Algebraic & Geometric Topology

Volume 15 (2015)

Issue 6 (pages 3107–3729)

msp

ALGEBRAIC & GEOMETRIC TOPOLOGY

msp.org/agt

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

The subscription price for 2015 is US $350/year for the electronic version, and $575/year (+$60, if shipping outside the US) for print and electronic. Subscriptions, requests for back issues and changes of subscribers 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) at Mathematical Sciences Publishers, c/o Department of Mathematics, University of California, 798 Evans Hall #3840, Berkeley, CA 94720-3840, is published 6 times per year and continuously online. Periodical rate postage paid at Berkeley, CA 94704, and additional mailing offices. POSTMASTER: send address changes to Mathematical Sciences Publishers, c/o Department of Mathematics, University of California, 798 Evans Hall #3840, Berkeley, CA 94720-3840.

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

PUBLISHED BY

mathematical sciences publishers

nonprofit scientific publishing

http://msp.org/

msp

Algebraic & Geometric Topology 15 (2015) 3107–3153

Universality of multiplicative infinite loop space machines

DAVID GEPNER
MORITZ GROTH
THOMAS NIKOLAUS

We establish a canonical and unique tensor product for commutative monoids and groups in an ∞–category $\mathcal{C}$ which generalizes the ordinary tensor product of abelian groups. Using this tensor product we show that $\mathbb{E}_n$–(semi)ring objects in $\mathcal{C}$ give rise to $\mathbb{E}_n$–ring spectrum objects in $\mathcal{C}$. In the case that $\mathcal{C}$ is the ∞–category of spaces this produces a multiplicative infinite loop space machine which can be applied to the algebraic K–theory of rings and ring spectra.

The main tool we use to establish these results is the theory of smashing localizations of presentable ∞–categories. In particular, we identify preadditive and additive ∞–categories as the local objects for certain smashing localizations. A central theme is the stability of algebraic structures under basechange; for example, we show $\mathrm{Ring}(\mathcal{D}\otimes\mathcal{C}) \simeq \mathrm{Ring}(\mathcal{D})\otimes\mathcal{C}$. Lastly, we also consider these algebraic structures from the perspective of Lawvere algebraic theories in ∞–categories.

55P48; 55P43, 19D23

0 Introduction

The Grothendieck group $\mathrm{K}_0(M)$ of a commutative monoid M, also known as the group completion, is the universal abelian group which receives a monoid map from M. It was a major insight of Quillen that higher algebraic K–groups can be defined as the homotopy groups of a certain spectrum which admits a similar description: more precisely, from the perspective of higher category theory, the algebraic K–theory spectrum of a ring R can be understood as the group completion of the groupoid of projective R–modules, viewed as a symmetric monoidal category with respect to the coproduct.

When R is commutative, the algebraic K–groups inherit a multiplication which stems from the tensor product of R–modules. Just as the K–groups arise as homotopy groups of the K–theory spectrum, it is essential for computational and theoretical purposes to understand the multiplication on these groups as coming from a highly structured multiplication on the K–theory spectrum itself. Unfortunately it turned out to be hard

Published: 12 January 2016 DOI: 10.2140/agt.2015.15.3107

to construct such a multiplication directly, partly because for a long time the proper framework to deal with multiplicative structures on spectra was missing. Important work on this question was pioneered by May [22], and the general theory of homotopy coherent algebraic structures goes back at least to Boardman and Vogt [8], May [21], and Segal [28].

It was first shown by May that the group completion functor from $\mathbb{E}_\infty$–spaces to spectra preserves multiplicative structure [22]; see also the more recent accounts [23; 24; 25]. Since then, several authors have given alternative constructions of multiplicative structure on K–theory spectra: most notably, Elmendorf and Mandell promote the infinite loop space machine of Segal to a multifunctor in [12] and in [13] they extend the K–theory functor from symmetric monoidal categories to symmetric multicategories (aka coloured operads), and Baas, Dundas, Richter and Rognes show how to correct the failure of the "phony multiplication" on the Grayson–Quillen $S^{-1}S$–construction in [2], as identified by Thomason [29].

All of these approaches are very carefully crafted and involve for example the intricacies of specific pairs of operads or indexing categories. Here we take a different approach to multiplicative infinite loop space theory, replacing the topological and combinatorial constructions of specific machines by the use of *universal properties*. The main advantage of our approach is that we get strong uniqueness results, which follow for free from the universal properties. The price we pay is that we use the extensive machinery of ∞–categories and argue in the abstract, without the aid of concrete models. Similar results for the case of Waldhausen K–theory, also using the language of ∞–categories, have been obtained by Barwick in a recent paper [3].

In this paper we choose to use the language of (presentable) ∞–categories. But we emphasize the fact that every combinatorial model category gives rise to a presentable ∞–category, and that all presentable ∞–categories arise in this way. Moreover the study of presentable ∞–categories is basically the same as the study of combinatorial model categories, so that in principle all our results could also be formulated in the setting of model categories.

Let us begin by mentioning one of our main results. Associated to an ∞–category $\mathcal{C}$ are the ∞–categories $\mathcal{C}_*$ of pointed objects in $\mathcal{C}$, $\mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{C})$ of commutative monoids in $\mathcal{C}$, $\mathrm{Grp}_{\mathbb{E}_\infty}(\mathcal{C})$ of commutative groups in $\mathcal{C}$, and $\mathrm{Sp}(\mathcal{C})$ of spectrum objects in $\mathcal{C}$. For these ∞–categories we establish the following:

Theorem 5-1 *Let $\mathcal{C}^\otimes$ be a closed symmetric monoidal structure on a presentable ∞–category $\mathcal{C}$. The ∞–categories $\mathcal{C}_*$, $\mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{C})$, $\mathrm{Grp}_{\mathbb{E}_\infty}(\mathcal{C})$, and $\mathrm{Sp}(\mathcal{C})$ all admit closed symmetric monoidal structures, which are uniquely determined by the requirement that* *the respective free functors from $\mathcal{C}$ are symmetric monoidal. Moreover, each of the following free functors also extends uniquely to a symmetric monoidal functor*

$$\mathcal{C}_* \to \mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{C}) \to \mathrm{Grp}_{\mathbb{E}_\infty}(\mathcal{C}) \to \mathrm{Sp}(\mathcal{C}).$$

Note that these symmetric monoidal structures allow us to talk about $\mathbb{E}_n$–(semi)ring objects and $\mathbb{E}_n$–ring spectrum objects in $\mathcal{C}$. Before we sketch the general ideas involved in the proof, it is worth indicating what this theorem amounts to for specific choices of $\mathcal{C}$.

(i) If $\mathcal{C}$ is the ordinary category of sets, then the symmetric monoidal structures of Theorem 5-1 recover for instance the tensor product of abelian monoids and abelian groups. This also reestablishes the easy result that the group completion functor K_0 is symmetric monoidal.

(ii) In the case of the 2–category Cat of ordinary categories, functors, and natural isomorphisms we obtain a symmetric monoidal structure on the 2–category of symmetric monoidal categories SymMonCat. The symmetric monoidal structure on $\mathrm{SymMonCat} \simeq \mathrm{Mon}_{\mathbb{E}_\infty}(\mathrm{Cat})$ has been the subject of confusion in the past due to the fact that SymMonCat only has the desired symmetric monoidal structure when considered as a 2–category and not as a 1–category. In this case, $\mathbb{E}_n$–(semi)ring objects are *$\mathbb{E}_n$–(semi)ring categories* (sometimes also called *rig categories*), important examples of which are given by the bipermutative categories of May [23]. We also obtain higher categorical analogues of this picture using Cat_n and Cat_∞.[1]

(iii) Finally, and most importantly for this paper, we consider Theorem 5-1 in the special case of the ∞–category $\mathcal{S}$ of spaces (which can be obtained from the model category of spaces or simplicial sets). That way we get canonical monoidal structures on $\mathbb{E}_\infty$–spaces and grouplike $\mathbb{E}_\infty$–spaces. The resulting $\mathbb{E}_n$–algebras are *$\mathbb{E}_n$–(semi)ring spaces*; more precisely, they are an ∞–categorical analogue of the $\mathbb{E}_n$–(semi)ring spaces of May; see, for example, May [25]. Moreover, we obtain unique multiplicative structures on the group completion functor $\mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{S}) \to \mathrm{Grp}_{\mathbb{E}_\infty}(\mathcal{S})$ and the *delooping functor* $\mathrm{Grp}_{\mathbb{E}_\infty}(\mathcal{S}) \to \mathrm{Sp}$ which assigns a spectrum to a grouplike $\mathbb{E}_\infty$–space. In particular, the spectrum associated to an $\mathbb{E}_n$–(semi)ring space is an $\mathbb{E}_n$–ring spectrum, which amounts to *multiplicative infinite loop space theory*.

These facts can be assembled together in Section 8 to obtain a new description of the multiplicative structure on the algebraic K–theory functor $\mathrm{K}\colon \mathrm{SymMonCat} \to \mathrm{Sp}$ and its

[1] Interestingly, we have equivalences $\mathrm{Grp}_{\mathbb{E}_\infty}(\mathrm{Cat}_n) \simeq \mathrm{Grp}_{\mathbb{E}_\infty}(\mathrm{Gpd}_n)$ and $\mathrm{Sp}(\mathrm{Cat}_n) \simeq \mathrm{Sp}(\mathrm{Gpd}_n)$, and the latter is trivial unless $n = \infty$; more generally, $\mathrm{Sp}(\mathcal{C})$ is trivial for any n–category $\mathcal{C}$ if n is finite.

∞–categorical variant $K\colon \mathrm{SymMonCat}_\infty \to \mathrm{Sp}$. In particular, the algebraic K–theory of an $\mathbb{E}_n$–semiring (∞–)category is canonically an $\mathbb{E}_n$–ring spectrum. By a *recognition principle* for $\mathbb{E}_n$–semiring (∞–)categories, this applies to many examples of interest. More precisely, we show in Theorem 8-8 that these semiring ∞–categories can be obtained from $\mathbb{E}_n$–monoidal ∞–categories with coproducts such that the monoidal structure preserves coproducts in each variable separately. For instance ordinary closed monoidal, braided monoidal, or symmetric monoidal categories admit the structure of $\mathbb{E}_n$–semiring categories for $n = 1, 2, \infty$, respectively in which the *addition* is given by the coproduct and the *multiplication* is given by the tensor product. More specific examples are given by (∞–)categories of modules over ordinary commutative rings or $\mathbb{E}_n$–ring spectra.[2]

One central idea to prove Theorem 5-1 as stated above, which is also of independent interest, is to identify the assignments

$$\mathcal{C} \mapsto \mathcal{C}_*, \quad \mathcal{C} \mapsto \mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{C}), \quad \mathcal{C} \mapsto \mathrm{Grp}_{\mathbb{E}_\infty}(\mathcal{C}), \quad \mathcal{C} \mapsto \mathrm{Sp}(\mathcal{C}) \tag{0-1}$$

as universal constructions. The first and the last case have already been thoroughly discussed by Lurie [20], where it is shown that, in the world of presentable ∞–categories, $\mathcal{C}_*$ is the *free pointed* ∞–category on $\mathcal{C}$ and $\mathrm{Sp}(\mathcal{C})$ is the *free stable* ∞–category on $\mathcal{C}$. We extend this picture by introducing *preadditive* and *additive* ∞–categories; see also Toën and Vezzosi [30] and Joyal [15]. These notions are obtained by imposing additional exactness conditions on pointed ∞–categories, just as is done in the case of ordinary categories. In fact, a presentable ∞–category $\mathcal{C}$ is (pre)additive if and only if its homotopy category $\mathrm{Ho}(\mathcal{C})$ is (pre)additive in the sense of ordinary category theory. We show that, again in the framework of presentable ∞–categories, $\mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{C})$ is the *free preadditive* ∞–category on $\mathcal{C}$ and that $\mathrm{Grp}_{\mathbb{E}_\infty}(\mathcal{C})$ is the *free additive* ∞–category on $\mathcal{C}$ (Corollary 4-9).

As an application of this description as free categories one can deduce the existence and uniqueness of the functors

$$\mathcal{C} \to \mathcal{C}_* \to \mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{C}) \to \mathrm{Grp}_{\mathbb{E}_\infty}(\mathcal{C}) \to \mathrm{Sp}(\mathcal{C}_*)$$

from the fact that every stable ∞–category is additive, every additive ∞–category is preadditive and every preadditive ∞–category is pointed. More abstractly, the assignments (0-1) give rise to endofunctors of the ∞–category Pr^{L} of presentable ∞–categories and left adjoint functors. The aforementioned universal properties are equivalent to the observation that these endofunctors are localizations (in the sense

[2] But note that the ∞–category of modules for an $\mathbb{E}_n$–ring spectrum is only an $\mathbb{E}_{n-1}$–semiring ∞–category.

of Bousfield) of Pr^{L} with local objects the pointed, preadditive, additive, and stable presentable ∞–categories, respectively.

A second main theme of the paper is the stability of algebraic structures under basechange. For example we show that we have equivalences

$$\mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{C} \otimes \mathcal{D}) \simeq \mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{C}) \otimes \mathcal{D}, \quad \mathrm{Ring}_{\mathbb{E}_n}(\mathcal{C} \otimes \mathcal{D}) \simeq \mathrm{Ring}_{\mathbb{E}_n}(\mathcal{C}) \otimes \mathcal{D},$$

where $\otimes$ denotes the tensor product on Pr^{L} as constructed by Lurie [20] (Corollary 4-7 and Proposition 7-7). Such basechange properties are satisfied by many endofunctors of Pr^{L} which arise when considering algebraic structures of certain kinds, eg $\mathcal{C} \mapsto \mathrm{Alg}_{\mathbb{T}}(\mathcal{C})$ for a Lawvere algebraic theory $\mathbb{T}$. We give a brief account of algebraic theories in Appendix B.

A key insight here is to consider endofunctors of Pr^{L} which satisfy both properties: namely, they are simultaneously localizations and satisfy basechange. In keeping with the terminology of stable homotopy theory we refer to such functors as *smashing localizations* of Pr^{L}. The endofunctors $(-)_*, \mathrm{Mon}_{\mathbb{E}_\infty}, \mathrm{Grp}_{\mathbb{E}_\infty}$ and Sp from (0-1) are the main examples treated in this paper. Then the proof of Theorem 5-1 follows as a special case of the general theory of smashing localizations $L\colon \mathrm{Pr}^{\mathrm{L}} \to \mathrm{Pr}^{\mathrm{L}}$. For example we prove that if $\mathcal{C} \in \mathrm{Pr}^{\mathrm{L}}$ is closed symmetric monoidal, then the ∞–category $L\mathcal{C}$ admits a *unique* closed symmetric monoidal structure such that the localization map $\mathcal{C} \to L\mathcal{C}$ is a symmetric monoidal functor (Proposition 3-9).

Organization of the paper In Section 1, we recall the definition of the ∞–category of monoid and group objects in an ∞–category. They form the generic examples of (pre)additive ∞–categories which we introduce in Section 2. In Section 3, we study smashing localizations of Pr^{L}, which turns out to be the central notion needed to deduce many of the subsequent results in this paper. We then show, in Section 4, that the formation of commutative monoids and groups in presentable ∞–categories are examples of smashing localizations of Pr^{L}, and we identify these localizations with the free (pre)additive ∞–category functor. This leads to the existence of the canonical symmetric monoidal structures described in Section 5, and the next Section 6 is devoted to studying the functoriality of these structures. Then in Section 7 we consider ∞–categories of (semi)ring objects in a closed symmetric monoidal presentable ∞–category; these are used in Section 8 to show that the algebraic K–theory of an $\mathbb{E}_n$–semiring ∞–category is an $\mathbb{E}_n$–ring spectrum. Finally, in Appendix A we show a relation of functors with comonoids, and in Appendix B we consider monoid, group, and ring objects from the perspective of Lawvere algebraic theories.

Conventions We freely use the language of ∞–categories throughout this paper. In particular, we adopt the notational conventions of Lurie [19; 20] and provide more specific references where necessary.

Acknowledgements We would like to thank Ulrich Bunke for suggesting that we work out these results in the setting of ∞–categories and for carefully reading a previous draft. We'd also like to thank Peter May, Tony Elmendorf and Mike Mandell for helpful comments and discussions. Finally, we are grateful to an anonymous referee for a number of useful suggestions.

1 ∞–categories of commutative monoids and groups

Given an ∞–category $\mathcal{C}$ with finite products, we may form the ∞–category $\mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{C})$ of $\mathbb{E}_\infty$–monoids in $\mathcal{C}$. By definition, an $\mathbb{E}_\infty$–monoid $M\in\mathcal{C}$ is a functor $M\colon \mathrm{N}(\mathrm{Fin}_*)\to \mathcal{C}$ such that the morphisms $M(\langle n\rangle)\to M(\langle 1\rangle)$ induced by the inert maps $\rho^i\colon \langle n\rangle \to \langle 1\rangle$ exhibit $M(\langle n\rangle)$ as an n–fold power of $M(\langle 1\rangle)$ in $\mathcal{C}$; see [20, 2.1.1.8, 2.4.2.1, 2.4.2.2] for details. In the terminology of [28], M is called a *special* Γ–object of $\mathcal{C}$. In what follows we will sometimes abuse notation and also use the same name for the underlying object of such an $\mathbb{E}_\infty$–monoid. Given an $\mathbb{E}_\infty$–monoid M, we obtain a (coherently associative and commutative) multiplication map

$$m\colon M\times M\to M,$$

uniquely determined up to a contractible space of choices.

We use the term $\mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{C})$ to denote the ∞–category of $\mathbb{E}_\infty$–monoids in $\mathcal{C}$ with respect to the cartesian product. If $\mathcal{C}$ is an ∞–category equipped with a symmetric monoidal structure which is not necessarily the cartesian product, we write $\mathrm{Alg}_{\mathbb{E}_\infty}(\mathcal{C})$ for the ∞–category of $\mathbb{E}_\infty$–algebras in $\mathcal{C}$; if the symmetric monoidal structure on $\mathcal{C}$ happens to be the cartesian product, then we have an equivalence $\mathrm{Mod}(\mathcal{C})\simeq \mathrm{Alg}(\mathcal{C})$.

Proposition 1-1 *Let $\mathcal{C}$ be an ∞–category with finite products and let M be an $\mathbb{E}_\infty$–monoid in $\mathcal{C}$. Then the following conditions are equivalent:*

(i) *The $\mathbb{E}_\infty$–monoid M admits an* inversion *map, ie, there is a map $i\colon M\to M$ such that the composition*

$$M\xrightarrow{\Delta} M\times M\xrightarrow{\mathrm{id}\times i} M\times M\xrightarrow{m} M$$

is homotopic to the constant functor at the unit.

(ii) *The commutative monoid object of* $\mathrm{Ho}(\mathcal{C})$ *underlying the* $\mathbb{E}_\infty$*–monoid* M *is a* group *object.*

(iii) *The* shear map $s\colon M \times M \to M \times M$, *defined as the projection* $\mathrm{pr}_1\colon M \times M \to M$ *on the first factor and the multiplication* $m\colon M \times M \to M$ *on the second factor, is an equivalence.*

(iv) *The special* Γ*–object* $M\colon \mathrm{N}(\mathrm{Fin}_*) \to \mathcal{C}$ *is* very special *(again in the terminology of [28]).*

Proof This follows immediately from the fact that $\mathcal{C} \to \mathrm{N}(\mathrm{Ho}(\mathcal{C}))$ is conservative and preserves products. □

Definition 1-2 Let $\mathcal{C}$ be an ∞–category with finite products. An object $M \in \mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{C})$ is called an $\mathbb{E}_\infty$*–group* in $\mathcal{C}$ if it satisfies the equivalent conditions of Proposition 1-1. We write $\mathrm{Grp}_{\mathbb{E}_\infty}(\mathcal{C})$ for the full subcategory of $\mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{C})$ consisting of the $\mathbb{E}_\infty$–groups.

Remark 1-3 There are similar equivalent characterizations as in the proposition for $\mathbb{E}_n$–monoids, $n \geq 1$. In fact, they can be applied more generally to algebras for monochromatic ∞–operads $\mathcal{O}$ equipped with a morphism $\mathbb{E}_1 \to \mathcal{O}$. In this case, these characterizations serve as a definition of $\mathcal{O}$*–groups*. Since an ordinary monoid having right-inverses is a group, we can use the fact that every morphism in $\mathrm{Ho}(\mathcal{C})$ lifts to a morphism in $\mathcal{C}$ to conclude that also the characterizations (i) and (iii) are equivalent to their respective two-sided variants, but in characterization (iv) one must instead use (very) special simplicial objects in $\mathcal{C}$.

Remark 1-4 Recall [20, Remark 5.2.6.9] that an $\mathbb{E}_n$–monoid object M of an ∞–topos $\mathcal{C}$ is said to be *grouplike* if (the sheaf) $\pi_0 M$ is a group object. In more general situations, such as for instance $\mathcal{C} = \mathrm{Cat}_\infty$, the correct π_0 is unclear, and in any case the resulting notion of "grouplike monoid" may not agree with that of "group".

Remark 1-5 In our definition of a group object we force the inversion morphism to be an actual morphism of the underlying objects in $\mathcal{C}$. In many situations, however, there is a natural inversion which is naturally only an anti-morphism. For example, this is the case in a tensor category with tensor inverses, or in the category of Poisson Lie groups. This suggests that there should be a notion of group object with such an anti-inversion morphism. It would be interesting to study such a notion, though we will not need this.

Given two ∞–categories $\mathcal{C}$ and $\mathcal{D}$ with finite products, we write $\mathrm{Fun}^{\Pi}(\mathcal{C},\mathcal{D})$ for the ∞–category of finite product preserving functors from $\mathcal{C}$ to $\mathcal{D}$. If $\mathcal{C}$ and $\mathcal{D}$ are complete, we write $\mathrm{Fun}^{\mathrm{R}}(\mathcal{C},\mathcal{D})$ for the ∞–category of limit preserving functors. In this situation, the ∞–category $\mathrm{Fun}^{\mathrm{R}}(\mathcal{C},\mathcal{D})$ is also complete and limits in $\mathrm{Fun}^{\mathrm{R}}(\mathcal{C},\mathcal{D})$ are formed pointwise in $\mathcal{D}$. This follows from the corresponding statement for $\mathrm{Fun}(\mathcal{C},\mathcal{D})$ and from the fact that such a pointwise limit of functors is again limit preserving.

Lemma 1-6 *If $\mathcal{C}$ and $\mathcal{D}$ are ∞–categories with finite products, then $\mathrm{Fun}^{\Pi}(\mathcal{C},\mathcal{D})$ also has finite products and we have canonical equivalences*

$$\mathrm{Mon}_{\mathbb{E}_\infty}(\mathrm{Fun}^{\Pi}(\mathcal{C},\mathcal{D}))\simeq \mathrm{Fun}^{\Pi}(\mathcal{C},\mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{D})),$$
$$\mathrm{Grp}_{\mathbb{E}_\infty}(\mathrm{Fun}^{\Pi}(\mathcal{C},\mathcal{D}))\simeq \mathrm{Fun}^{\Pi}(\mathcal{C},\mathrm{Grp}_{\mathbb{E}_\infty}(\mathcal{D})).$$

If $\mathcal{C}$ and $\mathcal{D}$ are complete, then so is $\mathrm{Fun}^{\mathrm{R}}(\mathcal{C},\mathcal{D})$, and we have canonical equivalences

$$\mathrm{Mon}_{\mathbb{E}_\infty}(\mathrm{Fun}^{\mathrm{R}}(\mathcal{C},\mathcal{D}))\simeq \mathrm{Fun}^{\mathrm{R}}(\mathcal{C},\mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{D})),$$
$$\mathrm{Grp}_{\mathbb{E}_\infty}(\mathrm{Fun}^{\mathrm{R}}(\mathcal{C},\mathcal{D}))\simeq \mathrm{Fun}^{\mathrm{R}}(\mathcal{C},\mathrm{Grp}_{\mathbb{E}_\infty}(\mathcal{D})).$$

Proof We only give the proof of the second case, as the first one is entirely analogous. As recalled above, an $\mathbb{E}_\infty$–monoid in an ∞–category $\mathcal{E}$ is given by a functor $M\colon \mathrm{N}(\mathrm{Fin}_*)\to\mathcal{E}$ satisfying the usual Segal condition, ie, the inert maps $\langle n\rangle\to\langle 1\rangle$ exhibit $M(\langle n\rangle)$ as the n–fold power of $M(\langle 1\rangle)$. We denote the full subcategory spanned by such functors by

$$\mathrm{Fun}^{\times}(\mathrm{N}(\mathrm{Fin}_*),\mathcal{E})\subseteq \mathrm{Fun}(\mathrm{N}(\mathrm{Fin}_*),\mathcal{E}).$$

Using this notation, we obtain a fully faithful inclusion

$$\begin{aligned}\mathrm{Mon}_{\mathbb{E}_\infty}(\mathrm{Fun}^{\mathrm{R}}(\mathcal{C},\mathcal{D}))&\simeq \mathrm{Fun}^{\times}(\mathrm{N}(\mathrm{Fin}_*),\mathrm{Fun}^{\mathrm{R}}(\mathcal{C},\mathcal{D}))\\&\subseteq \mathrm{Fun}(\mathrm{N}(\mathrm{Fin}_*),\mathrm{Fun}(\mathcal{C},\mathcal{D}))\simeq \mathrm{Fun}(\mathrm{N}(\mathrm{Fin}_*)\times\mathcal{C},\mathcal{D})\end{aligned}$$

whose essential image consists of those functors F such that $F(-,C)\colon \mathrm{N}(\mathrm{Fin}_*)\to\mathcal{D}$ is special for all $C\in\mathcal{C}$ and such that $F(\langle n\rangle,-)\colon\mathcal{C}\to\mathcal{D}$ preserves limits for all $\langle n\rangle\in\mathrm{N}(\mathrm{Fin}_*)$. This follows from the fact that limits in $\mathrm{Fun}^{\mathrm{R}}(\mathcal{C},\mathcal{D})$ are formed pointwise, as remarked above. In a similar vein, we obtain a fully faithful inclusion

$$\begin{aligned}\mathrm{Fun}^{\mathrm{R}}(\mathcal{C},\mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{D}))&\simeq \mathrm{Fun}^{\mathrm{R}}(\mathcal{C},\mathrm{Fun}^{\times}(\mathrm{N}(\mathrm{Fin}_*),\mathcal{D}))\\&\subseteq \mathrm{Fun}(\mathcal{C},\mathrm{Fun}(\mathrm{N}(\mathrm{Fin}_*),\mathcal{D}))\\&\simeq \mathrm{Fun}(\mathcal{C}\times\mathrm{N}(\mathrm{Fin}_*),\mathcal{D})\simeq \mathrm{Fun}(\mathrm{N}(\mathrm{Fin}_*)\times\mathcal{C},\mathcal{D})\end{aligned}$$

with the same essential image, concluding the proof for the case of monoids. The proof for the case of groups works exactly the same. In fact, using characterization (4) of Proposition 1-1, it suffices to replace special Γ–objects by very special Γ–objects. □

2 Preadditive and additive ∞–categories

An ∞–category is preadditive if finite coproducts and products exist and are equivalent. More precisely, we have the following definition.

Definition 2-1 An ∞–category $\mathcal{C}$ is *preadditive* if it is pointed, admits finite coproducts and finite products, and the canonical morphism $C_1 \sqcup C_2 \to C_1 \times C_2$ is an equivalence for all objects $C_1, C_2 \in \mathcal{C}$. In this case any such object will be denoted by $C_1 \oplus C_2$ and will be referred to as a *biproduct* of C_1 and C_2.

Let us collect a few immediate examples and closure properties of preadditive ∞–categories.

Example 2-2 An ordinary category $\mathcal{C}$ is preadditive if and only if $\mathrm{N}(\mathcal{C})$ is a preadditive ∞–category. Products and opposites of preadditive ∞–categories are preadditive. Clearly any ∞–category equivalent to a preadditive one is again preadditive. Finally, if $\mathcal{C}$ is a preadditive ∞–category and K is any simplicial set, then $\mathrm{Fun}(K, \mathcal{C})$ is preadditive. This follows immediately from the fact that (co)limits in functor categories are calculated pointwise [19, Corollary 5.1.2.3].

We will obtain more examples of preadditive ∞–categories from the following proposition, which gives a connection to Section 1.

Proposition 2-3 *Let $\mathcal{C}$ be an ∞–category with finite coproducts and products. Then the following are equivalent:*

(i) *The ∞–category $\mathcal{C}$ is preadditive.*

(ii) *The homotopy category $\mathrm{Ho}(\mathcal{C})$ is preadditive.*

(iii) *The ∞–operad $\mathcal{C}^{\sqcup} \to \mathrm{N}(\mathrm{Fin}_*)$ as constructed in [20, Construction 2.4.3.1] is cartesian [20, Definition 2.4.0.1].*

(iv) *The forgetful functor $\mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{C}) \to \mathcal{C}$ is an equivalence.*

Moreover, $\mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{C})$ is preadditive if $\mathcal{C}$ has finite products.

Proof Let us begin by proving that the first two statements are equivalent. The direction (i)$\Rightarrow$(ii) follows from the fact that the functor $\gamma \colon \mathcal{C} \to \mathrm{N}(\mathrm{Ho}(\mathcal{C}))$ preserves finite (co)products. For the converse direction, let us recall that a morphism in $\mathcal{C}$ is an equivalence if and only if γ sends it to an isomorphism. Now, by our assumption on $\mathrm{Ho}(\mathcal{C})$, the canonical map $C_1 \sqcup C_2 \to C_1 \times C_2$ in $\mathcal{C}$ is mapped to an isomorphism under γ and is hence an equivalence.

To show (i) $\Rightarrow$ (iii) we only need to check that the symmetric monoidal structure $\mathcal{C}^{\sqcup} \to \mathrm{N}(\mathrm{Fin}_*)$ exhibits finite tensor products (in this case the disjoint union) as products. But this follows directly from (i).

Now assume (iii) holds. Then by [20, Corollary 2.4.1.8] there exists an equivalence of symmetric monoidal structures $\mathcal{C}^{\sqcup} \simeq \mathcal{C}^{\times}$. Thus we get an induced equivalence

$$\mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{C}) \simeq \mathrm{Alg}_{\mathbb{E}_\infty}(\mathcal{C}^{\times}) \simeq \mathrm{Alg}_{\mathbb{E}_\infty}(\mathcal{C}^{\sqcup})$$

compatible with the forgetful functors to $\mathcal{C}$. But for the latter symmetric monoidal structure the forgetful functor $\mathrm{Alg}_{\mathbb{E}_\infty}(\mathcal{C}^{\sqcup}) \to \mathcal{C}$ always induces an equivalence, as shown in [20, Corollary 2.4.3.10].

Finally, assume (iv) holds. Then in order to show that $\mathcal{C}$ is preadditive it suffices to show that $\mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{C})$ is preadditive. To see that $\mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{C})$ is preadditive we note that limits in $\mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{C})$ are formed as the limits of the underlying objects of $\mathcal{C}$. In particular, the underlying object of the product in $\mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{C})$ is given by the product of the underlying objects. Coproducts are more complicated, but it is shown in [20, Proposition 3.2.4.7] that the underlying object of the coproduct is formed by the tensor product of the underlying objects, ie, by the product of the underlying objects in our case. Thus, the underlying object of the coproduct and the product are equivalent. But, by assumption, $\mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{C}) \to \mathcal{C}$ is fully faithful, so that we already have such an equivalence in $\mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{C})$. This implies (i) and concludes the proof. □

Corollary 2-4 *Let $\mathcal{C}$ and $\mathcal{D}$ be ∞–categories with finite products and suppose that either $\mathcal{C}$ or $\mathcal{D}$ is preadditive. Then the ∞–category $\mathrm{Fun}^{\Pi}(\mathcal{C},\mathcal{D})$ is preadditive.*

Proof If $\mathcal{D}$ is preadditive, then $\mathrm{Fun}(\mathcal{C},\mathcal{D})$ is also preadditive, and clearly $\mathrm{Fun}^{\Pi}(\mathcal{C},\mathcal{D}) \subseteq \mathrm{Fun}(\mathcal{C},\mathcal{D})$ is stable under products. In particular, given two product preserving functors $f,g\colon \mathcal{C} \to \mathcal{D}$, the pointwise product $f \times g\colon \mathcal{C} \to \mathcal{D}$ again lies in $\mathrm{Fun}^{\Pi}(\mathcal{C},\mathcal{D})$. Since (co)limits in $\mathrm{Fun}(\mathcal{C},\mathcal{D})$ are calculated pointwise [19, Corollary 5.1.2.3], we can use the preadditivity of $\mathcal{D}$ to conclude that $f \times g$ is also the coproduct $f \sqcup g$ of f and g in $\mathrm{Fun}(\mathcal{C},\mathcal{D})$, and hence, a posteriori, also the coproduct in $\mathrm{Fun}^{\Pi}(\mathcal{C},\mathcal{D})$. A similar reasoning yields a zero object in $\mathrm{Fun}^{\Pi}(\mathcal{C},\mathcal{D})$, and we conclude that $\mathrm{Fun}^{\Pi}(\mathcal{C},\mathcal{D})$ is preadditive.

The case in which $\mathcal{C}$ is preadditive is slightly more involved. Recall that a product preserving functor $f\colon \mathcal{C} \to \mathcal{D}$ induces a functor $\mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{C}) \to \mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{D})$ (simply by composing a special Γ–object in $\mathcal{C}$ with f). Since products in ∞–categories of $\mathbb{E}_\infty$–monoids are calculated in the underlying ∞–categories, this induced functor preserves products. Thus, we obtain a functor

$$\mathrm{Fun}^{\Pi}(\mathcal{C},\mathcal{D}) \to \mathrm{Fun}^{\Pi}(\mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{C}), \mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{D})).$$

By Proposition 2-3 we know that $\mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{D})$ is preadditive. The first part of this proof implies the same for $\mathrm{Fun}^{\Pi}(\mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{C}), \mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{D}))$, and hence we are done if we can show that the above functor is an equivalence. A functor in the reverse direction is given by composition with the equivalence $\mathcal{C} \simeq \mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{C})$ (use Proposition 2-3 again) and with $\mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{D}) \to \mathcal{D}$. It is easy to check that the resulting endofunctor of $\mathrm{Fun}^{\Pi}(\mathcal{C}, \mathcal{D})$ is equivalent to the identity, as is also the case for the other composition. □

Corollary 2-5 *Let $\mathcal{C}$ be an ∞–category with finite products and let $\mathcal{D}$ be a preadditive ∞–category.*

(i) *The ∞–category $\mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{C})$ is preadditive.*

(ii) *The forgetful functor $\mathrm{Mon}_{\mathbb{E}_\infty}(\mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{C})) \to \mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{C})$ is an equivalence.*

(iii) *There is an equivalence $\mathrm{Fun}^{\Pi}(\mathcal{D}, \mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{C})) \simeq \mathrm{Fun}^{\Pi}(\mathcal{D}, \mathcal{C})$.*

Proof The first assertion is a consequence of the proof of Proposition 2-3. The second follows immediately from that same proposition, while the last statement is implied by Lemma 1-6 and the observation that $\mathrm{Fun}^{\Pi}(\mathcal{D}, \mathcal{C})$ is preadditive whenever $\mathcal{D}$ is as guaranteed by Corollary 2-4. □

We now establish basically the analogous results for additive ∞–categories. As it is very similar to the case of preadditive ∞–categories, we leave out some of the details. Parallel to ordinary category theory, we introduce additive ∞–categories by imposing an additional exactness condition on preadditive ∞–categories. Let $\mathcal{C}$ be a preadditive ∞–category and let A be an object of $\mathcal{C}$. We know from Proposition 2-3 that A can be canonically endowed with the structure of an $\mathbb{E}_\infty$–monoid, and it is shown in [20, Section 2.4.3] that this structure is given by the fold map $\nabla\colon A \oplus A \to A$. The *shear map*

$$s\colon A \oplus A \to A \oplus A$$

is the projection $pr_1\colon A \oplus A \to A$ on the first factor and the fold map $\nabla\colon A \oplus A \to A$ on the second.

Definition 2-6 A preadditive ∞–category $\mathcal{C}$ is *additive* if, for every object $A \in \mathcal{C}$, the shear map $s\colon A \oplus A \xrightarrow{\sim} A \oplus A$ is an equivalence.

Examples 2-7 An ordinary category $\mathcal{C}$ is additive if and only if $\mathrm{N}(\mathcal{C})$ is an additive ∞–category. Products and opposites of additive ∞–categories are additive. If $\mathcal{C}$ is an additive ∞–category, then any ∞–category equivalent to $\mathcal{C}$ is additive ∞–category, and any functor ∞–category $\mathrm{Fun}(K, \mathcal{C})$ is additive.

