now suppose for each  $i = 1, \dots, N - 1$ , the sequence satisfies the convexity or hinge condition. We check the conditions on an angle structure, Definition 4.1. Condition (2) holds by our definition of  $x_i$  and  $y_i$ : by hypothesis we require  $x_i + y_i + z_i = \pi$ .

Condition (3) follows from Lemmas 4.10 and 4.11. These lemmas prove that, given our definitions of  $x_i$  and  $y_i$  in terms of  $z_{i-1}, z_i$ , and  $z_{i+1}$ , the sum of dihedral angles around every interior edge of the layered solid torus is  $2\pi$ .

As for condition (1), by hypothesis each  $z_i \in (0, \pi)$  for  $i = 1, \dots, N - 1$ . It remains to show that  $x_i$  and  $y_i$  lie in  $(0, \pi)$  for  $i = 1, \dots, N - 1$ . In the *LL* or *RR* case, we have noted that  $x_i$  and  $y_i$  are as in (4.13). Thus we need

$$0 < \frac{1}{2}(z_{i-1} - 2z_i + z_{i+1}) < \pi \quad \text{and} \quad 0 < \pi - \frac{1}{2}(z_{i-1} + z_{i+1}) < \pi.$$

These give four inequalities. When  $i = 2, \dots, N - 1$ , three of the inequalities are automatically satisfied when  $z_{i-1}, z_i, z_{i+1} \in (0, \pi)$  or when  $i = N - 1$  and  $z_N = 0$ . The final inequality holds if and only if  $z_{i-1} + z_{i+1} > 2z_i$ , which is the convexity condition.

When  $i = 1$ , the inequalities become

$$0 < \frac{1}{2}((\pi + \theta_p) - 2(\pi - \theta_r) + z_2) < \pi \quad \text{and} \quad 0 < \pi - \frac{1}{2}((\pi + \theta_p) + z_2) < \pi.$$

These give four inequalities, one of which is automatically true for  $0 < \pi + \theta_p < \pi$ , and the other three all hold if and only if

$$\pi - \theta_p - 2\theta_r < z_2 < \pi - \theta_p.$$

For  $2 \leq i \leq N - 1$  in the *RL* or *LR* case,  $x_i$  and  $y_i$  are as in (4.14). Thus we require

$$0 < \frac{1}{2}(z_{i-1} - z_i - z_{i+1}) < \pi \quad \text{and} \quad 0 < \pi - \frac{1}{2}(z_{i-1} + z_i - z_{i+1}) < \pi.$$

Again this gives four inequalities, two of which are automatic for  $z_{i-1}, z_i, z_{i+1} \in (0, \pi)$ , or when  $i = N - 1$ , for  $z_N = 0$ . The other two inequalities that must be satisfied are  $z_{i-1} > z_{i+1} - z_i$  and  $z_{i-1} > z_{i+1} + z_i$ . Both hold if and only if  $z_{i-1} > z_{i+1} + z_i$ .

This proves the if and only if statement of the lemma.

Now suppose we have an angle structure. At this point, we know all the inequalities of the lemma must hold, plus an extra one:  $z_{i-1} > z_{i+1}$ . However, the hinge and convexity equations imply that the sequence is strictly decreasing: the proof is by a downward induction starting at  $z_N = 0$ .  $\square$

**Lemma 4.15** *Suppose all tetrahedra are glued via *RR* or *LL* and never a hinge *RL* or *LR*. Then there exists a sequence satisfying the previous lemma if and only if*

$$i(m, p)\theta_p + i(m, q)\theta_q + i(m, r)\theta_r > 2\pi.$$

**Proof** Suppose first that such a sequence holds.

We claim that the convexity condition implies that  $z_{N-k} < z_{N-(k+1)}k/(k+1)$  for  $k = 1, \dots, N - 1$ . This can be seen by induction. When  $k = 1$ , the inequality

$z_{N-2} + z_N > 2z_{N-1}$  implies  $z_{N-1} < \frac{1}{2}z_{N-2}$ . Assuming  $z_{N-(j-1)} < z_{N-j}(j-1)/j$ , then  $z_{N-(j-1)} + z_{N-(j+1)} > 2z_{N-j}$  implies  $z_{N-j}(j-1)/j + z_{N-(j+1)} > 2z_{N-j}$ , which implies  $jz_{N-(j+1)} > (j+1)z_{N-j}$ , as desired.

Now observe that when  $k = N - 1$ , the inequality is  $Nz_1 < (N - 1)z_0$ , which then becomes  $N(\pi - \theta_r) < (N - 1)(\pi + \theta_p)$ . Simplifying, we obtain

$$\pi < (N - 1)\theta_p + N\theta_r \iff 2\pi < \theta_q + N\theta_p + (N + 1)\theta_r,$$

using  $\theta_p + \theta_q + \theta_r = \pi$ .

Suppose that the tetrahedra are all glued in the pattern  $LL \dots L$ . Apply an isometry to  $\mathbb{H}^2$  so that the triangle  $(p, q, r)$  maps to  $(0, 1/0, -1)$ . Then the slope  $m$  is mapped to the slope  $N/1 \in \mathbb{Q}$ , and the geometric intersection numbers satisfy  $i(1/0, N) = 1$ ,  $i(0, N) = N$ , and  $i(-1, N) = N + 1$ . Because applying an isometry of  $\mathbb{H}^2$  preserves intersection numbers, it follows that the inequality holds above if and only if

$$i(m, p)\theta_p + i(m, q)\theta_q + i(m, r)\theta_r > 2\pi.$$

The argument in the case that all tetrahedra are glued in the pattern  $RR \dots R$  is similar. It follows that if a sequence  $(z_0, \dots, z_N)$  exists, then the inequality holds.

Conversely, suppose the inequality holds. Set  $z_0 = \pi + \theta_p$  and  $z_1 = \pi - \theta_r$ . Choose  $z_2$  such that

$$\max\{0, 2z_1 - z_0 = \pi - 2\theta_r - \theta_p\} < z_2 < \min\{z_1 = \pi - \theta_r, \pi - \theta_p\}.$$

Inductively, choose a decreasing sequence  $z_k$  such that  $z_k > 2z_{k-1} - z_{k-2}$  and  $z_k \in (0, \pi)$ . We need to ensure we can choose the sequence all the way to  $z_{N-1}$  and set  $z_N = 0$ . Note by this choice of  $z_k$ , we have

$$z_k > \pi - k\theta_r - (k - 1)\theta_p,$$

so, when  $k = N$ ,

$$z_N > \pi - N\theta_r - (N - 1)\theta_p.$$

But, as above, the inequality on  $\theta_p, \theta_q$ , and  $\theta_r$  is equivalent to

$$\pi - N\theta_r - (N - 1)\theta_p < 0.$$

Thus we may set  $z_N = 0$  and satisfy all the required conditions. □

**Lemma 4.16** *Suppose there exists a hinge  $RL$  or  $LR$  in the sequence of labels of  $\Delta_1, \dots, \Delta_N$ . Then the inequality*

$$i(m, p)\theta_p + i(m, q)\theta_q + i(m, r)\theta_r > 2\pi$$

*is satisfied for every  $\theta_p, \theta_q$ , and  $\theta_r$  as in Proposition 4.3.*

**Proof** If there exists a hinge, it is not in the first two labels by choice of  $p$  and  $r$ . Suppose first that the first two labels are  $RR$ . Apply an isometry to  $\mathbb{H}^2$  taking  $(p, q, r)$  to  $(0, 1/0, -1)$ . Then the first two steps in the Farey graph move from triangle  $(0, 1/0, -1)$  to  $(0, 1, 1/2)$ . There may be some additional instances of  $R$  in the sequence. Starting at  $(0, 1/0, -1)$  and stepping through  $n$  initial labels  $R$  in the Farey graph puts  $\gamma$  in the triangle  $(0, 1/(n-1), 1/n)$ . At this point, the path  $\gamma$  goes left, crossing the edge  $(1/(n-1), 1/n)$ . Because  $\gamma$  never returns to an edge, this means that the slope  $m$  lies between  $1/(n-1)$  and  $1/n$  in the Farey complex. Write  $m = a/b$  in lowest terms. The set of rational numbers between  $1/(n-1)$  and  $1/n$  in the Farey complex can be obtained inductively by summing numerators and denominators of  $1/(n-1)$ ,  $1/n$  and other rationals obtained in this manner. Since  $a/b$  lies in this range,  $a \geq 2$  and  $b \geq 2n - 1 > 2$ .

Now, note that for  $(p, q, r) = (0, 1/0, -1)$ , we have  $i(a/b, p) = a$ ,  $i(a/b, q) = b$ , and  $i(a/b, r) = a + b$ . Thus

$$i(m, p)\theta_p + i(m, q)\theta_q + i(m, r)\theta_r = a(\theta_p + \theta_r) + b(\theta_q + \theta_r).$$

Because  $\theta_p + \theta_q + \theta_r = \pi$  and  $-\pi < \theta_p, \theta_q < \pi$ , both  $\theta_p + \theta_r = \pi - \theta_q$  and  $\theta_q + \theta_r = \pi - \theta_p$  are positive. Thus

$$a(\theta_p + \theta_r) + b(\theta_q + \theta_r) \geq \min\{a, b\}(\theta_p + \theta_q + 2\theta_r) = \min\{a, b\}(\pi + \theta_r) > 2\pi.$$

Since intersection numbers are unchanged under isometry of  $\mathbb{H}^2$ , this proves the result when the first two labels are  $RR$ .

The case that the first two labels are  $LL$  is similar. □

**Lemma 4.17** *Suppose there exists a hinge  $RL$  or  $LR$ . Then there exists a sequence satisfying Lemma 4.12.*

**Proof** Let  $h \in \{2, 3, \dots, N - 1\}$  be the smallest index such that  $\Delta_h$  is a hinge of the form  $RL$  or  $LR$ . Set  $z_0 = \pi + \theta_p$  and  $z_1 = \pi - \theta_r$ . We can choose inductively a positive decreasing sequence  $z_k$  such that  $z_2 < \pi - \theta_p$ , each  $z_k \in (0, \pi)$ , and  $z_k > 2z_{k-1} - z_{k-2}$  for  $2 \leq k \leq h$ .

The rest of the sequence  $z_i$  is constructed backwards from  $i = N$  to  $i = h$ . Consider a sequence  $z'_i$ . Set  $z'_N = 0$  and  $z'_{N-1} = 1$ . For each  $i$  such that  $N - 2 \geq i \geq h + 1$ , inductively choose  $z'_i$  such that  $z'_i > z'_{i+1} + z'_{i+2}$  or  $z'_i + z'_{i+2} > 2z'_{i+1}$ , depending on whether  $\Delta_{i+1}$  has a different label ( $L$  or  $R$ ) from  $\Delta_i$  or not, respectively. Observe,  $z'_i$  must be greater than  $z'_{i+1}$  for each  $i$ .

Choose  $\epsilon$  such that

$$0 < \epsilon < \frac{z_{h-1} - z_h}{z'_{h+1}}.$$

Set  $z_i = \epsilon z'_i$  for  $h + 1 \leq i \leq N$ . We need  $z_h$  to satisfy the hinge condition

$$z_h < z_{h-1} - z_{h+1} \quad \text{or} \quad z_h < z_{h-1} - \epsilon z'_{h+1}.$$

This holds by our choice of  $\epsilon$ .

Finally, we need each  $z_i$  to lie in  $(0, \pi)$ , for  $h + 1 \leq i \leq N - 1$ . Observe that  $z_{h-1} < \pi$ , so  $0 < \epsilon < \pi/z'_{h+1}$ . Then  $z_i = \epsilon z'_i < \pi z'_i/z'_{h+1}$ . For  $h + 1 \leq i \leq N - 1$ , we know  $z'_i \leq z'_{h+1}$ , hence  $z'_i$  is strictly less than  $\pi$ , as desired. Because  $z'_i$  is at least  $z'_{N-1} > 0$ ,  $z_i = \epsilon z'_i > 0$ . Thus we have found a sequence satisfying Lemma 4.12.  $\square$

**Proof of Proposition 4.3** Suppose

$$i(m, p)\theta_p + i(m, q)\theta_q + i(m, r)\theta_r \leq 2\pi.$$

By Lemma 4.16, there is no hinge  $RL$  or  $LR$  in the sequence of labels of tetrahedra making the layered solid torus. By Lemma 4.15, there does not exist a sequence satisfying Lemma 4.12. But such a sequence is required in an angle structure on a layered solid torus, so there is no angle structure in this case.

Now suppose  $i(m, p)\theta_p + i(m, q)\theta_q + i(m, r)\theta_r > 2\pi$ . Then Lemmas 4.17 and 4.15 imply there exists a sequence satisfying Lemma 4.12. It follows from that lemma that there exists an angle structure.  $\square$

**4.2 Volume maximisation**

We now show that the volume functional on the space of angle structures takes its maximum on the interior. This is essentially [16, Proposition 15], but we extract slightly more information from the proof.

**Lemma 4.18** *Suppose the volume functional on the space of angle structures on a layered solid torus takes its maximum on the boundary. Then the corresponding structure consists only of flat tetrahedra, with angles  $(x_i, y_i, z_i)$  a permutation of  $(0, 0, \pi)$  for each  $i = 1, \dots, N - 1$ .*

**Proof** By work of Rivin [28], if the volume functional takes its maximum on the boundary of the space of angle structures, then any tetrahedron with an angle 0 must also have an angle  $\pi$ . Thus those tetrahedra that do not have all angles strictly within  $(0, \pi)$  must have angles  $(x_i, y_i, z_i)$  a permutation of  $(0, 0, \pi)$ ; this is a flat tetrahedron.

By Lemma 4.12, a point on the boundary of the space of angle structures corresponds to a sequence  $(z_0 = \pi + \theta_p, z_1 = \pi - \theta_r, z_2, \dots, z_{N-1}, z_N = 0)$  satisfying the hinge and convexity equations, except the strict inequalities will be replaced by weak inequalities. This must be a nonincreasing sequence.

Suppose the  $i^{\text{th}}$  tetrahedron is the first flat tetrahedron. Then  $z_i \in \{0, \pi\}$  but  $z_{i-1} \in (0, \pi)$  unless  $i - 1 = 0$ . If  $z_i = \pi$ , then convexity implies  $z_j = \pi$  for  $j = i + 1, \dots, h$ , where  $h$  is the next hinge index. The hinge condition then implies that all later  $z_j$  equal 0. Similarly, if  $z_i = 0$  then all later  $z_j$  equal 0.

Now consider  $z_{i-1}$ . We have  $x_i, y_i, z_i, z_{i+1} \in \{0, \pi\}$ . Thus by one of Lemmas 4.10, 4.9, or 4.11, depending on the index  $i$ , we have  $z_{i-1} = 2\pi$ . But  $0 < z_{i-1} < \pi$  unless  $i - 1 = 0$ . So  $i - 1 = 0$ . Then the first flat tetrahedron is the first tetrahedron, so the entire solid torus consists of flat tetrahedra.  $\square$

**Corollary 4.19** *Suppose the set of angle structures as in Proposition 4.3 is nonempty. Then the volume functional takes its maximum on the interior of such angle structures.*

The following follows immediately from the Casson–Rivin theorem, Theorem 4.2.

**Corollary 4.20** *For slopes  $p, q, r$ , and  $m$  as in Proposition 4.3, and any angles  $\theta_p, \theta_q$ , and  $\theta_r$  satisfying (4.4), there exists a geometric triangulation of the layered solid torus  $T$  of that proposition with exterior dihedral angles  $\theta_p, \theta_q$ , and  $\theta_r$ .*

## 5 Dehn filling

In this section, we complete the proof of Theorem 5.5.

Let  $s$  be a slope, and let  $\text{len}(s)$  denote the Euclidean length of a geodesic representative of  $s$  on a horospherical cusp torus.

The following is a consequence of Thurston’s hyperbolic Dehn filling theorem [29]. The version below can be proved assuming a geometric triangulation exists for  $M$ , with ideas in Benedetti and Petronio [3], using methods of Neumann and Zagier [23].

**Theorem 5.1** (hyperbolic Dehn filling theorem) *Let  $M$  be a hyperbolic 3–manifold with a geometric ideal triangulation such that exactly two ideal tetrahedra,  $\Delta$  and  $\Delta'$ , meet a cusp of  $M$ . Let  $s$  be a slope on this cusp. Then for all but finitely many choices of  $s$ , the Dehn filled manifold  $M(s)$  admits a complete hyperbolic structure, obtained by deforming the triangulation of  $M$ , and taking the completion of the resulting structure. The tips of the tetrahedra  $\Delta$  and  $\Delta'$  spin asymptotically along the geodesic core of the filling solid torus of  $M(s)$ . As  $\text{len}(s)$  goes to infinity, the cross-ratios of the tetrahedra of  $M(s)$  become uniformly close to those of  $M$ .*

In particular, since  $M$  admits a geometric ideal triangulation, the cross-ratios of its tetrahedra have strictly positive imaginary part. This is an open condition. Thus for  $s$  large enough, the triangulation of  $M(s)$  also has cross-ratios with strictly positive imaginary part. It follows that the incomplete, spun triangulation of  $M(s)$  is built of geometric tetrahedra. However, we are not interested in incomplete triangulations. We will use the incomplete spun triangulation to build a complete geometric ideal triangulation.

### 5.1 Dehn filling and spun triangulations

The following proposition is essentially Proposition 8 of Guéritaud–Schleimer [16].

**Proposition 5.2** *Let  $X$  be a solid torus with boundary  $\partial X$  a punctured torus that is triangulated by two ideal triangles. Let  $m \subset \partial X$  be the meridian of  $X$ . The following are equivalent.*

- (1) *A complete hyperbolic structure on  $X$  is obtained by taking the completion of a spun triangulation consisting of two tetrahedra  $\Delta$  and  $\Delta'$ , where one face of  $\Delta$  and one face of  $\Delta'$  form the two ideal triangles making up  $\partial X$ .*
- (2) *The exterior dihedral angles  $a, b$ , and  $c$  on the edges of the triangulation satisfy  $a, b \in (-\pi, \pi)$ ,  $c \in (0, \pi)$ , and  $a + b + c = \pi$ , and also*

$$n_a a + n_b b + n_c c > 2\pi,$$

*where  $n_a, n_b$ , and  $n_c$  denote the number of times the meridian  $m \subset \partial X$  of  $X$  crosses the edge with angle  $a, b$ , and  $c$ , respectively.*

*Moreover, if the hyperbolic structure exists on  $X$ , then it is unique.*

**Remark 5.3** Proposition 8 of [16] is not quite stated the same as Proposition 5.2, but an almost identical proof gives the result claimed here. One difference is that in [16], they restrict to  $a, b, c \in [0, \pi)$ . However, this restriction is not required for the proof. What is required, if these form exterior dihedral angles of a solid torus as claimed, is that  $a + b + c = \pi$  and  $a, b, c \in (-\pi, \pi)$ . These conditions follow from considering the Euclidean geometry of a horospherical neighbourhood of the puncture on  $\partial X$ . Moreover, the condition  $a + b + c = \pi$  forces one of  $a, b$ , and  $c$  to be strictly positive; we let this angle be denoted by  $c$ .

Now in the proof of [16, Propostion 8], it is shown that angle structures can be put onto  $\Delta$  and  $\Delta'$  to form the spun triangulation of  $X$  if and only if  $n_a a + n_b b + n_c c > 2\pi$ . In the case the inequality holds, it is shown that the volume functional takes its maximum on the interior of the space of such angle structures, meaning there exists a hyperbolic structure, and that structure is unique by the Casson–Rivin theorem, Theorem 4.2.

**Theorem 5.4** *Let  $L$  be a hyperbolic fully augmented link with  $n \geq 2$  crossing circles. Then there exist constants  $A_1, \dots, A_n$  such that if  $M$  is a manifold obtained by Dehn filling the crossing circle cusps of  $S^3 - L$  along slopes  $s_1, \dots, s_n$  whose lengths satisfy  $\text{len}(s_i) \geq A_i$  for each  $i = 1, \dots, n$ , then  $M$  admits a geometric triangulation. Allowing some collection of  $s_i = \infty$ , ie leaving some crossing circle cusps unfilled, also admits a geometric triangulation.*

**Proof** By Proposition 2.4,  $S^3 - L$  admits a geometric ideal triangulation with the property that each crossing circle meets exactly two ideal tetrahedra. By Theorem 5.1, for any sufficiently long slope, the Dehn filling along that slope is obtained by taking the completion of a spun triangulation consisting of two tetrahedra. In particular, for the  $j^{\text{th}}$  twist region, there exists  $A_j$  such that if the length of the slope is at least  $A_j$ , then Dehn filling yields a manifold with spun triangulation. This can be repeated sequentially for each crossing circle, giving constants  $A_1, \dots, A_n$ .

For each crossing circle, consider the two tetrahedra that spin around the core of the Dehn filled solid torus. These two tetrahedra together form a spun triangulation of a solid torus. By Proposition 5.2, this torus is unique, and the exterior dihedral angles  $a, b$ , and  $c$  must satisfy  $n_a a + n_b b + n_c c > 2\pi$ , where  $n_a, n_b$ , and  $n_c$  denote the number of times the meridian meets the edge on the boundary with corresponding dihedral angle. Then Corollary 4.20 implies there exists a corresponding layered solid torus with the same dihedral angles along slopes on the boundary, and the same meridian, with a complete, geometric hyperbolic structure.



Because dihedral angles agree, each spun solid torus can be removed and replaced by the layered solid torus by isometry. The result is a geometric ideal triangulation of the Dehn filling of  $S^3 - L$ .  $\square$

**Theorem 5.5** *For every  $n \geq 2$ , there exists a constant  $A_n$  depending on  $n$ , such that if  $K$  is a link in  $S^3$  with a prime, twist-reduced diagram with  $n$  twist regions, and at least  $A_n$  crossings in each twist region, then  $S^3 - K$  admits a geometric triangulation.*

**Proof** If  $K$  has a prime, twist-reduced diagram with  $n \geq 2$  twist regions, then  $S^3 - K$  is obtained by Dehn filling a hyperbolic fully augmented link  $L$  with  $n$  crossing circles, where the Dehn filling is along slopes on each crossing circle. Let  $m_j$  be the number of crossings in the  $j^{\text{th}}$  twist region. Then the length of the  $j^{\text{th}}$  slope is at least  $\sqrt{m_j^2 + 1}$  by [11, Theorem 3.10].

Note that for fixed  $n$ , there are only finitely many fully augmented links with  $n$  crossing circles. Fix one of these fully augmented links; call it  $L_k$ . By Theorem 5.4, there exist constants  $A_{k,1}, \dots, A_{k,n}$  such that if the slope on the  $j^{\text{th}}$  crossing circle of  $L_k$  has length at least  $A_{k,j}$ , for  $j = 1, \dots, n$ , then the Dehn filling admits a geometric triangulation. Consider  $A_n = \max\{A_{k,j}\}$ , where the maximum is taken over all links  $L_k$  with  $n$  crossing circles. Then provided the number of crossings in each twist region of  $K$  is at least  $A_n$ , the length of each slope on each crossing circle will be at least  $A_n$ , which implies the Dehn filling yields a geometric triangulation.  $\square$

## 6 Borromean rings and related links

In the previous section, we completed the proof of Theorem 5.5, which is unfortunately not effective: the constants  $A_1, \dots, A_n$  are unknown. In this section, by restricting the fully augmented links we consider, we are able to prove an effective result, giving an explicit family of hyperbolic 3-manifolds with geometric triangulations. This is similar in spirit to Section 5 of [16], in which Guéritaud and Schleimer show a similar result for Dehn filling one cusp of the Whitehead link. We extend first to the Borromean rings, which is a fully augmented link with two crossing circle cusps, and to the two other fully augmented links with exactly two crossing circle cusps.

The augmented links we consider next are shown in Figure 13. The link on the left of Figure 13 shows a fully augmented link with three link components; this is ambient isotopic to the Borromean rings. There are two different fully augmented links obtained



Figure 13: A picture of the Borromean rings as a fully augmented link, and the other two fully augmented links with exactly two crossing circles.

by inserting half-twists into the crossing circles of the Borromean rings shown; these are the links in the middle and right of that figure.

Following the procedure for decomposing fully augmented links into polyhedra as in Section 2, we find that all three links in Figure 13 decompose into two ideal octahedra; the decomposition for the middle link is exactly the illustration shown in Figure 3.

**Lemma 6.1** *Let  $L$  be one of the three fully augmented links with exactly two crossing circles, as shown in Figure 13. Then  $M = S^3 - L$  has a decomposition into two regular ideal octahedra. Fix one of the two crossing circle cusps. The octahedra meet the fixed crossing circle cusp as follows.*

- *One vertex of each octahedron meets the crossing circle cusp. Taking such a vertex to infinity gives a square on  $\mathbb{R}^2$  in  $\partial\mathbb{H}^3$ . We may arrange that one square has corners at  $(0, 0)$ ,  $(1, 0)$ ,  $(1, 1)$ , and  $(0, 1)$  in  $\mathbb{R}^2$ , and the other has corners at  $(1, 0)$ ,  $(2, 0)$ ,  $(2, 1)$ , and  $(1, 1)$  in  $\mathbb{R}^2$ .*
- *When the crossing circle does not encircle a half-twist, then the arc running from  $(0, 0)$  to  $(0, 1)$  projects to a meridian of the crossing circle. When the crossing circle encircles a half-twist (single crossing), the arc from  $(0, 0)$  to  $(1, 1)$  projects to a meridian.*
- *In all cases, the arc from  $(0, 0)$  to  $(2, 0)$  projects to a longitude of the crossing circle, bounding a disc in  $S^3$ .*

**Proof** The lemma is proved by considering the decomposition. White faces become the circles shown in Figure 14, left. On each circle packing an  $x$  indicates a crossing circle cusp. Take one of these points to infinity to obtain the required square. Because the two polyhedra are glued along a white side, the squares line up side-by-side as claimed; see Figure 14, right. A longitude runs along two shaded faces, which runs along the base of both squares, as claimed.

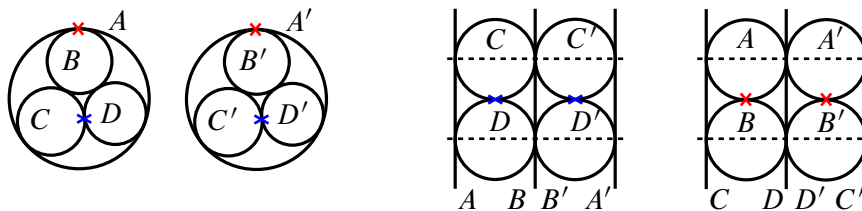


Figure 14: On the left are shown the two identical circle packings arising from the decomposition of the links of Figure 13. An  $x$  marks a crossing circle cusp in the two polyhedra. On the right, the two cusp neighbourhoods are shown, obtained by taking each pair of crossing circle cusps to infinity. Dashed lines indicate shaded faces.

When there is no half-twist, the shaded face running across the bottom of the square is glued to the shaded face running across the top of the same square, and hence the base of each square is glued to the top of the same square to form a fundamental domain for the cusp torus. A meridian runs along a white side of a square.

When there is a half-twist, a shaded face running across the bottom of the square on the left is glued to the shaded face running across the top of the square on the right, and so a shearing occurs; see [27, Proposition 3.2]. The result is that a meridian runs across the diagonal of a square, as claimed.  $\square$

**Lemma 6.2** *Let  $M$  be the complement of one of the three fully augmented links with exactly two crossing circles. Then  $M$  has a decomposition into exactly eight ideal tetrahedra, with four tetrahedra meeting each crossing circle cusp, two in each square of Lemma 6.1.*

The square bases are glued as follows. The square on the left of one cusp is glued to the square on the left of the other cusp by reflecting across the diagonal of negative slope. The other square, on the right of the first cusp, is glued to the square on the right of the other cusp by reflecting across the diagonal of positive slope, as shown in Figure 15.

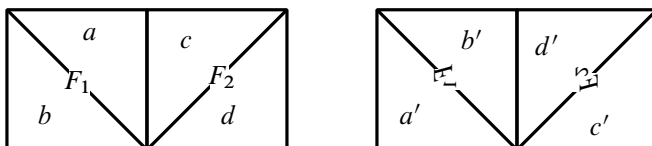


Figure 15: How to glue the bases of the crossing circle cusps of fully augmented links with two crossing circles. Note, we could choose the opposite diagonals instead.

**Proof** The gluing is obtained by considering squares at the base of the octahedra in Figure 14. For the cusp shown in Figure 14, middle right, one square base has as vertices the points of intersection of circles  $A \cap C$ ,  $A \cap D$ ,  $B \cap D$ , and  $B \cap C$  in anticlockwise order. This is glued to a square in the opposite cusp meeting the same points of intersection. Note that the points  $A \cap C$ ,  $A \cap D$ ,  $B \cap D$ , and  $B \cap C$  are now in clockwise order in the cusp on the right, with  $A \cap C$  and  $B \cap D$  in the same location in both. It follows that the squares are glued by a reflection in the negative diagonal.