The connection to $\mathbb{E}_\infty$–groups and hence to Section 1 is provided by the following analog of Proposition 2-3.

Proposition 2-8 *For an ∞–category $\mathcal{C}$ with finite products and coproducts, the following are equivalent:*

(i) *The ∞–category $\mathcal{C}$ is additive.*

(ii) *The homotopy category $\mathrm{Ho}(\mathcal{C})$ is additive.*

(iii) *The forgetful functor $\mathrm{Grp}_{\mathbb{E}_\infty}(\mathcal{C}) \to \mathcal{C}$ is an equivalence.*

Moreover, if $\mathcal{C}$ is an ∞–category with finite products, then $\mathrm{Grp}_{\mathbb{E}_\infty}(\mathcal{C})$ is additive.

Proof The proof of the equivalence of (i) and (iii) parallels the proof of Proposition 2-3. To see that (i) implies (iii) we note that by Proposition 2-3 we have an equivalence $\mathrm{Alg}_{\mathbb{E}_\infty}(\mathcal{C}^{\sqcup}) \simeq \mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{C}) \to \mathcal{C}$. But it is shown in [20, Section 2.4.3] that an inverse to this equivalence endows an object $A \in \mathcal{C}$ with the algebra structure given by the fold map $\nabla\colon A \oplus A \to A$. Now, the statement that such an algebra object is grouplike is equivalent to the shear map being an equivalence. Thus, invoking (i), we obtain an equivalence $\mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{C}) \simeq \mathrm{Grp}_{\mathbb{E}_\infty}(\mathcal{C})$, which gives (iii). Conversely, to see that (iii) implies (i), we need to show that $\mathrm{Grp}_{\mathbb{E}_\infty}(\mathcal{C})$ is additive. Preadditivity is clear and additivity follows from the characterization of groups given in Proposition 1-1. □

Corollary 2-9 *Let $\mathcal{C}$ and $\mathcal{D}$ be ∞–categories with finite products and suppose that either $\mathcal{C}$ or $\mathcal{D}$ is additive. Then the ∞–category $\mathrm{Fun}^{\Pi}(\mathcal{C}, \mathcal{D})$ is additive.*

Corollary 2-10 *Let $\mathcal{C}$ be an ∞–category with finite products and let $\mathcal{D}$ be an additive ∞–category.*

(i) *The ∞–category $\mathrm{Grp}_{\mathbb{E}_\infty}(\mathcal{C})$ is additive.*

(ii) *The forgetful functor $\mathrm{Grp}_{\mathbb{E}_\infty}(\mathrm{Grp}_{\mathbb{E}_\infty}(\mathcal{C})) \to \mathrm{Grp}_{\mathbb{E}_\infty}(\mathcal{C})$ is an equivalence.*

(iii) *There is an equivalence $\mathrm{Fun}^{\Pi}(\mathcal{D}, \mathrm{Grp}_{\mathbb{E}_\infty}(\mathcal{C})) \simeq \mathrm{Fun}^{\Pi}(\mathcal{D}, \mathcal{C})$.*

Remark 2-11 Corollaries 2-5 and 2-10 basically state that $\mathrm{Mon}_{\mathbb{E}_\infty}(-)$ and $\mathrm{Grp}_{\mathbb{E}_\infty}(-)$ are colocalizations of the ∞–category of ∞–categories with finite products and product preserving functors. Much of the remainder of the paper makes use of this observation, although we prefer to phrase things slightly differently: namely, $\mathrm{Mon}_{\mathbb{E}_\infty}(-)$ and $\mathrm{Grp}_{\mathbb{E}_\infty}(-)$ also induce colocalizations of Pr^{R}, which in turn (using the anti-equivalence between Pr^{L} and Pr^{R}) induce localizations of Pr^{L}. We have opted to state our results in term of localizations as we think they are slightly more intuitive from this perspective.

3 Smashing localizations

So far we have discussed ∞–categories with finite products. We now turn our attention to presentable ∞–categories. The primary purpose of this section is to review the notion of *smashing localizations*, which we then specialize to $\mathrm{Pr}^{\mathrm{L},\otimes}$ in order to deduce some important consequences which will play an essential role throughout the remainder of the paper.

Let $\mathcal{C}$ be an ∞–category. Recall that a *localization* of $\mathcal{C}$ is functor $L\colon \mathcal{C} \to \mathcal{D}$ which admits a fully faithful right adjoint $R\colon \mathcal{D} \to \mathcal{C}$. If $L\colon \mathcal{C} \to \mathcal{D}$ is a localization, then $\mathcal{D}$ is equivalent (via the fully faithful right adjoint) to a full subcategory $L\mathcal{C}$ of $\mathcal{C}$, called the subcategory of *local objects*. For this reason we typically identify localizations with reflective subcategories (ie, full subcategories such that the inclusion admits a left adjoint). We will also sometimes write L for the endofunctor of $\mathcal{C}$ obtained as the composite of $L\colon \mathcal{C} \to \mathcal{D}$ followed by the fully faithful right adjoint $R\colon \mathcal{D} \to \mathcal{C}$. Given such a localization, a map $X \to Y$ is a *local equivalence* if $LX \to LY$ is an equivalence.

Lemma 3-1 *Let $\mathcal{C}$ be an ∞–category and $M\colon \mathcal{C} \to \mathcal{C}$ an endofunctor equipped with a natural transformation $\eta\colon \mathrm{id} \to M$. Then M is equivalent to the composite $R \circ L$ of a localization $L\colon \mathcal{C} \to \mathcal{D}$ if and only if, for every object X of $\mathcal{C}$, the two obvious maps $M(X) \to M(M(X))$ are equivalences.*

Proof This is condition (3) of [19, Proposition 5.2.7.4]. □

If $\mathcal{C}$ has a symmetric monoidal structure $\mathcal{C}^{\otimes}$, then it is sometimes the case that a localization of $\mathcal{C}$ is given by *smashing* with a fixed object I of $\mathcal{C}$. In keeping with the terminology used in stable homotopy theory, we make the following definition.

Definition 3-2 Let $\mathcal{C}^{\otimes}$ be a symmetric monoidal ∞–category. We say that a localization $L\colon \mathcal{C} \to \mathcal{C}$ is *smashing* if it is of the form $L \simeq (-) \otimes I$ for some object I of $\mathcal{C}$.

Recall [20, Definition 4.8.2.1] that an *idempotent object* in $\mathcal{C}^{\otimes}$ is an object I together with a morphism from the tensor unit such that the two obvious maps $I \to I \otimes I$ are equivalences. It follows that the endofunctor of $\mathcal{C}$ given by tensoring with I is a localization [20, Proposition 4.8.2.4]. Conversely for a smashing localization $L \simeq (-) \otimes I$ the object I is necessarily an idempotent commutative algebra object of $\mathcal{C}$. In other words, showing that the functor $(-) \otimes I$ is a localization is the same as endowing I with the structure of an idempotent commutative algebra object of $\mathcal{C}$. This provides a one-to-one correspondence between smashing localizations and idempotent commutative algebra objects.

There are two obvious key features of smashing localizations: first, they preserve colimits (provided the tensor structure is compatible with colimits, which is always the case if it is closed), and second, they are symmetric monoidal in the sense of the following definition.

Definition 3-3 Let $\mathcal{C}^{\otimes}$ be a symmetric monoidal ∞–category equipped with a localization $L\colon \mathcal{C} \to \mathcal{D}$ of the underlying ∞–category $\mathcal{C}$. Then L is *compatible with the symmetric monoidal structure* (or simply *symmetric monoidal*) if, whenever $X \to Y$ is a local equivalence, then so is $X \otimes Z \to Y \otimes Z$ for any object Z of $\mathcal{C}$.

Given such a localization, the subcategory $\mathcal{D} \simeq L\mathcal{C}$ of local objects inherits a symmetric monoidal structure from that of $\mathcal{C}$. This is the content of the following lemma which also justifies the terminology *symmetric monoidal* localization. Identifying $\mathcal{D}$ with the full subcategory $L\mathcal{C}$ of local objects, let $R^{\otimes}\colon \mathcal{D}^{\otimes} \subseteq \mathcal{C}^{\otimes}$ be the inclusion of the full subcategory consisting of those objects $X_1 \oplus \cdots \oplus X_n$ such that each X_i is in $\mathcal{D}$.

Lemma 3-4 *Let $\mathcal{C}^{\otimes}$ be a symmetric monoidal ∞–category equipped with a symmetric monoidal localization $L\colon \mathcal{C} \to \mathcal{D}$. Then there is a symmetric monoidal structure $\mathcal{D}^{\otimes}$ on $\mathcal{D}$ such that L extends to a symmetric monoidal functor $L^{\otimes}\colon \mathcal{C}^{\otimes} \to \mathcal{D}^{\otimes}$ and such that the right adjoint $R^{\otimes}\colon \mathcal{D}^{\otimes} \to \mathcal{C}^{\otimes}$ is lax symmetric monoidal.*

Proof This is a special case of [20, Proposition 2.2.1.9]. $\square$

Remark 3-5 If $\mathcal{C}^{\otimes}$ is a closed symmetric monoidal ∞–category equipped with a symmetric monoidal localization $L\colon \mathcal{C} \to \mathcal{C}$. Then L is compatible with the closed structure in the sense that, for every pair of objects C and D of $\mathcal{C}$, the localization $C \to LC$ induces an equivalence

$$D^{LC} \simeq D^{C}$$

whenever D is local. This follows immediately from the definition.

Lemma 3-6 *Let $\mathcal{C}^{\otimes}$ be a symmetric monoidal ∞–category equipped with a symmetric monoidal localization $L\colon \mathcal{C} \to \mathcal{D}$, and let $R\colon \mathcal{D} \to \mathcal{C}$ denote the right adjoint of L. Then there is an induced localization $L'\colon \mathrm{Alg}_{\mathbb{E}_\infty}(\mathcal{C}) \to \mathrm{Alg}_{\mathbb{E}_\infty}(\mathcal{D})$ such that the diagram*

$$\begin{tikzcd} \mathrm{Alg}_{\mathbb{E}_\infty}(\mathcal{C}) \arrow[r, "L'"] \arrow[d] & \mathrm{Alg}_{\mathbb{E}_\infty}(\mathcal{D}) \arrow[d] \\ \mathcal{C} \arrow[r, "L"] & \mathcal{D} \end{tikzcd}$$

commutes. Moreover, given $A \in \mathrm{Alg}_{\mathbb{E}_\infty}(\mathcal{C})$*, there exists a unique commutative algebra structure on* RLA *such that unit map* $A \to RLA$ *extends to a morphism of commutative algebras.*

Proof By Lemma 3-4 above, we obtain maps $L'\colon \mathrm{Alg}_{\mathbb{E}_\infty}(\mathcal{C}) \to \mathrm{Alg}_{\mathbb{E}_\infty}(\mathcal{D})$ and $R'\colon \mathrm{Alg}_{\mathbb{E}_\infty}(\mathcal{D}) \to \mathrm{Alg}_{\mathbb{E}_\infty}(\mathcal{C})$ by composing sections $\mathbb{E}_\infty \to \mathcal{C}^\otimes$ with $L^\otimes$ and sections $\mathbb{E}_\infty \to \mathcal{D}^\otimes$ with $R^\otimes$, respectively. In a similar fashion we also obtain unit and counit transformations such that the counit is an equivalence. It follows that L' is a localization.

For the second assertion, we know already that $R'L'A$ comes with a canonical commutative algebra map $\eta'\colon A \to R'L'A$, the adjunction unit evaluated at A, and that this map extends the adjunction unit $\eta\colon A \to RLA$ of the underlying objects. If $\eta''\colon A \to R'B$ is a second such map of commutative algebras, then the universality of η' implies that η'' factors essentially uniquely as

$$\phi \circ \eta'\colon A \to R'L'A \to R'B.$$

Since the underlying map of ϕ is an identity, if follows that ϕ itself is an equivalence since $\mathrm{Alg}_{\mathbb{E}_\infty}(\mathcal{C}) \to \mathcal{C}$ is conservative. We can now conclude since the space of reflections of a fixed object in a full subcategory is contractible if non-empty. □

Remark 3-7 The second part of the lemma implies that RLA can be turned into an $\mathbb{E}_\infty$–algebra such that the unit map $A \to RLA$ can be enhanced to a morphism of $\mathbb{E}_\infty$–algebras. Moreover, the space of such enhancements is contractible. In particular, if RLA is endowed with two different $\mathbb{E}_\infty$–algebra structures, then the identity morphism of the underlying objects in $\mathcal{D}$ can be essentially uniquely turned into an equivalence of these two $\mathbb{E}_\infty$–algebras compatible with the localizations. We will apply this in Section 5 to smashing localizations on Pr^{L}.

Now we specialize to the case of the (very large) ∞–category Pr^{L} of presentable ∞–categories and colimit-preserving functors. We will write $\mathcal{C}$, $\mathcal{D}$, etc for objects of Pr^{L}. Recall that Pr^{L} admits a closed symmetric monoidal structure which is uniquely characterized as follows: given presentable ∞–categories $\mathcal{C}$ and $\mathcal{D}$, their tensor product $\mathcal{C} \otimes \mathcal{D}$ corepresents the functor $\mathrm{Pr}^{\mathrm{L}} \to \widehat{\mathrm{Cat}}_\infty$ which sends $\mathcal{E}$ to

$$\mathrm{Fun}^{\mathrm{L,L}}(\mathcal{C} \times \mathcal{D}, \mathcal{E}) \subseteq \mathrm{Fun}(\mathcal{C} \times \mathcal{D}, \mathcal{E}),$$

the full subcategory consisting of those functors $F\colon \mathcal{C} \times \mathcal{D} \to \mathcal{E}$ which preserve colimits separately in each variable. The unit of this monoidal structure on Pr^{L} is the

∞–category $\mathcal{S}$ of spaces, as follows from the fact that $\mathrm{Fun}^{\mathrm{L}}(\mathcal{S},\mathcal{C}) \simeq \mathcal{C}$ [20, Example 6.3.1.19]. Moreover, by [20, Proposition 4.8.1.16] this tensor product admits the description

$$\mathcal{C} \otimes \mathcal{D} \simeq \mathrm{Fun}^{\mathrm{R}}(\mathcal{C}^{\mathrm{op}}, \mathcal{D}).$$

Recall that $\mathrm{Fun}^{\mathrm{L}}(\mathcal{C},\mathcal{D})$ is presentable ([19, Propositon 5.5.3.8]). It is immediate from the definition of $\mathcal{C}\otimes\mathcal{D}$ as a corepresenting object that the symmetric monoidal structure on Pr^{L} is closed, with right adjoint to $\mathcal{C}\otimes(-)\colon \mathrm{Pr}^{\mathrm{L}} \to \mathrm{Pr}^{\mathrm{L}}$ given by $\mathrm{Fun}^{\mathrm{L}}(\mathcal{C},-)\colon \mathrm{Pr}^{\mathrm{L}} \to \mathrm{Pr}^{\mathrm{L}}$. Lastly, the (possibly large) mapping spaces in Pr^{L} are given by the formula

$$\mathrm{Map}_{\mathrm{Pr}^{\mathrm{L}}}(\mathcal{C},\mathcal{D}) \simeq \mathrm{Fun}^{\mathrm{L}}(\mathcal{C},\mathcal{D})^{\sim},$$

the maximal subgroupoid. This description will be applied in Section 4 to our context of monoids and groups.

Proposition 3-8 *Let $L\colon \mathrm{Pr}^{\mathrm{L}} \to \mathrm{Pr}^{\mathrm{L}}$ be a smashing localization or, more generally, a symmetric monoidal localization, and let $\mathcal{C}$ and $\mathcal{D}$ be presentable ∞–categories such that $\mathcal{D}$ is in the essential image of L.*

(i) *The map $\mathrm{Fun}^{\mathrm{L}}(L\mathcal{C},\mathcal{D}) \to \mathrm{Fun}^{\mathrm{L}}(\mathcal{C},\mathcal{D})$ induced by the localization $\mathcal{C} \to L\mathcal{C}$ is an equivalence.*

(ii) *If L is smashing, then the localization of the (very large) ∞–category Pr^{L} of presentable ∞–categories and colimit-preserving functors is equivalent to the ∞–category of modules over $L\mathcal{S} \in \mathrm{Alg}_{\mathbb{E}_\infty}(\mathrm{Pr}^{\mathrm{L}})$:*

$$L\mathrm{Pr}^{\mathrm{L}} \simeq \mathrm{Mod}_{L\mathcal{S}}(\mathrm{Pr}^{\mathrm{L}}).$$

(iii) *Given a second symmetric monoidal localization $L'\colon \mathrm{Pr}^{\mathrm{L}} \to \mathrm{Pr}^{\mathrm{L}}$ such that $L'\mathrm{Pr}^{\mathrm{L}} \subseteq L\mathrm{Pr}^{\mathrm{L}}$, then the canonical morphism $L\mathcal{C} \to L'\mathcal{C}$ induces an equivalence $\mathrm{Fun}^{\mathrm{L}}(L'\mathcal{C},\mathcal{D}) \to \mathrm{Fun}^{\mathrm{L}}(L\mathcal{C},\mathcal{D})$ for every L'–local $\mathcal{D}$.*

Proof The first statement follows from Remark 3-5 and the second from [20, Proposition 4.8.2.10]. Finally, the third one follows immediately from the first and the two-out-of-three property of equivalences. □

Let us now consider a presentable ∞–category endowed with a closed symmetric monoidal structure $\mathcal{C}^{\otimes}$. In this context the closedness is equivalent to the fact that the monoidal structure preserves colimits separately in each variable, ie, $\mathcal{C}^{\otimes}$ is essentially just a commutative algebra object in Pr^{L} [20, Remark 4.8.1.9].

Proposition 3-9 *Let $L\colon \mathrm{Pr}^{\mathrm{L}} \to \mathrm{Pr}^{\mathrm{L}}$ be a smashing localization or, more generally, a symmetric monoidal localization. Let $\mathcal{C}^{\otimes}$ and $\mathcal{D}^{\otimes}$ be closed symmetric monoidal presentable ∞–categories.*

(i) *The ∞–category $L\mathcal{C}$ admits a* unique *closed symmetric monoidal structure such that the localization map $\mathcal{C} \to L\mathcal{C}$ is a symmetric monoidal functor.*

(ii) *The map $\mathrm{Fun}^{\mathrm{L},\otimes}(L\mathcal{C}, \mathcal{D}) \to \mathrm{Fun}^{\mathrm{L},\otimes}(\mathcal{C}, \mathcal{D})$ induced by the localization $\mathcal{C} \to L\mathcal{C}$ is an equivalence whenever $\mathcal{D}$ is L–local.*

(iii) *Given a second symmetric monoidal localization $L'\colon \mathrm{Pr}^{\mathrm{L}} \to \mathrm{Pr}^{\mathrm{L}}$ such that $L'\mathrm{Pr}^{\mathrm{L}} \subseteq L\mathrm{Pr}^{\mathrm{L}}$, the induced morphism $L\mathcal{C} \to L'\mathcal{C}$ admits a unique symmetric monoidal structure. In particular, for every L'–local $\mathcal{D}$ the induced map $\mathrm{Fun}^{\mathrm{L},\otimes}(L'\mathcal{C}, \mathcal{D}) \to \mathrm{Fun}^{\mathrm{L},\otimes}(L\mathcal{C}, \mathcal{D})$ is an equivalence.*

Proof Statement (i) follows from Lemma 3-6, which induces an equivalence

$$\mathrm{Fun}^{\mathrm{L},\otimes}(L\mathcal{C}, \mathcal{D}^K) \to \mathrm{Fun}^{\mathrm{L},\otimes}(\mathcal{C}, \mathcal{D}^K)$$

on underlying ∞–groupoids for any simplicial set K such that $\mathcal{D}^K$ is local. Then (ii) follows from the fact that $\mathrm{Alg}_{\mathbb{E}_\infty}(\mathrm{Pr}^{\mathrm{L}})$ is cotensored over Cat_∞ in such a way that $\mathcal{D}^K$ is local whenever $\mathcal{D}$ is local; indeed, the cotensor $\mathcal{D}^K$ is given by the internal mapping object $\mathrm{Fun}^{\mathrm{L}}(\mathcal{P}(K), \mathcal{D})$, and this is a local object since $(-) \otimes \mathcal{P}(K)$ preserves local equivalences by assumption. Finally, (iii) is obtained by the same argument as (i) after replacing Pr^{L} with $L\mathrm{Pr}^{\mathrm{L}}$, which has an induced closed symmetric monoidal structure, $L\colon \mathrm{Pr}^{\mathrm{L}} \to \mathrm{Pr}^{\mathrm{L}}$ with the functor $L\mathrm{Pr}^{\mathrm{L}} \to L\mathrm{Pr}^{\mathrm{L}}$ induced by the composite $\mathrm{Pr}^{\mathrm{L}} \to L'\mathrm{Pr}^{\mathrm{L}} \subseteq L\mathrm{Pr}^{\mathrm{L}}$, and $\mathcal{C}$ with $L\mathcal{C}$, which also inherits a closed symmetric monoidal structure. □

We shall see in the next section that formation of ∞–categories of commutative monoid and group objects in a presentable ∞–category $\mathcal{C}$ are instances of smashing localizations of Pr^{L}. For the moment, it is worth mentioning that there are other well-known examples of smashing localizations of Pr^{L}. The most obvious one is the functor which associates to a presentable ∞–category $\mathcal{C}$ its ∞–category $\mathcal{C}_*$ of *pointed objects*; the fact that this is a smashing localization follows from the formula

$$\mathcal{C}_* \simeq \mathcal{C} \otimes \mathcal{S}_*$$

and the fact that $\mathcal{S}_*$ is an idempotent object of Pr^{L} [20, Proposition 4.8.2.11]. An important feature of $\mathcal{S}_*$ is that it is symmetric monoidal under the *smash product*, which is uniquely characterized by the requirement that the unit map $\mathcal{S} \to \mathcal{S}_*$ is symmetric monoidal. A further example of a smashing localization which is central to this paper is the passage from a presentable ∞–category $\mathcal{C}$ to the ∞–category $\mathrm{Sp}(\mathcal{C})$ of spectrum objects in $\mathcal{C}$ [20, Proposition 4.8.2.18].

4 Commutative monoids and groups as smashing localizations

In this section we show that the passage to ∞–categories of commutative monoids or groups are instances of smashing localizations of Pr^{L}.

Proposition 4-1 *Given a presentable ∞–category $\mathcal{C}$, then also the ∞–categories $\mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{C})$ and $\mathrm{Grp}_{\mathbb{E}_\infty}(\mathcal{C})$ are presentable.*

Proof By definition the ∞–categories $\mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{C})$ and $\mathrm{Grp}_{\mathbb{E}_\infty}(\mathcal{C})$ are full subcategories of the presentable ∞–category $\mathrm{Fun}(\mathrm{N}(\mathrm{Fin}_*), \mathcal{C})$. Therefore, it suffices to show that the monoids and groups, respectively, are precisely the S–local objects for a small collection S of morphisms in $\mathrm{Fun}(\mathrm{N}(\mathrm{Fin}_*), \mathcal{C})$ [19, Proposition 5.5.4.15]. We will give the details for the case of monoids and leave the case of groups to the reader.

In order to define S we first note that the evaluation functors

$$\mathrm{ev}_{\langle n\rangle}\colon \mathrm{Fun}(\mathrm{N}(\mathrm{Fin}_*), \mathcal{C}) \to \mathcal{C}$$

admit left adjoints $F_{\langle n\rangle}\colon \mathcal{C} \to \mathrm{Fun}(\mathrm{N}(\mathrm{Fin}_*), \mathcal{C})$. Now, $M \in \mathrm{Fun}(\mathrm{N}(\mathrm{Fin}_*), \mathcal{C})$ belongs to $\mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{C})$ if for every $n \in \mathbb{N}$ the morphism $M(\langle n\rangle) \to \prod M(\langle 1\rangle)$ is an equivalence in $\mathcal{C}$, and this is the case if and only if for every $C \in \mathcal{C}$ the morphism

$$\mathrm{Map}_{\mathcal{C}}(C, M(\langle n\rangle)) \to \prod \mathrm{Map}_{\mathcal{C}}(C, M(\langle 1\rangle)) \tag{4-2}$$

is an equivalence of spaces. Since $\mathcal{C}$ is accessible it suffices to check this for objects in $\mathcal{C}^\kappa$, the essentially small subcategory of κ–compact objects for some regular cardinal κ. Now we use the equivalences

$$\mathrm{Map}_{\mathcal{C}}(C, M(\langle n\rangle)) \simeq \mathrm{Map}_{\mathrm{Fun}(\mathrm{N}(\mathrm{Fin}_*),\mathcal{C})}(F_{\langle n\rangle}(C), M),$$

$$\prod \mathrm{Map}_{\mathcal{C}}(C, M(\langle 1\rangle)) \simeq \mathrm{Map}_{\mathrm{Fun}(\mathrm{N}(\mathrm{Fin}_*),\mathcal{C})}\left(\coprod F_{\langle 1\rangle}(C), M\right)$$

and see that the morphism (4-2) is induced by a morphism $\phi_{n,C}\colon \coprod_n F_{\langle 1\rangle}(C) \to F_{\langle n\rangle}(C)$ in $\mathrm{Fun}(\mathrm{N}(\mathrm{Fin}_*), \mathcal{C})$. Thus we may take S to consist of the $\phi_{n,C}$, where C ranges over any small collections of objects of $\mathcal{C}$ which contains a representative of each equivalence class of object in $\mathcal{C}^\kappa$. □

Remark 4-3 The proof for groups is similar, though we have to add more maps to the set S to account for the *very* special condition. This tells us in particular that $\mathrm{Grp}_{\mathbb{E}_\infty}(\mathcal{C})$ is a reflective subcategory of $\mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{C})$.

Corollary 4-4 *Let $\mathcal{C}$ be a presentable ∞–category. Then there are functors*

$$\mathcal{C} \to \mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{C}) \to \mathrm{Grp}_{\mathbb{E}_\infty}(\mathcal{C}),$$

which are left adjoint to the respective forgetful functors.

Proof Since limits in $\mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{C})$ and $\mathrm{Grp}_{\mathbb{E}_\infty}(\mathcal{C})$ are computed as the limits of the underlying objects, this follows from the adjoint functor theorem. □

Remark 4-5 Let $\mathcal{C}$ be a presentable ∞–category. The functor

$$\mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{C}) \to \mathrm{Grp}_{\mathbb{E}_\infty}(\mathcal{C}),$$

left adjoint to the forgetful functor $\mathrm{Grp}_{\mathbb{E}_\infty}(\mathcal{C}) \to \mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{C})$, is called the *group completion*. Thus, in the framework of ∞–categories, the group completion $\mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{C}) \to \mathrm{Grp}_{\mathbb{E}_\infty}(\mathcal{C})$ has the expected universal property, defining a left adjoint to the forgetful functor $\mathrm{Grp}_{\mathbb{E}_\infty}(\mathcal{C}) \to \mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{C})$.

The following theorem, while straightforward to prove, is central.

Theorem 4-6 *The assignments $\mathcal{C} \mapsto \mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{C})$ and $\mathcal{C} \mapsto \mathrm{Grp}_{\mathbb{E}_\infty}(\mathcal{C})$ refine to smashing localizations of Pr^{L}. Thus, we have, in particular, equivalences of ∞–categories*

$$\mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{C}) \simeq \mathcal{C} \otimes \mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{S}) \quad \text{and} \quad \mathrm{Grp}_{\mathbb{E}_\infty}(\mathcal{C}) \simeq \mathcal{C} \otimes \mathrm{Grp}_{\mathbb{E}_\infty}(\mathcal{S}).$$

The local objects are precisely the preadditive presentable ∞–categories and the additive presentable ∞–categories, respectively.

Proof The description of the tensor product of presentable ∞–categories together with Lemma 1-6 gives us the chain of equivalences

$$\begin{aligned}\mathcal{C} \otimes \mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{D}) \simeq \mathrm{Fun}^{\mathrm{R}}(\mathcal{C}^{\mathrm{op}}, \mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{D})) &\simeq \mathrm{Mon}_{\mathbb{E}_\infty}(\mathrm{Fun}^{\mathrm{R}}(\mathcal{C}^{\mathrm{op}}, \mathcal{D})) \\ &\simeq \mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{C} \otimes \mathcal{D}).\end{aligned}$$

In particular, we have $\mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{C}) \simeq \mathcal{C} \otimes \mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{S})$. The fact that $\mathrm{Mon}_{\mathbb{E}_\infty}$ is a localization follows from Corollary 2-5. The local objects are precisely the presentable ∞–categories $\mathcal{C}$ for which the canonical functor is an equivalence $\mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{C}) \simeq \mathcal{C}$, hence by Proposition 2-3 precisely the preadditive ∞–categories. The case of groups is established along the same lines. □

As a consequence we obtain the following result.

Corollary 4-7 *Let $\mathcal{C}$ and $\mathcal{D}$ be presentable ∞–category. Then there are canonical equivalences*

$$\mathcal{C} \otimes \mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{D}) \simeq \mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{C} \otimes \mathcal{D}) \simeq \mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{C}) \otimes \mathcal{D},$$
$$\mathcal{C} \otimes \mathrm{Grp}_{\mathbb{E}_\infty}(\mathcal{D}) \simeq \mathrm{Grp}_{\mathbb{E}_\infty}(\mathcal{C} \otimes \mathcal{D}) \simeq \mathrm{Grp}_{\mathbb{E}_\infty}(\mathcal{C}) \otimes \mathcal{D}.$$

Let us denote the full subcategories of Pr^{L} spanned by the preadditive and additive ∞–categories respectively by

$$\mathrm{Pr}^{\mathrm{L}}_{\mathrm{Pre}} \subseteq \mathrm{Pr}^{\mathrm{L}} \quad \text{and} \quad \mathrm{Pr}^{\mathrm{L}}_{\mathrm{Add}} \subseteq \mathrm{Pr}^{\mathrm{L}}.$$

Then Proposition 3-8 specializes to the following two corollaries.

Corollary 4-8 *The forgetful functors*

$$\mathrm{Mod}_{\mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{S})}(\mathrm{Pr}^{\mathrm{L}}) \to \mathrm{Pr}^{\mathrm{L}} \quad \textit{and} \quad \mathrm{Mod}_{\mathrm{Grp}_{\mathbb{E}_\infty}(\mathcal{S})}(\mathrm{Pr}^{\mathrm{L}}) \to \mathrm{Pr}^{\mathrm{L}}$$

induce equivalences of ∞–categories

$$\mathrm{Mod}_{\mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{S})}(\mathrm{Pr}^{\mathrm{L}}) \simeq \mathrm{Pr}^{\mathrm{L}}_{\mathrm{Pre}} \quad \textit{and} \quad \mathrm{Mod}_{\mathrm{Grp}_{\mathbb{E}_\infty}(\mathcal{S})}(\mathrm{Pr}^{\mathrm{L}}) \simeq \mathrm{Pr}^{\mathrm{L}}_{\mathrm{Add}}.$$

Corollary 4-9 *Let $\mathcal{C}$ and $\mathcal{D}$ be presentable ∞–categories.*

(i) *If $\mathcal{D}$ is preadditive then the free $\mathbb{E}_\infty$–monoid functor $\mathcal{C} \to \mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{C})$ induces an equivalence of ∞–categories*

$$\mathrm{Fun}^{\mathrm{L}}(\mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{C}), \mathcal{D}) \xrightarrow{\simeq} \mathrm{Fun}^{\mathrm{L}}(\mathcal{C}, \mathcal{D}),$$

exhibiting $\mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{C})$ *as the* free preadditive presentable ∞–category generated by $\mathcal{C}$. *In particular, we have canonical equivalences*

$$\mathrm{Fun}^{\mathrm{L}}(\mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{S}), \mathcal{D}) \xrightarrow{\simeq} \mathrm{Fun}^{\mathrm{L}}(\mathcal{S}, \mathcal{D}) \xrightarrow{\simeq} \mathcal{D}, u$$

exhibiting $\mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{S})$ *as the free preadditive presentable ∞–category on one generator.*

(ii) *If $\mathcal{D}$ is additive then the free $\mathbb{E}_\infty$–group functor $\mathcal{C} \to \mathrm{Grp}_{\mathbb{E}_\infty}(\mathcal{C})$ induces an equivalence of ∞–categories*

$$\mathrm{Fun}^{\mathrm{L}}(\mathrm{Grp}_{\mathbb{E}_\infty}(\mathcal{C}), \mathcal{D}) \xrightarrow{\simeq} \mathrm{Fun}^{\mathrm{L}}(\mathcal{C}, \mathcal{D}),$$

exhibiting $\mathrm{Grp}_{\mathbb{E}_\infty}(\mathcal{C})$ *as the* free additive presentable ∞–category generated by $\mathcal{C}$. *In particular, the free $\mathbb{E}_\infty$–group functor $\mathcal{S} \to \mathrm{Grp}_{\mathbb{E}_\infty}(\mathcal{S})$ induces canonical equivalences*

$$\mathrm{Fun}^{\mathrm{L}}(\mathrm{Grp}_{\mathbb{E}_\infty}(\mathcal{S}), \mathcal{D}) \xrightarrow{\simeq} \mathrm{Fun}^{\mathrm{L}}(\mathcal{S}, \mathcal{D}) \xrightarrow{\simeq} \mathcal{D},$$

exhibiting $\mathrm{Grp}_{\mathbb{E}_\infty}(\mathcal{S})$ *as the free additive, presentable ∞–category on one generator.*

The results of this section give us a refined picture of the stabilization process of presentable ∞–categories as we describe it in the next corollary (we will obtain a further *monoidal* refinement in Corollary 5-5). In [20, Chapter 1] it is shown that the stabilization of a presentable ∞–category $\mathcal{C}$ is given by the ∞–category $\mathrm{Sp}(\mathcal{C})$ of *spectrum objects* in $\mathcal{C}$, which is to say the limit

$$\mathrm{Sp}(\mathcal{C}) \simeq \lim\{\mathcal{C}_* \xleftarrow{\Omega} \mathcal{C}_* \xleftarrow{\Omega} \mathcal{C}_* \xleftarrow{\Omega} \cdots\},$$

taken in the ∞–category of (not necessarily small) ∞–categories, or equivalently in the ∞–category Pr^{R} of presentable ∞–categories by [19, Theorem 5.5.3.18]. Alternatively, $\mathrm{Sp}(\mathcal{C})$ is equivalent to the ∞–category of reduced excisive functors

$$\mathrm{Sp}(\mathcal{C}) \simeq \mathrm{Exc}_*(\mathcal{S}_*^{\mathrm{fin}}, \mathcal{C});$$

see [20, Section 1.4.2] for details. Recall from [20, Proposition 1.4.4.4] that for such a $\mathcal{C}$ the ∞–category $\mathrm{Sp}(\mathcal{C})$ is related to $\mathcal{C}$ by the suspension spectrum adjunction $(\Sigma^\infty_+, \Omega^\infty_-)\colon \mathcal{C} \rightleftarrows \mathrm{Sp}(\mathcal{C})$.

Corollary 4-10 *The stabilization of presentable ∞–categories $\mathrm{Pr}^{\mathrm{L}} \to \mathrm{Pr}^{\mathrm{L}}_{\mathrm{St}}$ factors as a composition of adjunctions*

$$\mathrm{Pr}^{\mathrm{L}} \rightleftarrows \mathrm{Pr}^{\mathrm{L}}_{\mathrm{Pt}} \rightleftarrows \mathrm{Pr}^{\mathrm{L}}_{\mathrm{Pre}} \rightleftarrows \mathrm{Pr}^{\mathrm{L}}_{\mathrm{Add}} \rightleftarrows \mathrm{Pr}^{\mathrm{L}}_{\mathrm{St}}.$$

In particular, if $\mathcal{C}$ is a presentable ∞–category, then $\Sigma^\infty_+\colon \mathcal{C} \to \mathrm{Sp}(\mathcal{C})$ factors as a composition of left adjoints

$$\Sigma^\infty_+\colon \mathcal{C} \to \mathcal{C}_* \to \mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{C}) \to \mathrm{Grp}_{\mathbb{E}_\infty}(\mathcal{C}) \to \mathrm{Sp}(\mathcal{C}),$$

each of which is uniquely determined by the fact that it commutes with the corresponding free functors from $\mathcal{C}$.