The other square of Figure 14, middle right, has as vertices the points of intersection  $A' \cap C'$ ,  $B' \cap C'$ ,  $B' \cap D'$ , and  $A' \cap D'$  in anticlockwise order. It is glued to a square on the right with the same vertices, but now  $A' \cap C'$ ,  $B' \cap C'$ ,  $B' \cap D'$ , and  $A' \cap D'$  are in clockwise order on the right, with  $A' \cap C'$  and  $B' \cap D'$  in the same location. Thus the squares are glued by a reflection in the positive diagonal.

To triangulate, choose both positive diagonals or both negative diagonals in the cusp of Figure 14, middle right. This splits the two squares into four triangles; either choice of diagonal will do, but we will choose the same diagonal in each square (as opposed to Figure 15 where different diagonals are marked). There are four tetrahedra lying over the four triangles in this cusp. Under the gluing, the diagonals and the squares are preserved, so the four triangles are mapped to four triangles in the second cusp. The four additional tetrahedra lie over these four triangles in the second cusp.  $\square$

## 7 Doubling layered solid tori

When the boundary of  $M$  is a once-punctured torus triangulated by just two ideal triangles, then we may glue a layered solid torus to  $\partial M$  to perform Dehn filling. In the case of the Borromean rings, the boundary of our manifold is a twice-punctured torus triangulated by four ideal triangles, in symmetric pairs. To perform Dehn filling, we need to modify the construction. This modification essentially appears at the end of Guéritaud and Schleimer [16] when they consider the Whitehead link. However, the construction applies much more generally than the Whitehead link application, so we walk through it carefully.

There are two different modifications required, depending on the slope we wish to Dehn fill. Consider the cover  $\mathbb{R}^2$  of the twice-punctured torus obtained by putting punctures at integral points  $\mathbb{Z}^2 \subset \mathbb{R}^2$ . Assume first that there are no half-twists, so a meridian  $\mu$  of slope  $1/0$  lifts to run from  $(0, 0)$  to  $(0, 1)$ . A longitude  $\lambda$  of slope  $0/1$  lifts to run

from  $(0, 0)$  to  $(2, 0)$ . Then any slope  $m = \ell/k = \ell\mu + k\lambda$  on the torus lifts to an arc beginning at  $(0, 0)$  and ending at  $(2k, \ell)$ .

The two modifications depend on whether  $\ell$  is even or odd. If  $\ell$  is odd, the lift of the slope  $\ell/k$  will only meet the points of  $\mathbb{Z}^2$ , which are lifts of punctures, at its endpoints. In this case, we will take a double cover of a layered solid torus.

**Lemma 7.1** *Suppose  $m = \ell/k = \ell\mu + k\lambda$  is a slope on the torus (with generators  $\mu, \lambda$  as above) such that  $\ell$  is odd, and  $\ell/k \notin \{1/0, \pm 1\}$ .*

*Consider first the layered solid torus  $X$ , constructed as follows. Begin in the Farey triangle with vertices  $(1/0, 0/1, \pm 1/1)$  and step to the triangle with slope  $\ell/2k$ , building the corresponding layered solid torus  $X$  as in Section 3.*

*Let  $Y$  be the double cover of  $X$ . Then  $Y$  satisfies the following properties.*

- *The boundary of  $Y$  is a twice-punctured torus, triangulated by four ideal triangles (in two symmetric pairs), lifting to give a triangulation of the cover  $\mathbb{R}^2$ . The basis slope  $\lambda$  lifts to run from  $(0, 0)$  to  $(2, 0)$  in  $\mathbb{R}^2$ , and projects to run twice around the slope  $0/1$  in  $\partial X$ . The slope  $\mu$  lifts to run from  $(0, 0)$  to  $(0, 1)$  in  $\mathbb{R}^2$ . Diagonals of the triangulation of  $\partial Y$  have positive or negative slope, depending on whether  $m$  is positive or negative.*
- *The meridian of  $Y$  is the slope  $m = \ell\mu + k\lambda$ .*

**Proof** Let  $X$  denote the layered solid torus with a boundary triangulation that includes the slopes  $0/1$  and  $1/0$  and a diagonal  $\pm 1/1$ , with sign agreeing with the sign of  $m$ , and with meridian  $\ell/2k$ . Note that since  $\ell$  is odd and  $\ell \notin \{1/0, \pm 1\}$ , we have that  $\ell/2k \notin \{0, 1/0, \pm 1, \pm 2, \pm 1/2\}$ , which were the excluded slopes for building a layered solid torus in Section 3.

Let  $Y$  denote the double cover of  $X$ . The double cover of a solid torus is a solid torus, and the once-punctured torus boundary lifts to a twice-punctured torus, with triangles lifting to triangles. We need to show that the slopes behave as claimed.

First, the slope  $1/0$  and the meridian  $\ell/2k$  of  $X$  have geometric intersection number  $|1 \cdot 2k - \ell \cdot 0| = |2k|$ , which is even, and thus  $1/0$  is homotopic to an even power of the core of the solid torus  $X$ . Thus in the double cover  $Y$ , the slope  $1/0$  lifts to a closed curve. As  $1/0$  is an edge of a triangle on  $\partial X$ , it will remain an edge of a triangle on  $\partial Y$ , and lift to a generator of the fundamental group denoted by  $\mu$ . We may take this to run from  $(0, 0)$  to  $(0, 1)$  in  $\mathbb{R}^2$ .

Next, the curve  $0/1$  meets  $\ell/2k$  a total of  $|0 \cdot 2k - \ell \cdot 1| = |\ell|$  times on  $\partial X$ , which is odd. Therefore it lifts to an arc rather than a closed curve on  $\partial Y$ , with endpoints on distinct punctures. Thus a second generator of the fundamental group of  $\partial Y$  is given by taking two lifts of  $0/1$ , end to end. Denote this generator by  $\lambda$ . Its lift runs from  $(0, 0)$  to  $(2, 0)$  in  $\mathbb{R}^2$ .

Finally we check that the meridian of  $Y$  is the slope  $m$ , written in terms of  $\mu$  and  $\lambda$  as claimed. In  $X$ , the curve  $\ell/2k$  bounds a disc. This lifts to bound a disc in  $Y$  as well. However, note the lift runs  $\ell$  times along  $\mu$  and  $k$  times along  $\lambda$ . Thus the meridian slope is as claimed.  $\square$

If  $\ell$  is even, say  $\ell = 2s$  for some integer  $s$ , the lift of  $m = \ell/k$  to  $\mathbb{R}^2$  is an arc running from  $(0, 0)$  through  $(k, s) \in \mathbb{Z}^2$  to  $(2k, 2s) \in \mathbb{Z}^2$ . Thus it meets a lift of a puncture in its interior. In this case, taking a double cover of a layered solid torus will not suffice. Instead, we need to give a different construction.

**Construction 7.2** Let  $m = \ell/k$  be a slope such that  $\ell$  is even, say  $\ell = 2s$ , and  $m \notin \{0/1, \pm 2/1\}$ . Let  $(T_0, \dots, T_N)$  be a sequence of triangles in the Farey triangulation where  $T_0$  is a triangle with slopes  $0/1$ ,  $1/0$ , and either  $1/1$  or  $-1/1$ , with sign agreeing with the sign of  $m$ , and  $T_N$  is a triangle with slopes  $u$ ,  $t$ , and  $s/k$ .

Start with the triangulation of the twice-punctured torus consisting of two side-by-side copies of the slopes of  $T_0$ . More precisely, fill  $\mathbb{R}^2 - \mathbb{Z}^2$  with unit squares with diagonals matching that of  $T_0$ , and quotient by  $(x, y) \mapsto (x + 2, y)$  and  $(x, y) \mapsto (x, y + 1)$ .

Inductively, for the  $j^{\text{th}}$  step across an edge in the Farey triangulation, attach two ideal tetrahedra to the twice-punctured torus, effecting two identical diagonal exchanges with the slopes of  $T_{j-1}$ , and producing a triangulation of a space homotopy equivalent to the product of the interval and the twice-punctured torus, with one boundary triangulated by two side-by-side copies of  $T_0$ , and the other triangulated by two side-by-side copies of  $T_j$ . Note that so far, this is identical to the procedure for the layered solid torus, only we are taking two copies of each tetrahedron instead of just one; see Figure 16.

This time, continue until  $j = N$ , so one boundary is labelled by slopes  $u$ ,  $t$ , and  $s/k$ , repeated twice in each of two parallelograms lying side-by-side. Now obtain a solid torus as follows. First, identify the two slopes  $s/k$  in the two boundary triangles. Then fill the remaining space with a single tetrahedron whose four faces are glued to the inner faces.

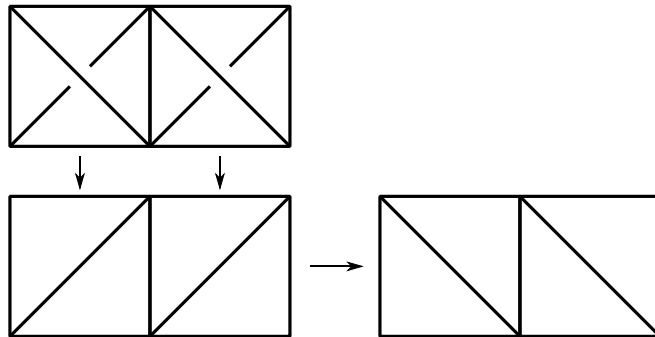


Figure 16: To create a solid torus with boundary a 2-punctured torus, at each step layer two identical tetrahedra onto the current boundary triangulation, effecting a diagonal exchange.

**Remark 7.3** To construct a layered solid torus in Section 3, we needed to exclude slopes in  $T_0$  and  $T_1$  in the Farey graph. It is no longer necessary to exclude slopes in  $T_1$  for the previous construction, because of the addition of extra tetrahedra corresponding to  $j = N$  and a final tetrahedron after identifying diagonals of the tetrahedra corresponding to  $j = N$ . We still exclude slopes in  $T_0$ .

**Lemma 7.4** Suppose  $m = \ell/k = \ell\mu + k\lambda$  is such that  $m = \ell/k \notin \{0/1, \pm 2/1\}$  and  $\ell$  is even. Then the triangulated space  $Y$  constructed as above, by taking a sequence of side-by-side tetrahedra and attaching a final tetrahedron at the core, forms a solid torus satisfying:

- The boundary of  $Y$  is a twice-punctured torus, triangulated by four triangles (in symmetric pairs), with basis slopes  $\mu$  running over one edge of a triangle, lifting to run from  $(0, 0)$  to  $(0, 1)$  in  $\mathbb{R}^2$ , and  $\lambda$  running over two edges (and two punctures), lifting to run from  $(0, 0)$  to  $(2, 0)$  in  $\mathbb{R}^2$ . Diagonal edges of the triangulation have positive or negative slope, where the sign is determined by the sign of  $m$ .
- The meridian of the solid torus  $Y$  is the slope  $m = \ell\mu + k\lambda$ .

**Proof** Generators for the fundamental group of the boundary torus consist of edges  $1/0$  in the original triangulation, and two copies of  $0/1$  by construction. Denote the first generator by  $\mu$  and the second, consisting of two edges, by  $\lambda$ .

The meridian of the solid torus is the curve that is homotopically trivial. This is the curve formed by pinching together two edges of slope  $s/k$  on the inside boundary of

the triangulated space. Thus it runs twice over this edge. In terms of the generators  $\mu$  and  $\lambda$ , the curve running twice over the edge  $s/k$  of the innermost twice-punctured torus has slope  $2s\mu + k\lambda$  or  $\ell\mu + k\lambda$ .  $\square$

**Remark 7.5** (symmetry of Construction 7.2) Notice that the solid torus  $Y$  from Construction 7.2 will admit an involution. This involution takes the innermost tetrahedron to itself (setwise), and for each of the other tetrahedra in the construction, it swaps the two tetrahedra that were layered together, corresponding to the same triangle in the Farey graph. We will give this solid torus an angle structure. We will choose the angles to be preserved under this involution. Thus, although tetrahedra are layered in pairs, and although there are two punctures on the boundary, the two cusp triangulations will be identical, with angles on any tetrahedron agreeing with the angles on its image under the involution.

**Lemma 7.6** *Let  $V$  be a solid torus with twice-punctured torus boundary, and with triangulation either as in Lemma 7.1 or Lemma 7.4, depending on whether the meridian  $m = \ell\mu + k\lambda$  has  $\ell$  even or odd. We also assume  $m \notin \{0/1, 1/0, \pm 1/1, \pm 2/1\}$ .*

*Let  $\{\theta_p, \theta_q, \theta_r\}$  be exterior dihedral angles along edges of the twice-punctured torus, where each is repeated twice symmetrically, such that*

$$0 < \theta_r < \pi, \quad -\pi < \theta_p, \theta_q < \pi, \quad \text{and} \quad \theta_p + \theta_q + \theta_r = \pi.$$

*Suppose also that in the case  $\ell$  is odd, intersection numbers satisfy*

$$i(p, \ell/2k)\theta_p + i(q, \ell/2k)\theta_q + i(r, \ell/2k)\theta_r > 2\pi.$$

*Then there exists an angle structure on the triangulated solid torus with these exterior angles. Conversely, if an angle structure exists and  $\ell$  is odd, then the intersection numbers satisfy the above equation.*

**Proof** The case of the double cover of a layered solid torus follows from the same result for usual layered solid tori, Proposition 4.3. In this case, the angle structure exists for the layered solid torus; lift the angles to the double cover to obtain the result. This gives the proof when  $\ell$  is odd.

In the case that  $\ell$  is even, we work with the side-by-side solid torus. For every pair of tetrahedra layered on at the  $i^{\text{th}}$  step of Construction 7.2, where  $1 \leq i \leq N$ , label the dihedral angles of both tetrahedra by  $x_i$ ,  $y_i$ , and  $z_i$  with  $x_i + y_i + z_i = \pi$ . Similarly for the last tetrahedron, label its angles  $x_{N+1}$ ,  $y_{N+1}$ , and  $z_{N+1}$  with



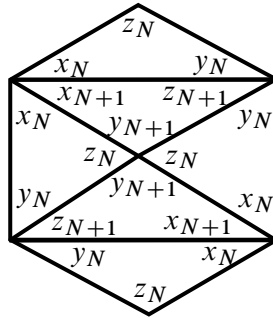


Figure 17: The innermost triangles of one of the two hexagons forming a cusp triangulation of a side-by-side solid torus.

$x_{N+1} + y_{N+1} + z_{N+1} = \pi$ . As in Section 4, we build an angle structure on the side-by-side layered solid torus by finding a sequence  $(z_0 = \pi + \theta_p, z_1 = \pi - \theta_r, \dots, z_N, z_{N+1})$  with  $z_i \in (0, \pi)$ .

The side-by-side layered solid torus has two ideal vertices. The cusp triangulation of each of the two ideal vertices is still a sequence of hexagons, and these will be identical because of the symmetry of the solid torus. The cusp triangulations are constructed exactly as in the case of the layered solid torus for the first  $N$  steps; these agree with Figure 11. But the cusp triangulation differs from that of the usual layered solid torus at the innermost hexagon, where the last tetrahedron is attached.

Because cusp triangulations agree before the last step, the dihedral angles  $x_i, y_i,$  and  $z_i$  for  $2 \leq i \leq N$  satisfy the same conditions of Lemma 4.10. Similarly,  $z_0 = \pi + \theta_p, z_1 = \pi - \theta_r, x_1,$  and  $y_1$  satisfy the conditions of Lemma 4.9, with the same cusp pictures. Therefore, just as in Lemma 4.12, for  $1 \leq i \leq N$  the sequence satisfies

$$\begin{cases} z_{i-1} > z_i + z_{i+1} & \text{if } \Delta_i \text{ and } \Delta_{i+1} \text{ are labelled } RL \text{ or } LR \text{ (hinge condition),} \\ z_{i-1} + z_{i+1} > 2z_i & \text{if } \Delta_i \text{ and } \Delta_{i+1} \text{ are labelled } RR \text{ or } LL \text{ (convexity condition).} \end{cases}$$

Moreover, for the first two tetrahedra,

$$\pi - \theta_p - 2\theta_r < z_2 < \pi - \theta_p.$$

Recall that to make the slope  $m$  trivial in the side-by-side case, we identify the two edges of slope  $(\ell/2)/k$  in the pair of tetrahedra at the  $N^{\text{th}}$  step. This corresponds to identifying a pair of opposite vertices in the innermost hexagon; see Figure 17.

As in the proof of Lemma 4.11, the sum of the interior angles of the hexagon is  $4\pi$ . This gives the equation

$$(2x_N + x_{N+1}) + (2y_N + z_{N+1}) + z_N = 2\pi \quad \text{or} \quad z_N = z_{N+1} + x_{N+1}.$$

Since  $x_{N+1} > 0$ , this implies that  $0 < z_{N+1} < z_N$ . Thus for angle structure to exist on the side-by-side layered solid torus, we require that  $0 < z_{N+1} < z_N$ . Finally, note again that by a downward induction, convexity and hinge equations imply that the sequence is strictly decreasing.

As in Lemmas 4.15 and 4.17, we find a sequence  $(z_0, z_1, \dots, z_N, z_{N+1})$  satisfying the above conditions. In the layered solid torus case, we had to split into two cases, depending on whether or not a hinge existed. When there was no hinge, we needed a convex sequence with the last term equal to zero. However, in this case, we need a convex sequence with  $z_{N+1} > 0$ , which can always be arranged. Thus we only need to show that a sequence satisfying the above requirements exists.

To find such a sequence, the same argument of Lemma 4.17 can be used. Namely, let  $h \in \{2, 3, \dots, N+1\}$  be the smallest index such that  $\Delta_h$  is a hinge of the form  $RL$  or  $LR$ , or set  $h = N+1$  if no such index exists. Set  $z_0 = \pi + \theta_p$  and  $z_1 = \pi - \theta_r$ . We can choose inductively a positive decreasing sequence  $z_k$  such that  $z_k > 2z_{k-1} - z_{k-2}$  for  $1 \leq k \leq h$ .

The rest of the sequence  $z_i$  is constructed backwards from  $i = N+1$  to  $i = h$ . Consider a sequence  $z'_i$ . Set  $1 > z'_{N+1} > 0$  and  $z'_N = 1$ . For each  $i$  such that  $N-1 \geq i \geq h+1$ , inductively choose  $z'_i$  such that  $z'_i > z'_{i+1} + z'_{i+2}$  or  $z'_i + z'_{i+2} > 2z'_{i+1}$ , depending on whether  $\Delta_{i+1}$  is a hinge or not.

Now choose  $\epsilon$  such that  $0 < \epsilon < (z_{h-1} - z_h)/z'_{h+1}$ . Set  $z_i = \epsilon z'_i$  for  $h+1 \leq i \leq N+1$ . The sequence  $z_i$  satisfies the required inequalities for  $i < h$  and  $i > h$ . Because of our choice of  $\epsilon$ , it also satisfies the hinge condition  $z_h < z_{h-1} - z_{h+1}$  or  $z_h < z_{h-1} - \epsilon z'_{h+1}$ . Thus we have found a sequence giving an angle structure.  $\square$

**Lemma 7.7** *Let  $V$  be the triangulated solid torus of Lemma 7.6. In particular, the meridian  $m$  of  $V$  is not one of  $\{0/1, 1/0, \pm 1/1, \pm 2/1\}$ . If the volume functional takes its maximum on the boundary of the space of angle structures, then all tetrahedra in the solid torus must be flat. Thus the volume functional is maximised in the interior.*

**Proof** Suppose the volume functional is maximised on the boundary for the double cover  $V$  of a layered solid torus. There is a symmetry of the triangulated solid torus  $V$  swapping triangles in pairs, changing the basepoint of the double covering. If the volume is maximised at a point in which angles are not preserved by this symmetry, then applying the symmetry gives two distinct maxima, contradicting the fact that the volume functional is convex. Thus a point of maximum for the double cover descends to a maximal point for the original layered solid torus. By Lemma 4.18, if the volume

functional is maximised on the boundary, all tetrahedra are flat. Thus the maximum is in the interior in this case.

For the side-by-side triangulation, a point on the boundary of the space of angle structures corresponds to a sequence  $(z_0, z_1, \dots, z_{N+1})$  satisfying hinge and convexity equations, but now with weak inequalities. This must be a nonincreasing sequence. Suppose the  $i^{\text{th}}$  tetrahedron is the first flat tetrahedron. Then  $z_i \in \{0, \pi\}$  but  $z_{i-1} \in (0, \pi)$  unless  $i - 1 = 0$ . If  $z_i = \pi$ , then convexity implies  $z_j = \pi$  for  $j = i + 1, \dots, h$ , where  $h$  is the next hinge index. The hinge condition then implies that all later  $z_j = 0$ . Similarly, if  $z_i = 0$  then all later  $z_j = 0$ .

Now consider  $z_{i-1}$ . We have  $x_i, y_i, z_i, z_{i+1} \in \{0, \pi\}$ . Thus by one of the formulas determining angles, as in Lemma 4.10,  $z_{i-1} = 2\pi$ . But  $0 < z_{i-1} < \pi$  unless  $i - 1 = 0$ . If  $i - 1 = 0$ , the first flat tetrahedron is the first tetrahedron, so the entire solid torus consists of flat tetrahedra.

There is one final thing to check. Above, we have restricted to angle structures in which a tetrahedron and its image under the involution preserving the side-by-side solid torus are given the same angles. We have shown that under this restriction, volume is maximised in the interior. However, note that if the volume were maximised for an angle structure on  $Y$  that did not have symmetric angles, then applying the involution would give a distinct angle structure on  $Y$  with the same volume, contradicting the fact that the volume functional has a unique maximum. Hence the maximum of the volume functional must occur at an angle structure that is preserved under our involution.  $\square$

**Lemma 7.8** *Let  $V$  be the triangulated solid torus of Lemma 7.6. In particular, the meridian  $m$  of  $V$  is not one of  $\{0/1, 1/0, \pm 1/1, \pm 2/1\}$ . Suppose  $V$  is a subset of a triangulation of a 3-manifold  $M$  such that the volume functional on  $M$  is maximised at a point in which a tetrahedron of  $V$  is flat. Then the exterior dihedral angles  $(\theta_p, \theta_q, \theta_r)$  of  $V$  are equal to one of  $(\pi, -\pi, \pi)$ ,  $(-\pi, \pi, \pi)$ ,  $(\pi, 0, 0)$ , or  $(0, \pi, 0)$ . Conversely, if the exterior dihedral angles satisfies one of these choices, then all tetrahedra in  $V$  are flat.*

**Proof** By Lemma 7.7, if one of the tetrahedra in the solid torus is flat then all of the tetrahedra in the solid torus are flat.

Next observe that the exterior dihedral angles of the solid torus satisfy  $\theta_r = \pi - z_1$ ,  $\theta_q = \pi - (z_2 + 2x_1)$  or  $\theta_q = \pi - (z_2 + 2y_1)$ , and  $\theta_p = \pi - \theta_q - \theta_r$ , where  $z_1, x_1$ , and  $y_1$  are the dihedral angles of the outermost tetrahedron in the flat layered solid torus, and  $z_2$  is an angle in the next outermost tetrahedron. Exactly one of  $x_1, y_1$ , or  $z_1$  is  $\pi$ , and the other two are 0.

If  $z_1 = 0$ , then  $z_2 = 0$  because the  $z_i$  form a nonincreasing nonnegative sequence. Then  $\theta_r = \pi$  and  $\theta_q$  is either  $\pi$  or  $-\pi$ , depending on  $x_1$  and  $y_1$ . Since  $\theta_p + \theta_q + \theta_r = \pi$ , this implies that  $(\theta_p, \theta_q, \theta_r)$  equals  $(\pi, -\pi, \pi)$  or  $(-\pi, \pi, \pi)$ .

Now suppose  $z_1 = \pi$ . Then  $x_1$  and  $y_1$  equal 0, and  $z_2$  is 0 or  $\pi$ . In this case,  $\theta_r = \pi - z_1 = 0$ ,  $\theta_q$  is 0 or  $\pi$  depending on  $z_2$ , and  $\theta_p = \pi - \theta_r - \theta_q$  is  $\pi$  if  $\theta_q = 0$  or 0 if  $\theta_q = \pi$ . Therefore,  $(\theta_p, \theta_q, \theta_r)$  equals  $(\pi, 0, 0)$ , or  $(0, \pi, 0)$ .

Conversely, suppose the exterior dihedral angles are as in the lemma. Then  $z_1 = \pi - \theta_r$  must be 0 or  $\pi$ . In either case, the angles of the outermost tetrahedron must then be a permutation of  $(0, 0, \pi)$ , and hence that tetrahedron is flat. By Lemma 7.7, if any tetrahedron is flat, all tetrahedra are flat.  $\square$

### 7.1 The case of a half-twist

In the case that a crossing circle encircles a half-twist, the cusp is still formed from two squares, but a meridian of the crossing circle cusp lifts to  $\mathbb{R}^2$  to run from  $(0, 0)$  to  $(1, 1)$ .

Suppose first that  $m$  is positive. Then apply a homeomorphism to the fully augmented link complement that reverses the direction of the crossing. The meridian  $\mu = 1/0$  of this new link complement lifts to run from  $(0, 0)$  to  $(-1, 1)$ , and the longitude  $\lambda = 0/1$  to run from  $(0, 0)$  to  $(2, 0)$ . The vertical line from  $(0, 0)$  to  $(0, 1)$  is a lift of the slope  $1/1$ . Now given  $\ell$  and  $k$  relatively prime, perform the construction of the solid torus in Lemma 7.1 or Lemma 7.4 depending on whether  $\ell$  is even or odd. Only now, lift  $\mu$  to the curve running from  $(0, 0)$  to  $(-1, 1)$  in  $\mathbb{R}^2$ . As the lift is purely topological, this gives a solid torus that can be used to perform the Dehn filling just as before. Moreover, Lemmas 7.6 and 7.7 still hold with their proofs unchanged in this case.

If  $m$  is negative, then the meridian  $\mu = 1/0$  of the crossing circle cusp lifts to run from  $(0, 0)$  to  $(1, 1)$ , the longitude  $\lambda = 0/1$  lifts to run from  $(0, 0)$  to  $(2, 0)$ , and the slope  $-1/1$  lifts to run from  $(0, 0)$  to  $(0, 1)$ . Given  $m = \ell/k$ , again perform the construction of the solid torus of Lemma 7.1 or Lemma 7.4, depending on whether  $\ell$  is even or odd, and lift  $\mu$  to the curve running from  $(0, 0)$  to  $(1, 1)$  in  $\mathbb{R}^2$ . Again Lemmas 7.6 and 7.7 hold with proofs unchanged to give the required angle structures, with volume maximised in the interior. Thus the construction works equally well with or without a half-twist.

### 7.2 A vertical construction

Finally, the above work is sufficient to perform all Dehn fillings on the family of fully augmented links with exactly two twist regions, shown in Figure 13. However, in

Section 9, we will need to extend this construction to obtain a solid torus in which  $\mu = 1/0$  lifts to run from  $(0, 0)$  to  $(0, 2)$  and  $\lambda = 0/1$  lifts to run from  $(0, 0)$  to  $(1, 0)$ . We treat that case in this subsection.

Note that above, we constructed a solid torus by taking side-by-side tetrahedra, ie stacking identical tetrahedra horizontally, in the  $x$ -axis direction. More precisely, we tiled all of  $\mathbb{R}^2 - \mathbb{Z}^2$  by unit squares cut through by a diagonal, layered on tetrahedra coming from a walk in the Farey graph, and then took the quotient by  $(2\mathbb{Z}, \mathbb{Z})$ .