Proof This follows from Corollary 4-9 and the corresponding corollary for the functor $(-)_+\colon \mathcal{C} \to \mathcal{C}_*$ together with the facts that $\mathrm{Sp}(\mathcal{C})$ is additive [20, Corollary 1.4.2.17 and Remark 1.1.3.5], $\mathrm{Grp}_{\mathbb{E}_\infty}(\mathcal{C})$ is preadditive (even additive by Corollary 2-10), and $\mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{C})$ is pointed (in fact, preadditive by Corollary 2-5). For the second statement, it suffices to use Proposition 3-8. □

5 Canonical symmetric monoidal structures

Let us now assume that $\mathcal{C}$ is a presentable ∞–category endowed with a closed symmetric monoidal structure $\mathcal{C}^{\otimes}$. In this section we specialize the general results from Section 3 (or more specifically Proposition 3-9) to the localizations $(-)_*$, $\mathrm{Mon}_{\mathbb{E}_\infty}(-)$,

$\mathrm{Grp}_{\mathbb{E}_\infty}(-)$, and $\mathrm{Sp}(-)$. The two cases of $\mathcal{C}_*$ and $\mathrm{Sp}(\mathcal{C})$ are already essentially covered in [20, Section 4.8.2], but since these results are not stated explicitly, we include them here for the sake of completeness.

Theorem 5-1 *Let $\mathcal{C}^\otimes$ be a closed symmetric monoidal structure on a presentable ∞–category $\mathcal{C}$. The ∞–categories $\mathcal{C}_*$, $\mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{C})$, $\mathrm{Grp}_{\mathbb{E}_\infty}(\mathcal{C})$, and $\mathrm{Sp}(\mathcal{C})$ all admit closed symmetric monoidal structures, which are uniquely determined by the requirement that the respective free functors from $\mathcal{C}$ are symmetric monoidal. Moreover, each of the functors*

$$\mathcal{C}_* \to \mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{C}) \to \mathrm{Grp}_{\mathbb{E}_\infty}(\mathcal{C}) \to \mathrm{Sp}(\mathcal{C})$$

uniquely extends to a symmetric monoidal functor.

Proof This follows directly from the fact that the localizations are smashing using Proposition 3-9. □

From now on, when considered as symmetric monoidal ∞–categories, these ∞–categories are always endowed with the canonical monoidal structures of the theorem.

Warning 5-2 The reader should not confuse the two symmetric monoidal structures on $\mathcal{C}$ that are used in the above construction. The first one is the cartesian structure $\mathcal{C}^\times$ which is used to define the ∞–category $\mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{C})$ of $\mathbb{E}_\infty$–monoids. The second one is the closed symmetric monoidal structure $\mathcal{C}^\otimes$ which induces a monoidal structure on $\mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{C})$ as described in the theorem. In applications, these two monoidal structures on $\mathcal{C}$ often agree, which amounts to assuming that $\mathcal{C}$ is cartesian closed. This is the case in the most important examples, namely ∞–topoi (such as $\mathcal{S}$) and Cat_∞.

Example 5-3 (i) The (nerve of the) category Set of sets is a cartesian closed presentable ∞–category, and $\mathrm{Grp}_{\mathbb{E}_\infty}(\mathrm{Set})$ is just the (nerve of the) category Ab of abelian groups. The free functor $\mathrm{Set} \to \mathrm{Ab}$ can then of course be turned into a symmetric monoidal functor with respect to the cartesian product on Set and the usual tensor product on Ab. Thus, in this very special case, the theorem reproduces the classical tensor product of abelian groups.

(ii) The ∞–category $\mathcal{S}$ of spaces is a cartesian closed presentable ∞–category. The ∞–category $\mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{S})$ of $\mathbb{E}_\infty$–spaces hence comes with a canonical closed symmetric monoidal structure, as does the ∞–category $\mathrm{Grp}_{\mathbb{E}_\infty}(\mathcal{S})$ of grouplike $\mathbb{E}_\infty$–spaces. Since the latter ∞–category is equivalent to the ∞–category of connective spectra [20, Remark 5.2.6.26], the canonical symmetric monoidal structure on $\mathrm{Grp}_{\mathbb{E}_\infty}(\mathcal{S})$ agrees with the smash product of connective spectra.

(iii) Let Cat denote the cartesian closed presentable ∞–category of small ordinary categories (this is actually a 2–category, in the sense of [19, Section 2.3.4]). Thus, the ∞–category SymMonCat $\simeq \mathrm{Mon}_{\mathbb{E}_\infty}(\mathrm{Cat})$ of small symmetric monoidal categories admits a canonical closed symmetric monoidal structure such that the free functor Cat $\to$ SymMonCat can be promoted to a symmetric monoidal functor in a unique way. This structure on SymMonCat has been explicitly constructed and discussed in the literature; see [14] and the more explicit [26]. In fact, this tensor product is slightly subtle since, at least to the knowledge of the authors, it can not be realized as a symmetric monoidal structure on the 1–category of small categories (as opposed to the 2–category Cat).

(iv) The ∞–category Cat_∞ of small ∞–categories is a cartesian closed presentable ∞–category. Thus, as an ∞–categorical variant of the previous example, we obtain a canonical closed symmetric monoidal structure on the ∞–category $\mathrm{SymMonCat}_\infty$ of small symmetric monoidal ∞–categories.

We have already seen that, for presentable ∞–categories $\mathcal{C}$, the passage to commutative monoids and commutative groups has a universal property (Corollary 4-9). In the case of closed symmetric monoidal presentable ∞–categories we now obtain a refined universal property for the symmetric monoidal structures of Theorem 5-1. For convenience, we also collect the analogous results for the passage to pointed objects and spectrum objects.

Proposition 5-4 *Suppose $\mathcal{C}$ and $\mathcal{D}$ are closed symmetric monoidal presentable ∞–categories.*

(i) *If $\mathcal{D}$ is pointed then the symmetric monoidal functor $\mathcal{C} \to \mathcal{C}_*$ induces an equivalence of ∞–categories*

$$\mathrm{Fun}^{\mathrm{L},\otimes}(\mathcal{C}_*, \mathcal{D}) \to \mathrm{Fun}^{\mathrm{L},\otimes}(\mathcal{C}, \mathcal{D}).$$

(ii) *If $\mathcal{D}$ is preadditive then the symmetric monoidal functor $\mathcal{C} \to \mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{C})$ induces an equivalence of ∞–categories*

$$\mathrm{Fun}^{\mathrm{L},\otimes}(\mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{C}), \mathcal{D}) \to \mathrm{Fun}^{\mathrm{L},\otimes}(\mathcal{C}, \mathcal{D}).$$

(iii) *If $\mathcal{D}$ is additive then the symmetric monoidal functor $\mathcal{C} \to \mathrm{Grp}_{\mathbb{E}_\infty}(\mathcal{C})$ induces an equivalence of ∞–categories*

$$\mathrm{Fun}^{\mathrm{L},\otimes}(\mathrm{Grp}_{\mathbb{E}_\infty}(\mathcal{C}), \mathcal{D}) \to \mathrm{Fun}^{\mathrm{L},\otimes}(\mathcal{C}, \mathcal{D}).$$

(iv) *If $\mathcal{D}$ is stable then the symmetric monoidal functor $\mathcal{C} \to \mathrm{Sp}(\mathcal{C})$ induces an equivalence of ∞–categories*

$$\mathrm{Fun}^{\mathrm{L},\otimes}(\mathrm{Sp}(\mathcal{C}), \mathcal{D}) \to \mathrm{Fun}^{\mathrm{L},\otimes}(\mathcal{C}, \mathcal{D}).$$

Proof This follows immediately from the second statement of Proposition 3-9. □

Here is the monoidal refinement of the stabilization process which is now an immediate consequence of the third statement of Proposition 3-9.

Corollary 5-5 (i) *Let $\mathcal{C}$ and $\mathcal{D}$ be closed symmetric monoidal presentable ∞–categories and let us consider a symmetric monoidal left adjoint $F\colon \mathcal{C} \to \mathcal{D}$. In the following commutative diagram, each of the functors induced by F admits a symmetric monoidal structure:*

$$\begin{tikzcd}
\mathcal{C} \arrow[r] \arrow[d] & \mathcal{C}_* \arrow[r] \arrow[d] & \mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{C}) \arrow[r] \arrow[d] & \mathrm{Grp}_{\mathbb{E}_\infty}(\mathcal{C}) \arrow[r] \arrow[d] & \mathrm{Sp}(\mathcal{C}) \arrow[d] \\
\mathcal{D} \arrow[r] & \mathcal{D}_* \arrow[r] & \mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{D}) \arrow[r] & \mathrm{Grp}_{\mathbb{E}_\infty}(\mathcal{D}) \arrow[r] & \mathrm{Sp}(\mathcal{D}_*)
\end{tikzcd}$$

Moreover, these symmetric monoidal structures are uniquely characterized by the fact that the functors commute with the free functors from $\mathcal{C}$.

(ii) *The stabilization of presentable ∞–categories $\mathrm{Pr}^{\mathrm{L}} \to \mathrm{Pr}^{\mathrm{L}}_{\mathrm{St}}$ admits a symmetric monoidal refinement $\mathrm{Pr}^{\mathrm{L},\otimes} \to \mathrm{Pr}^{\mathrm{L},\otimes}_{\mathrm{St}}$ which factors as a composition of adjunctions*

$$\mathrm{Pr}^{\mathrm{L},\otimes} \rightleftarrows \mathrm{Pr}^{\mathrm{L},\otimes}_{\mathrm{Pt}} \rightleftarrows \mathrm{Pr}^{\mathrm{L},\otimes}_{\mathrm{Pre}} \rightleftarrows \mathrm{Pr}^{\mathrm{L},\otimes}_{\mathrm{Add}} \rightleftarrows \mathrm{Pr}^{\mathrm{L},\otimes}_{\mathrm{St}}.$$

Remark 5-6 (i) One can use the theory of Γ–objects in $\mathcal{C}$ to obtain a more concrete description of the tensor product on $\mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{C})$ and $\mathrm{Grp}_{\mathbb{E}_\infty}(\mathcal{C})$ as the convolution product; see [20, Corollary 4.8.1.12] for the case in which $\mathcal{C}$ is the ∞–category of spaces.

(ii) The uniqueness of the symmetric monoidal structures can be used to compare our results to existing ones. Every simplicial combinatorial, monoidal model category leads to a presentable, closed symmetric monoidal ∞–category. Thus for the monoidal model category of Γ–spaces as discussed in [27] it follows immediately that the symmetric monoidal structure on the underlying ∞–category has to agree with our structure. The same applies to the model structure on Γ–objects in any nice model category, for example in presheaves as discussed in [4].

6 More functoriality

In Section 4 we saw that for presentable ∞–categories the passages to commutative monoids and groups are smashing localizations and hence, in particular, define functors

$$\mathrm{Mon}_{\mathbb{E}_\infty}(-), \mathrm{Grp}_{\mathbb{E}_\infty}(-)\colon \mathrm{Pr}^{\mathrm{L}} \to \mathrm{Pr}^{\mathrm{L}}.$$

But this passage allows for more functoriality. In fact, a product-preserving functor $F\colon \mathcal{C} \to \mathcal{D}$ induces functors

$$\underline{F}\colon \mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{C}) \to \mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{D}) \quad \text{and} \quad \underline{F}\colon \mathrm{Grp}_{\mathbb{E}_\infty}(\mathcal{C}) \to \mathrm{Grp}_{\mathbb{E}_\infty}(\mathcal{D})$$

simply by post-composing the respective (very) special Γ–objects with F. The main goal of this section is to establish Corollary 6-6, which states that under certain mild assumptions these extensions themselves are lax symmetric monoidal with respect to the canonical symmetric monoidal structures established in Theorem 5-1. This corollary will be needed in our applications to algebraic K–theory in Section 8. We begin by comparing these two potentially different functorialities of the assignments $\mathcal{C} \mapsto \mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{C})$ and $\mathcal{C} \mapsto \mathrm{Grp}_{\mathbb{E}_\infty}(\mathcal{C})$.

Lemma 6-1 *Let $L\colon \mathcal{C} \to \mathcal{D}$ be a functor of presentable ∞–categories with right adjoint $R\colon \mathcal{D} \to \mathcal{C}$.*

(i) *If $L\colon \mathcal{C} \to \mathcal{D}$ is product-preserving and if products in $\mathcal{C}$ and $\mathcal{D}$ commute with countable colimits, then the functors*

$$\mathrm{Mon}_{\mathbb{E}_\infty}(L)\colon \mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{C}) \to \mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{D}) \quad \textit{and} \quad \underline{L}\colon \mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{C}) \to \mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{D})$$

described above are equivalent.

(ii) *The canonical extension $\underline{R}\colon \mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{D}) \to \mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{C})$ is right adjoint to the functor $\mathrm{Mon}_{\mathbb{E}_\infty}(L)$.*

The corresponding two statements for $\mathbb{E}_\infty$–groups hold as well.

Proof For the first claim we must show that if L preserves products then the two functors agree. This follows if we can show that $\underline{L}$ is a left adjoint and the diagram

$$\begin{tikzcd} \mathcal{C}_{\mathrm{Fr}} \arrow[r, "L"] \arrow[d] & \mathcal{D} \arrow[d, "\mathrm{Fr}"] \\ \mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{C})\_\underline{L} \arrow[r] & \mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{D}) \end{tikzcd} \tag{6-2}$$

commutes in $\widehat{\mathrm{Cat}}_\infty$. To see that $\underline{L}$ is left adjoint we observe that it commutes with sifted colimits, as they are detected by the forgetful functors $\mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{C}) \to \mathcal{C}$ and $\mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{D}) \to \mathcal{D}$, and also that it commutes with coproducts, as coproducts in $\mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{C})$ and $\mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{D})$ are given by the tensor product which is preserved by L. To conclude this part of the proof it suffices to show that there is an equivalence $\mathrm{Fr} \circ L \simeq$

$\underline{L}\circ\mathrm{Fr}$. For this, we consider the mate of the equivalence $L\circ U\simeq U\circ\underline{L}\colon \mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{C})\to\mathcal{D}$, ie, we form the following pasting with the respective adjunction morphisms:

$$\begin{tikzcd}
\mathcal{C} \arrow[r, "\mathrm{Fr}"] \arrow[dr, bend right, "="'] & \mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{C}) \arrow[r, "\underline{L}"] \arrow[d, "U"'] \arrow[dr, phantom, "\simeq\Nearrow"] & \mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{D}) \arrow[d, "U"] \arrow[dr, bend left, "="] \arrow[dr, phantom, "\Nearrow"'] & \\
\arrow[ur, phantom, "\Nearrow"] & \mathcal{C} \arrow[r, "L"'] & \mathcal{D} \arrow[r, "\mathrm{Fr}"'] & \mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{D})
\end{tikzcd}$$

In order to show that the resulting transformation

$$\mathrm{Fr}\circ L\to\mathrm{Fr}\circ L\circ U\circ\mathrm{Fr}\simeq\mathrm{Fr}\circ U\circ\underline{L}\circ\mathrm{Fr}\to\underline{L}\circ\mathrm{Fr}$$

is an equivalence, it is enough to check that this is the case after applying the forgetful functor $U\colon \mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{D})\to\mathcal{D}$. But this follows from the explicit description of the free functors as

$$\mathrm{Fr}(C)\simeq\coprod_n C^n/\Sigma_n$$

(see [20, Example 3.1.3.14]) and by unraveling the definitions of $\underline{L}$ and the adjunction morphisms.

To prove the second statement we first remark that $\underline{R}$ has a left adjoint since it preserves all limits and filtered colimits which are formed in the underlying ∞–category. Moreover, any such left adjoint has to make diagram (6-2) commute since this is the case for the corresponding diagram of right adjoints. By the above, this left adjoint has to coincide with $\mathrm{Mon}_{\mathbb{E}_\infty}(L)$. The proof for the case of groups is completely parallel. □

This lemma can be applied to adjunctions between cartesian closed presentable ∞–categories.

Lemma 6-3 *Let $\mathcal{C}$ and $\mathcal{D}$ be closed symmetric monoidal presentable ∞–categories, let $L\colon\mathcal{C}\to\mathcal{D}$ be a symmetric monoidal left adjoint functor and let $R\colon\mathcal{D}\to\mathcal{C}$ be right adjoint to L.*

(i) *The functors $\underline{R}\colon \mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{D})\to\mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{C})$ and $\underline{R}\colon \mathrm{Grp}_{\mathbb{E}_\infty}(\mathcal{D})\to\mathrm{Grp}_{\mathbb{E}_\infty}(\mathcal{C})$ have canonical lax symmetric monoidal structures.*

(ii) *If $\mathcal{C}$ and $\mathcal{D}$ are cartesian closed, then the canonical extensions $\underline{L}\colon \mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{C})\to\mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{D})$ and $\underline{L}\colon \mathrm{Grp}_{\mathbb{E}_\infty}(\mathcal{C})\to\mathrm{Grp}_{\mathbb{E}_\infty}(\mathcal{D})$ both admit structures of symmetric monoidal functors which are determined up to a contractible space of choices by* *the fact that the following diagrams commute.*

$$\begin{tikzcd} \mathcal{C} \arrow[r] \arrow[d, "L"'] & \mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{C}) \arrow[d, "\underline{L}"] \\ \mathcal{D} \arrow[r] & \mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{D}) \end{tikzcd} \qquad \begin{tikzcd} \mathcal{C} \arrow[r] \arrow[d, "L"'] & \mathrm{Grp}_{\mathbb{E}_\infty}(\mathcal{C}) \arrow[d, "\underline{L}"] \\ \mathcal{D} \arrow[r] & \mathrm{Grp}_{\mathbb{E}_\infty}(\mathcal{D}) \end{tikzcd}$$

Proof Corollary 5-5 tells us that $\mathrm{Mon}_{\mathbb{E}_\infty}(L)$ is canonically symmetric monoidal, and the right adjoint of a symmetric monoidal functor always inherits a canonical lax symmetric monoidal structure [20, Corollary 7.3.2.7]. Together with Lemma 6-1 this establishes the first part. The second part is an immediate consequence of Corollary 5-5 and Lemma 6-1, and again the case of groups is entirely analogous. $\square$

Lemma 6-4 *Suppose $F\colon \mathcal{C} \to \mathcal{D}$ is an accessible functor between presentable ∞–categories.*

(i) *We can factor $F \simeq L \circ R$ where R is a right adjoint and L is a left adjoint functor.*

(ii) *If $\mathcal{C}$ and $\mathcal{D}$ are closed symmetric monoidal, then the factorization can be chosen such that L and the left adjoint to R are symmetric monoidal (this means of course that the intermediate ∞–category is symmetric monoidal as well). In particular, R itself is lax symmetric monoidal.*

(iii) *If F preserves products and $\mathcal{D}$ is cartesian closed, then L can be chosen to preserve products.*

Proof Choose κ sufficiently large such that both $\mathcal{C}$ and $\mathcal{D}$ are κ–compactly generated and F preserves κ–filtered colimits. Then the restricted Yoneda embedding $R\colon \mathcal{C} \to \mathcal{P}(\mathcal{C}^\kappa)$ preserves limits and κ–filtered colimits, and therefore admits a left adjoint. Similarly, the functor $L\colon \mathcal{P}(\mathcal{C}^\kappa) \to \mathcal{D}$ induced (under colimits) by the composite $\mathcal{C}^\kappa \to \mathcal{C} \to \mathcal{D}$ preserves all colimits, and therefore admits a right adjoint. Since F is equivalent to the composite $L \circ R$, this completes the proof of the first claim.

Now, if in addition $\mathcal{C}$ and $\mathcal{D}$ are closed symmetric monoidal, then it follows from the universal property of the convolution product [20, Proposition 4.8.1.10] that L is symmetric monoidal and also that the left adjoint $\mathcal{P}(\mathcal{C}^\kappa) \to \mathcal{C}$ of R is symmetric monoidal, completing the proof of the second claim (the fact that R is lax symmetric monoidal again follows from [20, Corollary 7.3.2.7]).

Finally, if F preserves products, then L preserves products of representables $\mathcal{C}^\kappa$, and if $\mathcal{D}$ is cartesian closed then products commute with colimits in both variables. Hence L preserves products. $\square$

Proposition 6-5 *Suppose $\mathcal{C}$ and $\mathcal{D}$ are closed symmetric monoidal presentable ∞–categories and let $F\colon \mathcal{C} \to \mathcal{D}$ be product-preserving, symmetric monoidal, and accessible. If $\mathcal{D}$ is also cartesian closed then the functors $\underline{F}\colon \mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{C}) \to \mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{D})$ and $\underline{F}\colon \mathrm{Grp}_{\mathbb{E}_\infty}(\mathcal{C}) \to \mathrm{Grp}_{\mathbb{E}_\infty}(\mathcal{D})$ admit lax symmetric monoidal structures.*

Proof Factor F according to Lemma 6-4 and apply Lemma 6-3. □

Corollary 6-6 *Let $\mathcal{C}$ and $\mathcal{D}$ be cartesian closed presentable ∞–categories and let $F\colon \mathcal{C} \to \mathcal{D}$ be product-preserving and accessible. Then the canonical extensions $\underline{F}\colon \mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{C}) \to \mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{D})$ and $\underline{F}\colon \mathrm{Grp}_{\mathbb{E}_\infty}(\mathcal{C}) \to \mathrm{Grp}_{\mathbb{E}_\infty}(\mathcal{D})$ are lax symmetric monoidal.*

7 ∞–categories of semirings and rings

In this section we will use the results of Section 5 to define and study semiring (aka *rig*) and ring objects in suitable ∞–categories. We know by Theorem 5-1 that given a closed symmetric monoidal presentable ∞–category $\mathcal{C}$, there are canonical closed symmetric monoidal structures on $\mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{C})$ and $\mathrm{Grp}_{\mathbb{E}_\infty}(\mathcal{C})$ which will respectively be denoted by

$$\mathrm{Mon}^{\otimes}_{\mathbb{E}_\infty\mathbb{E}_\infty}(\mathcal{C}) \quad \text{and} \quad \mathrm{Grp}^{\otimes}_{\mathbb{E}_\infty\mathbb{E}_\infty}(\mathcal{C}).$$

Definition 7-1 Let $\mathcal{C}$ be a closed symmetric monoidal presentable ∞–category and let $\mathcal{O}$ be an ∞–operad. The ∞–category $\mathrm{Rig}_{\mathcal{O}}(\mathcal{C})$ of *$\mathcal{O}$–semirings* in $\mathcal{C}$ and the ∞–category $\mathrm{Ring}_{\mathcal{O}}(\mathcal{C})$ of *$\mathcal{O}$–rings* in $\mathcal{C}$ are respectively defined as the ∞–categories of $\mathcal{O}$–algebras

$$\mathrm{Rig}_{\mathcal{O}}(\mathcal{C}) := \mathrm{Alg}_{\mathcal{O}}(\mathrm{Mon}^{\otimes}_{\mathbb{E}_\infty\mathbb{E}_\infty}(\mathcal{C})) \quad \text{and} \quad \mathrm{Ring}_{\mathcal{O}}(\mathcal{C}) := \mathrm{Alg}_{\mathcal{O}}(\mathrm{Grp}^{\otimes}_{\mathbb{E}_\infty\mathbb{E}_\infty}(\mathcal{C})).$$

In the case of ordinary categories and the associative or commutative operad, the alternative terminology *rig objects* is also used for what we call semiring objects, hence the notation. We will be mainly interested in the case of $\mathcal{O} = \mathbb{E}_n$ for $n = 1, 2, \ldots, \infty$. In the case $n = 1$, $\mathrm{Ring}_{\mathbb{E}_1}(\mathcal{C})$ is the ∞–category of *associative rings* in $\mathcal{C}$ and, in the case $n = \infty$, $\mathrm{Ring}_{\mathbb{E}_\infty}(\mathcal{C})$ is the ∞–category of *commutative rings* in $\mathcal{C}$. Similarly, there are ∞–categories of associative or commutative semirings in $\mathcal{C}$.

Let us take up again the examples of Section 5.

Example 7-2 (i) In the special case of the cartesian closed presentable ∞–category Set of sets, our notion of associative or commutative (semi)ring object coincides with the corresponding classical notion.

(ii) Since the ∞–category $\mathcal{S}$ of spaces is cartesian closed and presentable, we obtain, for each ∞–operad $\mathcal{O}$, the ∞–category $\mathrm{Rig}_{\mathcal{O}}(\mathcal{S})$ of *$\mathcal{O}$-rig spaces* and the ∞–category $\mathrm{Ring}_{\mathcal{O}}(\mathcal{S})$ of *$\mathcal{O}$-ring spaces*. For the special case of the operads $\mathcal{O} = \mathbb{E}_n$ for $n = 1, \dots, \infty$, the point-set analogue of these spaces were intensively studied by May and others using carefully chosen pairs of operads; see the recent articles [25; 23; 24] and the many references therein.

(iii) In the case of the cartesian closed presentable ∞–category Cat of ordinary small categories, we obtain the ∞–category $\mathrm{Rig}_{\mathcal{O}}\mathrm{Cat}$ of *$\mathcal{O}$-rig categories* and the ∞–category $\mathrm{Ring}_{\mathcal{O}}\mathrm{Cat}$ of *$\mathcal{O}$-ring categories*. Coherences for *lax* semiring categories have been studied by Laplaza [16; 17]; note that, in our case, all coherence morphisms must be invertible. It should be possible to obtain a precise comparison of our notion with these more classical ones, but we bypass this via a recognition principle (Corollary 8-9) for semiring ∞–categories which allows us to work directly with the examples of interest to us, without having to check coherences for distributors.

(iv) An ∞–categorical version of the previous example is obtained by considering the cartesian closed presentable ∞–category Cat_∞. Associated to it there is the ∞–category $\mathrm{Rig}_{\mathcal{O}}\mathrm{Cat}_\infty$ of *$\mathcal{O}$-semiring ∞–categories* and the ∞–category $\mathrm{Ring}_{\mathcal{O}}\mathrm{Cat}_\infty$ of *$\mathcal{O}$-ring ∞–categories*.

Remark 7-3 For a general closed symmetric monoidal presentable ∞–category $\mathcal{C}$ there are two potentially different symmetric monoidal structures playing a role in the notion of an $\mathcal{O}$–(semi)ring object. Thus it may be useful to provide an informal description of the structure given by an $\mathbb{E}_\infty$–semiring object in $\mathcal{C}$. It consists of an object $R \in \mathcal{C}$ together with an addition map $+\colon R \times R \to R$ and a multiplication map $\times\colon R \otimes R \to R$ such that both maps are coherently associative and commutative. Moreover, the multiplication has to distribute over the addition in a homotopy coherent fashion. In the case of an ordinary category with the Cartesian monoidal structure, our notion reduces to the usual one.

As in the case of commutative monoids and commutative groups, Theorem 5-1 also guarantees that the ∞–category $\mathrm{Sp}(\mathcal{C})$ of spectrum objects associated to a closed symmetric monoidal presentable ∞–category $\mathcal{C}$ has a canonical closed symmetric monoidal structure $\mathrm{Sp}^{\otimes}(\mathcal{C})$. This allows us to make the following definition.

Definition 7-4 Let $\mathcal{C}$ be a closed symmetric monoidal presentable ∞–category and let $\mathcal{O}$ be an ∞–operad. The ∞–category $\mathrm{RingSp}_{\mathcal{O}}(\mathcal{C})$ of *$\mathcal{O}$–ring spectrum objects in $\mathcal{C}$* is defined as

$$\mathrm{RingSp}_{\mathcal{O}}(\mathcal{C}) := \mathrm{Alg}_{\mathcal{O}}(\mathrm{Sp}^{\otimes}(\mathcal{C})).$$

Theorem 7-5 *Let $\mathcal{C}$ be a closed symmetric monoidal presentable ∞–category and let $\mathcal{O}$ be an ∞–operad. Then the group completion functor* $\mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{C}) \to \mathrm{Grp}_{\mathbb{E}_\infty}(\mathcal{C})$ *and the associated spectrum functor* $\mathrm{Grp}_{\mathbb{E}_\infty}(\mathcal{C}) \to \mathrm{Sp}(\mathcal{C})$ *refine to functors*

$$\mathrm{Rig}_{\mathcal{O}}(\mathcal{C}) \to \mathrm{Ring}_{\mathcal{O}}(\mathcal{C}) \quad \textit{and} \quad \mathrm{Ring}_{\mathcal{O}}(\mathcal{C}) \to \mathrm{RingSp}_{\mathcal{O}}(\mathcal{C}),$$

called the ring completion *and the* associated ring spectrum functor, *respectively.*

Proof This is clear since the group completion $\mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{C}) \to \mathrm{Grp}_{\mathbb{E}_\infty}(\mathcal{C})$ and also the associated spectrum functor $\mathrm{Grp}_{\mathbb{E}_\infty}(\mathcal{C}) \to \mathrm{Sp}(\mathcal{C})$ are symmetric monoidal as shown in Theorem 5-1. □

Example 7-6 (i) In the special case of the ∞–category Set of sets this reduces to the usual ring completion of associative or commutative semirings.

(ii) From an ∞–operad $\mathcal{O}$, we get an associated ring completion functor $\mathrm{Rig}_{\mathcal{O}}(\mathcal{S}) \to \mathrm{Ring}_{\mathcal{O}}(\mathcal{S})$ from $\mathcal{O}$–rig spaces to $\mathcal{O}$–ring spaces and an associated ring spectrum functor $\mathrm{Ring}_{\mathcal{O}}(\mathcal{S}) \to \mathrm{RingSp}_{\mathcal{O}}(\mathcal{S})$ from $\mathcal{O}$–ring spaces to $\mathcal{O}$–ring spectra. The latter ∞–category will also be written $\mathrm{RingSp}_{\mathcal{O}}$.

(iii) Let us again consider the cartesian closed presentable ∞–category Cat of ordinary small categories. Then for each ∞–operad $\mathcal{O}$, we obtain a ring completion functor $\mathrm{Rig}_{\mathcal{O}}\mathrm{Cat} \to \mathrm{Ring}_{\mathcal{O}}\mathrm{Cat}$ from $\mathcal{O}$–rig categories to $\mathcal{O}$–ring categories.

(iv) Again, we immediately obtain an ∞–categorical refinement of the previous example. For each ∞–operad $\mathcal{O}$, we obtain a ring completion functor $\mathrm{Rig}_{\mathcal{O}}\mathrm{Cat}_\infty \to \mathrm{Ring}_{\mathcal{O}}\mathrm{Cat}_\infty$ from $\mathcal{O}$–rig ∞–categories to $\mathcal{O}$–ring ∞–categories. Using explicit models, a similar construction was obtained by Baas, Dundas, Richter and Rognes [2].

Theorem 7-5 shows that semirings can be used to produce highly structured ring spectra. Unfortunately, the definition of a semiring object is a bit indirect, so in practice it is often difficult to write down explicit examples of such objects. Theorem 8-8 provides a natural class of semirings in the case of the cartesian closed ∞–category $\mathcal{C} = \mathrm{Cat}_\infty$. Moreover, this is the class that is of most interest in applications to algebraic K–theory, as we discuss in Section 8.

We conclude this section with a base-change result (similar to Corollary 4-7) which sheds some light on the definition of semiring and ring object. This result will also be needed in Appendix B where we show $\mathbb{E}_n$–(semi)rings to be *algebraic*.

Proposition 7-7 *Let $\mathcal{C}$ be a cartesian closed presentable ∞–category and $\mathcal{O}$ an ∞–operad. Then we have equivalences*

$$\mathrm{Rig}_{\mathcal{O}}(\mathcal{C}) \simeq \mathcal{C} \otimes \mathrm{Rig}_{\mathcal{O}}(\mathcal{S}) \quad \textit{and} \quad \mathrm{Ring}_{\mathcal{O}}(\mathcal{C}) \simeq \mathcal{C} \otimes \mathrm{Ring}_{\mathcal{O}}(\mathcal{S}).$$

Proof We show more generally that, for $\mathcal{D}$ any closed symmetric monoidal presentable ∞–category, there exists a canonical equivalence

$$\mathrm{Alg}_{\mathcal{O}}(\mathcal{D}) \otimes \mathcal{C} \to \mathrm{Alg}_{\mathcal{O}}(\mathcal{D} \otimes \mathcal{C}). \tag{7-8}$$

Then, taking $\mathcal{D}$ to be $\mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{S})$, using Theorem 4-6, we obtain the desired chain of equivalences

$$\begin{aligned}\mathrm{Rig}_{\mathcal{O}}(\mathcal{C}) \simeq \mathrm{Alg}_{\mathcal{O}}(\mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{C})) &\simeq \mathrm{Alg}_{\mathcal{O}}(\mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{S}) \otimes \mathcal{C}) \\ &\simeq \mathrm{Alg}_{\mathcal{O}}(\mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{S})) \otimes \mathcal{C} \simeq \mathrm{Rig}_{\mathcal{O}}(\mathcal{S}) \otimes \mathcal{C}.\end{aligned}$$

In the case of rings we get an analogous chain of equivalences.

To show (7-8), first consider the case in which $\mathcal{C} = \mathcal{P}(\mathcal{C}_0)$ is the ∞–category of presheaves of spaces on a (small) ∞–category $\mathcal{C}_0$. In this case, we have that $\mathcal{D} \otimes \mathcal{C} \simeq \mathrm{Fun}(\mathcal{C}_0^{\mathrm{op}}, \mathcal{D})$, so that

$$\mathrm{Alg}_{\mathcal{O}}(\mathcal{D}) \otimes \mathcal{C} \simeq \mathrm{Fun}(\mathcal{C}_0^{\mathrm{op}}, \mathrm{Alg}_{\mathcal{O}}(\mathcal{D})) \simeq \mathrm{Alg}_{\mathcal{O}}(\mathrm{Fun}(\mathcal{C}_0^{\mathrm{op}}, \mathcal{D})) \simeq \mathrm{Alg}_{\mathcal{O}}(\mathcal{D} \otimes \mathcal{C}).$$

A general cartesian closed presentable ∞–category $\mathcal{C}$ is a full symmetric monoidal subcategory of some $\mathcal{P}(\mathcal{C}_0)$, say for $\mathcal{C}_0$ the full subcategory of κ–compact objects in $\mathcal{C}$ for a sufficiently large regular cardinal κ. Since $\mathcal{D} \otimes \mathcal{C} \simeq \mathrm{Fun}^{\mathrm{R}}(\mathcal{C}^{\mathrm{op}}, \mathcal{D})$, we see that $\mathcal{D} \otimes \mathcal{C}$ is a full symmetric monoidal subcategory of $\mathcal{D} \otimes \mathcal{P}(\mathcal{C}_0)$, and similarly with $\mathcal{D}$ replaced by $\mathrm{Alg}_{\mathcal{O}}(\mathcal{D})$. Thus it suffices to show that $\mathrm{Alg}_{\mathcal{O}}(\mathcal{D}) \otimes \mathcal{C}$ and $\mathrm{Alg}_{\mathcal{O}}(\mathcal{D} \otimes \mathcal{C})$ define equivalent full subcategories of $\mathrm{Alg}_{\mathcal{O}}(\mathcal{D}) \otimes \mathcal{P}(\mathcal{C}_0) \simeq \mathrm{Alg}_{\mathcal{O}}(\mathcal{D} \otimes \mathcal{P}(\mathcal{C}_0))$.

If $\mathcal{O}$ is monochromatic (ie if there exists an essentially surjective functor $\Delta^0 \to \mathcal{O}^{\otimes}_{\langle 1 \rangle}$), then an object of $\mathrm{Alg}_{\mathcal{O}}(\mathcal{D} \otimes \mathcal{P}(\mathcal{C}_0))$ lies in the full subcategory $\mathrm{Alg}_{\mathcal{O}}(\mathcal{D} \otimes \mathcal{C})$ if and only if the projection to $\mathcal{D} \otimes \mathcal{P}(\mathcal{C}_0)$ factors through $\mathcal{D} \otimes \mathcal{C}$. For arbitrary $\mathcal{O}$, an object of $\mathrm{Alg}_{\mathcal{O}}(\mathcal{D} \otimes \mathcal{P}(\mathcal{C}_0))$ lies in the full subcategory $\mathrm{Alg}_{\mathcal{O}}(\mathcal{D} \otimes \mathcal{C})$ precisely when the restriction along any full monochromatic suboperad $\mathcal{O}' \to \mathcal{O}$ satisfies this same condition. As the analogous results for $\mathrm{Alg}_{\mathcal{O}}(\mathcal{D}) \otimes \mathcal{C}$ hold by the same argument, we see that $\mathrm{Alg}_{\mathcal{O}}(\mathcal{D}) \otimes \mathcal{C}$ and $\mathrm{Alg}_{\mathcal{O}}(\mathcal{D} \otimes \mathcal{C})$ define equivalent full subcategories of $\mathrm{Alg}_{\mathcal{O}}(\mathcal{D}) \otimes \mathcal{P}(\mathcal{C}_0) \simeq \mathrm{Alg}_{\mathcal{O}}(\mathcal{D} \otimes \mathcal{P}(\mathcal{C}_0))$. □

8 Multiplicative infinite loop space theory

In this section we apply the results of the previous section to some specific ∞–categories; namely, we consider the ∞–categories $\mathcal{S}$ of spaces, the ∞–category Cat of ordinary categories (really a 2–category, but we regard it as an ∞–category), and the ∞–category Cat_∞ of ∞–categories. Let us emphasize that, as a special case of Theorem 7-5, the group completion and the associated spectrum functor

$$\mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{S}) \to \mathrm{Grp}_{\mathbb{E}_\infty}(\mathcal{S}) \to \mathrm{Sp}$$

refine to functors

$$\mathrm{Rig}_{\mathcal{O}}(\mathcal{S}) \to \mathrm{Ring}_{\mathcal{O}}(\mathcal{S}) \to \mathrm{RingSp}_{\mathcal{O}}.$$

This gives us not only a way of obtaining (highly structured) ring spectra, but it also allows us to identify certain spectra as ring spectra.