This construction could instead have been done by taking identical tetrahedra stacked in the  $y$ -axis direction. That is, quotient out by  $(\mathbb{Z}, 2\mathbb{Z})$ . All the results above immediately hold for this construction. In particular:

**Lemma 7.9** *Let  $m = \ell/k$  be such that  $k$  is even, and  $m \notin \{1/0, \pm 1/2\}$ . The “vertical side-by-side” solid torus, constructed by layering on tetrahedra in a path from  $T_0 = (0/1, \pm 1/1, 1/0)$  to  $T_N = (u, v, \ell/(k/2))$ , has the following properties:*

- (1) *Its boundary is a twice-punctured torus, triangulated by four ideal triangles in two symmetric pairs, with basis slopes  $\mu$  running over two edges of a triangle, lifting to run from  $(0, 0)$  to  $(0, 2)$  in  $\mathbb{R}^2$ , and  $\lambda$  running over one edge, lifting to run from  $(0, 0)$  to  $(1, 0)$  in  $\mathbb{R}^2$ .*
- (2) *The meridian of the solid torus is the slope  $m = \ell\mu + k\lambda$ .*
- (3) *The triangulated solid torus admits an angle structure, with volume functional taking its maximum in the interior.*
- (4) *In a volume-maximising structure, if there is one flat tetrahedron then all tetrahedra must be flat. Moreover, all tetrahedra are flat if and only if the exterior dihedral angles  $(\theta_p, \theta_q, \theta_r)$  are one of  $(\pi, -\pi, \pi)$ ,  $(-\pi, \pi, \pi)$ ,  $(\pi, 0, 0)$ , or  $(0, \pi, 0)$ .*

Similarly, if  $\ell/k$  is a slope such that  $k$  is odd, we may take a double cover of a layered solid torus to produce a solid torus whose boundary is a twice-punctured torus, only now with a fundamental domain that consists of two squares stacked vertically rather than horizontally:

**Lemma 7.10** *Let  $m = \ell/k$  be such that  $k$  is odd and  $m \notin \{0/1, \pm 1/1\}$ . Then the (vertical) double cover  $Y$  of the layered solid torus  $X$  constructed from a walk in the Farey graph from  $T_0 = (0/1, 1/0, \pm 1/1)$  to slope  $(2\ell)/k$  has the following properties:*

- *The boundary of  $Y$  is a twice-punctured torus, triangulated by four ideal triangles (in two symmetric pairs), lifting to give a triangulation of the cover  $\mathbb{R}^2$ . The*

basis slope  $\mu$  lifts to run from  $(0, 0)$  to  $(0, 2)$  in  $\mathbb{R}^2$ , and projects to run twice around the slope  $1/0$  in  $\partial X$ . The basis slope  $\lambda$  lifts to run from  $(0, 0)$  to  $(1, 0)$ .

- The meridian of  $Y$  is the slope  $m = \ell\mu + k\lambda$ .

**Proof** Let  $X$  be a layered solid torus with meridian slope  $2\ell/k$ , where  $k$  is odd. Let  $Y$  be the (vertical double) cover of  $X$ .

The slope from  $(0, 0)$  to  $(0, 1)$  is a generator in the solid torus  $X$ . It meets the meridian  $2\ell/k$  of  $X$  a total of  $|1 \cdot k - 0 \cdot 2\ell| = |k|$  times, which is odd. Therefore the curve from  $(0, 0)$  to  $(0, 1)$  lifts to an arc in  $\partial Y$ . Thus a generator of the fundamental group of  $\partial Y$  is given by taking two lifts of the curve from  $(0, 0)$  to  $(0, 1)$ , lined up end-to-end. Denote the resulting closed curve in  $Y$  by  $\mu$ . Its lift runs from  $(0, 0)$  to  $(0, 2)$  in  $\mathbb{R}^2$ .

The slope from  $(0, 0)$  to  $(1, 0)$  is a generator in the solid torus  $X$ . This curve and the meridian of  $X$ , of slope  $2\ell/k$ , have geometric intersection number  $|0 \cdot k - 2 \cdot \ell| = |2\ell|$ , which is even. Thus the curve from  $(0, 0)$  to  $(1, 0)$  is homotopic to an even power of the core of  $X$ . Therefore a second generator of the fundamental group of  $\partial Y$  is given by taking the lift of the curve from  $(0, 0)$  to  $(1, 0)$ . Denote this generator by  $\lambda$ . Its lift runs from  $(0, 0)$  to  $(1, 0)$  in  $\mathbb{R}^2$ .

The meridian  $2\ell/k$  of  $X$  lifts to bound a disc in  $Y$ . Note that the lift runs  $2\ell$  times along  $\mu$  and  $k$  times along  $\lambda$ . So in the basis for  $Y$  the meridian has the form  $\ell/k = \ell\mu + k\lambda$ .  $\square$

## 8 Dehn filling the Borromean rings

In this section we finish the proof that triangulations of Dehn fillings of the crossing circles of the Borromean rings are geometric, for appropriate choices of slopes, and similarly for the other fully augmented link complements shown in Figure 13.

**Lemma 8.1** *Let  $M$  be one of the fully augmented link complements with exactly two crossing circles. Let  $m_1$  and  $m_2$  be slopes such that*

$$m_1, m_2 \notin \{0/1, 1/0, \pm 1/1, \pm 2/1\}.$$

*Then the Dehn filling of  $M$  on its crossing circle cusps along slopes  $m_1$  and  $m_2$ , denoted by  $M(m_1, m_2)$ , admits a topological triangulation built by gluing together two triangulated solid tori that are both double covers of layered solid tori, one double cover and one solid torus with the side-by-side construction of Lemma 7.4, or two solid tori of that form.*

**Proof** The slopes  $m_1$  and  $m_2$  on the two crossing circle cusps each determine a triangulation of a solid torus by Lemma 7.1 or Lemma 7.4. To perform Dehn filling, we remove interiors of all edges, faces, and tetrahedra meeting a crossing circle cusp, and replace them by one of these two triangulated solid tori.

Removing interiors of edges, faces, and tetrahedra meeting the two crossing circle cusps removes all but two squares from the manifold, namely the squares shown in Figure 15, identified as shown in Lemma 6.2. We wish to attach triangulated solid tori to these squares. There are two cases to consider.

**Case 1** (the triangulations on the boundary of the solid tori agree) If  $M$  is the Borromean rings complement, this is the case that the slopes  $m_1$  and  $m_2$  are both positive, or the case that the slopes  $m_1$  and  $m_2$  are both negative. In this case, the squares of Figure 15 are triangulated by the same diagonals. If  $M$  has one or two half-twists, again the sign of  $m_1$  and  $m_2$  determine the diagonals. This is the case that the choice of triangulation gives the same diagonals.

In this case, the corresponding solid tori have boundary triangulation that matches the triangulation on the squares. Moreover, when we attach the two solid tori to opposite sides of the squares, following the gluing instructions of Figure 15, their triangulations match the given triangulations, giving a topological triangulation of the Dehn filling.

**Case 2** (the triangulations disagree) For example, this is the case that  $m_1$  and  $m_2$  have opposite signs when  $M$  is the Borromean rings complement.

The triangulations of the solid tori will induce triangulations of the two squares with opposite diagonals. To glue these together, we add two identical extra tetrahedra between the two squares, with the tetrahedra effecting a diagonal exchange on the squares.

Because both tetrahedra are attached along exactly two faces to the second solid torus, in fact they form an extra layer on the solid torus, equivalent to changing the initial triangulation on the boundary in the construction from  $(0, 1, 1/0)$  to  $(0, -1, 1/0)$ , or vice versa. Thus we glue these two tetrahedra to the second solid torus. Then the second solid torus has the form of Lemma 7.1 or Lemma 7.4, except with triangulation on its boundary consisting of diagonals having opposite sign from the slope.  $\square$

**Lemma 8.2** Suppose the exterior dihedral angles on the  $i^{\text{th}}$  solid torus are denoted by  $\alpha_i$  for the diagonal edges, and  $\theta_i$  for the horizontal edges, for  $i = 1, 2$ . If an angle structure exists on  $M(m_1, m_2)$  then  $\alpha_1 = -\alpha_2$  and

$$\theta_1 = \theta_2 + \alpha_2, \quad \theta_2 = \alpha_1 + \theta_1.$$

**Proof** Angle structures for each of the triangulated solid tori will come from Lemma 7.6, once we decide on exterior dihedral angles. Angle structures on the solid tori will induce an angle structure on the entire manifold if and only if the edge equations are satisfied for edges that lie on the two squares, on the boundaries of the solid tori.

Each solid torus has six edges on its boundary, and we assign three angles, giving a pair of symmetric edges the same angle. On the first solid torus, the horizontal edges all have the same exterior angle, denoted by  $\theta_1$ . The diagonal edges have the same exterior angle, denoted by  $\alpha_1$ , and the vertical edges have the same exterior angle, which must be  $\pi - \alpha_1 - \theta_1$ . Similarly denote the horizontal, diagonal, and vertical exterior angles on the second solid torus by  $\theta_2$ ,  $\alpha_2$ , and  $\pi - \theta_2 - \alpha_2$ , respectively.

Diagonal edges are glued to diagonal edges. Thus the sum of interior angles satisfies

$$(\pi - \alpha_1) + (\pi - \alpha_2) = 2\pi \quad \text{or} \quad \alpha_1 = -\alpha_2.$$

Horizontal edges with exterior angle  $\theta_1$  are identified to vertical edges with exterior angle  $\pi - \theta_2 - \alpha_2$ ; both vertical edges in the second solid torus lie in this edge class. Similarly, horizontal edges with exterior angle  $\theta_2$  are identified to the vertical edges of the first solid torus. Thus the sum of interior angles around these two edge classes satisfy

$$2(\pi - \theta_1) + 2(\theta_2 + \alpha_2) = 2\pi \quad \text{or} \quad \theta_1 = \theta_2 + \alpha_2,$$

and

$$2(\pi - \theta_2) + 2(\theta_1 + \alpha_1) = 2\pi \quad \text{or} \quad \theta_2 = \theta_1 + \alpha_1. \quad \square$$

**Lemma 8.3** *Let  $M$  denote the complement of a fully augmented link with two crossing circles, as in Figure 13, and let  $M(m_1, m_2)$  denote its Dehn filling along slopes  $m_1$  and  $m_2$  on the crossing circle cusps. If  $m_1$  and  $m_2$  satisfy*

$$m_1, m_2 \notin \{0/1, 1/0, \pm 1/1/, \pm 2/1\},$$

*then  $M(m_1, m_2)$  admits an angle structure.*

**Proof** If an angle structure exists, it must satisfy the equations of Lemma 8.2. In addition, there is an angle structure on each of two solid tori, with exterior angles denoted by  $\theta_{p_1}$ ,  $\theta_{q_1}$ , and  $\theta_{r_1}$  in the first solid torus, satisfying

$$(8.4) \quad \theta_{p_1} + \theta_{q_1} + \theta_{r_1} = \pi, \quad -\pi < \theta_{p_1}, \theta_{q_1} < \pi, \quad 0 < \theta_{r_1} < \pi,$$

and denoted by  $\theta_{p_2}$ ,  $\theta_{q_2}$ , and  $\theta_{r_2}$  in the second solid torus, satisfying similar conditions.



In addition, in the special case that  $m_i = \ell_i/k_i$  and  $\ell_i$  is odd, and moreover the path in the Farey graph from the initial triangulation has no hinges, then we require the slopes  $p_i, q_i, r_i$  and  $\ell_i/(2k_i)$  to satisfy the intersection condition (4.5).

Provided we can find any angles that simultaneously satisfy all the above, we will have proved the lemma.

The difficulty is that the angles  $\theta_{p_i}, \theta_{q_i},$  and  $\theta_{r_i}$  have different relationships with angles  $\alpha_i, \theta_i$  and  $\pi - \alpha_i - \theta_i$  depending on whether  $m_1$  and  $m_2$  have the same or different sign, and on whether a crossing circle has a half-twist.

Suppose first that there are no half-twists, ie  $M$  is the Borromean rings complement, and that  $m_1$  and  $m_2$  have the same sign; for concreteness, say they are both positive. In this case, we build the two solid tori corresponding to  $m_1$  and  $m_2$  by starting in the triangle in the Farey graph with vertices 0, 1, and  $1/0$ , and go up. In this case, 1 is not covered in the first step, so  $r = 0$  or  $r = 1/0$ , and 1 is the slope  $p$  or  $q$ . If both solid tori have hinges, then the intersection condition will either be automatically satisfied for the layered solid tori we construct, or it is unnecessary for the side-by-side tori we construct. So the more difficult remaining case is when  $m_1 = \ell_1/k_1$  with  $\ell_1$  odd, and there are no hinges in the path from the triangle  $(0, 1, 1/0)$  to  $\ell_1/(2k_1)$ , and similarly for  $\ell_2/(2k_2)$ .

No hinges means the slope  $\ell_i/(2k_i)$  is of the form  $1/2n$  for  $n$  a positive integer, or of the form  $n/1$ ; the second is impossible because  $1 \neq 2k_i$ . Thus the path in the Farey graph consists only of copies of  $L$ , and goes from  $(0, 1, 1/0)$  to  $1/(2k_i)$ . Then we know that  $r_i$ , the first slope covered, corresponds to  $1/0$ , which is the vertical edge in the initial triangulation, labelled with exterior angle  $\pi - \alpha_i - \theta_i$ . The slope  $p_i$ , the second slope covered, corresponds to  $1/1$ , which is the diagonal edge in the initial triangulation, labelled with exterior angle  $\alpha_i$ . Thus the slope  $q_i$  corresponds to 0, the horizontal edge, with exterior angle  $\theta_i$ . The required equations then become  $\theta_i = \theta_{q_i}, \alpha_i = \theta_{p_i},$  and  $\pi - \theta_i - \alpha_i = \theta_{r_i},$  satisfying (8.4), as well as intersection conditions

$$i(1/(2k_i), 1/1)\alpha_i + i(1/(2k_i), 0/1)\theta_i + i(1/(2k_i), 1/0)(\pi - \alpha_i - \theta_i) > 2\pi$$

or

$$(2k_i - 1)\alpha_i + \theta_i + 2k_i(\pi - \alpha_i - \theta_i) > 2\pi.$$

There are many solutions to these equations. For example, set  $\alpha_1 = \pi/6$  and  $\theta_1 = 5\pi/9$ , so  $\pi - \alpha_1 - \theta_1 = 5\pi/18$ , and  $\alpha_2 = -\alpha_1 = -\pi/6$  and  $\theta_2 = \alpha_1 + \theta_1 = 13\pi/18$ , so  $\pi - \alpha_2 - \theta_2 = 4\pi/9$ . This gives an angle structure, as desired.

The case that both  $m_1$  and  $m_2$  are negative is similar.

Next suppose there are no half-twists, but  $m_1$  and  $m_2$  have opposite signs; say  $m_1 > 0$  and  $m_2 < 0$ . Then we insert an extra tetrahedron onto the first solid torus. Thus the construction of this solid torus now starts at the Farey triangle  $(0, -1, 1/0)$  and immediately crosses into the triangle  $(0, 1, 1/0)$ . It follows that  $r = -1$ , corresponding to exterior angle  $\alpha_1$ . The only case in which the intersection condition comes up is if the path in the Farey triangulation only steps  $L$ , and the slope  $m_1$  equals  $1/(2k_1)$ . Then in this case,  $p = 1/0$ , corresponding to exterior angle  $\pi - \alpha_1 - \theta_1$ , and  $q = 0$ , corresponding to exterior angle  $\theta_1$ . The intersection condition is similar to above, the only difference is that we need to ensure we have a solution in which  $\alpha_1$  now lies strictly between  $0$  and  $\pi$ . But notice we already found such a solution in the previous case. Thus the same angles in the previous case still work to give an angle structure in this case. Notice that  $\alpha_2 < 0$ , but because  $m_2 < 0$  as well,  $\alpha_2$  does not correspond to the exterior angle on slope  $r_2$ , the first slope covered, and so  $-\pi < \alpha_2 < 0$  works in this case.