Recall that the group completion functor $\mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{S}) \to \mathrm{Grp}_{\mathbb{E}_\infty}(\mathcal{S}) \to \mathrm{Sp}$ plays an important role in algebraic K–theory. The input data for algebraic K–theory is often a symmetric monoidal category $\mathcal{M}$; as a primary example, we have the category $\mathcal{M} = \mathrm{Proj}_R$ of finitely generated projective modules over a ring R, which is symmetric monoidal under the direct sum $\oplus$ (which is the coproduct). In any case, given such a category $\mathcal{M}$, we form the subcategory of isomorphisms $\mathcal{M}^\sim$ and pass to the geometric realization $|\mathcal{M}^\sim|$. That way we obtain an $\mathbb{E}_\infty$–space $|\mathcal{M}^\sim|$, ie, an object of $\mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{S})$. The *algebraic* K*–theory spectrum* $\mathrm{K}(\mathcal{M})$ is then defined to be the spectrum associated to the group completion of $|\mathcal{M}^\sim|$; see eg [28]. In other words, (direct sum) algebraic K–theory is defined as the composition

$$\text{(8-1)} \qquad \mathrm{K}\colon \mathrm{SymMonCat} \xrightarrow{(-)^\sim} \mathrm{SymMonCat} \xrightarrow{|-|} \mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{S}) \to \mathrm{Grp}_{\mathbb{E}_\infty}(\mathcal{S}) \to \mathrm{Sp}.$$

It is a result of May [22], with refinements by Elmendorf and Mandell [12] and Bass, Dundas, Richter and Rognes [2], that this functor respects multiplicative structures, in the appropriate sense. Our methods give an even more refined result.

Proposition 8-2 *The algebraic* K*–theory functor* $\mathrm{K}\colon \mathrm{SymMonCat} \to \mathrm{Sp}$ *is lax symmetric monoidal. In particular, it induces a functor* $\mathrm{Rig}_{\mathcal{O}}\mathrm{Cat} \to \mathrm{RingSp}_{\mathcal{O}}$ *for any* ∞*–operad* $\mathcal{O}$.

Proof The last two functors in the composition (8-1) are symmetric monoidal by Theorem 5-1. The remaining two functors $(-)^\sim\colon \mathrm{SymMonCat} \to \mathrm{SymMonCat}$ and $|-|\colon \mathrm{SymMonCat} \to \mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{S})$ are the canonical extensions of the product preserving functors $(-)^\sim\colon \mathrm{Cat} \to \mathrm{Cat}$ and $|-|\colon \mathrm{Cat} \to \mathcal{S}$ respectively. Since these latter functors are accessible, Corollary 6-6 implies that their canonical extensions are lax symmetric monoidal, concluding the proof. $\square$

We now have the tools necessary to establish corresponding results in the ∞–categorical case. Note that the composition of the first two functors in (8-1) is the same as the composition of the nerve $\mathrm{SymMonCat} \to \mathrm{SymMonCat}_\infty$ followed by the functor $(-)^\sim\colon \mathrm{SymMonCat}_\infty \to \mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{S})$, which sends a symmetric monoidal ∞–category to its maximal subgroupoid, and of course is again symmetric monoidal. This allows us to recover the algebraic K–theory of a symmetric monoidal category $\mathcal{M}$ by an application of the following ∞–categorical version of algebraic K–theory to the nerve of $\mathcal{M}$.

Definition 8-3 Let $\mathcal{M}$ be a symmetric monoidal ∞–category. The *algebraic* K–*theory spectrum* $\mathrm{K}(\mathcal{M})$ is the spectrum associated to the group completion of $\mathcal{M}^\sim$. Thus, the algebraic K–theory functor is defined as the composition

$$\mathrm{K}\colon \mathrm{SymMonCat}_\infty \xrightarrow{(-)^\sim} \mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{S}) \longrightarrow \mathrm{Grp}_{\mathbb{E}_\infty}(\mathcal{S}) \longrightarrow \mathrm{Sp}. \tag{8-4}$$

Remark 8-5 Strictly speaking, this is the *direct sum* K*–theory*, since it does not take into account a potential exactness (or Waldhausen) structure on the symmetric monoidal ∞–categories in question. Nevertheless, in many cases of interest, eg that of a connective ring spectrum R, the algebraic K–theory of R, defined in terms of Waldhausen's $S_\bullet$ construction applied to the stable ∞–category of R–modules (which agrees with the K–theory of any suitable model category of R–modules; see [6] for details), is computed as the direct sum K–theory of the symmetric monoidal ∞–category Proj_R of finitely generated projective R–modules [20, Definition 7.2.2.4].

For more sophisticated versions of K–theory, the situation is slightly more complicated but entirely analogous. In [7] it is shown that the algebraic K–theory $\mathrm{K}\colon \mathrm{Cat}_\infty^{\mathrm{perf}} \to \mathrm{Sp}$ of small idempotent-complete stable ∞–categories is a lax symmetric monoidal functor, as is the nonconnective version; the methods employed to do so are similar to the ones used in the present paper, in that K is shown to be the tensor unit in a symmetric monoidal ∞–category of all additive (respectively, localizing) functors $\mathrm{Cat}_\infty^{\mathrm{perf}} \to \mathrm{Sp}$, so that the commutative algebra structure ultimately relies on the existence of an idempotent object in an appropriate symmetric monoidal ∞–category. The case of general Waldhausen ∞–categories is treated in [3], where it is shown that the algebraic K–theory $\mathrm{K}\colon \mathrm{Wald}_\infty \to \mathrm{Sp}$ of Waldhausen ∞–categories is again a lax symmetric monoidal functor.

As already mentioned, the ∞–categorical algebraic K–theory

$$\mathrm{K}\colon \mathrm{SymMonCat}_\infty \to \mathrm{Sp}$$

applied to nerves of ordinary symmetric monoidal categories recovers the 1–categorical algebraic K–theory $\mathrm{K}\colon \mathrm{SymMonCat} \to \mathrm{Sp}$. Note, however, that the inclusion of symmetric monoidal 1–categories into symmetric monoidal ∞–categories given by the nerve functor does not commute with the tensor products. In fact, the tensor product $\mathrm{N}(\mathcal{C}) \otimes \mathrm{N}(\mathcal{D})$ of the nerves of two symmetric monoidal 1–categories $\mathcal{C}$, $\mathcal{D}$ need not again be (the nerve of) a symmetric monoidal 1–category; rather, one can show that $\mathrm{N}(\mathcal{C} \otimes \mathcal{D})$ is the 1–categorical truncation of $\mathrm{N}(\mathcal{C}) \otimes \mathrm{N}(\mathcal{D})$.

Theorem 8-6 *The algebraic* K*–theory functor* $\mathrm{K}\colon \mathrm{SymMonCat}_\infty \to \mathrm{Sp}$ *is lax symmetric monoidal. In particular, it refines to a functor* $\mathrm{Rig}_{\mathcal{O}}(\mathrm{Cat}_\infty) \to \mathrm{RingSp}_{\mathcal{O}}$ *for any* ∞*–operad* $\mathcal{O}$*.*

Proof The proof is almost the same as in the 1–categorical case. The last two functors in the defining composition (8-4) are symmetric monoidal by Theorem 5-1. The remaining functor $(-)^\sim\colon \mathrm{SymMonCat}_\infty \to \mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{S})$ is the canonical extension of the accessible, product preserving functors $(-)^\sim\colon \mathrm{Cat}_\infty \to \mathcal{S}$. Thus, Corollary 6-6 implies that this canonical extension is lax symmetric monoidal as intended. □

Remark 8-7 The K–theory functor is defined as the composition (8-4) of lax symmetric monoidal functors. We know that the last two of these (namely, the group completion and the associated spectrum functor) are actually symmetric monoidal. Thus, one might wonder whether also the first functor (and hence the K–theory functor) is symmetric monoidal as well. This is not the case, as the following counterexample shows.

We begin by recalling from [20, Remark 2.1.3.10] that the ∞–category $\mathrm{Mon}_{\mathbb{E}_0}(\mathrm{Cat}_\infty)$ is equivalent to $(\mathrm{Cat}_\infty)_{\Delta^0/}$. Thus, an object in $\mathrm{Mon}_{\mathbb{E}_0}(\mathrm{Cat}_\infty)$ is just an ∞–category $\mathcal{C}$ together with a chosen object $x \in \mathcal{C}$. The fact that an ordinary monoid gives rise to a category with one object (which is hence distinguished) admits the following ∞–categorical variant. There is a functor

$$\mathrm{B}\colon \mathrm{Mon}_{\mathbb{E}_1}(\mathcal{S}) \to \mathrm{Mon}_{\mathbb{E}_0}(\mathrm{Cat}_\infty)$$

which is left adjoint to the functor which sends $x\colon \Delta^0 \to \mathcal{C}$ to the endomorphism monoid $\mathrm{End}_{\mathcal{C}}(x)$ of the distinguished object. Similarly, there is a functor

$$\mathrm{B}\colon \mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{S}) \to \mathrm{Mon}_{\mathbb{E}_\infty}(\mathrm{Cat}_\infty)$$

which is left adjoint to the functor which sends a symmetric monoidal ∞–category to the $\mathbb{E}_\infty$–monoid of endomorphisms of the monoidal unit (we are also using the fact that $\mathbb{E}_n \otimes \mathbb{E}_\infty \simeq \mathbb{E}_\infty$ for $n = 0, 1$).

Now, let $\mathcal{F} = \mathrm{Fr}(\Delta^0)$ denote the free symmetric monoidal ∞–category on the point, which is to say the nerve of the groupoid of finite sets and isomorphisms. We claim that, for any symmetric monoidal ∞–groupoid $\mathcal{C}$,

$$(\mathrm{B}\mathcal{F}) \otimes \mathcal{C} \simeq \mathrm{B}\mathcal{C}.$$

This is clearly true if $\mathcal{C} = \mathcal{F}$, and the general formula follows by the observation that both sides commute with colimits in the $\mathcal{C}$ variable and the fact that every symmetric monoidal ∞–groupoid is an iterated colimit of $\mathcal{F}$. But the groupoid core $(\mathrm{B}\mathcal{F})^{\sim}$ is trivial. Thus, $\mathrm{K}(\mathrm{B}\mathcal{F}) \otimes \mathrm{K}(\mathcal{C}) = 0$ for every $\mathcal{C}$. On the other hand, taking $\mathcal{C} = \mathbb{Z}$, we have that $(\mathrm{B}\mathcal{C})^{\sim} \simeq \mathrm{B}\mathcal{C}$, so $\mathrm{K}(\mathrm{B}\mathcal{C}) \simeq \Sigma\mathrm{H}\mathbb{Z}$, the suspension of the Eilenberg–MacLane spectrum.

We have the following *recognition principle* for semiring ∞–categories.

Theorem 8-8 *Let $\mathcal{C}$ be an $\mathbb{E}_n$–monoidal ∞–category with coproducts such that the monoidal product*

$$\otimes\colon \mathcal{C} \times \mathcal{C} \to \mathcal{C}$$

preserves coproducts separately in each variable. Then $(\mathcal{C}, \sqcup, \otimes)$ is canonically an object of $\mathrm{Rig}_{\mathbb{E}_n}(\mathrm{Cat}_\infty)$.

Proof Let $\mathrm{Cat}_\infty^\Sigma$ be the ∞–category of ∞–categories which admit finite coproducts and coproduct preserving functor. There is a fully faithful functor

$$\mathrm{Cat}_\infty^\Sigma \to \mathrm{SymMonCat}_\infty$$

given by considering an ∞–category with coproducts as a cocartesian symmetric monoidal ∞–category; see [20, Variant 2.4.3.12]. We want to show that this functor naturally extends to a lax symmetric monoidal functor, essentially by the construction of the tensor product on $\mathrm{Cat}_\infty^\Sigma$ of [20, Corollary 4.8.1.4] . From this the claim follows, since an $\mathbb{E}_n$–algebra in $\mathrm{Cat}_\infty^\Sigma$ is the same as an $\mathbb{E}_n$–monoidal ∞–category such that the tensor product preserves finite coproducts in each variable separately.

The first thing we want to observe is that the ∞–category $\mathrm{Cat}_\infty^\Sigma$ is preadditive. To see this, note that $\mathrm{Cat}_\infty^\Sigma$ has finite coproducts and products, because $\mathrm{Cat}_\infty^\Sigma$ is presentable; this follows from [20, Lemma 4.8.4.2] by taking $\mathcal{K}$ to be the collection of finite sets. It remains to check that the product $\mathcal{C} \times \mathcal{D}$, which is calculated as the product in $\mathrm{Ho}(\mathrm{Cat}_\infty)$, satisfies the universal property of the coproduct in $\mathrm{Ho}(\mathrm{Cat}_\infty^\Sigma)$. Given a third ∞–category with finite coproducts $\mathcal{E}$, we note that any pair of coproduct preserving functors $f\colon \mathcal{C} \to \mathcal{E}$ and $g\colon \mathcal{D} \to \mathcal{E}$ extends to the coproduct preserving functor

$$\mathcal{C} \times \mathcal{D} \xrightarrow{f \times g} \mathcal{E} \times \mathcal{E} \xrightarrow{\sqcup} \mathcal{E}.$$

Moreover, this extension is unique up to homotopy, because $(c,d) \cong (c,\varnothing) \sqcup (\varnothing,d)$ for any $(c,d) \in \mathcal{C} \times \mathcal{D}$.

Using [20, Proposition 4.8.1.10] again, the inclusion functor $i\colon \mathrm{Cat}_\infty^\Sigma \to \mathrm{Cat}_\infty$ admits a left adjoint L which is symmetric monoidal. By Proposition 5-4 the functor L extends to a left adjoint functor

$$L'\colon \mathrm{SymMonCat}_\infty \simeq \mathrm{Mon}_{\mathbb{E}_\infty}(\mathrm{Cat}_\infty) \to \mathrm{Mon}_{\mathbb{E}_\infty}(\mathrm{Cat}_\infty^\Sigma) \simeq \mathrm{Cat}_\infty^\Sigma.$$

The right adjoint of this functor can be described as the functor

$$\mathrm{Mon}_{\mathbb{E}_\infty}(i)\colon \mathrm{Cat}_\infty^\Sigma \simeq \mathrm{Mon}_{\mathbb{E}_\infty}(\mathrm{Cat}_\infty^\Sigma) \to \mathrm{Mon}_{\mathbb{E}_\infty}(\mathrm{Cat}_\infty).$$

We can now conclude that $\mathrm{Mon}_{\mathbb{E}_\infty}(i)$ is lax symmetric monoidal since it is right adjoint to a symmetric monoidal functor. It remains to show that $\mathrm{Mon}_{\mathbb{E}_\infty}(i)$ is the desired functor. This is obvious. □

Corollary 8-9 *If $\mathcal{C}$ is an ordinary monoidal category with coproducts such that $\otimes\colon \mathcal{C} \times \mathcal{C} \to \mathcal{C}$ preserves coproducts in each variable separately, then $(\mathcal{C}, \sqcup, \otimes)$ is canonically an object of $\mathrm{Rig}_{\mathbb{E}_1}(\mathrm{Cat}) \subset \mathrm{Rig}_{\mathbb{E}_1}(\mathrm{Cat}_\infty)$. If $\mathcal{C}$ is moreover braided or symmetric monoidal then $(\mathcal{C}, \sqcup, \otimes)$ is an object of $\mathrm{Rig}_{\mathbb{E}_2}(\mathrm{Cat})$ or $\mathrm{Rig}_{\mathbb{E}_\infty}(\mathrm{Cat})$ respectively.*

Proof We only need the identification of the $\mathbb{E}_n$–monoids in Cat with the respective monoidal categories. This has been given in [20, Example 5.1.2.4]. □

Corollary 8-10 *Let $\mathcal{C}$ be an $\mathbb{E}_n$–monoidal ∞–category with coproducts such that $\otimes\colon \mathcal{C} \times \mathcal{C} \to \mathcal{C}$ preserves coproducts in each variable separately. Then the largest Kan complex $\mathcal{C}^\sim$ inside of $\mathcal{C}$ together with $\sqcup$ and $\otimes$ is an object of $\mathrm{Rig}_{\mathbb{E}_n}(\mathcal{S}) \subseteq \mathrm{Rig}_{\mathbb{E}_n}(\mathrm{Cat}_\infty)$.*

Proof The functor $(-)^\sim\colon \mathrm{Cat}_\infty \to \mathcal{S} \subset \mathrm{Cat}_\infty$ preserves products and is accessible. Thus we can apply Corollary 6-6 to deduce that the induced functor $\mathrm{Mon}_{\mathbb{E}_\infty}(\mathrm{Cat}_\infty) \to \mathrm{Mon}_{\mathbb{E}_\infty}(\mathrm{Cat}_\infty)$ is lax symmetric monoidal. But this implies that we obtain a further functor $\mathrm{Rig}_{\mathbb{E}_n}(\mathrm{Cat}_\infty) \to \mathrm{Rig}_{\mathbb{E}_n}(\mathrm{Cat}_\infty)$ which preserves the underlying object of Cat_∞. Now apply this functor to the semiring ∞–category of Theorem 8-8. □

Example 8-11 (i) For an ordinary commutative ring R, let Mod_R denote the (ordinary) category of R–modules. Then Mod_R and the ∞–groupoid $\mathrm{Mod}_R^\sim$, equipped with the operations $\oplus$ and $\otimes_R$, form $\mathbb{E}_\infty$–semiring categories. The same applies to the category of sheaves on schemes and other variants.

(ii) For an $\mathbb{E}_n$–ring spectrum R, the ∞–category Mod_R of (left) R–modules is a $\mathbb{E}_{n-1}$–monoidal ∞–category by [20, Section 4.8 or Proposition 7.1.2.6]. Since the tensor product preserves coproducts in each variable we conclude that Mod_R, together with the coproduct $\oplus$ and tensor product $\otimes_R$, is an $\mathbb{E}_{n-1}$–semiring ∞–category.

Now we want to apply this to identify certain spectra as $\mathbb{E}_\infty$–ring spectra. For a connective $\mathbb{E}_{n+1}$–ring spectrum R the ∞–category Proj_R of finitely generated projective R–modules is an $\mathbb{E}_n$–semiring. The K–theory spectrum $\mathrm{K}(R)$ can then be defined as $\mathrm{K}(\mathrm{Proj}_R)$. This definition is actually equivalent to the definition using Waldhausen categories: for the variant which uses finitely generated free R–modules in place of projective, this is shown in [11, Chapter VI.7], and for the general case this follows from [5, Section 4].

Corollary 8-12 *For a connective $\mathbb{E}_{n+1}$–ring spectrum R, the algebraic* K*–theory spectrum $\mathrm{K}(R)$ of R is an $\mathbb{E}_n$–ring spectrum.*

We also have the following proposition, which states roughly that group completion of monoidal ∞–categories not only inverts objects, but arrows as well. It also shows why it is necessary to discard all non-invertible morphisms *before* group completion.

Proposition 8-13 *The underlying ∞–category of an $\mathbb{E}_\infty$–group object of Cat_∞ is an ∞–groupoid. More precisely, the group completion functor $\mathrm{Mon}_{\mathbb{E}_\infty}(\mathrm{Cat}_\infty) \to \mathrm{Grp}_{\mathbb{E}_\infty}(\mathrm{Cat}_\infty)$ factors through the groupoid completion*

$$\mathrm{Mon}_{\mathbb{E}_\infty}(\mathrm{Cat}_\infty) \to \mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{S}) \to \mathrm{Grp}_{\mathbb{E}_\infty}(\mathcal{S}) \to \mathrm{Grp}_{\mathbb{E}_\infty}(\mathrm{Cat}_\infty)$$

and induces an equivalence $\mathrm{Grp}_{\mathbb{E}_\infty}(\mathcal{S}) \simeq \mathrm{Grp}_{\mathbb{E}_\infty}(\mathrm{Cat}_\infty)$.

Proof Let $\mathcal{C}$ be an $\mathbb{E}_\infty$–group object of Cat_∞. Then the underlying ∞–category of $\mathcal{C}$ is an ∞–groupoid precisely if its homotopy category $\mathrm{Ho}(\mathcal{C})$ is a groupoid. Thus it suffices to show that $\mathrm{Ho}(\mathcal{C})$ is a groupoid. But since $\mathrm{Ho}(\mathcal{C})$ is a group object in Cat, this reduces the proof of the proposition to ordinary categories $\mathcal{C}$.

A group object $\mathcal{C}$ in categories is a symmetric monoidal category $(\mathcal{C}, \otimes)$ together with an inversion functor $I\colon \mathcal{C} \to \mathcal{C}$ as in to Proposition 1-1. We clearly have $I^2 \simeq \mathrm{id}$. As a first step we show that all endomorphisms of the tensor unit $\mathbb{1}$ in $\mathcal{C}$ are automorphisms. This follows from the Eckman–Hilton argument since $\hom_{\mathcal{C}}(\mathbb{1}, \mathbb{1})$ carries two commuting monoid structures (composition and tensoring), and as one of these is a group structure the other must also be as well. It follows that all endomorphisms in $\mathcal{C}$ are automorphisms by the identification

$$I(x)\otimes -\colon \hom_{\mathcal{C}}(x,x)\cong \hom_{\mathcal{C}}(\mathbb{1},\mathbb{1}).$$

Finally, to show that $\mathcal{C}$ is a groupoid, it now suffices to show that for every morphisms $f\colon x\to y$ in $\mathcal{C}$ there is a morphism $g\colon y\to x$ in $\mathcal{C}$. By tensoring with $I(y)$ we see that we may assume that $y=\mathbb{1}$. Then we have $I(f)\colon I(x)\to\mathbb{1}$, and therefore, using the usual identifications, $g:=I(f)\otimes x\colon \mathbb{1}\to x$. □

Remark 8-14 Proposition 8-13 is closely related to our comment in Remark 1-5. More precisely, the reason that group completion produces groupoids lies in our definition of group objects. We could have alternatively stipulated that a symmetric monoidal ∞–category $\mathcal{C}$ (considered as an object in $\mathrm{Mon}_{\mathbb{E}_\infty}(\mathrm{Cat}_\infty)$) is a *group* if every object of $\mathcal{C}$ is tensor invertible (or even just dualizable). Let us denote the ∞–category of symmetric monoidal categories satisfying this weaker group condition by $\mathrm{Grp}^{\tau}_{\mathbb{E}_\infty}(\mathrm{Cat}_\infty)$. Then we have strict inclusions

$$\mathrm{Grp}_{\mathbb{E}_\infty}(\mathrm{Cat}_\infty)\subset \mathrm{Grp}^{\tau}_{\mathbb{E}_\infty}(\mathrm{Cat}_\infty)\subset \mathrm{Mon}_{\mathbb{E}_\infty}(\mathrm{Cat}_\infty)$$

of reflective subcategories. The reflection from $\mathrm{Mon}_{\mathbb{E}_\infty}(\mathrm{Cat}_\infty)$ to $\mathrm{Grp}^{\tau}_{\mathbb{E}_\infty}(\mathrm{Cat}_\infty)$ is closely related to Quillen's $S^{-1}S$ construction.

Appendix A: Comonoids

In this short section we establish additional universal mapping properties for $\mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{S})$ and $\mathrm{Grp}_{\mathbb{E}_\infty}(\mathcal{S})$ respectively. This gives a characterization of these ∞–categories among all presentable ∞–categories and not only among the (pre)additive ones. We write $\mathrm{Fun}^{\mathrm{RAd}}(\mathcal{C},\mathcal{D})$ for the ∞–category of right adjoint functors from $\mathcal{C}$ to $\mathcal{D}$, which is a full subcategory of $\mathrm{Fun}(\mathcal{C},\mathcal{D})$.

Lemma A-1 *If $\mathcal{C}$ and $\mathcal{D}$ are presentable, then we have canonical equivalences*

$$\mathrm{Mon}_{\mathbb{E}_\infty}(\mathrm{Fun}^{\mathrm{RAd}}(\mathcal{C},\mathcal{D}))\simeq \mathrm{Fun}^{\mathrm{RAd}}(\mathcal{C},\mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{D})),$$
$$\mathrm{Grp}_{\mathbb{E}_\infty}(\mathrm{Fun}^{\mathrm{RAd}}(\mathcal{C},\mathcal{D}))\simeq \mathrm{Fun}^{\mathrm{RAd}}(\mathcal{C},\mathrm{Grp}_{\mathbb{E}_\infty}(\mathcal{D})).$$

Proof We note that right adjoint functors between presentable ∞–categories can be described as accessible functors that preserve limits. Then the proof works exactly the same as the proof of Lemma 1-6. □

Definition A-2 Let $\mathcal{C}$ be an ∞–category with finite coproducts. We define the ∞–categories of *comonoids* and *cogroups* in $\mathcal{C}$ to be the respective ∞–categories

$$\mathrm{coMon}_{\mathbb{E}_\infty}(\mathcal{C}) = \mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{C}^{\mathrm{op}})^{\mathrm{op}} \quad \text{and} \quad \mathrm{coGrp}_{\mathbb{E}_\infty}(\mathcal{C}) = \mathrm{Grp}_{\mathbb{E}_\infty}(\mathcal{C}^{\mathrm{op}})^{\mathrm{op}}.$$

Remark A-3 The comonoids as defined above are comonoids for the coproduct as tensor product. This is a structure which is often rather trivial. For example in the ∞–category $\mathcal{S}$ of spaces (or in the ordinary category of sets) there is exactly one comonoid in the sense above, namely the empty set $\varnothing$.

Proposition A-4 *Let $\mathcal{C}$ and $\mathcal{D}$ be presentable ∞–categories. Then there are natural equivalences*

$$\mathrm{Fun}^{\mathrm{L}}(\mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{C}),\mathcal{D}) \simeq \mathrm{coMon}_{\mathbb{E}_\infty}(\mathrm{Fun}^{\mathrm{L}}(\mathcal{C},\mathcal{D}))$$
$$\mathrm{Fun}^{\mathrm{L}}(\mathrm{Grp}_{\mathbb{E}_\infty}(\mathcal{C}),\mathcal{D}) \simeq \mathrm{coGrp}_{\mathbb{E}_\infty}(\mathrm{Fun}^{\mathrm{L}}(\mathcal{C},\mathcal{D})).$$

In particular, for a presentable ∞–category $\mathcal{D}$ we have natural equivalences

$$\mathrm{Fun}^{\mathrm{L}}(\mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{S}),\mathcal{D}) \simeq \mathrm{coMon}_{\mathbb{E}_\infty}(\mathcal{D}) \quad \textit{and} \quad \mathrm{Fun}^{\mathrm{L}}(\mathrm{Grp}_{\mathbb{E}_\infty}(\mathcal{S}),\mathcal{D}) \simeq \mathrm{coGrp}_{\mathbb{E}_\infty}(\mathcal{D}).$$

Proof Let us recall that given two ∞–categories $\mathcal{E}$ and $\mathcal{F}$, there is an equivalence of categories $\mathrm{Fun}^{\mathrm{L}}(\mathcal{E},\mathcal{F})$ and $\mathrm{Fun}^{\mathrm{RAd}}(\mathcal{F},\mathcal{E})^{\mathrm{op}}$ [19, Proposition 5.2.6.2]. The adjoint functor theorem [19, Corollary 5.5.2.9] together with Lemma 1-6 then yields the following chain of equivalences:

$$\begin{aligned}
\mathrm{Fun}^{\mathrm{L}}(\mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{C}),\mathcal{D}) &\simeq \mathrm{Fun}^{\mathrm{RAd}}(\mathcal{D},\mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{C}))^{\mathrm{op}}\\
&\simeq \mathrm{Mon}_{\mathbb{E}_\infty}(\mathrm{Fun}^{\mathrm{RAd}}(\mathcal{D},\mathcal{C}))^{\mathrm{op}}\\
&\simeq \mathrm{Mon}_{\mathbb{E}_\infty}(\mathrm{Fun}^{\mathrm{L}}(\mathcal{C},\mathcal{D})^{\mathrm{op}})^{\mathrm{op}}\\
&= \mathrm{coMon}_{\mathbb{E}_\infty}(\mathrm{Fun}^{\mathrm{L}}(\mathcal{C},\mathcal{D})).
\end{aligned}$$

In the special case of $\mathcal{C}=\mathcal{S}$ we can use the universal property of ∞–categories of presheaves [19, Theorem 5.1.5.6] to extend the above chain of equivalences by

$$\mathrm{coMon}_{\mathbb{E}_\infty}(\mathrm{Fun}^{\mathrm{L}}(\mathcal{S},\mathcal{D})) \simeq \mathrm{coMon}_{\mathbb{E}_\infty}(\mathcal{D}).$$

This settles the case of monoids and the case of groups works the same. □

Appendix B: Algebraic theories and monadic functors

In this section we give a short discussion of Lawvere algebraic theories in ∞–categories and show that our examples are algebraic. For other treatments of ∞–categorical algebraic theories; see [9; 10], [15, Section 32] and [19, Section 5.5.8]. We write Fin for the category of finite sets.

Definition B-1 An *algebraic theory* is an ∞–category $\mathbb{T}$ with finite products and a distinguished object $1_{\mathbb{T}}$, such that the unique product-preserving functor $\mathrm{N}(\mathrm{Fin})^{\mathrm{op}} \to \mathbb{T}$ which sends the singleton to $1_{\mathbb{T}}$ is essentially surjective. A *morphism of algebraic theories* is a functor which preserves products and the distinguished object. We write $\mathrm{Th} \subseteq (\mathrm{Cat}_{\infty}^{\Pi})_*$ for the ∞–category of theories and morphisms thereof.

This is the obvious generalization of algebraic theories, as defined by Lawvere [18], to ∞–categories.

Definition B-2 Let $\mathcal{C}$ be an ∞–category with finite products. An *algebra* in $\mathcal{C}$ for an algebraic theory $\mathbb{T}$ is a finite product preserving functor $\mathbb{T} \to \mathcal{C}$. We write $\mathrm{Alg}_{\mathbb{T}}(\mathcal{C})$ for the ∞–category of algebras of $\mathbb{T}$ in $\mathcal{C}$, ie, for the full subcategory of $\mathrm{Fun}(\mathbb{T}, \mathcal{C})$ spanned by the algebras.

The notation $\mathrm{Alg}_{\mathbb{T}}(\mathcal{C})$ should not be confused with the definition of algebra for an ∞–operad used previously in the paper. If $\mathcal{C}$ is a presentable ∞–category and $\mathbb{T}$ a theory, then $\mathrm{Alg}_{\mathbb{T}}(\mathcal{C})$ is again presentable. This follows since $\mathrm{Alg}_{\mathbb{T}}(\mathcal{C})$ is an accessible localization of the presentable ∞–category $\mathrm{Fun}(\mathbb{T}, \mathcal{C})$ (the proof is similar to that of Proposition 4-1 which takes care of the case of commutative monoids). Applying the adjoint functor theorem we also get that the forgetful functor $\mathrm{Alg}_{\mathbb{T}}(\mathcal{C}) \to \mathcal{C}$, ie the evaluation at the distinguished object $1_{\mathbb{T}}$, has a left adjoint.

Proposition B-3 *Let $\mathcal{C}$ be a presentable ∞–category and $\mathbb{T}$ a theory. Then we have an equivalence*

$$\mathrm{Alg}_{\mathbb{T}}(\mathcal{C}) \simeq \mathcal{C} \otimes \mathrm{Alg}_{\mathbb{T}}(\mathcal{S}).$$

Proof The same proof as for Lemma 1-6 shows that we have an equivalence

$$\mathrm{Alg}_{\mathbb{T}}(\mathrm{Fun}^{\mathrm{R}}(\mathcal{C}^{\mathrm{op}}, \mathcal{S})) \simeq \mathrm{Fun}^{\mathrm{R}}(\mathcal{C}^{\mathrm{op}}, \mathrm{Alg}_{\mathbb{T}}(\mathcal{S})).$$

This then implies the claim since we have $\mathcal{C} \otimes \mathcal{D} \simeq \mathrm{Fun}^{\mathrm{R}}(\mathcal{C}^{\mathrm{op}}, \mathcal{D})$ for any presentable ∞–category $\mathcal{D}$. □

A *monad* on an ∞–category $\mathcal{C}$ is an algebra M in the monoidal ∞–category $\mathrm{Fun}(\mathcal{C}, \mathcal{C})$ of endofunctors; see [20, Chapter 4.7] for details. Any such monad $M \in \mathrm{Alg}(\mathrm{Fun}(\mathcal{C}, \mathcal{C}))$ admits an ∞–category of *modules* which we denote $\mathrm{Alg}_M(\mathcal{C})$. This ∞–category comes equipped with a forgetful functor $\mathrm{Alg}_M(\mathcal{C}) \to \mathcal{C}$ which is a right adjoint (again, this ∞–category should not be confused with the ∞–category of algebras for an ∞–operad). Thus, given an arbitrary right adjoint functor $U\colon \mathcal{D} \to \mathcal{C}$, it is natural to ask whether this functor is equivalent to the forgetful functor from modules over a monad on $\mathcal{C}$. In this case the corresponding monad is uniquely determined as the composition $M = U \circ F$, where F is a left adjoint of U. The functors U for which this is the case are called *monadic*.

The Barr–Beck theorem (also called Beck's monadicity theorem) gives necessary and sufficient conditions for a functor U to be monadic. The conditions are that U is conservative (ie, reflects equivalences) and that U preserves U–split geometric realizations [20, Theorem 4.7.4.5]. We will not need to discuss here what U–split means exactly since in our cases all geometric realizations will be preserved.

Proposition B-4 *Let $\mathcal{C}$ be a presentable ∞–category and let $\mathbb{T}$ be a theory. Then the forgetful functor $\mathrm{Alg}_{\mathbb{T}}(\mathcal{C}) \to \mathcal{C}$ is monadic and preserves sifted colimits.*

Proof We will show that the evaluation $\mathrm{Fun}^{\Pi}(\mathbb{T}, \mathcal{C}) \to \mathcal{C}$ is conservative and preserves sifted colimits. The result then follows immediately from the monadicity theorem. That the functor is conservative is clear, so it remains to check the sifted colimit condition. But the inclusion of the finite product preserving functors

$$\mathrm{Fun}^{\Pi}(\mathbb{T}, \mathcal{C}) \to \mathrm{Fun}(\mathbb{T}, \mathcal{C})$$

preserves sifted colimits by (4) of [19, Proposition 5.5.8.10], and as colimits in functor ∞–categories are computed pointwise the evaluation $\mathrm{Fun}(\mathbb{T}, \mathcal{C}) \to \mathcal{C}$ also preserves sifted colimits. □

We will obtain a converse to the previous proposition in the case of the ∞–category of spaces; namely, in this case we will identify algebraic theories with certain monads. To this end, note that an arbitrary monadic functor $U\colon \mathrm{Alg}_M(\mathcal{S}) \to \mathcal{S}$ defines a theory $\mathbb{T}_M$ by

$$\mathbb{T}_M := (\mathrm{Alg}^{\mathrm{ff}}_M(\mathcal{S}))^{\mathrm{op}},$$

where $\mathrm{Alg}^{\mathrm{ff}}_M(\mathcal{S}) \subseteq \mathrm{Alg}_M(\mathcal{S})$ is the full subcategory spanned by the free M–algebras on finite sets (which we abusively refer to as *finite free* algebras, and should not to be confused with more general free algebras on finite or finitely presented spaces). There is a canonical functor

$$R\colon \mathrm{Alg}_M(\mathcal{S}) \to \mathrm{Alg}_{\mathbb{T}_M}(\mathcal{S})$$

from modules for M to models to the associated theory $\mathbb{T}_M$, which is just the restriction of the Yoneda embedding to the full subcategory $\mathrm{Alg}^{\mathrm{ff}}_M(\mathcal{S})$.