In the case that there is one half-twist, the half-twist changes the names of the slopes in the framing: the meridian of the unfilled manifold is now a diagonal and the longitude runs over two horizontal segments. However, we still assign the same exterior angles  $\alpha_1$  to the diagonal and  $\theta_1$  to the horizontal. In fact, this gives the same required equations as above, both in the case of  $m_1$  and  $m_2$  having the same sign, and  $m_1$  and  $m_2$  having opposite signs, and so the same choices of angles will give an angle structure.

Finally, when there are two half-twists, again we change the framing on both solid tori, but ensuring the triangulations match up will again give the same required equations, and so the solution above always gives an angle structure.  $\square$

**Lemma 8.5** *Let  $M$  denote the complement of one of the fully augmented links with two crossing circles, shown in Figure 13, and let  $M(m_1, m_2)$  denote its Dehn filling along slopes  $m_1$  and  $m_2$  on the crossing circle cusps. Suppose  $m_1$  and  $m_2$  satisfy*

$$m_1, m_2 \notin \{0/1, 1/0, \pm 1/1/, \pm 2/1\}.$$

*Then, for the space of angle structures on the triangulation of  $M(m_1, m_2)$  from above, the volume functional takes a maximum in the interior.*

**Proof** Consider a point on the boundary of the space of angle structures. Because it is on the boundary, it contains a flat tetrahedron, with angles  $0, 0$ , and  $\pi$ . Because the triangulation of  $M(m_1, m_2)$  is built of two triangulated solid tori, one of the tetrahedra in one of the solid tori is flat. Then by Lemma 7.7, all of the tetrahedra in this solid torus are flat.

By Lemma 7.8, we know that the exterior dihedral angles of the flat solid torus  $(\theta_p, \theta_q, \theta_r)$  are  $(\pi, -\pi, \pi)$ ,  $(-\pi, \pi, \pi)$ ,  $(\pi, 0, 0)$ , or  $(0, \pi, 0)$ .

Suppose that  $(\theta_p, \theta_q, \theta_r) = (\pi, -\pi, \pi)$  or  $(-\pi, \pi, \pi)$ . In either case, this implies that  $\alpha_1 = \pm\pi$ , so  $-\alpha_2 = \alpha_1 = \pm\pi$ , and  $\theta_2 = \alpha_1 + \theta_1 = 0$  (it cannot be  $2\pi$  since we restrict to exterior angles between  $-\pi$  and  $\pi$ ), by Lemma 8.2. Then the third angle satisfies  $\pi - \alpha_2 - \theta_2 = 0$ . So the exterior dihedral angles of the second solid torus are  $(\pi, 0, 0)$ . It follows from Lemma 7.8 that the second solid torus must also be flat. Thus such an angle structure has zero volume, and cannot maximise volume.

Now suppose that  $(\theta_p, \theta_q, \theta_r) = (0, 0, \pi)$ , up to permutation. Then  $\alpha_1$  is 0 or  $\pi$ , so  $\alpha_2 = -\alpha_1$  is 0 or  $-\pi$ , and  $\theta_2 = \alpha_1 + \theta_1$  is 0 or  $\pi$ . In any case, the exterior dihedral angles must all be either 0 or  $\pm\pi$ , which again implies that the second layered solid torus is flat. As before the angle structure cannot maximise volume. □

**Theorem 8.6** *Let  $L$  be a fully augmented link with exactly two crossing circles, as in Figure 13. Let  $M$  be the manifold obtained by Dehn filling the crossing circles of  $S^3 - L$  along slopes  $m_1, m_2 \in (\mathbb{Q} \cup \{1/0\}) - \{0, 1/0, \pm 1, \pm 2\}$ . Then  $M$  admits a geometric triangulation.*

**Proof** By Lemma 8.1,  $M(m_1, m_2)$  admits a topological triangulation. By Lemma 8.3, this triangulation admits an angle structure. By Lemma 8.5, the volume functional takes its maximum on the interior of the space of angle structures. Then the Casson–Rivin theorem, Theorem 4.2, implies that the triangulation is geometric. □

## 9 Fully augmented 2–bridge links

In this section we consider links obtained by fully augmenting the standard diagram of a 2–bridge link, which we call fully augmented 2–bridge links for short. These admit a decomposition into two identical, totally geodesic, right-angled ideal polyhedra as in Section 2. In this case, the polyhedra have a particularly nice form: they are built by gluing finitely many regular ideal octahedra. The construction is illustrated carefully in [26, Section 4]. We review it briefly here.

A 2–bridge link has one of two forms, depending on whether there are an even or odd number of twist regions; see Figure 18. As before, we augment each twist region with a crossing circle, and remove all even pairs of crossings from the corresponding twist region, leaving one or zero crossings encircled by each crossing circle. When there is

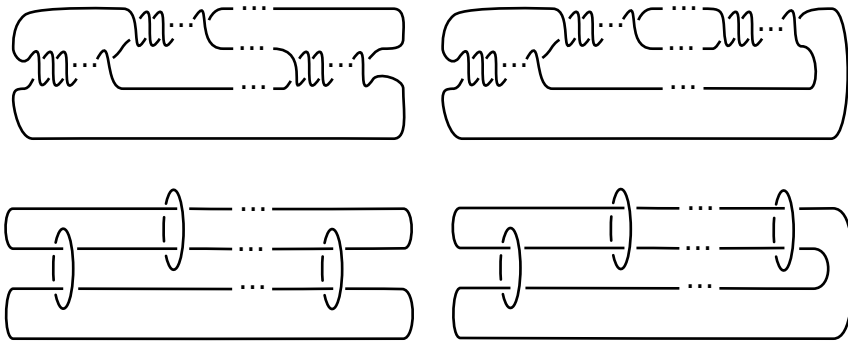


Figure 18: Top: the two forms of a 2-bridge link. Bottom: the forms of the fully augmented 2-bridge link (without half-twists).

one crossing, we say the crossing circle has a half-twist. In fact, we will not consider half-twists here, so assume the fully augmented 2-bridge link has no half-twists.

To obtain the polyhedra, cut the fully augmented 2-bridge link along the geodesic surface of the projection plane. This cuts each of the 2-punctured disks bounded by a crossing circle in half. The 3-sphere is cut into two pieces, one above and one below the projection plane. For each half, we cut open half discs and flatten them in the projection plane. Lastly, shrink the link components to ideal vertices; see Figure 19, left. The circle packing giving the polyhedral decomposition of a fully augmented 2-bridge link is shown in Figure 19, right.

**Lemma 9.1** *The cusp shapes of any fully augmented 2-bridge link complement with no half-twists consist of a  $1 \times 2$  block of squares if the cusp is the first or the last in the diagram, or a  $2 \times 2$  block of squares for all other cusps.*

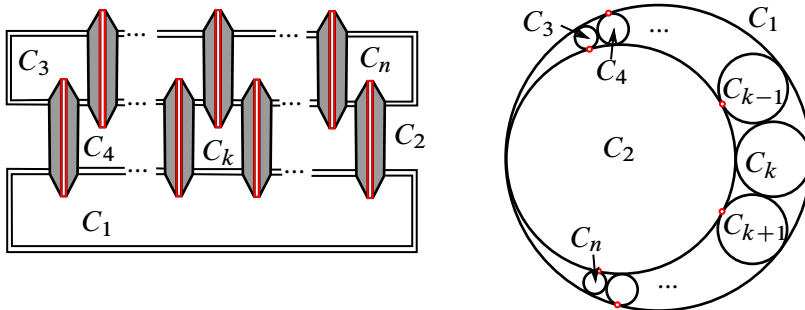


Figure 19: How to decompose a fully augmented 2-bridge link into two polyhedra. On the right is the corresponding circle packing.

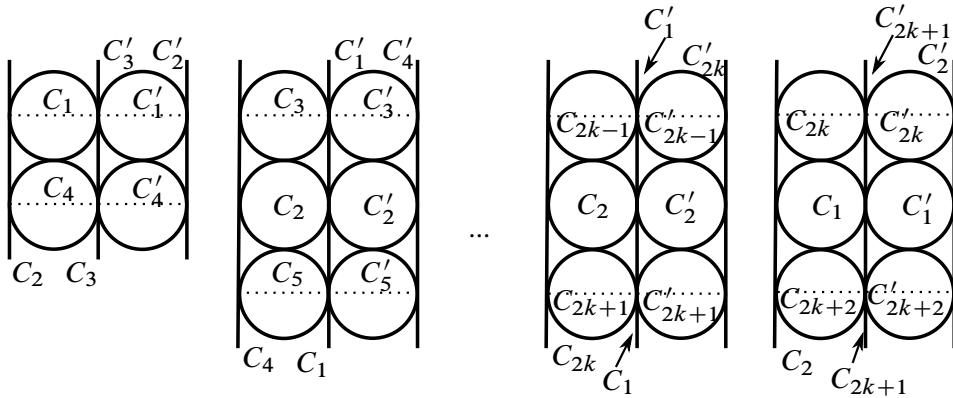


Figure 20: Left to right: the first cusp, the next cusp, the  $(2k)^{\text{th}}$  cusp, and the  $(2k+1)^{\text{th}}$  cusp.

**Proof** The fully augmented 2-bridge link has a circle packing built of two tangent circles, labelled  $C_1$  and  $C_2$ , and a string of circles  $C_3, \dots, C_n$ , each tangent to both  $C_1$  and  $C_2$ , and  $C_j$  tangent to  $C_{j-1}$  and  $C_{j+1}$  for  $j = 4, \dots, n - 1$ , as in Figure 19, right. This circle packing describes each of the two polyhedra making up the link complement.

We need to determine the cusp shapes of the crossing circle cusps. In each polyhedron the crossing circle cusps correspond to tangencies in the circle packing between  $C_2$  and  $C_3$ , between  $C_4$  and  $C_1$ , and more generally, between  $C_2$  and  $C_{2k+1}$ , and between  $C_1$  and  $C_{2k}$ . We take each of these tangent points to infinity to determine the cusp shape. There are two cases.

**Case 1** Consider the first and last crossing circles, corresponding to tangencies of  $C_2$  and  $C_3$ , and of either  $C_2$  and  $C_n$  or  $C_1$  and  $C_n$  if  $n$  is odd or even, respectively.

Take the point of tangency to infinity. In the case of  $C_2$  and  $C_3$ , there are two circles,  $C_1$  and  $C_4$ , tangent to both  $C_2$  and  $C_3$ , and tangent to each other. Thus the circle packing forms a square, similar to the case of the Borromean rings. When we glue across white faces, we glue an identical square coming from the second polyhedron, and the cusp becomes a  $1 \times 2$  rectangle; see Figure 20, left. The case of  $C_n$  is similar.

**Case 2** For tangencies between circles  $C_1$  and  $C_{2k}$  or  $C_2$  and  $C_{2k+1}$ , where the circles  $C_{2k}$  or  $C_{2k+1}$  are not the first or last such circles, when we take the point of tangency to infinity we see a pattern as shown in Figure 20, right. That is, in the first case, circles  $C_1$  and  $C_{2k}$  become parallel lines. Between them are three circles tangent

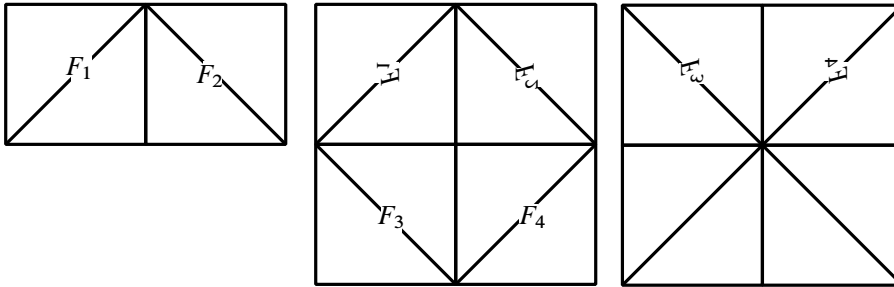


Figure 21: How the square faces of pyramids glue to each other.

to both parallel lines, namely  $C_{2k-1}$ ,  $C_2$ , and  $C_{2k+1}$ . These circles form the cusp circle packing. Again glue a white face to another white face to obtain the full cusp shape, shown in Figure 20, middle right. In this case, the cusp shape is a  $2 \times 2$  rectangle. The case of  $C_2$  and  $C_{2k+1}$  is similar, and is illustrated in Figure 20, far right.  $\square$

Notice that each crossing circle cusp is tiled by ideal pyramids over a square base. The first and last crossing circle cusps are tiled by two pyramids, the others by four.

The fully augmented 2-bridge link is obtained by gluing all these pyramids together according to the following pattern.

**Lemma 9.2** *Consider the fully augmented 2-bridge link, with no half-twists, and cusp shapes as in Lemma 9.1. The gluing is as follows:*

- The first cusp, which is a  $1 \times 2$  rectangle, is glued to the top half of the second cusp, with left side gluing by a reflection in the diagonal of positive slope, and right side gluing by reflection in the diagonal of negative slope.
- The bottom half of the second cusp, another  $1 \times 2$  rectangle, is glued to the top half of the third cusp, with the left side gluing by reflection in the negative diagonal, and the right side gluing by reflection in the positive diagonal.
- Inductively, the  $1 \times 2$  bottom half of the  $k^{\text{th}}$  cusp glues to the  $1 \times 2$  top half of the  $(k+1)^{\text{th}}$  cusp, with left side gluing by reflection in one diagonal, and the right side gluing by reflection in the other diagonal; see Figure 21. Importantly, diagonals glue to diagonals, and horizontal lines glue to vertical and vice-versa.
- Finally, on the last  $2 \times 2$  rectangle, the bottom  $1 \times 2$  rectangle glues to the final crossing circle cusp, a  $1 \times 2$  rectangle, again with vertical edges gluing to horizontal and horizontal to vertical, and diagonals gluing to diagonals.