Definition B-5 A monadic functor $U\colon \mathrm{Alg}_M(\mathcal{S}) \to \mathcal{S}$ is called *algebraic* if the restricted Yoneda embedding $R\colon \mathrm{Alg}_M(\mathcal{S}) \to \mathrm{Alg}_{\mathbb{T}_M}(\mathcal{S})$ is an equivalence of ∞–categories over $\mathcal{S}$. We also say that a monad M on spaces is *algebraic* if the associated forgetful functor $U\colon \mathrm{Alg}_M(\mathcal{S}) \to \mathcal{S}$ is algebraic.

The main result of this section is Theorem B-7, which provides necessary and sufficient conditions for a monadic functor to spaces to be algebraic. As preparation, we first collect the following result, a straightforward generalization of a well-known result in ordinary category theory.

Proposition B-6 *Let $\mathcal{C}$ be a presentable ∞–category and let $M\colon \mathcal{C} \to \mathcal{C}$ be a monad which commutes with κ–filtered colimits for some infinite regular cardinal κ. Then $\mathrm{Alg}_M(\mathcal{C})$ is a presentable ∞–category.*

Proof To begin with let us choose a regular cardinal κ such that $\mathcal{C}$ is κ–compactly generated and M commutes with κ–filtered colimits. Let $\mathcal{C}^\kappa \subseteq \mathcal{C}$ and $\mathrm{Alg}_M(\mathcal{C})^\kappa \subseteq \mathrm{Alg}_M(\mathcal{C})$ be the respective full subcategories spanned by the κ–compact objects. We claim that there is an equivalence $\mathrm{Ind}_\kappa(\mathrm{Alg}_M(\mathcal{C})^\kappa) \simeq \mathrm{Alg}_M(\mathcal{C})$. Since $\mathrm{Alg}_M(\mathcal{C})$ admits κ–filtered colimits, the inclusion $\mathrm{Alg}_M(\mathcal{C})^\kappa \subseteq \mathrm{Alg}_M(\mathcal{C})$ induces a functor

$$\phi\colon \mathrm{Ind}_\kappa(\mathrm{Alg}_M(\mathcal{C})^\kappa) \to \mathrm{Alg}_M(\mathcal{C})$$

which we want to show is an equivalence. The fully faithfulness of ϕ is a special case of the following: if $\mathcal{D}$ is an ∞–category with κ–filtered colimits, then the inclusion $\mathcal{D}^\kappa \subseteq \mathcal{D}$ of the κ–compact objects induces a fully faithful functor $\mathrm{Ind}_\kappa(\mathcal{D}^\kappa) \to \mathcal{D}$. Thus it remains to show that ϕ is essentially surjective.

Because M commutes with κ–filtered colimits, we see that, if $X \in \mathcal{C}^\kappa$, then $FX \in \mathrm{Alg}_M(\mathcal{C})^\kappa$, where $F\colon \mathcal{C} \to \mathrm{Alg}_M(\mathcal{C})$ denotes a left adjoint to the forgetful functor $\mathrm{Alg}_M(\mathcal{C}) \to \mathcal{C}$. Since the forgetful functor $\mathrm{Alg}_M(\mathcal{C}) \to \mathcal{C}$ is conservative and $\mathcal{C}$ is κ–compactly generated, a map $f\colon A \to B$ of M–modules is an equivalence if and only if

$$\mathrm{map}_{\mathrm{Alg}_M(\mathcal{C})}(FX, A) \to \mathrm{map}_{\mathrm{Alg}_M(\mathcal{C})}(FX, B)$$

is an equivalence for all $X \in \mathcal{C}^\kappa$. We apply this criterion to the map $\mathrm{colim}_{A' \in \mathrm{Alg}_M(\mathcal{C})^\kappa_{/A}} A' \to A$, whose domain is a κ–filtered colimit, in order to obtain the essential surjectivity of ϕ. We first show that, for any $X \in \mathcal{C}^\kappa$, the induced map

$$\mathrm{colim}_{\mathrm{Alg}_M(\mathcal{C})^\kappa_{/A}} \pi_0 \mathrm{map}(FX, A') \to \pi_0 \mathrm{map}(FX, A)$$

is an isomorphism. Indeed, it is surjective because any (homotopy class of the) map $FX \to A$ is the image of the identity map $FX \to A'$ for $A' = FX$, which is by construction a κ–compact object of $\mathrm{Alg}_M(\mathcal{C})$. Similarly, injectivity follows because given any two maps $f, g\colon FX \to A'$, the fact that $\mathrm{Alg}_M(\mathcal{C})^\kappa_{/A}$ is κ–filtered implies that there exists an $A'' \to A$ which coequalizes f and g. Replacing X by $K \otimes X$ for some finite simplicial set K, and noting that $K \otimes X$ is a κ–compact object of $\mathcal{C}$ since K is finite, we obtain an isomorphism

$$\pi_0 \mathrm{map}(K, \operatorname{colim} \mathrm{map}(FX, A')) \cong \pi_0 \mathrm{map}(K, \mathrm{map}(FX, A)).$$

It follows that $\operatorname{colim} \mathrm{map}(FX, A') \to \mathrm{map}(FX, A)$ is a homotopy equivalence, as desired. $\square$

Theorem B-7 *A monadic functor $U\colon \mathrm{Alg}_M(\mathcal{S}) \to \mathcal{S}$ is algebraic if and only if it preserves sifted colimits.*

Proof This is a necessary condition since the forgetful functor $\mathrm{Alg}_{\mathbb{T}_M}(\mathcal{S}) \to \mathcal{S}$ preserves sifted colimits (see Proposition B-4). Thus, suppose that U preserves sifted colimits; we must show that R is an equivalence. Note that $\mathrm{Alg}_M(\mathcal{S})$ is presentable by Proposition B-6, and $\mathrm{Alg}_{\mathbb{T}_M}(\mathcal{S})$ is presentable as an accessible localization of the presentable ∞–category $\mathrm{Fun}(\mathbb{T}, \mathcal{S})$. Because $\mathrm{Alg}^{\mathrm{ff}}_M(\mathcal{S}) \subseteq \mathrm{Alg}_M(\mathcal{S})$ is a subcategory of compact projective objects, R preserves sifted colimits, and clearly R also preserves small limits. Thus R admits a left adjoint L.

We now check that the adjunction counit $LR \to \mathrm{id}$ is an equivalence. Since R is conservative, as both the projections down to $\mathcal{S}$ are conservative, this will also imply that the unit $\mathrm{id} \to RL$ is an equivalence. Observe that both functors commute with sifted colimits and spaces is freely generated under sifted colimits by the finite sets $\langle n \rangle$, it is enough to check the counit equivalence on objects of the form $F\langle n \rangle$. Now, $RF\langle n \rangle = \widehat{F}\langle n \rangle$, the functor represented by $\widehat{F}\langle n \rangle$, so we must show that we have an equivalence $L\widehat{F}\langle n \rangle \to F\langle n \rangle$. Let $A \in \mathrm{Alg}_M(\mathcal{S})$ and consider the map

$$\mathrm{map}(F\langle n \rangle, A) \to \mathrm{map}(L\hat{F}\langle n \rangle, A).$$

The left-hand side can be identified with $\mathrm{map}(F\langle n \rangle, A) \simeq U(A)^n$. Similarly, the right-hand side is

$$\mathrm{map}(L\widehat{F}\langle n \rangle, A) \simeq \mathrm{map}(\widehat{F}\langle n \rangle, RA) \simeq \mathrm{map}(\widehat{F}\langle 1 \rangle, RA)^n \simeq U(A)^n$$

where we used in the last step that R is compatible with the forgetful functors to $\mathcal{S}$. $\square$

Finally, we wish to apply the results of this section to the study of semirings and rings in ∞–categories. We begin by showing that semirings and rings are algebraic over spaces.

Proposition B-8 *The functors $\mathrm{Rig}_{\mathbb{E}_n}(\mathcal{S}) \to \mathcal{S}$ and $\mathrm{Ring}_{\mathbb{E}_n}(\mathcal{S}) \to \mathcal{S}$ are monadic and algebraic over $\mathcal{S}$.*

Proof We claim that the functors $\mathrm{Rig}_{\mathbb{E}_n}(\mathcal{S}) \to \mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{S}) \to \mathcal{S}$ and $\mathrm{Ring}_{\mathbb{E}_n}(\mathcal{S}) \to \mathrm{Grp}_{\mathbb{E}_\infty}(\mathcal{S}) \to \mathcal{S}$ all preserve sifted colimits and reflect equivalences. Then the monadicity follows from the Barr–Beck theorem [20, Theorem 4.7.4.5], and the algebraicity from Theorem B-7.

To see that this claim is true note that three of the four functors, namely $\mathrm{Rig}_{\mathbb{E}_n}(\mathcal{S}) \to \mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{S})$, $\mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{S}) \to \mathcal{S}$, and $\mathrm{Ring}_{\mathbb{E}_n}(\mathcal{S}) \to \mathrm{Grp}_{\mathbb{E}_\infty}(\mathcal{S})$, are forgetful functors from ∞–categories of algebras over an ∞–operad. These forgetful functors are always conservative and for suitable monoidal structures they also preserve sifted colimits [20, Corollary 3.2.3.2]. Thus we only have to establish the same properties for $\mathrm{Grp}_{\mathbb{E}_\infty}(\mathcal{S}) \to \mathcal{S}$. It is easy to see that this functor is conservative since $\mathrm{Grp}_{\mathbb{E}_\infty}(\mathcal{S})$ is a full subcategory of $\mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{S})$ and the given functor factors over the conservative functor $\mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{S}) \to \mathcal{S}$. It remains to show that $\mathrm{Grp}_{\mathbb{E}_\infty}(\mathcal{S}) \to \mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{S})$ preserves sifted colimits. But for an $\mathbb{E}_\infty$–monoid in the ∞–category of spaces being a group object is equivalent to being grouplike. Thus, via the left adjoint functor π_0 it reduces to the statement that the sifted colimit of groups formed in the category of monoids is again a group. And this result is a special case of [1, Proposition 9.3]. □

Definition B-9 We denote the algebraic theory corresponding to the functor

$$\mathrm{Rig}_{\mathbb{E}_n}(\mathcal{S}) \to \mathcal{S}$$

by $\mathbb{T}_{\mathbb{E}_n\text{–Rig}}$ and call it the *theory of* $\mathbb{E}_n$*–semirings*. Accordingly we denote the algebraic theory corresponding to the functor $\mathrm{Ring}_{\mathbb{E}_n}(\mathcal{S}) \to \mathcal{S}$ by $\mathbb{T}_{\mathbb{E}_n\text{–Ring}}$ and call it the *theory of* $\mathbb{E}_n$*–rings*.

Proposition B-10 *Let $\mathcal{C}$ be a cartesian closed, presentable ∞–category. Then we have equivalences*

$$\mathrm{Rig}_{\mathbb{E}_n}(\mathcal{C}) \simeq \mathrm{Alg}_{\mathbb{T}_{\mathbb{E}_n\text{-Rig}}}(\mathcal{C}) \quad \text{and} \quad \mathrm{Ring}_{\mathbb{E}_n}(\mathcal{C}) \simeq \mathrm{Alg}_{\mathbb{T}_{\mathbb{E}_n\text{-Ring}}}(\mathcal{C}).$$

Proof For $\mathcal{C} = \mathcal{S}$ the ∞–category of spaces the statement is true by definition of $\mathbb{T}_{\mathbb{E}_n\text{–Rig}}$ and $\mathbb{T}_{\mathbb{E}_n\text{–Ring}}$. The general case follows from the base change formulas given in Propositions 7-7 and B-3. □

Remark B-11 (i) Theories of $\mathbb{E}_\infty$–semirings and rings have also been constructed in [9] by the use of spans and distributive laws. These two approaches do agree.

(ii) The theory approach of semirings and rings gives a way of defining ring objects in a much broader generality. One only needs an ∞–category $\mathcal{C}$ with finite products. In this way we can drop the assumption that $\mathcal{C}$ is presentable and cartesian closed. However in this case semiring and ring objects do not admit a nice description in terms of a tensor product on monoids. It is also impossible to apply this to different tensor products than the cartesian one.

(iii) We showed in Corollary 6-6 that an accessible, product preserving functor $F\colon \mathcal{C} \to \mathcal{D}$ between cartesian closed symmetric monoidal categories extends to a lax symmetric monoidal functor $\mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{C}) \to \mathrm{Mon}_{\mathbb{E}_\infty}(\mathcal{D})$. This means that F extends to functors $\mathrm{Rig}_{\mathbb{E}_n}(\mathcal{C}) \to \mathrm{Rig}_{\mathbb{E}_n}(\mathcal{D})$ and $\mathrm{Ring}_{\mathbb{E}_n}(\mathcal{C}) \to \mathrm{Ring}_{\mathbb{E}_n}(\mathcal{D})$. Therefore we may drop the assumption that F is accessible and conclude that any product preserving functor $\mathcal{C} \to \mathcal{D}$ extends to functors $\mathrm{Rig}_{\mathbb{E}_n}(\mathcal{C}) \to \mathrm{Rig}_{\mathbb{E}_n}(\mathcal{D})$ and $\mathrm{Ring}_{\mathbb{E}_n}(\mathcal{C}) \to \mathrm{Ring}_{\mathbb{E}_n}(\mathcal{D})$.

References

[1] **J Adámek**, **J Rosický**, **E M Vitale**, *Algebraic theories: A categorical introduction to general algebra*, Cambridge Tracts in Mathematics 184, Cambridge Univ. Press (2011) MR2757312

[2] **N A Baas**, **B I Dundas**, **B Richter**, **J Rognes**, *Ring completion of rig categories*, J. Reine Angew. Math. 674 (2013) 43–80 MR3010546

[3] **C Barwick**, *Multiplicative structures on algebraic K–theory*, preprint (2013) `arXiv:1304.4867`

[4] **H S Bergsaker**, *Homotopy theory of presheaves of Γ–spaces*, Homology, Homotopy Appl. 11 (2009) 35–60 MR2486652

[5] **A Blumberg**, **D Gepner**, **G Tabuada**, *K–theory of endomorphisms via noncommutative motives* (2013) `arXiv:1302.1214`

[6] **A J Blumberg**, **D Gepner**, **G Tabuada**, *A universal characterization of higher algebraic K–theory*, Geom. Topol. 17 (2013) 733–838 MR3070515

[7] **A J Blumberg**, **D Gepner**, **G Tabuada**, *Uniqueness of the multiplicative cyclotomic trace*, Adv. Math. 260 (2014) 191–232 MR3209352

[8] **J M Boardman**, **R M Vogt**, *Homotopy invariant algebraic structures on topological spaces*, Lecture Notes in Mathematics 347, Springer, Berlin (1973) MR0420609

[9] **J Cranch**, *Algebraic theories and $(\infty, 1)$–categories*, PhD thesis, University of Sheffield (2010) `arXiv:1011.3243`

[10] **J Cranch**, *Algebraic theories, span diagrams, and commutative monoids in homotopy theory*, preprint (2011) `arXiv:1109.1598`

[11] **A D Elmendorf**, **I Kriz**, **M A Mandell**, **J P May**, *Rings, modules, and algebras in stable homotopy theory*, Mathematical Surveys and Monographs 47, Amer. Math. Soc. (1997) MR1417719

[12] **A D Elmendorf**, **M A Mandell**, *Rings, modules, and algebras in infinite loop space theory*, Adv. Math. 205 (2006) 163–228 MR2254311

[13] **A D Elmendorf**, **M A Mandell**, *Permutative categories, multicategories and algebraic K–theory*, Algebr. Geom. Topol. 9 (2009) 2391–2441 MR2558315

[14] **M Hyland**, **J Power**, *Pseudo-commutative monads and pseudo-closed 2–categories*, J. Pure Appl. Algebra 175 (2002) 141–185 MR1935977

[15] **A Joyal**, *Notes on quasi-categories* (2008) Available at `http://www.math.uchicago.edu/~may/IMA/Joyal.pdf`

[16] **M L Laplaza**, *Coherence for distributivity*, from: "Coherence in categories", (S Mac Lane, editor), Lecture Notes in Math. 281, Springer, Berlin (1972) 29–65 MR0335598

[17] **M L Laplaza**, *A new result of coherence for distributivity*, from: "Coherence in categories", (S Mac Lane, editor), Lecture Notes in Math. 281, Springer, Berlin (1972) 214–235 MR0335599

[18] **F W Lawvere**, *Functorial semantics of algebraic theories*, Proc. Nat. Acad. Sci. USA 50 (1963) 869–872 MR0158921

[19] **J Lurie**, *Higher topos theory*, Annals of Mathematics Studies 170, Princeton Univ. Press (2009) MR2522659

[20] **J Lurie**, *Higher algebra* (2014) Available at `http://www.math.harvard.edu/~lurie/papers/higheralgebra.pdf`

[21] **J P May**, *The geometry of iterated loop spaces*, Lecture Notes in Mathematics 271, Springer, Berlin (1972) MR0420610

[22] **J P May**, *Multiplicative infinite loop space theory*, J. Pure Appl. Algebra 26 (1982) 1–69 MR669843

[23] **J P May**, *The construction of E_∞ ring spaces from bipermutative categories*, from: "New topological contexts for Galois theory and algebraic geometry", (A Baker, B Richter, editors), Geom. Topol. Monogr. 16 (2009) 283–330 MR2544392

[24] **J P May**, *What are E_∞ ring spaces good for?*, from: "New topological contexts for Galois theory and algebraic geometry", (A Baker, B Richter, editors), Geom. Topol. Monogr. 16 (2009) 331–365 MR2544393

[25] **J P May**, *What precisely are E_∞ ring spaces and E_∞ ring spectra?*, from: "New topological contexts for Galois theory and algebraic geometry", (A Baker, B Richter, editors), Geom. Topol. Monogr. 16 (2009) 215–282 MR2544391

[26] **V Schmitt**, *Tensor product for symmetric monoidal categories*, preprint (2007) `arXiv:0711.0324`

[27] **S Schwede**, *Stable homotopical algebra and Γ–spaces*, Math. Proc. Cambridge Philos. Soc. 126 (1999) 329–356 MR1670249

[28] **G Segal**, *Categories and cohomology theories*, Topology 13 (1974) 293–312 MR0353298

[29] **R W Thomason**, *Beware the phony multiplication on Quillen's $\mathcal{A}^{-1}\mathcal{A}$*, Proc. Amer. Math. Soc. 80 (1980) 569–573 MR587929

[30] **B Töen**, **G Vezzosi**, *Caractères de Chern, traces équivariantes et géométrie algébrique dérivée*, preprint (2009) `arXiv:0903.3292`

Department of Mathematics, Purdue University
150 N. University Street
West Lafayette, IN 47907, USA

Max-Planck-Institut für Mathematik
Vivatsgasse 7
D-53111 Bonn, Germany

Mathematisches Institut der Universität Bonn
Endenicher Allee 60
D-53115 Bonn, Germany

`dgepner@math.purdue.edu`, `mgroth@mpim-bonn.mpg.de`,
`nikolaus@math.uni-bonn.de`

Received: 2 October 2013 Revised: 9 October 2014

Algebraic & Geometric Topology 15 (2015) 3155–3213

Floer homology and splicing knot complements

EAMAN EFTEKHARY

We obtain a formula for the Heegaard Floer homology (hat theory) of the three-manifold $Y(K_1, K_2)$ obtained by splicing the complements of the knots $K_i \subset Y_i$, $i = 1, 2$, in terms of the knot Floer homology of K_1 and K_2. We also present a few applications. If h_n^i denotes the rank of the Heegaard Floer group $\widehat{\mathrm{HFK}}$ for the knot obtained by n–surgery over K_i, we show that the rank of $\widehat{\mathrm{HF}}(Y(K_1, K_2))$ is bounded below by

$$|(h_\infty^1 - h_1^1)(h_\infty^2 - h_1^2) - (h_0^1 - h_1^1)(h_0^2 - h_1^2)|.$$

We also show that if splicing the complement of a knot $K \subset Y$ with the trefoil complements gives a homology sphere L–space, then K is trivial and Y is a homology sphere L–space.

57M27; 57R58

1 Introduction

Heegaard Floer homology, introduced by Ozsváth and Szabó [12], has been the source of powerful techniques for the study of objects in low-dimensional topology. It is interesting to investigate whether Heegaard Floer homology can distinguish the standard sphere from other homology spheres. Since the Heegaard Floer groups of the connected sum of two homology spheres are obtained as the tensor product of the Heegaard Floer groups associated with the two pieces, the question is reduced to determining prime homology spheres with trivial Heegaard Floer groups. The Poincaré homology sphere $\Sigma(2, 3, 5)$ with either orientation is the unique known example of a non-trivial prime homology sphere Y with $\widehat{\mathrm{HF}}(Y; \mathbb{Z}) = \mathbb{Z}$. A conjecture of Ozsváth and Szabó predicts that this is in fact the only possible example.

In this paper we study the Heegaard Floer groups of a homology sphere Y which contains an incompressible torus. We may use the incompressible torus to decompose Y, fill out the torus boundary of each of the two pieces by gluing a solid torus, and obtain two new homology spheres, Y_1 and Y_2. By requiring Y_1 and Y_2 to be homology spheres the gluing of the solid tori is determined; the decomposition determines a knot K_i in Y_i, $i = 1, 2$, and $Y = Y(K_1, K_2)$ is obtained by splicing the complements of

Published: 12 January 2016 DOI: 10.2140/agt.2015.15.3155

K_1 and K_2 in Y_1 and Y_2, respectively. A formula is obtained for $\widehat{\mathrm{HF}}(Y;\mathbb{F})$ in terms of the knot Floer objects associated with $K_1 \subset Y_1$ and $K_2 \subset Y_2$, where $\mathbb{F}$ denotes the field $\mathbb{Z}/2\mathbb{Z}$ with two elements.

The more precise statement of the splicing formula obtained in this paper is as follows. Let $K \subset Y$ denote a null-homologous knot inside a three-manifold Y. For every $n \in \mathbb{Z} \cup \{\infty\}$ let $Y_n = Y_n(K)$ denote the three-manifold obtained by performing n–surgery on K and let $K_n \subset Y_n$ denote the knot in Y_n which is the core of the neighbourhood replaced for $\mathrm{nd}(K) \subset Y$ in constructing Y_n. Denote the homology group $\widehat{\mathrm{HFK}}(Y_n, K_n; \mathbb{F})$ by $\mathbb{H}_n(K)$ and its dimension as a vector space over $\mathbb{F}$ by $h_n(K)$. In particular, $\mathbb{H}_\infty(K) = \widehat{\mathrm{HFK}}(Y, K; \mathbb{F})$ and $\mathbb{H}_0(K) = \widehat{\mathrm{HFL}}(Y, K; \mathbb{F})$ are the knot Floer homology and the longitude Floer homology of K, respectively (see Ozsváth and Szabó [11] and Eftekhary [2]).

Choose a Heegaard diagram

$$H = (\Sigma,\ \boldsymbol{\alpha} = \{\alpha_1, \dots, \alpha_g\},\ \widehat{\boldsymbol{\beta}} = \{\beta_1, \dots, \beta_{g-1}\})$$

for the knot complement $Y \setminus K$, and let $\lambda_\bullet$ denote an oriented longitude which has framing coefficient $\bullet \in \mathbb{Z} \cup \{\infty\}$. One can choose the curves $\lambda_\bullet$ (which are disjoint from the curves in $\widehat{\boldsymbol{\beta}}$) so that the pairs $(\lambda_0, \lambda_\infty)$, $(\lambda_1, \lambda_\infty)$ and (λ_0, λ_1) have single intersection points in the Heegaard diagram. For $\bullet \in \{0, 1, \infty\}$ set

$$\boldsymbol{\beta}_\bullet = \{\beta_1^\bullet, \dots, \beta_{g-1}^\bullet, \lambda_\bullet\},$$

where $\beta_i^\bullet$ is an isotopic copy of the curve β_i. The pictures on the left-hand side and the right-hand side of Figure 1 illustrate two possible general arrangements for the curves λ_0, λ_1 and λ_∞. In Figure 1 and other figures in this paper, the surface orientation is chosen opposite from the standard orientation of the page in order to stay compatible with the orientation convention of [12].

The two Heegaard quadruples

$$(\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \boldsymbol{\beta}_\infty; u, v, w) \quad \text{and} \quad (\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \boldsymbol{\beta}_\infty; \bar{u}, \bar{v}, \bar{w})$$

obtained in this way then correspond to the exact triangles

$$\begin{tikzcd} \mathbb{H}_0(K) \arrow[dr, "\mathfrak{f}_\infty(K)"'] & & \mathbb{H}_\infty(K) \arrow[ll, "\mathfrak{f}_1(K)"'] \\ & \mathbb{H}_1(K) \arrow[ur, "\mathfrak{f}_0(K)"'] & \end{tikzcd} \quad \text{and} \quad \begin{tikzcd} \mathbb{H}_0 \arrow[dr, "\bar{\mathfrak{f}}_\infty(K)"'] & & \mathbb{H}_\infty(K) \arrow[ll, "\bar{\mathfrak{f}}_1(K)"'] \\ & \mathbb{H}_1(K) \arrow[ur, "\bar{\mathfrak{f}}_0(K)"'] & \end{tikzcd} \tag{1}$$

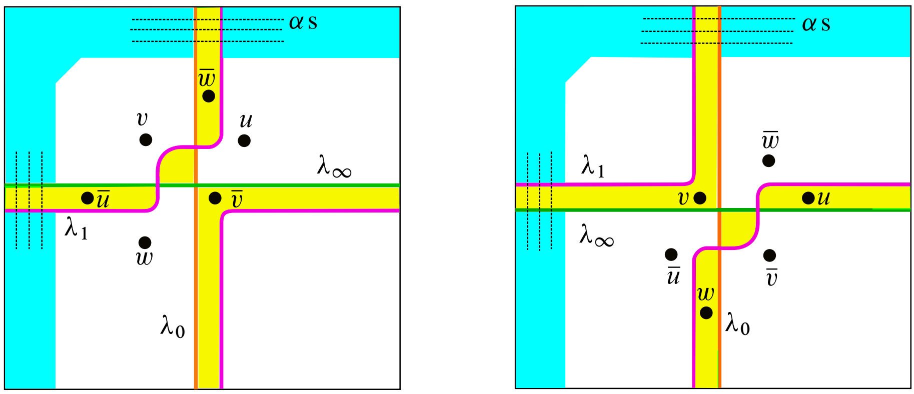


Figure 1: The curves λ_0 (orange), λ_1 (pink) and λ_∞ (green) and the punctures are chosen following one of the above two patterns. The punctures u, v and w are used to define $\mathfrak{f}_0$, $\mathfrak{f}_1$ and $\mathfrak{f}_\infty$, while $\overline{u}$, $\overline{v}$ and $\overline{w}$ are used to define $\bar{\mathfrak{f}}_0$, $\bar{\mathfrak{f}}_1$ and $\bar{\mathfrak{f}}_\infty$.

respectively. The ranks of both $\mathfrak{f}_\bullet(K)$ and $\bar{\mathfrak{f}}_\bullet(K)$ are equal to

$$a_\bullet(K) = \tfrac{1}{2}(h_0(K) + h_1(K) + h_\infty(K) - 2h_\bullet(K)), \quad \bullet \in \{0, 1, \infty\}.$$

The exactness of the above two triangles imply that the induced maps

$$\mathrm{Coker}(\mathfrak{f}_0(K)) \longrightarrow \mathrm{Ker}(\mathfrak{f}_\infty(K)) \quad \text{and} \quad \mathrm{Coker}(\bar{\mathfrak{f}}_0(K)) \longrightarrow \mathrm{Ker}(\bar{\mathfrak{f}}_\infty(K))$$

by $\mathfrak{f}_1(K)$ and $\bar{\mathfrak{f}}_1(K)$ are isomorphisms. Both the domain and the target of the aforementioned isomorphisms are of dimension $a_1(K)$. Take $\theta(K)\colon \mathbb{H}_0(K) \to \mathbb{H}_\infty(K)$ (resp. $\bar{\theta}(K)\colon \mathbb{H}_0(K) \to \mathbb{H}_\infty(K)$) to be an arbitrary extension of the inverse of the isomorphism induced by $\mathfrak{f}_1(K)$ (resp. $\bar{\mathfrak{f}}_1(K)$), so that the ranks of both $\theta(K)$ and $\bar{\theta}(K)$ are equal to $a_1(K)$.

Suppose that a pair of knots K_1 and K_2 is given. For every $\star, \bullet \in \{0, 1, \infty\}$ and $i = 1, 2$, set $\mathbb{H}^i_\bullet = \mathbb{H}_\bullet(K_i)$, $\mathbb{H}_{\star,\bullet} = \mathbb{H}^1_\star \otimes \mathbb{H}^2_\bullet$, $\mathfrak{f}^i_\bullet = \mathfrak{f}_\bullet(K_i)$, $\bar{\mathfrak{f}}^i_\bullet = \bar{\mathfrak{f}}_\bullet(K_i)$, $\theta^i = \theta(K_i)$ and $\bar{\theta}^i = \bar{\theta}(K_i)$. Consider the chain complex $(\square(K_1, K_2), d_\square)$ constructed as follows. The $\mathbb{F}$–module $\square(K_1, K_2)$ is the direct sum of the modules which appear on the vertices of the cube illustrated in Figure 2.

Each directed edge (including the dashed edges) in the aforementioned diagram determines a homomorphism from $\square(K_1, K_2)$ to itself, which is trivial on all summands except for the one which corresponds to its start point. The map takes the summand corresponding to its start point to the summand corresponding to its endpoint by the homomorphism which labels the directed edge. The differential $d_\square$ of the complex $\square(K_1, K_2)$ is defined to be the sum of the homomorphisms which correspond to the

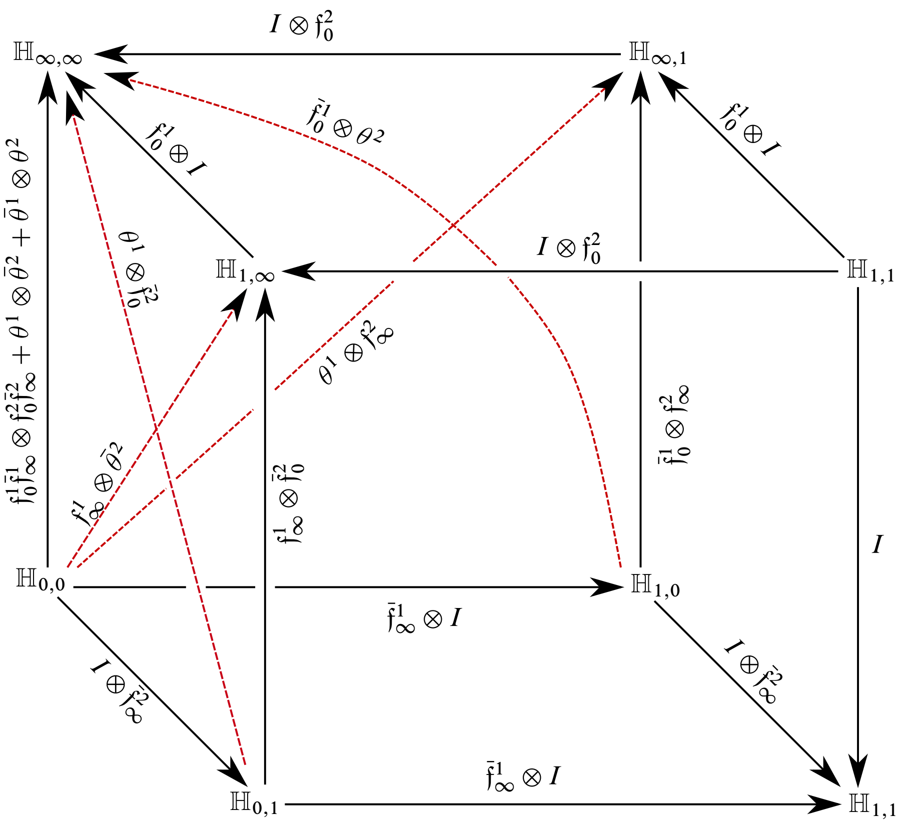


Figure 2: The above cube determines the chain complex $\square(K_1, K_2)$ and its differential $d_{\square}$

directed edges of the cube in Figure 2. One should of course make sure that $d_{\square} \circ d_{\square} = 0$. However, this follows quickly from the exactness of the triangle in (1).

Theorem 1.1 *With the above notation fixed, the Heegaard Floer homology of the three-manifold $Y(K_1, K_2)$ obtained by splicing the knot complements $Y_1 \setminus K_1$ and $Y_2 \setminus K_2$ is given by*

$$\widehat{\mathrm{HF}}(Y(K_1, K_2); \mathbb{F}) \simeq H_*(\square(K_1, K_2), d_{\square}).$$

We use the combinatorial description of Heegaard Floer homology by Sarkar and Wang [16], which is also adapted for knots in S^3 by Manolescu, Ozsváth and Sarkar [8] and Manolescu, Ozsváth, Szabó and Thurston [9]. These combinatorial descriptions help us avoid several technical issues that arise when one glues holomorphic curves.

For the knots $K_1 \subset Y_1$ and $K_2 \subset Y_2$ as above, define

$$\chi(K_1, K_2) := (h^1_\infty - h^1_1)(h^2_\infty - h^2_1) - (h^1_0 - h^1_1)(h^2_0 - h^2_1).$$

As a corollary of Theorem 1.1 we prove the following:

Corollary 1.2 *For $Y = Y(K_1, K_2)$ as above we have*

$$\mathrm{rnk}(\widehat{\mathrm{HF}}(Y; \mathbb{F})) \geq \max\{|\chi(K_1, K_2)|, |\chi(\overline{K_1}, \overline{K_2})|\},$$

where $\overline{K_i} \subset \overline{Y_i} = -Y_i$ denotes the mirror of K_i in the three-manifold $\overline{Y_i} = -Y_i$ for $i = 1, 2$.

When one of the two knots is the trefoil, the formula is simplified significantly. In particular, we prove the following corollary in Section 6:

Corollary 1.3 *Let R denote the right-handed trefoil. With the above notation fixed,*

$$\mathrm{rnk}\big(\widehat{\mathrm{HF}}(Y(R, K))\big) \leq h_0(K) + h_1(K),$$
$$\mathrm{rnk}\big(\widehat{\mathrm{HF}}(Y(R, K))\big) \geq 4\max\{h_0(K), h_1(K), h_\infty(K)\} - (h_0(K) + h_1(K) + 2h_\infty(K)).$$

Moreover, if K is non-trivial, $Y(R, K)$ is not an L–space.

It is shown by Hedden and Levine [5] that splicing non-trivial knots inside homology sphere L–spaces never produces an L–space. Meanwhile, the knot K in Corollary 1.3 lives in an arbitrary homology sphere. In this regard, Corollary 1.3 goes beyond the result of Hedden and Levine.

Remark 1.4 The splicing formula of Theorem 1.1 is different from the splicing formula from the original arXiv version of the paper. The results of a few other papers of the author are based on the splicing formula of this paper. The results of [3] remain unchanged, since the formula (17) presented in Section 5.1 which is used in [3] remains unchanged. The proof of the main theorem of [4] no longer goes through. Fixing the argument requires developing some technology, including a description of the bordered Floer homology for a knot complement only in terms of the knot chain complex associated with the knot. The modifications will appear in an upcoming revision of [4].

Acknowledgements

The author would like to thank Zoltán Szabó for his suggestions. In particular, it was his idea that the difficulties with gluing could disappear if the combinatorial approach of Sarkar and Wang [16] were used. He would also like to thank Matt Hedden and Adam Levine, as well as the anonymous referee for pointing out (several) mistakes in the earlier versions of this paper.

2 Graphs of chain complexes

2.1 Oriented graphs and chain complexes

Let G denote an oriented graph without oriented loops, which consists of a set $V(G)$ of vertices and a set

$$E(G) \subset V(G) \times V(G)$$

of directed edges. For every $e=(v_1,v_2)\in E(G)$ we let $v_s(e)=v_2$ and $v_t(e)=v_1$. The edge e is thus oriented from its starting vertex $v_s(e)$ towards its terminal vertex $v_t(e)$. The condition that G does not contain any oriented loops implies that there is no sequence $e_1,\dots,e_k \in E(G)$ with the property

$$v_t(e_i)=v_s(e_{i+1}), \quad i=1,\dots,k-1, \quad \text{and} \quad v_t(e_k)=v_s(e_1).$$

Definition 2.1 Let G denote an oriented graph without any oriented loops, as above. A collection $\{(C_v,d_v)\}_{v\in V(G)}$ of chain complexes, together with the chain maps

$$\{f_e\colon C_{v_s(e)} \to C_{v_t(e)} \mid e\in E(G)\}$$

is called a *graph of complexes* if, for every v_1, $v_2 \in V(G)$,

$$\sum_{\substack{e_1,e_2\in E(G)\\ v_s(e_1)=v_1,\, v_t(e_2)=v_2\\ v_t(e_1)=v_s(e_2)}} f_{e_2}\circ f_{e_1}=0. \tag{2}$$

Associated with a graph of complexes as above, write $C_G=\bigoplus_{v\in V(G)} C_v$ and define the differential $d_G\colon C_G\to C_G$ as follows. For $c\in C_v\subset C_G$, let

$$d_G(c)=\sum_{w\in V(G)} d_{G,w}(c),$$

where $d_{G,w}(c)\in C_w$ is defined by

$$d_{G,w}(c)=\begin{cases} d_v(c) & \text{if } w=v,\\ f_e(c) & \text{if there exists } e\in E(G) \text{ with } v_s(e)=v \text{ and } v_t(e)=w,\\ 0 & \text{otherwise.}\end{cases}$$

Definition 2.2 The chain complex (C_G,d_G) is called the chain complex *associated with* the graph G of chain complexes.

The condition (2) implies that $d_G\circ d_G=0$, ie that (C_G,d_G) is a chain complex, since each f_e is a chain map. The chain complex (C_G,d_G) is usually represented by drawing the oriented graph G, labelling each vertex $v \in V(G)$ by the chain complex (C_v, d_v) (or simply by C_v if there is no confusion) and labelling each oriented edge e by the chain map f_e.

Let G denote an oriented graph without any loops. It is then possible to label the vertices of G by $1, 2, \dots, n$ so that for each $e \in E(G)$ we have $v_s(e) < v_t(e)$ (as numbers in $\{1, \dots, n\}$). Correspondingly, the chain complexes associated with the vertices of G may be labelled $(C_1, d_1), \dots, (C_n, d_n)$. Let H denote the graph with vertices $1, \dots, n$ and edges

$$E(H) = \{(i, j) \mid i, j \in \{1, \dots, n\} \text{ and } i > j\}.$$

For $e \in E(H)$ let $g_e = f_e$ if $e \in E(G)$ and $g_e = 0$ otherwise. Associated with $\{C_i\}_{i \in V(H)}$ and $\{g_e\}_{e \in E(H)}$ we thus find the complex (C_H, d_H), which is identified with (C_G, d_G). In other words, we may always assume that the underlying graph in a graph of complexes is the complete oriented graph H. The condition (2) in this case is equivalent to

$$\sum_{i>k>j} g_{(i,k)} \circ g_{(k,j)} = 0 \quad \text{for all } i, j \in \{1, \dots, n\}.$$

2.2 Replacing chain complexes with their homology

When the ring of coefficients is $\mathbb{F} = \mathbb{Z}/2\mathbb{Z}$ we would like to replace each complex (C_i, d_i) in (C_H, d_H) with $(H_*(C_i, d_i), 0)$, at the expense of modifying the chain maps $\{g_e\}_{e \in E(H)}$ so that the homology of the chain complex associated with the graph of chain complexes remains intact. Let us begin with a lemma.

Lemma 2.3 *Suppose that a chain complex (C, d_C) is decomposed, as a vector space over $\mathbb{F}$, as $C \simeq A \oplus A \oplus B$ for some vector spaces A and B. Suppose that the differential d_C of C has the following block form in this decomposition:*

$$d_C = \begin{pmatrix} 0 & I_A & f_1 \\ 0 & 0 & f_2 \\ g_1 & g_2 & h \end{pmatrix}$$

Then $d_B = h + g_2 f_1 \colon B \to B$ is a differential and $H_*(C, d_C) = H_*(B, d_B)$.

Proof Since d_C is a differential, $f_1 g_2 = 0$ and the matrix

$$P = \begin{pmatrix} I & 0 & 0 \\ 0 & I & f_1 \\ g_2 & 0 & I \end{pmatrix}$$

is thus its own inverse. Since $d_C^2=0$ we get

$$P\begin{pmatrix}0 & I_A & f_1\\ 0 & 0 & f_2\\ g_1 & g_2 & h\end{pmatrix}P=\begin{pmatrix}0 & I_A & 0\\ 0 & 0 & 0\\ 0 & 0 & h+g_2f_1\end{pmatrix}.$$

This completes the proof of the lemma. □

We refer to the procedure which changes the chain complex (C,d_C) to the chain complex (B,d_B) as the *cancellation* of the two subspaces $A\oplus 0\oplus 0\simeq A$ and $0\oplus A\oplus 0\simeq A$ of C against each other.

The differential d_i of C_i may be used to decompose C_i as $A_i^1\oplus H_i\oplus A_i^2$, where A_i^1 and A_i^2 are two copies of the same $\mathbb{F}$–module A_i, so that d_i takes the form

$$d_i=\begin{pmatrix}0 & 0 & I_{A_i}\\ 0 & 0 & 0\\ 0 & 0 & 0\end{pmatrix},\quad i=1,\dots,n.$$

Note that $H_i=H_*(C_i,d_i)$ is in fact the homology of the complex C_i. In particular, $H_i\subset \mathrm{Ker}(d_i\colon C_i\to C_i)$. Since $d_{v_t(e)}\circ g_e=g_e\circ d_{v_s(e)}$, in this basis the matrix block presentation of g_e is of the form

$$g_e=\begin{pmatrix}M_e & P_e & N_e\\ 0 & G_e & Q_e\\ 0 & 0 & M_e\end{pmatrix}\quad \text{for all } e\in E(H).$$

Initially, the block presentation for d_H is of the form

$$d_H=\begin{pmatrix}d_1 & 0 & 0 & \dots & 0\\ g_{(2,1)} & d_2 & 0 & \dots & 0\\ g_{(3,1)} & g_{(3,2)} & d_3 & \dots & 0\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ g_{(n,1)} & g_{(n,2)} & g_{(n,3)} & \dots & d_n\end{pmatrix}.$$

Replacing the above 3×3 block presentations for $g_{(i,j)}$ and d_i, the homomorphism d_H takes a $3n\times 3n$ block presentation, where n of the block entries are the identity matrices corresponding to $d_1,\dots,d_n$. Lemma 2.3 may be used inductively to cancel A_i^1 against A_i^2 for $i=1,\dots,n$ and modify the remaining blocks correspondingly. Straightforward linear algebra implies the following lemma:

Lemma 2.4 *Fix the above notation and for $i,j\in\{1,\dots,n\}$ let*

$$h_{(i,j)}=G_{(i,j)}+\sum_{\ell\geq 1}\sum_{i>k_1>k_2>\dots>k_\ell>j}Q_{(i,k_1)}N_{(k_1,k_2)}\cdots N_{(k_{\ell-1},k_\ell)}P_{(k_\ell,j)}.$$

Then the homology of the chain complex associated with H that has complexes $\{(C_i,d_i)\}_{i=1}^n$ and chain maps $\{g_e\}_{e\in E(H)}$ is isomorphic to the homology of the chain complex associated with H that has complexes $\{(H_i,0)\}_{i=1}^n$ and homomorphisms $\{h_e\}_{e\in E(H)}$.

For $\ell=1$ set $k=k_1$. For $h_j\in \mathrm{Ker}(G_{(k,j)}\colon H_j\to H_k)$, $P_{(k,j)}(h_j)=d_k(a_k)$ for some $a_k\in A_k$. The element a_k may of course be modified by adding to a_k an element $h_k\in H_k$. From here, $Q_{(i,k)}P_{(k,j)}(h_j)$ is equal to $g_{(i,k)}(a_k)$ up to the addition of an element in $g_{(i,k)}(H_k)$. In particular, we find a natural well-defined map

$$\theta_{(i>k>j)}\colon \mathrm{Ker}(G_{(k,j)})\longrightarrow \mathrm{Coker}(G_{(i,k)})$$

and $Q_{(i,k)}P_{(k,j)}$ is an extension of $\theta_{(i>k>j)}$ to a homomorphism from H_j to H_k. It is however important to note that simultaneous replacement of the maps $Q_{(i,k)}P_{(k,j)}$ with arbitrary extensions of $\theta(i>k>j)$ in Lemma 2.3 is not a priori possible.

In this paper, we will face situations where each complex C_i is of the form $C_i^1\otimes C_i^2$ and each chain map $g_{(i,j)}\colon C_j\to C_i$ is of the form $g^1_{(i,j)}\otimes g^2_{(i,j)}$, where $g^1_{(i,j)}\colon C_j^1\to C_i^1$ and $g^2_{(i,j)}\colon C_j^2\to C_i^2$ are chain maps. In this situation, we may choose the decompositions $C_i^r=A_i^r\oplus H_i^r\oplus A_i^r$ for $r=1,2$. Subsequently, note that

$$H_i=H_i^1\otimes H_i^2 \quad\text{and}\quad A_i=(A_i^1\otimes A_i^2)\oplus(A_i^1\otimes H_i^2)\oplus(H_i^1\otimes A_i^2)\oplus(A_i^1\otimes A_i^2).$$

Moreover, corresponding to $g^r_{(i,j)}$ we obtain the blocks $M^r_{(i,j)}$, $N^r_{(i,j)}$, $P^r_{(i,j)}$, $Q^r_{(i,j)}$ and $G^r_{(i,j)}$ for $r=1,2$.

We close this section with a pair of simple lemmas addressing this situation.

Lemma 2.5 *In the situation above,*

$$\begin{aligned}Q_{(i,k)}P_{(k,j)}=Q^1_{(i,k)}P^1_{(k,j)}\otimes Q^2_{(i,k)}P^2_{(k,j)}+Q^1_{(i,k)}P^1_{(k,j)}\otimes G^2_{(i,k)}G^2_{(k,j)}&\\+G^1_{(i,k)}G^1_{(k,j)}\otimes Q^2_{(i,k)}P^2_{(k,j)}&.\end{aligned}$$

Proof Choose $h_j\in H_j=H_j^1\otimes H_j^2$. The image of h_j under $g_{(k,j)}$ is in

$$(A_k^1\oplus H_k^1\oplus 0)\otimes(A_k^2\oplus H_k^2\oplus 0)\subset C_k.$$

In particular, we find

$$P_{(k,j)}(h_j)\in(A_k^1\otimes A_k^2)\oplus(A_k^1\otimes H_k^2)\oplus(H_k^1\otimes A_k^2)\oplus 0\subset A_k.$$

For $h_j=h_j^1\otimes h_j^2$ the corresponding decomposition of $P_{(k,j)}(h_j)$ is of the form

$$\big(P^1_{(k,j)}(h_j^1)\otimes P^2_{(k,j)}(h_j^2),\,P^1_{(k,j)}(h_j^1)\otimes G^2_{(k,j)}(h_j^2),\,G^1_{(k,j)}(h_j^1)\otimes P^2_{(k,j)}(h_j^2),\,0\big).$$

Note that $Q_{(i,k)}P_{(k,j)}(h_j)$ is the $H_j = H_j^1 \otimes H_j^2$–component of $g_{(i,k)}P_{(k,j)}(h_j)$. The components in the above presentation are thus mapped by $Q_{(i,k)}P_{(k,j)}$ to

$$Q^1_{(i,k)}P^1_{(k,j)}(h^1_j) \otimes Q^2_{(i,k)}P^2_{(k,j)}(h^2_j), \quad Q^1_{(i,k)}P^1_{(k,j)}(h^1_j) \otimes G^2_{(i,k)}G^2_{(k,j)}(h^2_j),$$
$$G^1_{(i,k)}G^1_{(k,j)}(h^1_j) \otimes Q^2_{(i,k)}P^2_{(k,j)}(h^2_j) \quad \text{and} \quad 0,$$

respectively. This completes the proof of the lemma. □

An interesting particular case of the above lemma is when one of the chain maps $g^1_{(i,k)}$ or $g^1_{(k,j)}$ is the identity, where we find

$$g^1_{(i,k)} = \mathrm{Id} \implies Q_{(i,k)}P_{(k,j)}G^1_{(k,j)} \otimes Q^2_{(i,k)}P^2_{(k,j)},$$
$$g^1_{(k,j)} = \mathrm{Id} \implies Q_{(i,k)}P_{(k,j)} = G^1_{(i,k)} \otimes Q^2_{(i,k)}P^2_{(k,j)},$$

respectively.

Lemma 2.6 *With the above notation fixed, if $i > k > l > j$ and $g^1_{(i,k)}$ and $g^2_{(l,j)}$ are both the identity map, we find*

$$Q_{(i,k)}N_{(k,l)}P_{(l,i)} = (Q^1_{(k,l)}P^1_{(l,j)}) \otimes (Q^2_{(i,k)}P^2_{(k,l)}).$$

Proof Following the proof of Lemma 2.5, for $h_j = h^1_j \otimes h^2_j \in H^1_j \otimes H^2_j$ one finds

$$P_{(l,j)}(h_j) = (0, P_{(l,j)}(h^1_j) \otimes h^2_j, 0) \in A_l.$$

The image of this element of A_l under $g^1_{(k,l)} \otimes g^2_{(k,l)}$ is precisely

$$g^1_{(k,l)}P^1_{(l,j)}(h^1_j) \otimes g^2_{(k,l)}(h^2_j) \subset C^1_j \otimes (H^2_j \oplus A^2_j).$$

On the other hand, the domain of $C_{(i,k)}$ (where it is non-zero) is the subset

$$0 \oplus 0 \oplus (H^1_k \otimes A^2_k) \oplus 0 \subset A_k.$$

In other words, only the component of $g^1_{(k,l)}P^1_{(l,j)}(h^1_j) \otimes g^2_{(k,l)}(h^2_j)$ which lands in $H^2_k \otimes A^2_k$ survives under the map $Q_{(i,k)}$. The aforementioned component is precisely $Q^1_{(k,l)}P^1_{(l,j)}(h^1_j) \otimes P^2_{(k,l)}(h^2_j)$ and the image of this element under $Q_{(i,k)}$ is precisely $Q^1_{(k,l)}P^1_{(l,j)}(h^1_j) \otimes Q^2_{(i,k)}P^2_{(k,l)}(h^2_j)$. This completes the proof of the lemma. □

3 A pair of exact triangles

3.1 The chain maps

Let $K \subset Y$ denote a null-homologous knot and fix a Heegaard diagram

$$\widehat{H} = (\Sigma,\ \boldsymbol{\alpha} = \{\alpha_1, \dots, \alpha_g\},\ \widehat{\boldsymbol{\beta}} = \{\beta_1, \dots, \beta_{g-1}\})$$

for the knot complement $Y \setminus K$. Set $\boldsymbol{\beta}_\bullet = \{\beta_1^\bullet, \dots, \beta_{g-1}^\bullet, \lambda_\bullet\}$, where $\beta_i^\bullet$ is an isotopic copy of the curve β_i and $\lambda_\bullet$ is chosen so that the Heegaard triple $(\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}_\bullet)$ corresponds to the three-manifold obtained from Y by $\bullet$–surgery on the knot K. Choose the curves λ_0, λ_1 and λ_∞ so that each pair of them has a unique transverse intersection point. The orientation on K induces an orientation on the three curves λ_0, λ_1 and λ_∞.

We assume that the intersection pattern of λ_0, λ_1 and λ_∞ is one of the two patterns illustrated in Figure 1. This gives the Heegaard quadruples

$$H = (\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \boldsymbol{\beta}_\infty; u, v, w) \quad \text{and} \quad \overline{H} = (\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \boldsymbol{\beta}_\infty; \bar{u}, \bar{v}, \bar{w}).$$

Note that there is an identification $\widehat{\mathrm{CF}}(\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}_\bullet; u, v, w) = \widehat{\mathrm{CF}}(\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}_\bullet; \bar{u}, \bar{v}, \bar{w})$ for $\bullet \in \{0, 1, \infty\}$. Moreover, for $\bullet, \star \in \{0, 1, \infty\}$ the complexes $\widehat{\mathrm{CF}}(\Sigma, \boldsymbol{\beta}_\bullet, \boldsymbol{\beta}_\star; u, v, w)$ and $\widehat{\mathrm{CF}}(\Sigma, \boldsymbol{\beta}_\bullet, \boldsymbol{\beta}_\star; \bar{u}, \bar{v}, \bar{w})$ are identical and the corresponding homology group is $\widehat{\mathrm{HF}}(\#^{g-1}(S^1 \times S^2))$. The top generator $\Theta = \Theta_{\bullet,\star}$ in this Heegaard Floer homology group may be used to define two holomorphic triangle maps (see Ozsváth and Szabó [12] for more details on the definition of holomorphic triangle maps).

Definition 3.1 Associated with the Heegaard triples

$$H_\bullet = H \setminus \boldsymbol{\beta}_\bullet \quad \text{and} \quad \overline{H}_\bullet = \overline{H} \setminus \boldsymbol{\beta}_\bullet,$$

define the maps

$$\begin{aligned}
\phi(H_0),\ \phi(\overline{H}_0)&\colon \widehat{\mathrm{CF}}(\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}_1; u, v, w) \longrightarrow \widehat{\mathrm{CF}}(\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}_\infty; u, v, w),\\
\phi(H_1),\ \phi(\overline{H}_1)&\colon \widehat{\mathrm{CF}}(\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}_\infty; u, v, w) \longrightarrow \widehat{\mathrm{CF}}(\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}_0; u, v, w),\\
\phi(H_\infty),\ \phi(\overline{H}_\infty)&\colon \widehat{\mathrm{CF}}(\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}_0; u, v, w) \longrightarrow \widehat{\mathrm{CF}}(\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}_1; u, v, w)
\end{aligned}$$

to be the holomorphic triangle maps corresponding to the triply punctured Heegaard triples H_0, $\overline{H}_0$, H_1, $\overline{H}_1$, H_∞ and $\overline{H}_\infty$, respectively, defined using the top generators $\Theta_{\bullet,\star}$. Denote the induced maps in homology by $\phi_*(H_\bullet)$ and $\phi_*(\overline{H}_\bullet)$ and set

$$\mathfrak{f}_\bullet(K) := \phi_*(H_\bullet) \quad \text{and} \quad \bar{\mathfrak{f}}_\bullet(K) := \phi_*(\overline{H}_\bullet) \quad \text{for } \bullet \in \{0, 1, \infty\}.$$

3.2 Behaviour under Heegaard moves

The group $\widehat{\mathrm{HF}}(\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}_\bullet; u, v, w)$, denoted by $\mathbb{H}_\bullet(K)$, is independent of the particular Heegaard diagram used for the definition. We have thus defined the maps

$$\mathfrak{f}_0(K), \bar{\mathfrak{f}}_0(K)\colon \mathbb{H}_1(K) \to \mathbb{H}_\infty(K) \quad \text{and} \quad \mathfrak{f}_\infty(K), \bar{\mathfrak{f}}_\infty(K)\colon \mathbb{H}_0(K) \to \mathbb{H}_1(K).$$

The definition of the map $\mathfrak{f}_0(K)$ depends on a Heegaard triple $(\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_\infty; u, v, w)$ associated with the knot K. Changing H to another Heegaard triple changes $\mathbb{H}_1(K)$ and $\mathbb{H}_\infty(K)$ by an isomorphism which is determined by the corresponding Heegaard moves that change one Heegaard diagram to the other. We would now like to show that the corresponding change in the triangle maps $\mathfrak{f}_0(H)$ and $\bar{\mathfrak{f}}_0(H)$ respects the above isomorphisms. This justifies using the names $\mathfrak{f}_0(K)$ and $\bar{\mathfrak{f}}_0(K)$ for the above two homomorphisms. The same statement would be true for $\mathfrak{f}_\bullet(K)$ and $\bar{\mathfrak{f}}_\bullet(K)$.

Let $\{*\} = \{0, 1, \infty\} \setminus \{\bullet, \star\}$. Suppose that two marked Heegaard triples

$$H_* = (\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}_\bullet, \boldsymbol{\beta}_\star, u, v, w) \quad \text{and} \quad H'_* = (\Sigma', \boldsymbol{\alpha}', \boldsymbol{\beta}'_\bullet, \boldsymbol{\beta}'_\star, u', v', w')$$

correspond to the same knot $K \subset Y$ for a pair $(\bullet, \star) \in \{(\infty, 1), (1, 0)\}$. Similarly, one may consider the Heegaard diagrams $\overline{H}_*$ and $\overline{H}'_*$. Suppose furthermore that the maps

$$\imath_\bullet\colon \widehat{\mathrm{HF}}(\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}_\bullet; u, v, w) \longrightarrow \widehat{\mathrm{HF}}(\Sigma', \boldsymbol{\alpha}', \boldsymbol{\beta}'_\bullet; u', v', w'),$$
$$\imath_\star\colon \widehat{\mathrm{HF}}(\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}_\star; u, v, w) \longrightarrow \widehat{\mathrm{HF}}(\Sigma', \boldsymbol{\alpha}', \boldsymbol{\beta}'_\star; u', v', w')$$

are the isomorphisms of the corresponding Heegaard Floer homology groups associated with the Heegaard moves (and the change of almost complex structure) changing one Heegaard diagram to the other.

Theorem 3.2 *With the above notation fixed,*

$$\mathfrak{f}_*(H_*) \circ \imath_\bullet = \imath_\star \circ \mathfrak{f}_*(H'_*) \quad \textit{and} \quad \bar{\mathfrak{f}}_*(\overline{H}_*) \circ \imath_\bullet = \imath_\star \circ \bar{\mathfrak{f}}_*(\overline{H}'_*).$$

Proof The proof consists of some standard steps in Heegaard Floer theory, which are sketched below for the Heegaard moves.

Note that the first Heegaard triple may be changed to the second Heegaard triple by a sequence of Heegaard moves, supported in the complement of the marked points, of the following types:

- Changing the almost complex structure on the surface Σ.
- Isotopies of the curves in $\boldsymbol{\alpha}$ which are supported away from a neighbourhood U of $\lambda_\bullet \cap \lambda_\star$ containing the marked points u, v and w, so that the curves in each collection remain disjoint.

- Handle slides among the curves in $\boldsymbol{\alpha}$ supported away from U.
- Simultaneous handle slides among $\boldsymbol{\beta}_\bullet \setminus \{\lambda_\bullet\}$ and $\boldsymbol{\beta}_\star \setminus \{\lambda_\star\}$ supported away from U.
- Stabilization and destabilization of the Heegaard triple away from U.

The independence of the induced map in homology from the choice of the path of almost complex structures follows the corresponding argument of Ozsváth and Stipsicz [10]. Corresponding to each one of the above Heegaard moves, we obtain a holomorphic square map in the level of chain complexes, comprising of a chain homotopy map between the compositions of the chain maps we are interested in. More precisely, performing an isotopy or a handle slide in $\boldsymbol{\alpha}$ would result in a new set of simple closed curves, which may be denoted by $\boldsymbol{\alpha}'$, by slight abuse of notation. The punctured Heegaard 4–tuple

$$(\Sigma, \boldsymbol{\alpha}, \boldsymbol{\alpha}', \boldsymbol{\beta}_\bullet, \boldsymbol{\beta}_\star; u, v, w)$$

determines a homomorphism

$$\begin{aligned}\widehat{\Phi}\colon \widehat{\mathrm{CF}}(\Sigma, \boldsymbol{\alpha}, \boldsymbol{\alpha}'; u, v, w) \otimes \widehat{\mathrm{CF}}(\Sigma, \boldsymbol{\alpha}', \boldsymbol{\beta}_\bullet; u, v, w) \otimes \widehat{\mathrm{CF}}(\Sigma, \boldsymbol{\beta}_\bullet, \boldsymbol{\beta}_\star; u, v, w)\\ \longrightarrow \widehat{\mathrm{CF}}(\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}_\star; u, v, w),\end{aligned}$$

which is defined by counting holomorphic squares with Maslov index -1. Using the top closed elements in the complexes $\widehat{\mathrm{CF}}(\Sigma, \boldsymbol{\alpha}, \boldsymbol{\alpha}'; u, v, w)$ and $\widehat{\mathrm{CF}}(\Sigma, \boldsymbol{\beta}_\bullet, \boldsymbol{\beta}_\star; u, v, w)$, we obtain a corresponding map

$$\Phi\colon \widehat{\mathrm{CF}}(\Sigma, \boldsymbol{\alpha}', \boldsymbol{\beta}_\bullet; u, v, w) \longrightarrow \widehat{\mathrm{CF}}(\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}_\star; u, v, w).$$

Let us denote the differentials of the chain complexes

$$\widehat{\mathrm{CF}}(\Sigma, \boldsymbol{\alpha}', \boldsymbol{\beta}_\bullet; u, v, w) \quad \text{and} \quad \widehat{\mathrm{CF}}(\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}_\star; u, v, w)$$

by $d_{\alpha',\beta_\bullet}$ and $d_{\alpha,\beta_\star}$, respectively. The Heegaard triples $(\Sigma, \boldsymbol{\alpha}, \boldsymbol{\alpha}', \boldsymbol{\beta}_\bullet)$, $(\Sigma, \boldsymbol{\alpha}, \boldsymbol{\alpha}', \boldsymbol{\beta}_\star)$ determine chain equivalences

$$\begin{aligned}\imath(\boldsymbol{\alpha}, \boldsymbol{\alpha}', \boldsymbol{\beta}_\bullet)&\colon \widehat{\mathrm{CF}}(\Sigma, \boldsymbol{\alpha}', \boldsymbol{\beta}_\bullet; u, v, w) \longrightarrow \widehat{\mathrm{CF}}(\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}_\bullet; u, v, w),\\ \imath(\boldsymbol{\alpha}, \boldsymbol{\alpha}', \boldsymbol{\beta}_\star)&\colon \widehat{\mathrm{CF}}(\Sigma, \boldsymbol{\alpha}', \boldsymbol{\beta}_\star; u, v, w) \longrightarrow \widehat{\mathrm{CF}}(\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}_\star; u, v, w).\end{aligned}$$

Moreover, we obtain holomorphic triangle maps associated with the Heegaard triples $(\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}_\bullet, \boldsymbol{\beta}_\star)$ and $(\Sigma, \boldsymbol{\alpha}', \boldsymbol{\beta}_\bullet, \boldsymbol{\beta}_\star)$, which are denoted by

$$\begin{aligned}\phi(\boldsymbol{\alpha}, \boldsymbol{\beta}_\bullet, \boldsymbol{\beta}_\star)&\colon \widehat{\mathrm{CF}}(\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}_\bullet; u, v, w) \longrightarrow \widehat{\mathrm{CF}}(\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}_\star; u, v, w),\\ \phi(\boldsymbol{\alpha}', \boldsymbol{\beta}_\bullet, \boldsymbol{\beta}_\star)&\colon \widehat{\mathrm{CF}}(\Sigma, \boldsymbol{\alpha}', \boldsymbol{\beta}_\bullet; u, v, w) \longrightarrow \widehat{\mathrm{CF}}(\Sigma, \boldsymbol{\alpha}', \boldsymbol{\beta}_\star; u, v, w).\end{aligned}$$

Considering different types of degenerations for a square of Maslov index 0, we obtain the relation

$$d_{\alpha\beta_\star}\circ\Phi+\Phi\circ d_{\alpha'\beta_\bullet}=\imath(\boldsymbol{\alpha},\boldsymbol{\alpha}',\boldsymbol{\beta}_\star)\circ\phi(\boldsymbol{\alpha},\boldsymbol{\beta}_\bullet,\boldsymbol{\beta}_\star)+\phi(\boldsymbol{\alpha}',\boldsymbol{\beta}_\bullet,\boldsymbol{\beta}_\star)\circ\imath(\boldsymbol{\alpha},\boldsymbol{\alpha}',\boldsymbol{\beta}_\bullet).$$

The induced relation in homology gives the claim for the invariance of $\phi_*(H_*)$ under handle slides in α. The corresponding argument for $\phi_*(\overline{H}_*)$ is done by changing the marked points.

The invariance under handle slides among the β–curves is proved similarly, and we only highlight the important modifications. Let $\boldsymbol{\beta}'_\bullet$ and $\boldsymbol{\beta}'_\star$ be obtained from $\boldsymbol{\beta}_\bullet$ and $\boldsymbol{\beta}_\star$ by handle slides which correspond to a handle slide in $\widehat{\boldsymbol{\beta}}$. We thus have the following square of chain maps:

$$\begin{tikzcd}
\widehat{\mathrm{CF}}(\Sigma,\boldsymbol{\alpha},\boldsymbol{\beta}_\bullet;u,v,w) \arrow[r, "\phi(\boldsymbol{\alpha},\boldsymbol{\beta}_\bullet,\boldsymbol{\beta}_\star)"] \arrow[d, "\imath_\bullet(\boldsymbol{\alpha},\boldsymbol{\beta}_\bullet,\boldsymbol{\beta}'_\bullet)"'] \arrow[dr, "\phi(\alpha,\boldsymbol{\beta}_\bullet,\boldsymbol{\beta}'_\star)"] & \widehat{\mathrm{CF}}(\Sigma,\boldsymbol{\alpha},\boldsymbol{\beta}_\star;u,v,w) \arrow[d, "\imath_\bullet(\alpha,\boldsymbol{\beta}_\star,\boldsymbol{\beta}'_\star)"] \\
\widehat{\mathrm{CF}}(\Sigma,\boldsymbol{\alpha},\boldsymbol{\beta}'_\bullet;u,v,w) \arrow[r, "\phi(\boldsymbol{\alpha},\boldsymbol{\beta}'_\bullet,\boldsymbol{\beta}'_\star)"] & \widehat{\mathrm{CF}}(\Sigma,\boldsymbol{\alpha},\boldsymbol{\beta}'_\star;u,v,w),
\end{tikzcd}$$

while the quadruples $(\Sigma,\boldsymbol{\alpha},\boldsymbol{\beta}_\bullet,\boldsymbol{\beta}'_\bullet,\boldsymbol{\beta}'_\star;u,v,w)$ and $(\Sigma,\boldsymbol{\alpha},\boldsymbol{\beta}_\bullet,\boldsymbol{\beta}_\star,\boldsymbol{\beta}'_\star;u,v,w)$ determine a pair of holomorphic square maps

$$\Phi_1,\ \Phi_2\colon \widehat{\mathrm{CF}}(\alpha,\boldsymbol{\beta}_\bullet;u,v,w)\longrightarrow\widehat{\mathrm{CF}}(\boldsymbol{\alpha},\boldsymbol{\beta}'_\star;u,v,w).$$

Considering different possible degenerations of holomorphic squares of Maslov index 0 gives the relations

$$\begin{aligned}
d_{\alpha\beta'_\star}\circ\Phi_1+\Phi_1\circ d_{\alpha\beta_\bullet}&=\phi(\boldsymbol{\alpha},\boldsymbol{\beta}'_\bullet,\boldsymbol{\beta}'_\star)\circ\imath(\boldsymbol{\alpha},\boldsymbol{\beta}_\bullet,\boldsymbol{\beta}'_\bullet)+\phi(\boldsymbol{\alpha},\boldsymbol{\beta}_\bullet,\boldsymbol{\beta}'_\star),\\
d_{\alpha\beta'_\star}\circ\Phi_2+\Phi_2\circ d_{\alpha\beta_\bullet}&=\imath(\boldsymbol{\alpha},\boldsymbol{\beta}_\star,\boldsymbol{\beta}'_\star)\circ\phi(\boldsymbol{\alpha},\boldsymbol{\beta}_\bullet,\boldsymbol{\beta}_\star)+\phi(\boldsymbol{\alpha},\boldsymbol{\beta}_\bullet,\boldsymbol{\beta}'_\star).
\end{aligned}$$

If we set $\Phi=\Phi_1+\Phi_2$ we thus find

$$d_{\alpha\beta'_\star}\circ\Phi+\Phi\circ d_{\alpha\beta_\bullet}=\phi(\boldsymbol{\alpha},\boldsymbol{\beta}'_\bullet,\boldsymbol{\beta}'_\star)\circ\imath(\boldsymbol{\alpha},\boldsymbol{\beta}_\bullet,\boldsymbol{\beta}'_\bullet)+\imath(\boldsymbol{\alpha},\boldsymbol{\beta}_\star,\boldsymbol{\beta}'_\star)\circ\phi(\boldsymbol{\alpha},\boldsymbol{\beta}_\bullet,\boldsymbol{\beta}_\star),$$

which completes the proof of the invariance under handle slides of the β–curves for $\phi_*(H_*)$. The argument for $\phi(\overline{H}_*)$ is completely similar.

The proof of the invariance under stabilization and destabilization follows the general argument of [10] as well. □

Remark 3.3 This theorem should be compared with the naturality theorem of Ozsváth and Stipsicz [10].

Lemma 3.4 *With the above notation fixed, the triangles*

$$
\begin{tikzcd}
\mathbb{H}_0(K) \arrow[dr, "\mathfrak{f}_0(K)"'] & & \mathbb{H}_\infty(K) \arrow[ll, "\mathfrak{f}_1(K)"'] \\
& \mathbb{H}_1(K) \arrow[ur, "\mathfrak{f}_\infty(K)"'] &
\end{tikzcd}
\quad \text{and} \quad
\begin{tikzcd}
\mathbb{H}_0 \arrow[dr, "\bar{\mathfrak{f}}_0(K)"'] & & \mathbb{H}_\infty(K) \arrow[ll, "\bar{\mathfrak{f}}_1(K)"'] \\
& \mathbb{H}_1(K) \arrow[ur, "\bar{\mathfrak{f}}_\infty(K)"'] &
\end{tikzcd}
\tag{3}
$$

are both exact.