**Proof** As in the case of the link with two crossing circles, this result is obtained by observing the intersections of circles in the circle packings; refer to Figure 20. For the first cusp, the square on the left has ideal vertices  $C_1 \cap C_2$ ,  $C_2 \cap C_4$ ,  $C_3 \cap C_4$ , and  $C_3 \cap C_1$  in anticlockwise order. These map to the top square in the middle left, but note this list of vertices now runs in clockwise order, with  $C_2 \cap C_4$  and  $C_3 \cap C_1$  in the same locations. It follows that the gluing is a reflection in the diagonal of positive slope. For the square on the right, ideal vertices  $C'_1 \cap C'_3$ ,  $C'_3 \cap C'_4$ ,  $C'_2 \cap C'_4$ , and  $C'_2 \cap C'_1$  in anticlockwise order are mapped to the top right square in the middle left, only now in clockwise order. The gluing is a reflection in the diagonal of negative slope.

The rest of the squares are treated similarly to obtain the result. □

To Dehn fill, we remove the interiors of the pyramids meeting the cusp, leaving only the square base of each pyramid behind. We will then put in a triangulated solid torus in which the chosen slope bounds a meridian and the boundary is triangulated to match the triangulation on the original manifold.

We know how to fill in a solid torus in the case of the  $1 \times 2$  rectangles: we just take the double cover of the layered solid torus as in Lemma 7.1 or the side-by-side solid torus as in Lemma 7.4.

For the  $2 \times 2$  rectangle, we see in Lemma 9.3 that the double cover of one of the solid tori we have already encountered will always work.

**Lemma 9.3** *Suppose  $a/b$  is a slope, and  $a/b \notin \{0/1, 1/0, \pm 1/1\}$ .*

- (1) *If  $a$  is odd and  $b$  is even, then let  $Y$  be the double cover of the vertical side-by-side solid torus  $X$  from Lemma 7.9, constructed so that  $X$  has meridian  $a/2b$ .*
- (2) *If  $a$  is even and  $b$  is odd, or if  $a$  is odd and  $b$  is odd, then let  $Y$  be the double cover of the horizontal side-by-side solid torus  $X$  from Lemma 7.4, constructed so that  $X$  has meridian  $2a/b$ .*

*In either case,  $Y$  is a triangulated solid torus whose boundary consists of eight ideal triangles in four symmetric pairs, forming a  $2 \times 2$  square. The slope  $a/b$  is a meridian of  $Y$ .*

**Proof Case 1** Suppose  $a/b$  is of the form odd over even.

Let  $X$  be the triangulated solid torus that is a “vertical” side-by-side of a layered solid torus, constructed so that the meridian of  $X$  has slope  $a/2b := m_X$ , as in Lemma 7.9. Note that to build such a triangulation, we walk to the slope  $a/b$  in the Farey graph,

then identify edges of slope  $a/b$  in the final layered tetrahedra, and insert one more tetrahedron. In  $\mathbb{R}^2$ , the slope  $1/0 := \mu_X$  in  $X$  lifts to run from  $(0, 0)$  to  $(0, 2)$ , and  $0/1 := \lambda_X$  lifts to run from  $(0, 0)$  to  $(1, 0)$ . The meridian of  $X$ , the slope  $a/2b = m_X$ , lifts to run from  $(0, 0)$  to  $(2b, 2a)$ .

Let  $Y$  denote the (horizontal) double cover of  $X$ . We will show  $Y$  gives the required Dehn filling. To do so, we need to show that  $a/b$ , now written as a slope on  $Y$ , in the basis for  $Y$ , bounds a disc.

In the solid torus  $X$ , the slope  $\mu_X = 1/0$  lifts to the curve running from  $(0, 0)$  to  $(0, 2)$  in  $\mathbb{R}^2$ . The meridian slope  $m_X = a/2b$  meets the slope  $\mu_X$  a total of  $|a \cdot 0 - 2b \cdot 1| = |2b|$  times, which is even. Thus the curve  $\mu_X$  is homotopic to an even power of the core of  $X$ . Therefore it lifts to a generator of the fundamental group of  $\partial Y$ . Denote this generator by  $\mu_Y = 1/0$ . When we lift  $\partial Y$  to  $\mathbb{R}^2$ , the lift of  $\mu_Y$  still runs from  $(0, 0)$  to  $(0, 2)$ .

The meridian slope  $m_X = a/2b$  of  $X$  meets  $\lambda_X = 0/1$  a total of  $|2b \cdot 0 - a \cdot 1| = |a|$  times, which is odd. Thus a second generator of the fundamental group of  $\partial Y$  is given by taking two lifts of the curve  $\lambda_X = 0/1$ , lined up end-to-end. Denote this generator by  $\lambda_Y$ . When we lift  $\partial Y$  to  $\mathbb{R}^2$ ,  $\lambda_Y$  lifts to run from  $(0, 0)$  to  $(0, 2)$  in  $\mathbb{R}^2$ , ie twice the lift of the corresponding generator in  $\partial X$ .

The meridian  $m_X = a/2b$  of  $X$  lifts to bound a disc in  $Y$ . Note that the lift runs  $a$  times along  $\mu_Y$  and  $b$  times along  $\lambda_Y$ . This means the meridian of  $Y$  is the slope  $a/b$ , as desired.

**Case 2** Suppose  $a/b$  is of the form even over odd or is of the form odd over odd.

Let  $X$  be the triangulated solid torus that is a “horizontal” side-by-side solid torus, constructed so that the meridian of  $X$  has slope  $2a/b := m_X$ . Note that to build such a triangulation, we walk to the slope  $a/b$  in the Farey graph, as in Lemma 7.4. The solid torus  $X$  has generators of the fundamental group given by slopes  $\mu_X = 1/0$ , lifting to run from  $(0, 0)$  to  $(0, 1)$  in the  $\mathbb{R}^2$  cover of  $\partial X$ , and  $\lambda_X = 0/1$ , lifting to run from  $(0, 0)$  to  $(2, 0)$ . The meridian  $2a/b$  lifts to run from  $(0, 0)$  to  $(2b, 2a)$ .

Let  $Y$  denote the (vertical) double cover of  $X$ . We will show  $Y$  gives the required Dehn filling. To do so, we need to show that  $a/b$ , written in the basis for  $Y$ , bounds a disc.

In  $X$ ,  $\mu_X = 1/0$  meets the meridian slope  $m_X = 2a/b$  a total of  $|1 \cdot b - 0 \cdot 2a| = |b|$  times, which is odd. Thus  $\mu_X$  in  $X$  lifts to an arc in  $\partial Y$ . A generator of the fundamental group of  $\partial Y$  is given by taking two lifts of this curve, lined up end-to-end. Denote the resulting closed curve in  $Y$  by  $\mu_Y$ . Its lift runs from  $(0, 0)$  to  $(0, 2)$  in  $\mathbb{R}^2$ .



In  $X$ , the slope  $\lambda_X = 0/1$  meets the meridian slope  $m_X = 2a/b$  of  $X$  a total of  $|0 \cdot b - 1 \cdot 2a| = |2a|$  times, which is even. Thus the curve  $\lambda_X$  is homotopic to an even power of the core of  $X$ , and it lifts to a closed curve in  $Y$ . A second generator of the fundamental group of  $\partial Y$  is given by taking this lift of  $\lambda_X$ . Denote it by  $\lambda_Y$ . When we lift  $\partial Y$  to  $\mathbb{R}^2$ , the lift of  $\lambda_Y$  is the same as the lift of  $\lambda_X$ : it runs from  $(0, 0)$  to  $(0, 2)$  in  $\mathbb{R}^2$ . The meridian  $m_X = 2a/b$  of  $X$  lifts to bound a disc in  $Y$ . Note that the lift runs  $a$  times along  $\mu_Y$  and  $b$  times along  $\lambda_Y$ , hence has the slope  $a/b$ , as desired.  $\square$

**Lemma 9.4** *Let  $a/b \in \mathbb{Q} \cup \{1/0\}$  be such that  $a/b \notin \{0/1, 1/0, \pm 1/1\}$ . Let  $Y$  denote the triangulated solid torus with meridian  $a/b$  of Lemma 9.3. Let  $\{\theta_p, \theta_q, \theta_r\}$  be exterior dihedral angles along the boundary of the four-punctured torus forming the  $2 \times 2$  square, each repeated four times symmetrically, satisfying*

$$0 < \theta_r < \pi, \quad -\pi < \theta_p, \theta_q < \pi, \quad \text{and} \quad \theta_p + \theta_q + \theta_r = \pi.$$

*Then there exists an angle structure on the triangulated solid torus of Lemma 9.3 with these exterior angles.*

**Proof** This is automatic from Lemma 7.6 or Lemma 7.9: our solid torus is the double cover of a vertical or horizontal side-by-side solid torus  $X$ , with meridian of  $X$  not one of the slopes  $\{0/1, 1/0, \pm 1/1, \pm 1/2, \pm 2/1\}$ . For such solid tori, the angle structure exists, so it exists for the double cover by lifting angles.  $\square$

**Lemma 9.5** *Let  $T$  be the triangulated solid torus of Lemma 9.3. If the volume functional has its maximum in the boundary of the space of angle structures, then all tetrahedra of  $T$  must be flat. Hence, the volume functional takes its maximum in the interior.*

*Moreover, all tetrahedra are flat if and only if exterior angles  $(\theta_p, \theta_q, \theta_r)$  are one of  $(\pi, 0, 0), (0, \pi, 0), (-\pi, \pi, \pi)$  or  $(\pi, -\pi, \pi)$ .*

**Proof** This follows immediately from the similar fact for side-by-side solid tori, Lemmas 7.7 and 7.8, or in the vertical side-by-side case by Lemma 7.9.  $\square$

After removing pyramids from the fully augmented 2-bridge link, and putting in triangulated solid tori satisfying the above lemmas, we have a triangulation of a Dehn filling. To obtain an angle structure, we need gluing equations to be satisfied. Since we already know gluing equations inside the solid tori, we only need to ensure gluing equations hold on the boundaries of these solid tori where they glue to each other.

**Lemma 9.6** *Let  $L$  be a fully augmented 2-bridge link with  $n > 2$  crossing circles (and no half-twists). Let  $s_1, \dots, s_n$  be slopes that are all positive or all negative, and further*

$$s_1, s_n \notin \{0/1, 1/0, \pm 1/1, \pm 2/1\} \quad \text{and} \quad s_2, \dots, s_{n-1} \notin \{0/1, 1/0, \pm 1/1\}.$$

*Label horizontal edges in all crossing circle cusps by  $\theta_j$ , diagonals by  $\alpha_j$ , and verticals by  $\pi - \theta_j - \alpha_j$ . (These are exterior angles). Let  $T$  be the triangulation of the Dehn filling of  $S^3 - L$  along these slopes obtained by inserting solid tori as above. Then if there is an angle structure, exterior angles on the solid tori must satisfy*

**(9.7)** (Diagonal equations)  $\alpha_i = -\alpha_{i+1} \quad \text{for all } i,$

**(9.8)** (Interior equations)  $2(\theta_i + \alpha_i) - \theta_{i-1} - \theta_{i+1} = 0 \quad \text{for } 2 \leq i \leq n - 1,$

**(9.9)** (End equations)  $\theta_1 + \alpha_1 - \theta_2 = 0 \quad \text{and} \quad \theta_n + \alpha_n - \theta_{n-1} = 0.$

**Proof** This follows from the gluing description given above.

Diagonal edges map to diagonal edges, and these are the only edges in this edge class. Thus for all  $i$ ,  $(\pi - \alpha_i) + (\pi - \alpha_{i+1}) = 2\pi$ , which implies  $\alpha_i = -\alpha_{i+1}$ , giving (9.7).

For the first end equation, the vertical edges with angles  $\pi - \theta_1 - \alpha_1$  in the first  $1 \times 2$  cusp glue to the horizontal edges on the left side of the first  $2 \times 2$  cusp, labelled  $\theta_2$ . Note that both vertical edges and both horizontal edges are glued to the same edge class. Thus  $2(\pi - (\pi - \theta_1 - \alpha_1)) + 2(\pi - \theta_2) = 2\pi$ . This gives the first end equation in (9.9).

For the interior equations, the horizontal edge with angle  $\theta_{i-1}$  in the  $(i-1)^{\text{th}}$  cusp glues to both vertical edges in the  $i^{\text{th}}$  cusp, with angles  $\pi - \theta_i - \alpha_i$ . In turn, both vertical edges in the  $i^{\text{th}}$  cusp glue to the horizontal edge with angle  $\pi - \theta_{i+1}$  in the  $(i+1)^{\text{th}}$  cusp. Note that this is true for  $2 \leq i \leq n - 1$ . Thus we require that  $\pi - \theta_{i-1} + 2(\pi - (\pi - \theta_i - \alpha_i)) + \pi - \theta_{i+1} = 2\pi$  for  $2 \leq i \leq n - 1$ . This gives the interior equations in (9.8).

For the second end equation, both horizontal edges with angle  $\theta_{n-1}$  in the last  $2 \times 2$  cusp glue to both vertical edges with angle  $\pi - \theta_n - \alpha_n$  in the last  $1 \times 2$  cusp. Thus  $2(\pi - \theta_{n-1}) + 2(\pi - (\pi - \theta_n - \alpha_n)) = 2\pi$ , giving the second end equation in (9.9).  $\square$

**Lemma 9.10** *For the triangulation on the Dehn filling of the fully augmented 2-bridge link given above, the space of angle structures is nonempty.*

**Proof** Because the signs of all the slopes agree, say all are positive, the solid tori are constructed by starting in the Farey triangulation in the triangle  $(0, 1, 1/0)$ , and moving

either across the edge from 0 to 1 or from 1 to 1/0. In either case, the slope 1/1 cannot correspond to the slope covered first, so  $\alpha_j$  will never correspond to the slope  $\theta_{r_j}$ ; it will be  $\theta_{p_j}$  or  $\theta_{q_j}$ . Then set all  $\alpha_j = 0$ . This is in the required range of Lemma 9.4.

Because  $\alpha_j = 0$  for all  $j$ , the end equations imply  $\theta_1 = \theta_2$ , and the interior equations imply  $2\theta_2 = \theta_1 + \theta_3$ , hence  $\theta_3 = \theta_1$ . Inductively assume  $\theta_j = \theta_1$  for  $j \leq k$  and  $k \leq n - 1$ . Then  $2\theta_k = \theta_{k-1} + \theta_{k+1}$ , hence  $\theta_{k+1} = \theta_1$  as well. Finally, the end equations imply  $\theta_n = \theta_{n-1} = \theta_1$ . So all angles  $\theta_j$  equal  $\theta_1$ .