Proof The more general forms of exact triangles associated with pointed Heegaard diagrams are discussed by Alishahi and Eftekhary [1, Section 9], using a generalization of Lemma 4.4 of Ozsváth and Szabó [13]. The arguments are rather standard and are omitted from the paper. The only remark is that if the intersection pattern of λ_0, λ_1 and λ_∞ follows the left-hand side of Figure 1, the contributing holomorphic triangles for $(\Sigma, \boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \boldsymbol{\beta}_\infty; u, v, w)$ come in cancelling pairs, allowing us to follow the standard arguments. For the Heegaard triple $(\Sigma, \boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \boldsymbol{\beta}_\infty; \bar{u}, \bar{v}, \bar{w})$, however, there is a unique contributing triangle class, which corresponds to the small triangle bounded between the three curves, which implies that the corresponding triangle map takes $\Theta_{0,1} \otimes \Theta_{1,\infty}$ to $\Theta_{0,\infty}$. Nevertheless, the position of the punctures in this case implies that the map $\phi(\boldsymbol{\alpha}, \boldsymbol{\beta}_0, \boldsymbol{\beta}_\infty; \bar{u}, \bar{v}, \bar{w})$ that is defined using $\Theta_{0,1}$ is trivial (unlike $\phi(\boldsymbol{\alpha}, \boldsymbol{\beta}_\infty, \boldsymbol{\beta}_0; \bar{u}, \bar{v}, \bar{w})$). From here, the rest of the argument is standard. □

By exactness of the triangles in (3), $\mathrm{Ker}(\mathfrak{f}_\infty(K))$ is isomorphic to $\mathrm{Coker}(\mathfrak{f}_0(K))$ while $\mathrm{Ker}(\bar{\mathfrak{f}}_\infty(K))$ is isomorphic to $\mathrm{Coker}(\bar{\mathfrak{f}}_0(K))$. Furthermore, the first isomorphism is induced by the natural chain map $\mathfrak{f}_1(K)$ while the second isomorphism is induced by $\bar{\mathfrak{f}}_1(K)$. Let $\theta(K)\colon \mathbb{H}_0(K) \to \mathbb{H}_\infty(K)$ denote a map which has the same rank as $\mathfrak{f}_1(K)$ and induces the inverse of the isomorphism

$$\mathfrak{f}_1(K)\colon \mathrm{Ker}(\mathfrak{f}_\infty(K)) \longrightarrow \mathrm{Coker}(\mathfrak{f}_1(K)),$$

while $\bar{\theta}(K)\colon \mathbb{H}_0(K) \to \mathbb{H}_\infty(K)$ denotes a map which has the same rank as $\bar{\mathfrak{f}}_1(K)$ and induces the inverse of the isomorphism

$$\bar{\mathfrak{f}}_1(K)\colon \mathrm{Ker}(\mathfrak{f}_\infty(K)) \longrightarrow \mathrm{Coker}(\mathfrak{f}_1(K)).$$

The choice of the maps $\theta(K)$ and $\bar{\theta}(K)$ are of course not unique. If

$$\phi_\infty = \phi(\boldsymbol{\alpha}, \boldsymbol{\beta}_0, \boldsymbol{\beta}_1; u, v, w) \quad \text{and} \quad \phi_0 = \phi(\boldsymbol{\alpha}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_\infty; u, v, w)$$

denote the triangle maps associated with the punctured Heegaard triples

$$(\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}_0, \boldsymbol{\beta}_1; u, v, w) \quad \text{and} \quad (\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_\infty; u, v, w),$$

as above, the map $\theta(K)$ is in fact the correction term, in the sense of Lemma 2.4, associated with the sequence (or, in fact, graph of complexes)

$$\widehat{\mathrm{CF}}(\Sigma,\boldsymbol{\alpha},\boldsymbol{\beta}_0;u,v,w)\xrightarrow{\phi_\infty}\widehat{\mathrm{CF}}(\Sigma,\boldsymbol{\alpha},\boldsymbol{\beta}_1;u,v,w)\xrightarrow{\phi_0}\widehat{\mathrm{CF}}(\Sigma,\boldsymbol{\alpha},\boldsymbol{\beta}_\infty;u,v,w).$$

Similarly, $\bar{\theta}(K)$ corresponds to the sequence

$$\widehat{\mathrm{CF}}(\Sigma,\boldsymbol{\alpha},\boldsymbol{\beta}_0;\bar{u},\bar{v},\bar{w})\xrightarrow{\bar{\phi}_\infty}\widehat{\mathrm{CF}}(\Sigma,\boldsymbol{\alpha},\boldsymbol{\beta}_1;\bar{u},\bar{v},\bar{w})\xrightarrow{\bar{\phi}_0}\widehat{\mathrm{CF}}(\Sigma,\boldsymbol{\alpha},\boldsymbol{\beta}_\infty;\bar{u},\bar{v},\bar{w}),$$

where

$$\bar{\phi}_\infty=\phi(\boldsymbol{\alpha},\boldsymbol{\beta}_0,\boldsymbol{\beta}_1;\bar{u},\bar{v},\bar{w})\quad\text{and}\quad\bar{\phi}_0=\phi(\boldsymbol{\alpha},\boldsymbol{\beta}_1,\boldsymbol{\beta}_\infty;\bar{u},\bar{v},\bar{w}).$$

3.3 Some properties of the maps $\mathfrak{f}_\bullet(K)$ and $\bar{\mathfrak{f}}_\bullet(K)$

Our first observation is that changing the orientation of the knot K and, correspondingly that of K_1 and K_0, corresponds to changing the markings u, v, w with $\bar{u}$, $\bar{v}$, $\bar{w}$ in Figure 1. Suppose that $(\Sigma,\boldsymbol{\alpha},\boldsymbol{\beta};z_1,z_2)$ represents $K_\bullet$, meaning that an oriented longitude for $K_\bullet$ is constructed from gluing an oriented arc on Σ from z_1 to z_2 in the complement of $\boldsymbol{\alpha}$ and an oriented arc on Σ from z_2 to z_1 in the complement of $\boldsymbol{\beta}$. Then $(\Sigma,\boldsymbol{\alpha},\boldsymbol{\beta};z_2,z_1)$ is a Heegaard diagram for $-K_\bullet$ (the knot $K_\bullet$ with the reverse orientation) while $(-\Sigma,\boldsymbol{\beta},\boldsymbol{\alpha};z_2,z_1)$ is a Heegaard diagram for $K_\bullet$. The chain complexes associated with the above three Heegaard diagrams are identical. Heegaard moves give chain homotopy equivalences

$$\tau_\bullet(K)\colon \widehat{\mathrm{CF}}(\Sigma,\boldsymbol{\alpha},\boldsymbol{\beta};z_1,z_2)\longrightarrow\widehat{\mathrm{CF}}(-\Sigma,\boldsymbol{\beta},\boldsymbol{\alpha};z_2,z_1)=\widehat{\mathrm{CF}}(\Sigma,\boldsymbol{\alpha},\boldsymbol{\beta};z_1,z_2).$$

These chain homotopy equivalences induce the involutions

$$\tau_\bullet(K)\colon \mathbb{H}_\bullet(K)\longrightarrow\mathbb{H}_\bullet(K),\qquad \bullet\in\{0,1,\infty\}.$$

In terms of these isomorphisms,

$$\begin{aligned}\bar{\mathfrak{f}}_0(K)&=\tau_\infty(K)\circ\mathfrak{f}_0(K)\circ\tau_1(K),\\ \bar{\mathfrak{f}}_1(K)&=\tau_0(K)\circ\mathfrak{f}_1(K)\circ\tau_\infty(K),\\ \bar{\mathfrak{f}}_\infty(K)&=\tau_1(K)\circ\mathfrak{f}_\infty(K)\circ\tau_0(K).\end{aligned}\tag{4}$$

Note however, that the equality $\bar{\theta}(K)=\tau_\infty(K)\theta(K)\tau_0(K)$ is only satisfied for the induced maps from $\mathrm{Ker}(\bar{\mathfrak{f}}_\infty(K))$ to $\mathrm{Coker}(\bar{\mathfrak{f}}_0(K))$.

The exactness of the sequences in (3) implies that, in appropriate decompositions

$$
\begin{aligned}
\mathbb{H}_0(K) &= \frac{\mathbb{H}_0(K)}{\mathrm{Ker}(\mathfrak{f}_\infty(K))} \oplus \mathrm{Ker}(\mathfrak{f}_\infty(K)) =: \mathbb{A}_\infty(K) \oplus \mathbb{A}_1(K),\\
\mathbb{H}_1(K) &= \frac{\mathbb{H}_1(K)}{\mathrm{Ker}(\mathfrak{f}_0(K))} \oplus \mathrm{Ker}(\mathfrak{f}_0(K)) =: \mathbb{A}_0(K) \oplus \mathbb{A}_\infty(K),\\
\mathbb{H}_\infty(K) &= \frac{\mathbb{H}_\infty(K)}{\mathrm{Ker}(\mathfrak{f}_1(K))} \oplus \mathrm{Ker}(\mathfrak{f}_1(K)) =: \mathbb{A}_1(K) \oplus \mathbb{A}_0(K),
\end{aligned} \tag{5}
$$

we have

$$\mathfrak{f}_\bullet(K) = \begin{pmatrix} 0 & 0 \\ I_{a_\bullet(K)} & 0 \end{pmatrix},$$

where $a_\bullet(K)$ denotes the rank of $\mathbb{A}_\bullet(K)$ for every $\bullet \in \{0, 1, \infty\}$. In this basis we may present the matrices $\tau_\bullet(K)$ as

$$\tau_\bullet(K) = \begin{pmatrix} A_\bullet(K) & B_\bullet(K) \\ C_\bullet(K) & D_\bullet(K) \end{pmatrix}, \quad \bullet \in \{0, 1, \infty\}.$$

The map $B_0(K)$ corresponds to the induced map

$$\tau_0(K)\colon \mathrm{Ker}(\mathfrak{f}_\infty(K)) \longrightarrow \frac{\mathbb{H}_0(K)}{\mathrm{Ker}(\mathfrak{f}_\infty(K))}.$$

The decomposition $\mathbb{H}_0(K) = \mathbb{A}_\infty(K) \oplus \mathbb{A}_1(K)$ may be modified using a change of basis of the form $P_X = \left(\begin{smallmatrix} I & 0 \\ -X & I \end{smallmatrix}\right)$, which does not change the block presentations of the maps $\mathfrak{f}_\infty(K)$ and $\mathfrak{f}_1(K)$. In the new basis, $\tau_0(K)$ has the following presentation:

$$
\begin{aligned}
\tau_0(K) &= \begin{pmatrix} I & 0 \\ -X & I \end{pmatrix} \begin{pmatrix} A_0(K) & B_0(K) \\ C_0(K) & D_0(K) \end{pmatrix} \begin{pmatrix} I & 0 \\ -X & I \end{pmatrix}\\
&= \begin{pmatrix} A_0(K) - B_0(K)X & B_0(K) \\ \star & -XB_0(K) + D_0(K) \end{pmatrix}
\end{aligned}
$$

If $B_0(K)$ is injective we may thus assume that $D_0(K) = 0$, while if $B_0(K)$ is surjective we may assume that $A_0(K) = 0$. With similar reasoning, if $B_\bullet(K)$ is injective we may assume that $D_\bullet(K) = 0$, while if $B_\bullet(K)$ is surjective we may assume that $A_\bullet(K) = 0$.

In the above decompositions for $\mathbb{H}_\bullet(K)$, the map $\theta(K) : \mathbb{H}_0(K) \to \mathbb{H}_\infty(K)$ takes the form

$$\theta(K) = \begin{pmatrix} X & I \\ Z & Y \end{pmatrix},$$

since the induced map from $\mathbb{A}_1(K) \subset \mathbb{H}_0(K)$ to $\mathbb{A}_1(K) \subset \mathbb{H}_\infty(K)$ is the inverse of the map induced by $\mathfrak{f}_1(K)$, ie the identity. Moreover, since the rank of $\theta(K)$ is the same as the rank of $\mathfrak{f}_1(K)$, we conclude that $Z = YX$. Applying the change of basis

P_Y on $\mathbb{H}_0(K)$ and the corresponding change of basis P_X on $\mathbb{H}_\infty(K)$, $\theta(K)$ takes the form

$$\begin{pmatrix} I & 0 \\ -Y & I \end{pmatrix}\begin{pmatrix} X & I \\ YX & Y \end{pmatrix}\begin{pmatrix} I & 0 \\ -X & I \end{pmatrix} = \begin{pmatrix} 0 & I \\ 0 & 0 \end{pmatrix}.$$

It is thus possible to choose the above decompositions so that $\theta(K) = \left(\begin{smallmatrix} 0 & I \\ 0 & 0 \end{smallmatrix}\right)$. If this is the case, the 2×2 presentation of $\tau_\infty(K)\bar{\theta}(K)\tau_0(K)$ will be of the form

$$\tau_\infty(K)\bar{\theta}(K)\tau_0(K) = \begin{pmatrix} M & I \\ Q & P \end{pmatrix}$$

and, since the ranks of $\theta(K)$ and $\bar{\theta}(K)$ are the same, we find $Q = PM$.

3.4 Relative Spin^c structures

The vector spaces $\mathbb{H}_\infty(K)$ and $\mathbb{H}_1(K)$ are naturally decomposed by relative Spin^c classes in

$$\underline{\mathrm{Spin}}^c(Y, K) = \underline{\mathrm{Spin}}^c(Y_1(K), K_1) = \mathbb{Z},$$

where the identification with $\mathbb{Z}$ is made using the first Chern class (divided by 2). Similarly, the relative Spin^c classes corresponding to K_0 are identified with $\frac{1}{2} + \mathbb{Z}$. Thus,

$$\mathbb{H}_\bullet(K) = \bigoplus_{i\in\mathbb{Z}} \mathbb{H}_\bullet(K, i), \quad \bullet \in \{1, \infty\}, \qquad \text{and} \qquad \mathbb{H}_0(K) = \bigoplus_{j\in\frac{1}{2}+\mathbb{Z}} \mathbb{H}_0(K, j).$$

Note that $\tau_\bullet(K)$ takes $\mathbb{H}_\bullet(K, i)$ isomorphically to $\mathbb{H}_\bullet(K, -i)$ for $\bullet = 0, 1, \infty$.

Let $H_0 = (\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_\infty; u, v, w)$ be a Heegaard triple used for defining $\mathfrak{f}_0(K)$. If $\boldsymbol{x} \in \mathbb{T}_\alpha \cap \mathbb{T}_{\beta_1}$ and $\boldsymbol{y} \in \mathbb{T}_\alpha \cap \mathbb{T}_{\beta_\infty}$ are two generators connected by a triangle class $\Delta \in \pi_2(\boldsymbol{x}, \Theta_{1,\infty}, \boldsymbol{y})$ with $n_u(\Delta) = n_w(\Delta) = 0$ (as observed in the surgery exact sequences of [14]), then $c_1(\underline{\mathfrak{s}}_{u,w}(\boldsymbol{x})) = c_1(\underline{\mathfrak{s}}_{u,v}(\boldsymbol{y}))$. This observation, together with (4) imply that the maps $\mathfrak{f}_0(K)$ and $\bar{\mathfrak{f}}_0(K)$ are decomposed as

$$\mathfrak{f}_0(K) = \bigoplus_{i\in\mathbb{Z}} \mathfrak{f}_0(K, i), \quad \mathfrak{f}_0(K, i)\colon \mathbb{H}_1(K, i) \longrightarrow \mathbb{H}_\infty(K, i),$$
$$\bar{\mathfrak{f}}_0(K) = \bigoplus_{i\in\mathbb{Z}} \bar{\mathfrak{f}}_0(K, i), \quad \bar{\mathfrak{f}}_0(K, i)\colon \mathbb{H}_1(K, i) \longrightarrow \mathbb{H}_\infty(K, i).$$

The map $\bar{\mathfrak{f}}_\infty(K)\colon \mathbb{H}_0(K) \to \mathbb{H}_1(K)$ drops the Spin^c grading by $\frac{1}{2}$, while the map $\mathfrak{f}_\infty(K)\colon \mathbb{H}_1(K) \to \mathbb{H}_0(K)$ increases the Spin^c grading by $\frac{1}{2}$. The corresponding decompositions are thus

$$\mathfrak{f}_\infty(K) = \bigoplus_{i\in\mathbb{Z}} \mathfrak{f}_\infty(K,i), \quad \mathfrak{f}_\infty(K,i)\colon \mathbb{H}_0\big(K, i-\tfrac{1}{2}\big) \longrightarrow \mathbb{H}_1(K,i),$$
$$\bar{\mathfrak{f}}_\infty(K) = \bigoplus_{i\in\mathbb{Z}} \bar{\mathfrak{f}}_\infty(K,i), \quad \bar{\mathfrak{f}}_\infty(K,i)\colon \mathbb{H}_0\big(K, i+\tfrac{1}{2}\big) \longrightarrow \mathbb{H}_1(K,i).$$

In particular, for a knot K of genus g the maps $\bar{\mathfrak{f}}_\infty(K,g)$ and $\mathfrak{f}_\infty(K,-g)$ are trivial, since $\mathbb{H}_0\big(K, g+\frac{1}{2}\big) = \mathbb{H}_0\big(K, -g-\frac{1}{2}\big) = 0$ by [2, Theorem 3.2]. Moreover,

$$\mathfrak{f}_1(K) = \bigoplus_{i\in\mathbb{Z}} \mathfrak{f}_1(K,i), \quad \mathfrak{f}_1(K,i)\colon \mathbb{H}_\infty(K,i) \longrightarrow \mathbb{H}_0\big(K, i-\tfrac{1}{2}\big),$$
$$\bar{\mathfrak{f}}_1(K) = \bigoplus_{i\in\mathbb{Z}} \bar{\mathfrak{f}}_1(K,i), \quad \bar{\mathfrak{f}}_1(K,i)\colon \mathbb{H}_\infty(K,i) \longrightarrow \mathbb{H}_0\big(K, i+\tfrac{1}{2}\big).$$

Let us now assume that $(\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \boldsymbol{\beta}_\infty; u, v, w)$ is one of the Heegaard quadruples illustrated in Figure 1. If we drop the marked point u (resp. the marked point w) from the Heegaard diagram, associated with either of the two resulting punctured Heegaard quadruples we obtain a triangle of chain maps:

$$\begin{tikzcd}
\widehat{\mathrm{CF}}(\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}_\infty; v, w) \arrow[dr, "\phi(\boldsymbol{\alpha},\boldsymbol{\beta}_\infty,\boldsymbol{\beta}_0;v,w)"'] & & \widehat{\mathrm{CF}}(\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}_1; v, w) \arrow[ll, "\phi(\boldsymbol{\alpha},\boldsymbol{\beta}_1,\boldsymbol{\beta}_\infty;v,w)"'] \\
& \widehat{\mathrm{CF}}(\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}_0; v, w) \arrow[ur, "\phi(\boldsymbol{\alpha},\boldsymbol{\beta}_0,\boldsymbol{\beta}_1;v,w)"'] &
\end{tikzcd}$$

$$\begin{tikzcd}
\widehat{\mathrm{CF}}(\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}_1; u, v) \arrow[dr, "\phi(\boldsymbol{\alpha},\boldsymbol{\beta}_1,\boldsymbol{\beta}_\infty;u,v)"'] & & \widehat{\mathrm{CF}}(\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}_0; u, v) \arrow[ll, "\phi(\boldsymbol{\alpha},\boldsymbol{\beta}_0,\boldsymbol{\beta}_1;u,v)"'] \\
& \widehat{\mathrm{CF}}(\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}_\infty; u, v) \arrow[ur, "\phi(\boldsymbol{\alpha},\boldsymbol{\beta}_\infty,\boldsymbol{\beta}_0;u,v)"'] &
\end{tikzcd}$$

The domain of any holomorphic triangle which contributes to $\phi(\boldsymbol{\alpha}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_\infty; v, w)$ has coefficient 1 precisely at one of the base points u and $\bar{u}$, and coefficient 0 at the other one. In other words,

$$\begin{aligned}
\phi(\boldsymbol{\alpha}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_\infty; v, w) &= \phi(\boldsymbol{\alpha}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_\infty; u, v, w) + \phi(\boldsymbol{\alpha}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_\infty; \bar{u}, v, w) \\
&= \phi(\boldsymbol{\alpha}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_\infty; u, v, w) + \phi(\boldsymbol{\alpha}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_\infty; \bar{u}, \bar{v}, \bar{w}).
\end{aligned}$$

A similar argument implies that

$$\phi(\boldsymbol{\alpha}, \boldsymbol{\beta}_0, \boldsymbol{\beta}_1; u, v) = \phi(\boldsymbol{\alpha}, \boldsymbol{\beta}_0, \boldsymbol{\beta}_1; u, v, w) + \phi(\boldsymbol{\alpha}, \boldsymbol{\beta}_0, \boldsymbol{\beta}_1; \bar{u}, \bar{v}, \bar{w}).$$

We thus obtain the following two exact triangles, respectively:

$$
\begin{tikzcd}
\mathbb{H}_\infty(K) & & \mathbb{H}_1(K) \arrow[ll, "\mathfrak{f}_0+\bar{\mathfrak{f}}_0"'] \\
& \widehat{\mathrm{HF}}(Y_0(K)) \arrow[ur] &
\arrow[from=1-1, to=2-2]
\end{tikzcd}
\qquad
\begin{tikzcd}
\mathbb{H}_1(K) & & \mathbb{H}_0(K) \arrow[ll, "\mathfrak{f}_\infty+\bar{\mathfrak{f}}_\infty"'] \\
& \widehat{\mathrm{HF}}(Y) \arrow[ur] &
\arrow[from=1-1, to=2-2]
\end{tikzcd}
\tag{6}
$$

where $\mathfrak{f}_\bullet = \mathfrak{f}_\bullet(K)$ and $\bar{\mathfrak{f}}_\bullet = \bar{\mathfrak{f}}_\bullet(K)$. The exact triangles in (3) and (6) may be used to deduce the following conclusions regarding the ranks of the chain maps:

$$
\begin{aligned}
\mathrm{rnk}(\mathfrak{f}_\bullet(K)) = \mathrm{rnk}(\bar{\mathfrak{f}}_\bullet(K)) &= \tfrac{1}{2}(h_\infty(K)+h_1(K)+h_0(K)-2h_\bullet(K)),\\
\mathrm{rnk}(\mathfrak{f}_\bullet(K)+\bar{\mathfrak{f}}_\bullet(K)) &= \tfrac{1}{2}(h_\infty(K)+h_1(K)+h_0(K)-y_\bullet(K)-h_\bullet(K)),
\end{aligned}
\tag{7}
$$

where $h_\bullet(K)$ denotes the rank of $\mathbb{H}_\bullet(K)$ and $y_\bullet(K)$ denotes the rank of $\widehat{\mathrm{HF}}(Y_\bullet(K))$.

4 Combinatorial presentation of the exact triangles

4.1 Heegaard diagrams for knot complements

The aim of this subsection is to construct Heegaard diagrams of particular type associated with a knot K inside a three-manifold Y, so that the chain complexes $C_\bullet(K)$ and the chain maps $\mathfrak{f}_\bullet(K)$ and $\bar{\mathfrak{f}}_\bullet(K)$ may all be described combinatorially.

Let us assume that a framed longitude $\widehat{\lambda}$ for K is given as a simple closed curve on the torus boundary of $Y \setminus \mathrm{nd}(K)$. Together with the meridian $\widehat{\mu}$ of the knot K, $\widehat{\lambda}$ gives a parametrization of the boundary of $Y \setminus \mathrm{nd}(K)$. It also determines the three-manifold $Y_{\widehat{\lambda}}(K)$ obtained by surgery on K. The curves $\widehat{\mu}$ and $\widehat{\lambda}$ thus give $Y \setminus \mathrm{nd}(K)$ the structure of a bordered three-manifold. As such, we remind the reader that a *nice Heegaard diagram*

$$(\Sigma,\ \boldsymbol{\alpha} = \{\alpha_1, \dots, \alpha_g\},\ \widehat{\boldsymbol{\beta}} = \{\beta_1, \dots, \beta_{g-1}\}, \mu, \lambda; z)$$

for the bordered three-manifold determined by (Y, K) and $\widehat{\lambda}$ consists of a surface Σ of genus g, a g–tuple of disjoint simple closed curves $\boldsymbol{\alpha}$, a $(g-1)$–tuple of disjoint simple closed curves $\widehat{\boldsymbol{\beta}}$, a pair of simple closed curves μ and λ disjoint from $\widehat{\boldsymbol{\beta}}$ which intersect in a single transverse point, and a marked point z in the complement of all curves in Σ. The data satisfies the following conditions:

- The diagram $(\Sigma, \boldsymbol{\alpha}, \widehat{\boldsymbol{\beta}})$ corresponds to $Y \setminus \mathrm{nd}(K)$, while $(\Sigma, \boldsymbol{\alpha}, \widehat{\boldsymbol{\beta}} \cup \{\mu\})$ and $(\Sigma, \boldsymbol{\alpha}, \widehat{\boldsymbol{\beta}} \cup \{\lambda\})$ correspond to the three-manifolds Y and $Y_{\widehat{\lambda}}(K)$, respectively.

- All domains in $\Sigma \setminus (\boldsymbol{\alpha} \cup \widehat{\boldsymbol{\beta}} \cup \{\mu, \lambda\})$ are either bigons, triangles or rectangles, except for the domain D_z containing the marked point z, which is a $(2N+1)$–gon for some integer N. In particular, D_z contains the single intersection point of μ and λ as a corner.
- Every curve $\beta_i \in \widehat{\boldsymbol{\beta}}$ contains at least one of the $2N+1$ edges of D_z.

Nice Heegaard diagrams exist by Lipshitz, Ozsváth and Thurston [6, Proposition 8.2]. However, two remarks are necessary here. First, note that in the aforementioned proposition the roles of the α– and β–curves is the opposite of our convention. In particular, the curves μ and λ are α–curves in [6]. The second point is that the third condition above is a priori not guaranteed by [6, Proposition 8.2]. However, if β_i does not contain any of the edges of D_z, all neighbouring regions of β_i would be bigons or rectangles. Since β_i is homotopically non-trivial, a computation of the Euler characteristic for the neighbourhood of β_i (the union of all regions which are neighbours of β_i) implies that all neighbouring regions of β_i are rectangles. However, this in turn implies that, for some $j \neq i$, β_j is parallel (and thus homologous) to β_i, a contradiction. Thus, the third condition is also guaranteed by [6, Proposition 8.2].

The picture on the top of Figure 3 describes a surface $\widehat{\Sigma}_1$ of genus 4. The opposite edges of the rectangle are identified and the pairs of yellow and red circles are also glued together (using a horizontal reflection). The pair of green circles is identified using a vertical reflection. The solid red curves are labelled μ and λ, which meet in a single transverse point O. The green domains glue together and form a disk D on $\widehat{\Sigma}_1$. We set $\Sigma_1 = \widehat{\Sigma}_1 \setminus \mathrm{Int}(D)$. The dashed blue curves in Σ_1 correspond to the β–curves, while the solid black curves correspond to the α–curves. The α– and β–curves may have boundary in ∂D.

Lemma 4.1 *Let K be a knot inside a three-manifold Y together with an arbitrary framing. Then there is a nice Heegaard diagram $(\Sigma, \boldsymbol{\alpha}, \widehat{\boldsymbol{\beta}} \cup \{\mu, \lambda\}, z)$ for the corresponding bordered three-manifold with the following properties:*

- $\Sigma = \Sigma_1 \amalg_{\partial\Sigma_1 = \partial\Sigma_2} \Sigma_2$, *where $\widehat{\Sigma}_1$ is the surface of genus 4 illustrated in Figure 3 and Σ_2 is a surface with one boundary component.*
- *The arcs in $\boldsymbol{\alpha} \cap \Sigma_1$ are identified with the solid black curves in Figure 3, while the arcs in $\widehat{\boldsymbol{\beta}} \cap \Sigma_1$ are identified with the dashed blue curves in Figure 3.*
- *The curves μ and λ correspond to the bold red curves on Σ_1.*
- *The domains on Σ_1 which contain the bold markings belong to the connected component D_z in $\Sigma \setminus (\boldsymbol{\alpha} \cup \widehat{\boldsymbol{\beta}} \cup \lambda \cup \mu)$ which contains z.*

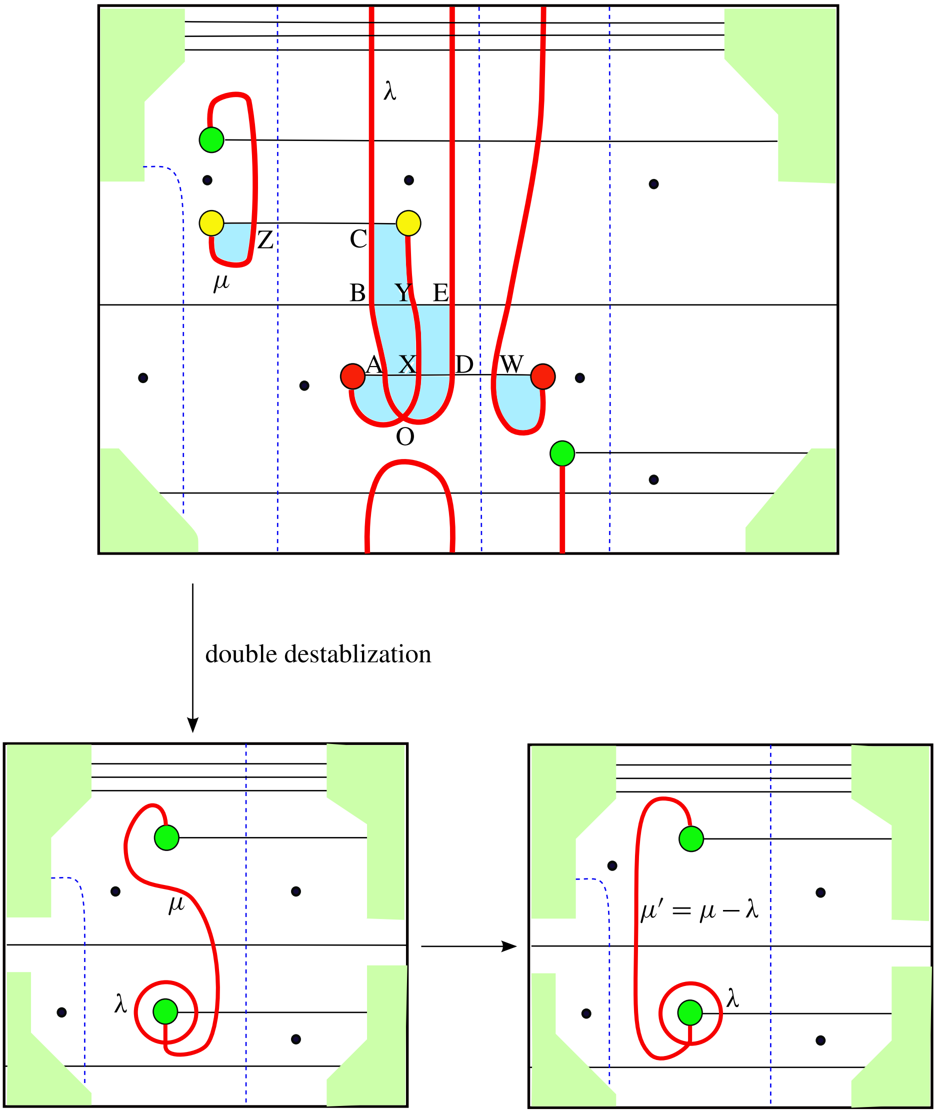


Figure 3: Special Heegaard diagrams for knot complements are the union of the genus-4 surface Σ_1 with boundary illustrated as the white part of the figure on top with another surface with boundary. The curves λ and μ are illustrated as bold red curves, while $\boldsymbol{\alpha} \cap \Sigma_1$ and $\widehat{\boldsymbol{\beta}} \cap \Sigma_1$ are denoted by black curves and dashed blue curves, respectively. The intersection of μ and λ is denoted by O and some of the intersection points in $\boldsymbol{\alpha} \cap (\lambda \cup \mu)$ are labelled (by A, B, C, D, E, X, Y, Z and W). Double destablization and a change in the framing (equivalently, in the parametrization of the boundary torus) gives the two Heegaard diagrams on the bottom of the figure.

Proof Destabilization on Σ_1 gives the equivalent Heegaard diagram, which locally looks like the surface on the lower left part of Figure 3. Changing μ to $\mu' = \mu - \lambda$ in the aforementioned diagram corresponds to changing the parametrization of the boundary. It is thus enough to show that every bordered three-manifold with torus boundary admits a nice Heegaard diagram which locally looks like the lower right side of Figure 3, so that every domain which meets the green region is either a bigon, a rectangle or contains the puncture. If this is the case, every domain in the Heegaard diagram illustrated on the upper side of Figure 3 is either a bigon, a triangle, a rectangle or contains the puncture. In other words, the diagram on the upper side of Figure 3 is nice.

Start with a nice bordered Heegaard diagram for $Y \setminus \mathrm{nd}(K)$ with parametrization given by μ' and λ, which exists by [6, Proposition 8.2]. Denote the intersection point of μ' and λ by O. Three of the four quadrants around O are triangles, while the last quadrant contains the marked point z. There is thus some curve α_i in $\boldsymbol{\alpha}$ which cuts μ' in the points D and A close to O and the curve λ in X and W (close to O), so that the picture around O on Σ is the one illustrated in part (a) of Figure 4. We may assume for simplicity that $i = g$. The three triangles are thus $[DOX]$, $[XOA]$ and $[AOW]$. There is a path γ disjoint from $\boldsymbol{\beta}_0 \cup \{\mu', \lambda\}$ which starts from the interior of the triangle $[AOW]$ and ends at the marked point z and passes only through the rectangles. One may add a 1–handle to Σ with attaching circles placed at the endpoints of γ. The core of this 1–handle may be added to $\boldsymbol{\alpha}$ as the curve α_{g+1} and the arc γ may be completed to a simple closed curve β_g by attaching its endpoints with an arc going over the 1–handle. This gives a stabilization of the previous Heegaard diagram. We may then handle slide α_{g+1} over α_g to obtain the Heegaard diagram illustrated in part (b) of Figure 4.

Next, we may add a 1–handle to the Heegaard diagram with attaching circles placed in the middle of the arcs $[OX]$ and $[OW]$. Denote the arc connecting the above two midpoints by δ. The curve μ' will be renamed β_{g+1}, the core of this handle will be replaced for μ', the curve λ will be modified by deleting the arc δ from it and replacing a corresponding arc which travels over the 1–handle, and, finally, the arc δ is completed to a simple closed curve α_{g+2} using the 1–handle. The new Heegaard diagram is illustrated in part (c) of Figure 4. This new Heegaard diagram corresponds to the same bordered three-manifold.

Next, we attach another 1–handle to the Heegaard diagram. The attaching circles are placed on λ on the two sides of the arc bounded between the intersection of α_{g+1} and λ and the intersection of μ' and λ. The aforementioned arc may be completed (by adding to it a segment which travels over the 1–handle) to a simple closed curve, which will be replaced for λ. The remainder of (the old) λ may also be completed (again by

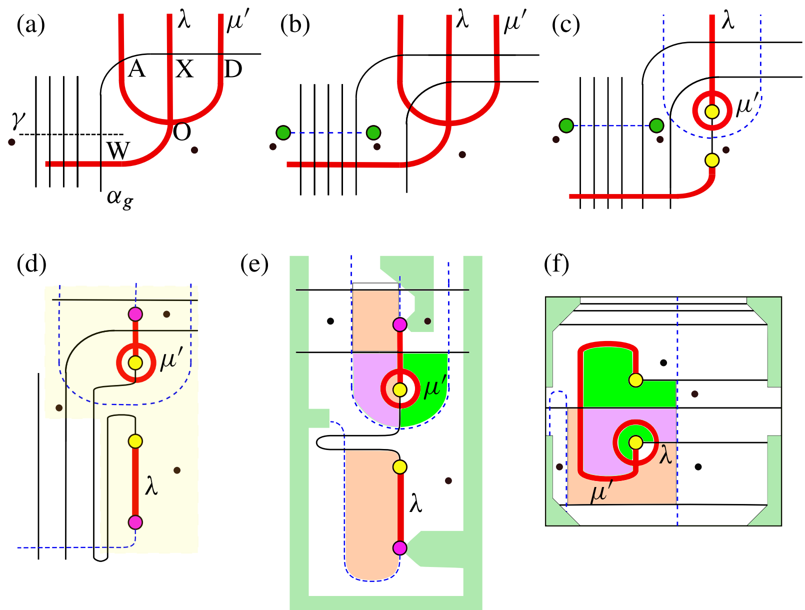


Figure 4: The α–curves are denoted by solid black lines, the β–curves are the dashed blue lines, and the curves μ' and λ are denoted by bold red lines. D_z is the domain containing bold circles. (a) In a nice Heegaard diagram, three of the quadrants around $O = \mu' \cap \lambda$ are triangles. Use an arc γ disjoint from $\widehat{\boldsymbol{\beta}} \cup \{\mu', \lambda\}$ to connect the triangle $[AOW]$ to z. The closest α–curve to O is α_g. (b) Attach a handle at the endpoints of γ, complete γ to a β–curve and slide the core of the handle over α_g to produce a new α–curve. (c) Attach a handle on λ at the two sides of O (the attaching circles are painted yellow). Rename μ' to β_{g+1} and replace the core of the handle for μ'. Push λ above the handle and complete the segment on μ' containing O to α_{g+2}. (d) Attach a handle on λ at the points illustrated by purple circles. The arcs on λ connecting the purple attaching circles to the yellow attaching circles may be completed to a closed curve, which will be replaced for λ. The complement of these two arcs on initial λ may be completed to a β–curve. The core of the 1–handle slides over α_g to produce the new α–curve. Finally, a finger move modifies α_{g+2}. (e)–(f) Re-draw the subsurface of genus 2 around the intersection of μ' and λ which was shaded in part (d).

adding to it a segment which travels over the 1–handle) to a simple closed curve, which will be denoted by β_{g+2}. One may slide the core of the new 1–handle over α_{g+1} to obtain α_{g+3}. Finally, we apply a finger move isotopy to α_{g+2} to create a pair of intersection points between α_{g+2} and β_{g+2}. The new Heegaard diagram (which still corresponds to the same bordered three-manifold) is illustrated in part (d) of Figure 4 and a subset of the diagram which lives on a subsurface of genus 2 is re-drawn in part (e) of the same picture, where a 7–gon and a pair of pentagons are painted orange, green and purple, respectively. One may then identify the aforementioned subsurface of genus 2 with the surface illustrated in part (f). To illustrate the correspondence, the domains corresponding to the 7–gon and the two pentagons are painted in the new picture with the relevant colour. This completes the proof of the lemma. □

Definition 4.2 For every knot $K \subset Y$ and every framing λ for K, the Heegaard diagrams of the type constructed in Lemma 4.1 are called *special* Heegaard diagrams.

4.2 A combinatorial description of $\mathfrak{f}_\bullet(K)$ and $\bar{\mathfrak{f}}_\bullet(K)$

Suppose that (Y, K) denotes a knot K inside a homology sphere Y. Let us assume that

$$(\Sigma,\ \boldsymbol{\alpha} = \{\alpha_1, \dots, \alpha_g\},\ \widehat{\boldsymbol{\beta}} = \{\beta_1, \dots, \beta_{g-1}\}, \mu, \lambda, z)$$

is a special Heegaard diagram for the bordered three-manifold determined by a zero-framed longitude for K inside Y. The picture around the intersection point O of the simple closed curves μ and λ is illustrated on the top of Figure 3.