Now by Lemma 9.4 in the case of the  $2 \times 2$  square, or by Lemma 7.6 in the case of the  $1 \times 2$  square, an angle structure exists on the solid tori. By choice of angles, these satisfy the gluing equations required in Lemma 9.6. So this gives an angle structure.  $\square$

**Lemma 9.11** *Volume is maximised in the interior of space of angle structures.*

**Proof** Suppose volume is not maximised in the interior. Then there is a flat tetrahedron, say in the  $i^{\text{th}}$  solid torus a tetrahedron is flat. By Lemma 9.5 the  $i^{\text{th}}$  solid torus must be a flat solid torus.

A solid torus is flat if and only if the exterior angles  $\alpha_i, \theta_i$ , and  $\pi - \alpha_i - \theta_i$  are  $(0, 0, \pi)$  or  $(\pi, -\pi, \pi)$ , up to permutation, by Lemma 9.5. There are five cases:

**Case 1** ( $\alpha_i = \theta_i = 0$ ) Here  $\alpha_j = 0$  for all  $j$  by the diagonal equations (9.7). As in the proof of Lemma 9.10, this implies that  $\theta_j = \theta_1$  for all  $j$ . In particular,  $\theta_1 = \theta_i = 0$ , so all  $\theta_j = 0$ , so all the solid tori are flat by Lemma 9.5.

**Case 2** ( $\alpha_i = 0$  and  $\theta_i = \pi$ ) Here  $2\theta_i = \theta_{i-1} + \theta_{i+1}$  by the interior equations (9.8). Since  $\theta_i = \pi$ , we have  $2\pi = \theta_{i-1} + \theta_{i+1}$ , which implies that both  $\theta_{i-1}$  and  $\theta_{i+1}$  are  $\pi$ . Hence all solid tori are flat by Lemma 9.5.

**Case 3** ( $\alpha_i = \pi$  and  $\theta_i = 0$ ) By the interior equations (9.8),  $\theta_{i+1} + \theta_{i-1} - 2\theta_i = 2\pi$ . Now  $\theta_i = 0$  gives  $\theta_{i-1} + \theta_{i+1} = 2\pi$ , which implies that  $\theta_{i-1}$  and  $\theta_{i+1}$  are  $\pi$ . Then the diagonal equations plus these results imply  $\alpha_{i-1} = -\pi$  and  $\theta_{i-1} = \pi$ . This is case 5. We show below that all tetrahedra are flat.

**Case 4** ( $\alpha_i = \pi$  and  $\theta_i = -\pi$ ) First suppose  $i$  is even, where  $1 < i \leq n$ . Then  $\alpha_j = \pi$  for  $j = 2k$  and  $\alpha_j = -\pi$  for  $j = 2k + 1$ . In particular,  $\alpha_1 = -\pi$ . By the end equations (9.9), we have  $\theta_2 = \theta_1 - \pi$ . By the interior equations (9.8), we have  $2\theta_j - \theta_{j-1} - \theta_{j+1} = 2\pi$  for  $j = 2k + 1$ , and  $2\theta_j - \theta_{j-1} - \theta_{j+1} = -2\pi$  for  $j = 2k$ . In particular, when  $j = 2k = 2$ , this implies  $\theta_1 = \theta_3$ .

Now inductively assume that  $\theta_{2j+1} = \theta_1$  for  $j \leq k$  and  $\theta_{2j} = \theta_1 - \pi$  for  $j \leq k$ , and  $2k + 3 \leq n - 1$ . Then the interior equations imply

$$2\theta_{2k+1} - \theta_{2k} - \theta_{2k+2} = 2\theta_1 - \theta_1 + \pi - \theta_{2k+2} = 2\pi,$$

and so  $\theta_{2k+2} = \theta_1 - \pi$ . Moreover,

$$2\theta_{2k+2} - \theta_{2k+1} - \theta_{2k+3} = 2(\theta_1 - \pi) - \theta_1 - \theta_{2k+3} = -2\pi,$$

thus  $\theta_{2k+3} = \theta_1$ . Finally, the end equation implies

$$\theta_n = \begin{cases} \theta_{n-1} - \pi = \theta_1 - \pi & \text{if } n \text{ even,} \\ \theta_{n-1} + \pi = (\theta_1 - \pi) + \pi = \theta_1 & \text{if } n \text{ odd.} \end{cases}$$

Since our fixed  $i$  is even, we have  $\theta_1 = \theta_i + \pi = -\pi + \pi = 0$ . This implies that  $\theta_j = 0$  for all  $j$  even and  $\theta_j = -\pi$  for all  $j$  odd. It follows that all tetrahedra are flat by Lemma 9.5.

Now suppose  $i$  is odd,  $\alpha_i = \pi$ , and  $\theta_i = -\pi$ . Then  $\alpha_j = -\pi$  for  $j$  even and  $\alpha_j = \pi$  for  $j$  odd. In particular,  $\alpha_1 = \pi$ . As above, the end and interior equations imply that  $\theta_j = \theta_1$  when  $j$  is odd, and  $\theta_j = \theta_1 + \pi$  when  $j$  is even. So again,  $-\pi = \theta_i = \theta_1 = \theta_j$  for all  $j$  odd, and  $0 = \theta_1 + \pi = \theta_j$  for all  $j$  even. Thus again all tetrahedra are flat.

**Case 5** ( $\alpha_i = -\pi$  and  $\theta_i = \pi$ ) This case is similar to case 4 above. When  $i$  is odd, one can show  $\theta_{2j+1} = \theta_1$  and  $\theta_{2j} = \theta_1 - \pi$  for all  $j$ , implying  $\theta_1 = \theta_i = \pi = \theta_{2j+1}$  and  $\theta_{2j} = 0$ . Thus all tetrahedra are flat.

When  $i$  is even, one can show  $\theta_{2j+1} = \theta_1$  and  $\theta_{2j} = \theta_1 + \pi$  for all  $j$ , implying  $\pi = \theta_i = \theta_1 + \pi$ , so  $\theta_1 = \theta_{2j+1} = 0$  and  $\theta_{2j} = 0 + \pi$ , so again all tetrahedra are flat.  $\square$

**Theorem 9.12** *Let  $L$  be a fully augmented 2-bridge link with  $n > 2$  crossing circles (and no half-twists). Let  $s_1, s_2, \dots, s_n \in \mathbb{Q} \cup \{1/0\}$  be slopes, one for each crossing circle, that are all positive or all negative. Suppose finally that  $s_1$  and  $s_n$  are the slopes on the crossing circles on either end of the diagram, and the slopes satisfy*

$$s_1, s_n \notin \{0/1, 1/0, \pm 1/1, \pm 2/1\} \quad \text{and} \quad s_2, \dots, s_{n-1} \notin \{0/1, 1/0, \pm 1/1\}.$$

*Then the manifold obtained by Dehn filling  $S^3 - L$  along these slopes on its crossing circles admits a geometric triangulation.*

**Proof** With these slopes, there exists a triangulated solid torus with meridian  $s_j$  and boundary triangulated by a number of triangles matching that on the crossing circle boundary. Topologically, the Dehn filling is given by triangulating the solid tori and gluing them together. By Lemma 9.10, the result admits an angle structure.

By Lemma 9.11, the volume is maximised in the interior of the angle structure. The Casson–Rivin Theorem, Theorem 4.2, then implies the triangulation is geometric.  $\square$

## References

- [1] **I Agol**, *Ideal triangulations of pseudo-Anosov mapping tori*, from “Topology and geometry in dimension three” (W Li, L Bartolini, J Johnson, F Luo, R Myers, JH Rubinstein, editors), Contemp. Math. 560, Amer. Math. Soc., Providence, RI (2011) 1–17 MR Zbl
- [2] **I Agol, D Thurston** (2004) MR Zbl Appendix to M Lackenby, *The volume of hyperbolic alternating link complements*, Proc. London Math. Soc. 88 (2004) 204–224
- [3] **R Benedetti, C Petronio**, *Lectures on hyperbolic geometry*, Springer (1992) MR Zbl
- [4] **R H Bing**, *An alternative proof that 3-manifolds can be triangulated*, Ann. of Math. 69 (1959) 37–65 MR Zbl
- [5] **Y-E Choi**, *Positively oriented ideal triangulations on hyperbolic three-manifolds*, Topology 43 (2004) 1345–1371 MR Zbl
- [6] **M Culler, N M Dunfield, J R Weeks**, *SnapPy, a computer program for studying the topology of 3-manifolds* Available at <http://snappy.computop.org>
- [7] **D B A Epstein, R C Penner**, *Euclidean decompositions of noncompact hyperbolic manifolds*, J. Differential Geom. 27 (1988) 67–80 MR Zbl
- [8] **D Futer, F Guéritaud**, *From angled triangulations to hyperbolic structures*, from “Interactions between hyperbolic geometry, quantum topology and number theory” (A Champanerkar, O Dasbach, E Kalfagianni, I Kofman, W Neumann, N Stoltzfus, editors), Contemp. Math. 541, Amer. Math. Soc., Providence, RI (2011) 159–182 MR Zbl
- [9] **D Futer, F Guéritaud**, *Explicit angle structures for veering triangulations*, Algebr. Geom. Topol. 13 (2013) 205–235 MR Zbl
- [10] **D Futer, E Kalfagianni, J S Purcell**, *Dehn filling, volume, and the Jones polynomial*, J. Differential Geom. 78 (2008) 429–464 MR Zbl
- [11] **D Futer, J S Purcell**, *Links with no exceptional surgeries*, Comment. Math. Helv. 82 (2007) 629–664 MR Zbl
- [12] **D Futer, S J Taylor, W Worden**, *Random veering triangulations are not geometric*, Groups Geom. Dyn. 14 (2020) 1077–1126 MR Zbl
- [13] **M Goerner**, *A census of hyperbolic platonic manifolds and augmented knotted trivalent graphs*, New York J. Math. 23 (2017) 527–553 MR Zbl
- [14] **M Goerner**, *Geodesic triangulations exist for cusped Platonic manifolds*, New York J. Math. 23 (2017) 1363–1367 MR Zbl
- [15] **F Guéritaud**, *On canonical triangulations of once-punctured torus bundles and two-bridge link complements*, Geom. Topol. 10 (2006) 1239–1284 MR With an appendix by David Futer

- [16] **F Guéritaud, S Schleimer**, *Canonical triangulations of Dehn fillings*, *Geom. Topol.* 14 (2010) 193–242 MR Zbl
- [17] **C D Hodgson, A Issa, H Segerman**, *Non-geometric veering triangulations*, *Exp. Math.* 25 (2016) 17–45 MR Zbl
- [18] **C D Hodgson, J H Rubinstein, H Segerman**, *Triangulations of hyperbolic 3–manifolds admitting strict angle structures*, *J. Topol.* 5 (2012) 887–908 MR Zbl
- [19] **C D Hodgson, J H Rubinstein, H Segerman, S Tillmann**, *Veering triangulations admit strict angle structures*, *Geom. Topol.* 15 (2011) 2073–2089 MR Zbl
- [20] **W Jaco, J H Rubinstein**, *Layered-triangulations of 3–manifolds*, preprint (2006) arXiv math/0603601
- [21] **F Luo, S Schleimer, S Tillmann**, *Geodesic ideal triangulations exist virtually*, *Proc. Amer. Math. Soc.* 136 (2008) 2625–2630 MR Zbl
- [22] **E E Moise**, *Affine structures in 3–manifolds, V: The triangulation theorem and Hauptvermutung*, *Ann. of Math.* 56 (1952) 96–114 MR Zbl
- [23] **W D Neumann, D Zagier**, *Volumes of hyperbolic three-manifolds*, *Topology* 24 (1985) 307–332 MR Zbl
- [24] **C Petronio, J Porti**, *Negatively oriented ideal triangulations and a proof of Thurston’s hyperbolic Dehn filling theorem*, *Expo. Math.* 18 (2000) 1–35 MR Zbl
- [25] **J S Purcell**, *Cusp shapes under cone deformation*, *J. Differential Geom.* 80 (2008) 453–500 MR Zbl
- [26] **J S Purcell**, *Slope lengths and generalized augmented links*, *Comm. Anal. Geom.* 16 (2008) 883–905 MR Zbl
- [27] **J S Purcell**, *An introduction to fully augmented links*, from “Interactions between hyperbolic geometry, quantum topology and number theory” (A Champanerkar, O Dasbach, E Kalfagianni, I Kofman, W Neumann, N Stoltzfus, editors), *Contemp. Math.* 541, Amer. Math. Soc., Providence, RI (2011) 205–220 MR Zbl
- [28] **I Rivin**, *Euclidean structures on simplicial surfaces and hyperbolic volume*, *Ann. of Math.* 139 (1994) 553–580 MR Zbl
- [29] **W P Thurston**, *The geometry and topology of three-manifolds*, lecture notes, Princeton University (1979) Available at <http://msri.org/publications/books/gt3m>

*School of Mathematics, Monash University  
Clayton, Australia*

Current address: *School of Mathematics and Statistics, The University of Sydney  
Sydney, Australia*

*School of Mathematics, Monash University  
Clayton, Australia*

sophie.ham@sydney.edu.au, jessica.purcell@monash.edu

Received: 13 May 2021      Revised: 25 August 2021

## 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 Bib $\TeX$  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](mailto:graphics@msp.org).

### Proofs

Page proofs will be made available to authors (or to the designated corresponding author) in PDF format. Failure to acknowledge the receipt of proofs or to return corrections within the requested deadline may cause publication to be postponed.

# ALGEBRAIC & GEOMETRIC TOPOLOGY

Volume 23    Issue 3 (pages 963–1462)    2023

---

- Projective naturality in Heegaard Floer homology 963  
MICHAEL GARTNER
- Geometrically bounding 3–manifolds, volume and Betti numbers 1055  
JIMING MA and FANGTING ZHENG
- Constrained knots in lens spaces 1097  
FAN YE
- Convexity in hierarchically hyperbolic spaces 1167  
JACOB RUSSELL, DAVIDE SPRIANO and HUNG CONG TRAN
- Finite presentations for stated skein algebras and lattice gauge field theory 1249  
JULIEN KORINMAN
- On the functoriality of  $\mathfrak{sl}_2$  tangle homology 1303  
ANNA BELIAKOVA, MATTHEW HOGANCAMP,  
KRZYSZTOF K PUTYRA and STEPHAN M WEHRLI
- Asymptotic translation lengths and normal generation for pseudo-Anosov monodromies of fibered 3–manifolds 1363  
HYUNGRYUL BAIK, EIKO KIN, HYUNSHIK SHIN and  
CHENXI WU
- Geometric triangulations and highly twisted links 1399  
SOPHIE L HAM and JESSICA S PURCELL