We introduce three auxiliary curves, denoted by λ_∞, λ_0 and λ_1, respectively, as in the Heegaard diagram illustrated in Figure 5. The Heegaard diagrams

$$H_\bullet = (\Sigma, \boldsymbol{\alpha}, \widehat{\boldsymbol{\beta}} \cup \{\lambda_\bullet\}; u, v, w) \quad \text{and} \quad \overline{H}_\bullet = (\Sigma, \boldsymbol{\alpha}, \widehat{\boldsymbol{\beta}} \cup \{\lambda_\bullet\}; \bar{u}, \bar{v}, \bar{w})$$

are (triply punctured) diagrams that correspond to the knot $K_\bullet \subset Y_\bullet(K)$ for $\bullet \in \{0, 1, \infty\}$ (note that two of the three punctures are placed in the same connected component of $\Sigma \setminus (\boldsymbol{\alpha} \cup \widehat{\boldsymbol{\beta}} \cup \lambda_\bullet)$ for $\bullet \in \{0, 1, \infty\}$). The above claim is checked by computing the intersection numbers of each $\lambda_\bullet$ with the simple closed curves μ and λ, since the curves are disjoint from $\widehat{\boldsymbol{\beta}}$. Each pair of these three curves intersect each other exactly once. Each of the three diagrams $H_\bullet$, $\overline{H}_\bullet$, $\bullet \in \{0, 1, \infty\}$, is a nice Heegaard diagram and they determine the chain complexes $C_\bullet = \widehat{\mathrm{CF}}(H_\bullet) = \widehat{\mathrm{CF}}(\overline{H}_\bullet)$. Denote the differential of the complex $C_\bullet$ by $d_\bullet$ for $\bullet \in \{0, 1, \infty\}$. The chain maps $\mathfrak{f}_\bullet(K)$ and $\bar{\mathfrak{f}}_\bullet(K)$ have a simple combinatorial description, which is discussed in the remainder of this section.

Fix the labelling of the intersection points of λ_0, λ_1, λ_∞, β_{g-1} and β_{g-2} with the curves in $\boldsymbol{\alpha}$ as in Figure 5. Let

$$\{P_0\} = \lambda_1 \cap \lambda_\infty, \quad \{P_1\} = \lambda_0 \cap \lambda_\infty \quad \text{and} \quad \{P_\infty\} = \lambda_0 \cap \lambda_1.$$

The Heegaard triple

$$(\Sigma, \boldsymbol{\alpha}, \widehat{\boldsymbol{\beta}} \cup \{\lambda_1\}, \widehat{\boldsymbol{\beta}} \cup \{\lambda_\infty\}; \bar{u}, \bar{v}, \bar{w})$$

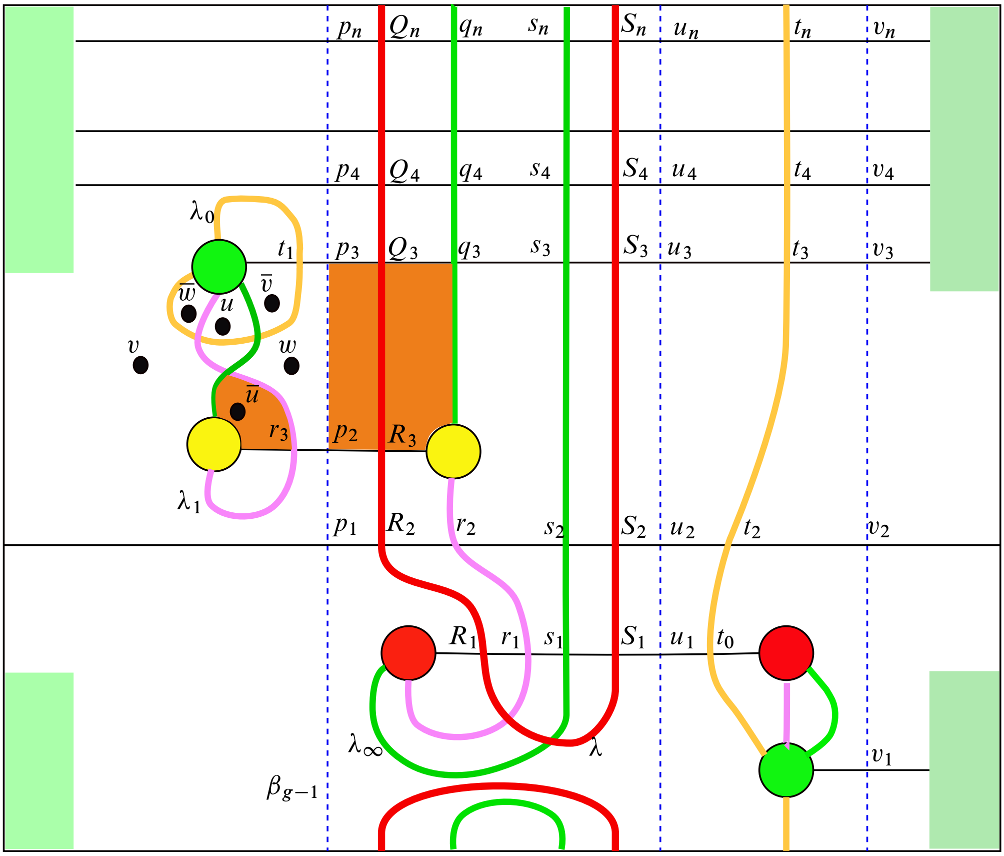


Figure 5: The curves in $\boldsymbol{\alpha}$ are denoted by solid black lines while the curves in $\widehat{\boldsymbol{\beta}}$ are denoted by dashed blue lines. Three simple closed curves λ_0, λ_1 and λ_∞ are denoted by bold red, purple and green lines, respectively. Six marked points u, v, w, $\overline{u}$, $\overline{v}$ and $\overline{w}$ are introduced close to the intersection points of these three curves. The intersection points on β_{g-1}, λ_∞ and λ_1 are labelled. Associated with $i \geq 3$ there is a pentagon with vertices at P_0, r_3, p_2, p_i and q_i. For $i = 3$ the pentagon is shaded orange in the picture.

determines a *combinatorial triangle map* $\bar{f}_0\colon C_1 \to C_\infty$ as follows. Let β_{g-1} be the β–curve which contains the intersection points $p_1, p_2, \ldots, p_n$ in Figure 5. Let $\boldsymbol{x} = (x_1, \ldots, x_g)$ be a generator of C_1 with $x_i \in \alpha_{\sigma(i)} \cap \beta_i$ for some $\sigma \in S_g$, $i = 1, \ldots, g-1$, and $x_g \in \lambda_1$. Define

$$\bar{f}_0(\boldsymbol{x}) := \begin{cases} (x_1, \ldots, x_{g-1}, s_i) & \text{if } x_g = r_i,\ i = 1, 2, \\ 0 & \text{otherwise.} \end{cases}$$

Similarly, the Heegaard triple

$$(\Sigma, \boldsymbol{\alpha}, \widehat{\boldsymbol{\beta}} \cup \{\lambda_1\}, \widehat{\boldsymbol{\beta}} \cup \{\lambda_\infty\}; u, v, w)$$

determines a combinatorial triangle map $f_0\colon C_1 \to C_\infty$ defined by

$$f_0(\boldsymbol{x}) := \begin{cases} (x_1, \dots, x_{g-2}, p_2, q_i) & \text{if } (x_{g-1}, x_g) = (p_i, r_3),\ i \geq 3, \\ (x_1, \dots, x_{g-2}, p_1, q_i) & \text{if } (x_{g-1}, x_g) = (p_i, r_2),\ i \geq 3, \\ 0 & \text{otherwise.} \end{cases}$$

The Heegaard triples

$$(\Sigma, \boldsymbol{\alpha}, \widehat{\boldsymbol{\beta}} \cup \{\lambda_0\}, \widehat{\boldsymbol{\beta}} \cup \{\lambda_1\}; u, v, w) \quad \text{and} \quad (\Sigma, \boldsymbol{\alpha}, \widehat{\boldsymbol{\beta}} \cup \{\lambda_0\}, \widehat{\boldsymbol{\beta}} \cup \{\lambda_1\}; \bar{u}, \bar{v}, \bar{w}),$$

correspond to the combinatorial triangle maps f_∞, $\bar{f}_\infty\colon C_0 \to C_1$. For a generator $\boldsymbol{x} = (x_1, \dots, x_g)$, these two maps are defined by setting

$$\bar{f}_\infty(\boldsymbol{x}) = \begin{cases} (x_1, \dots, x_{g-1}, r_1) & \text{if } x_g = t_0, \\ 0 & \text{otherwise,} \end{cases}$$

$$f_\infty(\boldsymbol{x}) = \begin{cases} (x_1, \dots, x_{g-2}, p_3, r_3) & \text{if } (x_{g-1}, x_g) = (p_2, t_1), \\ 0 & \text{otherwise.} \end{cases}$$

Lemma 4.3 *With the above notation fixed,* $f_0 \circ f_\infty = \bar{f}_0 \circ \bar{f}_\infty = 0$.

Proof This is trivial from the combinatorial definitions of f_0, $\bar{f}_0$, f_∞ and $\bar{f}_\infty$. □

Let

$$\Sigma \setminus (\boldsymbol{\alpha} \cup \widehat{\boldsymbol{\beta}} \cup \lambda_0 \cup \lambda_1) = \left(\coprod_{i=1}^{N} D_i \right) \cup D_{\bar{u}} \cup D_{\bar{v}} \cup D_{\bar{w}},$$

where $D_\bullet$ are the regions in the complement of these curves, with $D_{\bar{u}}$, $D_{\bar{v}}$ and $D_{\bar{w}}$ the regions containing the marked points $\bar{u}$, $\bar{v}$ and $\bar{w}$, respectively. We set

$$\beta_i^0 = \beta_i, \quad i = 1, \dots, g-1, \qquad \text{and} \qquad \boldsymbol{\beta} = \{\beta_1, \dots, \beta_g\} = \widehat{\boldsymbol{\beta}} \cup \{\lambda_0\}.$$

The construction of the Heegaard diagram implies the following properties:

- The regions $D_2, \dots, D_N$ are rectangles or bigons, while D_1 is a pentagon.
- One of the corners of the pentagon D_1 is the unique intersection point $P = P_\infty = \lambda_0 \cap \lambda_1$, and the three punctures $\bar{u}$, $\bar{v}$ and $\bar{w}$ are placed on three of the quadrants around P (other than the quadrant corresponding to D_1).
- All the neighbours of D_1 (the regions having an edge in common with D_1) are punctured.
- Each β–curve is adjacent to at least one of the punctured domains.

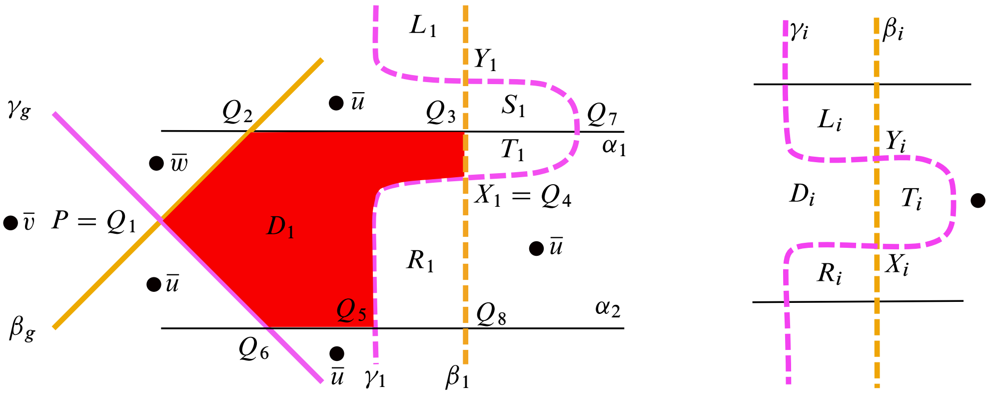


Figure 6: The region around the *pentagon* D_1 is illustrated on the left-hand side. The punctured domains are marked by solid circles inside them. The curves in $\boldsymbol{\beta} = \boldsymbol{\beta}_0$, $\boldsymbol{\gamma} = \boldsymbol{\beta}_1$ and $\boldsymbol{\alpha}$ have colours orange, pink and black, respectively. The pentagon is changed to a hexagon in the new Heegaard diagram, which is coloured red. The initial pentagon is the union of the hexagon D_1 with the triangle R_1. The right-hand side illustrates the labelling near the intersection of β_i with its Hamiltonian isotope γ_i.

The edges of the pentagon are five arcs: two of them are on λ_0 and λ_1, two of them are on the α–curves and one of them is on a β–curve, which is assumed to be β_1. The α–curve which cuts λ_0 in a corner of the pentagon is assumed to be α_1 and the other one is assumed to be α_2. Denote the vertices of the pentagon by $P = Q_1$, Q_2, Q_3, Q_8 and Q_6 in counter-clockwise order, so that Q_1 is the intersection point of λ_0 and λ_1, Q_2 is on the intersection of α_1 with λ_0, and Q_6 is the intersection point of λ_1 with α_2.

For $i = 2, \dots, g-1$, let $\beta_i^1 = \gamma_i$ be a parallel copy of β_i which is drawn very close to β_i and is slightly pushed to one of the punctured domains adjacent to β_i by a finger move, so that a pair of intersection points (denoted by X_i and Y_i) is created between these two curves (see the right-hand side picture in Figure 6). Let us assume that the small positively oriented disk connecting these two intersection points (with β_i on the left and γ_i on the right) goes from X_i to Y_i. In order to define γ_1, choose a parallel copy of β_1 and push it slightly over the intersection point of β_1 with α_1 to obtain γ_1, so that a pair of cancelling intersection points X_1 and Y_1 is created between γ_1 and β_1 on the two sides of the intersection point Q_3 of α_1 and β_1, and so that γ_1 slightly enters the punctured domain next to the β–edge of the pentagon. The local picture around D_1 looks like Figure 6, where this procedure is pictured. Let γ_g be the curve λ_1 and set $\boldsymbol{\beta}_1 = \boldsymbol{\gamma} = \{\gamma_1, \dots, \gamma_g\}$.

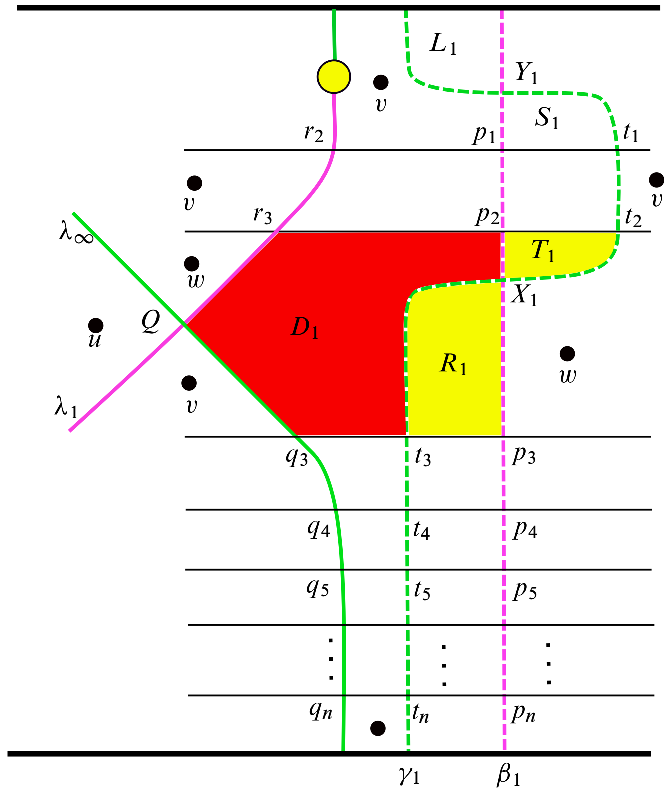


Figure 7: The region around the hexagon D_1 is illustrated. The labelling of the intersection points in the Heegaard diagram, as well as the labelling of some of the connected components in the complement of the curves, is illustrated. The curves in $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_\infty$ and $\boldsymbol{\alpha}$ have colours pink, green and black, respectively.

In order to construct β_i^∞ for $i = 2, \dots, g-1$, choose a parallel copy of $\gamma^i = \beta_i^1$ and, as this parallel copy enters the bigon T_i, push it into the neighbouring punctured domain by a finger move. The curve β_1^∞ is constructed as illustrated in Figure 7. We set

$$\boldsymbol{\beta}_\infty = \{\beta_1^\infty, \dots, \beta_{g-1}^\infty, \lambda_\infty\}.$$

Lemma 4.4 *The punctured Heegaard diagrams*

$$(\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}_\star, \boldsymbol{\beta}_\bullet; u, v, w) \quad \text{and} \quad (\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}_\star, \boldsymbol{\beta}_\bullet; \bar{u}, \bar{v}, \bar{w})$$

for $(\star, \bullet) \in \{(0,1), (1,\infty)\}$ do not contain any non-trivial, positive, triply periodic domains.

Proof Let $\mathcal{D}$ denote a positive, triply periodic domain in the Heegaard diagram $(\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_\infty; u, v, w)$. Thus,

$$\partial\mathcal{D} = \sum_{i=1}^{g} a_i\alpha_i + \sum_{i=1}^{g-1} b_i\beta_i^1 + \sum_{i=1}^{g-1} c_i\beta_i^\infty + b\lambda_1 + c\lambda_\infty.$$

Let $\mathcal{D}_i$ denote the doubly periodic domain with $\partial\mathcal{D}_i = \beta_i^1 - \beta_i^\infty$ for $i = 1, \dots, g-1$. Setting $\mathcal{D}' = \mathcal{D} - \sum_{i=1}^{g-1} b_i\mathcal{D}_i$, we find

$$\partial\mathcal{D}' = \sum_{i=1}^{g} a_i\alpha_i + \sum_{i=1}^{g-1} (c_i - b_i)\beta_i^\infty + b\lambda_1 + c\lambda_\infty.$$

Since the left-hand side is trivial in $H_1(Y \setminus \mathrm{nd}(K); \mathbb{Z})$, so is the right-hand side. This implies that $c = -b$. Let $\mathcal{D}_0$ denote the triply periodic domain in the punctured Heegaard triple $(\Sigma, \boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \boldsymbol{\beta}_\infty; u, v, w)$ with $\partial\mathcal{D}_0 = \lambda_1 - \lambda_0 - \lambda_\infty$. For $\mathcal{D}'' = \mathcal{D}' - b\mathcal{D}_0$ we thus obtain

$$\partial\mathcal{D}'' = \sum_{i=1}^{g} a_i\alpha_i + \sum_{i=1}^{g-1} (c_i - b_i)\beta_i^\infty + b\lambda_0.$$

In other words, $\mathcal{D}''$ is a doubly periodic domain for the nice (and hence weakly admissible) Heegaard diagram

$$(\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}_\infty \cup \{\lambda_0\} \setminus \{\lambda_\infty\}; u, v, w).$$

The coefficients of $\mathcal{D}''$ and all $\mathcal{D}_i$, $i = 1, \dots, g-1$, over the small triangle bounded between λ_0, λ_1 and λ_∞ is zero. In other words, the coefficient of

$$\mathcal{D} = \mathcal{D}'' + b\mathcal{D}_0 + \sum_{i=1}^{g-1} b_i\mathcal{D}_i$$

over this small triangle is b, which should thus be non-negative. Choosing this triangle sufficiently small we may thus assume that the total area of $b\mathcal{D}_0$ is negative unless $b = 0$.

One may choose the area form on the surface Σ so that all doubly periodic domains for the punctured Heegaard diagram $(\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}_\infty \cup \{\lambda_0\} \setminus \{\lambda_\infty\}; u, v, w)$ and all $\mathcal{D}_i$, $i = 1, \dots, g-1$, have zero total area. However, this implies that the total area of $\mathcal{D}$ is the same as the total area of $b\mathcal{D}_0$, which is at most zero. Since $\mathcal{D}$ is a positive domain, we conclude $\mathcal{D} = 0$. This completes the proof for the triple $(\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_\infty; u, v, w)$. The proof for the other triples is completely similar. □

The Heegaard diagrams

$$(\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}_\bullet; u, v, w) \quad \text{and} \quad (\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}_\bullet; \bar{u}, \bar{v}, \bar{w})$$

are nice so, by Sarkar and Wang [16], the differentials of the complexes

$$\widehat{\mathrm{CF}}(\Sigma,\boldsymbol{\alpha},\boldsymbol{\beta}_\bullet;u,v,w) \quad\text{and}\quad \widehat{\mathrm{CF}}(\Sigma,\boldsymbol{\alpha},\boldsymbol{\beta}_\bullet;\bar u,\bar v,\bar w)$$

are given by counts of bigons and rectangles.

Theorem 4.5 *Under the above identification of the chain complexes* $(C_\bullet, d_\bullet)$*,*

$$f_0=\phi(\Sigma,\boldsymbol{\alpha},\boldsymbol{\beta}_1,\boldsymbol{\beta}_\infty;u,v,w),\qquad \bar f_0=\phi(\Sigma,\boldsymbol{\alpha},\boldsymbol{\beta}_1,\boldsymbol{\beta}_\infty;\bar u,\bar v,\bar w),$$
$$f_\infty=\phi(\Sigma,\boldsymbol{\alpha},\boldsymbol{\beta}_0,\boldsymbol{\beta}_1;u,v,w),\qquad \bar f_\infty=\phi(\Sigma,\boldsymbol{\alpha},\boldsymbol{\beta}_0,\boldsymbol{\beta}_1;\bar u,\bar v,\bar w).$$

4.3 Proof of Theorem 4.5

A similar discussion is carried over in [3] (and in particular Theorem 2.3 from that paper). We repeat the proof, in most parts with more details, to keep the paper easier to read.

Proof We start by proving the statement for $\bar f_\infty$. Note that the top generator Θ of the Heegaard Floer homology group $\widehat{\mathrm{HF}}(\#^{g-1}\, S^1\times S^2)$ coming from the Heegaard diagram $(\Sigma,\boldsymbol{\beta},\boldsymbol{\gamma};\bar u,\bar v,\bar w)$ is the generator $\{P, X_1,\dots,X_{g-1}\}$.

Let $\boldsymbol{x}=(x_1,\dots,x_g)$ and $\boldsymbol{y}=(y_1,\dots,y_g)$ be generators with $x_i\in\alpha_{\sigma(i)}\cap\beta_i$ and $y_i\in\alpha_{\tau(i)}\cap\gamma_i$, with $\sigma,\tau\in S_g$. Let $\Delta\colon \mathbb{D}\to\mathrm{Sym}^g(\Sigma)$ be the homotopy class of a triangle in $\pi_2(\boldsymbol{x},\Theta,\boldsymbol{y})$, with Maslov index zero, so that it supports a holomorphic representative and remains disjoint from the punctures.

There are two types of domain in the complement $\Sigma\setminus(\boldsymbol{\alpha}\cup\boldsymbol{\beta}\cup\boldsymbol{\gamma})$ of the curves, the *large* domains and the *small* domains. The small domains are those created between the parallel pairs of curves γ_i and β_i $(i=1,\dots,g-1)$, and their area may be chosen arbitrarily small by choosing γ_i close enough to β_i. The large domains are the rest of the domains, which are in correspondence with the domains $D_\bullet$, $\bullet\in\{\bar u,\bar v,\bar w,1,\dots,N\}$, introduced above. We abuse the notation and still denote these new regions by $D_\bullet$.

Let us assume that the small bigon connecting X_i to Y_i is denoted by T_i and the region having the small interval $[X_i,Y_i]$ on β_i in common with T_i is D_i, $i=2,\dots,g-1$. Then there are two triangles with corners X_i and Y_i that have an edge in common with D_i, which will be denoted by R_i and L_i, respectively. For $i=1$, instead of these three regions we have four triangles with one corner being X_1 or Y_1, which will be denoted by R_1, T_1, S_1 and L_1, respectively (as they appear while we travel on β_1 from X_1 to Y_1; see Figure 6). We are implicitly assuming that the regions D_i for $i=1,\dots,g-1$ (as described above) are different, while it may happen that this is not the case. However, the argument we give below remains true in general and only needs notational corrections.

Let $\mathcal{D}=\mathcal{D}(\Delta)$ denote the domain (ie the 2–chain on Σ) associated with the triangle class Δ. Let $d_i\geq 0$ denote the coefficient of D_i in $\mathcal{D}$. Similarly, denote the coefficients of T_i, R_i and L_i by t_i, r_i and l_i, respectively. The coefficient of S_1 will be denoted by s_1. Of course, there are other regions which may appear in $\mathcal{D}$ with positive coefficient, but all such regions are bigons or rectangles. Since P appears in Θ and three of the corners around P are punctured, the coefficient d_1 is equal to 1.

Let $P=Q_1,Q_2,\dots,Q_6$ denote the corners of D_1 (now a hexagon) in counter-clockwise order (so $Q_4=X_1$). Since two opposite quadrants around each one of Q_2 and Q_6 are punctured, we have $x_g=Q_2$ and $y_g=Q_6$. Thus, Q_3 is not one of $x_1,\dots,x_g$ and Q_5 is not one of $y_1,\dots,y_g$. Considering the local coefficients around Q_3, we conclude that $t_1=1+s_1$. If Q_7 is the third corner of T_1 (other than Q_3 and Q_4), in order for $\mathcal{D}$ to be a non-negative domain we need $x_1=Q_7$ and the 4 local coefficients around Q_7 are forced to be $t_1=1+s_1$, s_1, 0 and 0 in the counter-clockwise order. Two opposite quadrants around Y_1 have zero coefficients in $\mathcal{D}$. Since Y_1 does not appear in Θ, this implies that $s_1=l_1=0$ (thus $t_1=1$). Similarly, considering the local coefficients around p_1 we conclude $r_1=1$. Since Q_5 is not among $y_1,\dots,y_g$, the local coefficients around Q_5 are 1, $r_1=1$, 0 and 0 in the counter-clockwise order. Let Q_8 be the third corner of R_1 other than Q_4 and Q_5. Since two opposite corners around Q_8 have zero coefficient and $r_1=1$, we have $x_1=Q_8$. Thus $\mathcal{D}=\mathcal{D}'+\mathcal{D}_1=\mathcal{D}'+(R_1+D_1+T_1)$, where $\mathcal{D}'$ is a non-negative 2–chain which is disjoint from $\mathcal{D}_1$ and $\mathcal{D}_1$ is a hexagon with five acute angles and one obtuse angle and with vertices $\{P,y_g,x_1,X_1,y_1,x_g\}$. The contribution of $\mathcal{D}_1$ to the index of Δ is zero, by Sarkar's formula [15].

By Sarkar's formula for the index of triangles [15],

$$\mu(\Delta)=e(\mathcal{D})+\mu_{\boldsymbol{x}}(\mathcal{D})+\mu_{\boldsymbol{y}}(\mathcal{D})+\mathfrak{b}(\mathcal{D}).\mathfrak{c}(\mathcal{D})-\tfrac{1}{2}g. \tag{8}$$

Here $e(\mathcal{D})$ is the Euler measure of the domain $\mathcal{D}$, $\mathfrak{b}(\mathcal{D})$ is the part of $\partial\mathcal{D}$ on the β–curves, and $\mathfrak{c}(\mathcal{D})$ is the part of $\partial\mathcal{D}$ on the γ–curves. Furthermore, $\mu_{\boldsymbol{x}}(\mathcal{D})$ and $\mu_{\boldsymbol{y}}(\mathcal{D})$ denote the local contributions of the intersection points included in $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively, to the corners of $\mathcal{D}$. We refer to [15] for more detailed definitions. Separating $\mathcal{D}_1$ — which has Maslov index 0 — from $\mathcal{D}$ we obtain the equality

$$\mu(\Delta)=e(\mathcal{D}_s)+e(\mathcal{D}_l)+\mu_{\boldsymbol{x}}(\mathcal{D}')+\mu_{\boldsymbol{y}}(\mathcal{D}')+\mathfrak{b}(\mathcal{D}').\mathfrak{c}(\mathcal{D}')-\tfrac{1}{2}(g-2).$$

Here $\mathcal{D}_s$ denotes the part of $\mathcal{D}'$ which uses the regions D_i, R_i, T_i and L_i for $i=2,\dots,g-1$ and $\mathcal{D}_l=\mathcal{D}'-\mathcal{D}_s$. Clearly, $e(\mathcal{D}_l)\geq 0$ and

$$\mathcal{D}_s=\sum_{i=2}^{g-1}(d_iD_i+t_iT_i+r_iR_i+l_iL_i).$$

Considering the local coefficients around X_i and Y_i, we conclude $r_i = l_i + 1$ and $d_i = t_i + l_i$. Having in mind that T_i are bigons, R_i and L_i are triangles and D_i are hexagons, this implies the following computation:

$$
\begin{aligned}
(9)\qquad e(\mathcal{D}_s) &= \sum_{i=2}^{g-1}((t_i+l_i)e(D_i) + t_i e(T_i) + (l_i+1)e(R_i) + l_i e(L_i)) \\
&= \sum_{i=2}^{g-1}\big((t_i+l_i)\big(-\tfrac{1}{2}\big) + t_i\big(\tfrac{1}{2}\big) + (l_i+1)\big(\tfrac{1}{4}\big) + l_i\big(\tfrac{1}{4}\big)\big) = \tfrac{1}{4}(g-2).
\end{aligned}
$$

The 1–chain $\mathfrak{b}(\mathcal{D}')$ is a union of 1–chains on $\beta_i, i = 2, \dots, g-1$, denoted by $\mathfrak{b}_i(\mathcal{D}')$. Similarly we have $\mathfrak{c}(\mathcal{D}') = \sum_{i=2}^{g-1} \mathfrak{c}_i(\mathcal{D}')$. It is clear that $\mathfrak{b}_i(\mathcal{D}')$ and $\mathfrak{c}_j(\mathcal{D}')$ are disjoint unless $i = j$. In this latter case, the only possible geometric intersections are at X_i and Y_i, where the intersection numbers are $(l_i + \frac{1}{2})(t_i - \frac{1}{2})$ and $-l_i t_i$, respectively. Thus,

$$
(10)\qquad \mathfrak{b}(\mathcal{D}').\mathfrak{c}(\mathcal{D}') = \sum_{i=2}^{g-1}\big(\big(l_i+\tfrac{1}{2}\big)\big(t_i-\tfrac{1}{2}\big) - l_i t_i\big) = -\tfrac{1}{4}(g-2) + \tfrac{1}{2}\sum_{i=2}^{g-1}(t_i - l_i).
$$

Let us now consider the coefficients around the intersection points x_i and y_i for $i = 2, \dots, g-1$. Since x_i is on β_i, there are non-negative integers a_i, b_i, c_i and e_i such that the local coefficients around x_i are a_i, b_i, $b_i + l_i + 1$ and $a_i + l_i$, and the local coefficients around y_i are c_i, e_i, $e_i + t_i - 1$ and $c_i + t_i$. Thus,

$$
(11)\qquad \mu_{\boldsymbol{x}}(\mathcal{D}') + \mu_{\boldsymbol{y}}(\mathcal{D}') = \tfrac{1}{2}\sum_{i=2}^{g-1}((a_i+b_i+c_i+e_i) + (l_i+t_i)).
$$

Combining (9), (10) and (11) and replacing for the terms in the definition of $\mu(\Delta)$, we obtain

$$
\begin{aligned}
0 = \mu(\Delta) &= e(\mathcal{D}_s) + e(\mathcal{D}_l) + \mu_{\boldsymbol{x}}(\mathcal{D}') + \mu_{\boldsymbol{y}}(\mathcal{D}') + \mathfrak{b}(\mathcal{D}').\mathfrak{c}(\mathcal{D}') - \tfrac{1}{2}(g-2) \\
&= e(\mathcal{D}_l) - \tfrac{1}{2}(g-2) + \tfrac{1}{2}\sum_{i=2}^{g-1}(a_i+b_i+c_i+e_i+2t_i) \\
&\geq \tfrac{1}{2}\sum_{i=2}^{g-1}(a_i+b_i+c_i+(e_i+t_i-1)+t_i).
\end{aligned}
$$

Note that $e_i + t_i - 1$ is the coefficient of one of the domains around y_i and is thus non-negative. The above inequality thus implies that $a_i = b_i = c_i = t_i = 0$ and $e_i = 1$ for $i = 2, \dots, g-1$. Thus, the coefficients on the two sides of γ_i either agree or differ by 1, and the coefficients on the two sides of β_i differ either by l_i or by $l_i + 1$. If we start from y_i, where on the left (or right) side of y_i the coefficients on the two sides of γ_i are zero, and travel on the α curve intersecting γ_i (ie orthogonal to γ_i) until we get to an intersection point with β_i, as we pass β_i the coefficient changes either to $-l_i$ or to $-l_i-1$. Since the latter is negative, the former happens and $l_i=0$. It is easy to see from here that x_i and y_i are the corresponding intersection points of β_i and γ_i with the same α–curve and that the domain $\mathcal{D}'$ is a union of obvious triangles which are disjoint from each other.

We conclude that the domain of Δ is the disjoint union of $g-2$ simple triangles with a hexagon with five acute angles and one obtuse angle. It is quite well known that the moduli space corresponding to this homotopy class contributes 1 to the triangle map for a generic path of almost complex structures. These are thus the only holomorphic triangles which contribute to the chain map $\bar{f}_\infty$ defined using the Heegaard triple $(\Sigma,\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma};\bar{u},\bar{v},\bar{w})$. Under the obvious identification of $\widehat{\mathrm{CF}}(\Sigma,\boldsymbol{\alpha},\boldsymbol{\gamma};\bar{u},\bar{v},\bar{w})$ with $\widehat{\mathrm{CF}}(\Sigma,\boldsymbol{\alpha},\boldsymbol{\beta}_1;\bar{u},\bar{v},\bar{w})$, this is just the map which replaces the pair $\{Q_2,Q_8\}$ with $\{Q_6,Q_7\}$. This completes the proof of Theorem 4.5 for $\bar{f}_\infty$.

The proofs of the other three claims are completely similar. In fact, the proofs of the statement of the theorem for $\bar{f}_0$ and f_∞ are even easier, since the domains which are not punctured in the corresponding Heegaard triple are all bigons, rectangles or triangles. We thus only need to use the second part of the above argument in these two cases (and the study of the neighbourhood of the hexagon is not needed). The proof of the claim for f_0 requires some more serious modification, which will be outlined below.

Note that the Heegaard triples $(\Sigma,\boldsymbol{\alpha},\boldsymbol{\beta}_1,\boldsymbol{\beta}_\infty;u,v,w)$ and

$$(\Sigma,\boldsymbol{\alpha},\boldsymbol{\beta}'=\widehat{\boldsymbol{\beta}}\cup\{\lambda_1\},\boldsymbol{\gamma}'=\{\beta_1^\infty,\gamma_2,\dots,\gamma_{g-1},\lambda_\infty\};u,v,w)$$

may be identified using a diffeomorphism of the surface Σ. It is thus enough to show that $f_0=\phi(\Sigma,\boldsymbol{\alpha},\boldsymbol{\beta}',\boldsymbol{\gamma}';u,v,w)$. This allows us to keep the same labelling for the points X_i, Y_i, $i=2,\dots,g-2$. For the intersection points on $\gamma_1'=\beta_1^\infty$ and β_1 as well as some of the intersection points on λ_∞ and λ_1, we use the labelling of Figure 7. We abuse the notation and denote the two intersection points between β_1 and γ_1' by X_1 and Y_1. Moreover, some of the regions in the neighbourhood of X_1 and Y_1 are labelled: again by abuse of notation, we denote these regions by D_1, R_1, L_1, S_1 and T_1 (see Figure 7). Let us use d_i, r_i, s_i, t_i and l_i to denote the coefficients of the domains D_i, R_i, S_i, T_i and L_i in the 2–chain $\mathcal{D}$ associated with a holomorphic triangle connecting $\boldsymbol{x}=(x_1,\dots,x_g)$, $\boldsymbol{y}=(y_1,\dots,y_g)$ and Θ that contributes to

$\phi(\Sigma,\boldsymbol{\alpha},\boldsymbol{\beta}',\boldsymbol{\gamma}';u,v,w)$. We assume that, for some elements σ, $\tau\in S_g$,

$$x_i\in\begin{cases}\beta_i\cap\alpha_{\sigma(i)} & \text{if } i=1,\dots,g-1,\\ \lambda_1\cap\alpha_{\sigma(g)} & \text{if } i=g,\end{cases}$$
$$y_i\in\begin{cases}\gamma_i\cap\alpha_{\tau(i)} & \text{if } i=2,\dots,g-1,\\ \gamma_1'\cap\alpha_{\tau(1)} & \text{if } i=1,\\ \lambda_0\cap\alpha_{\tau(g)} & \text{if } i=g.\end{cases}$$

The examination of the coefficients in Figure 7 implies the following:

- We have $d_1=r_1=t_1=1$ and $s_1=l_1=0$.
- Either $x_g=r_2$ and $y_1=t_1$, or $x_g=r_3$ and $y_1=t_2$.
- There are j, $k\in\{3,4,\dots,n\}$ such that $y_g=q_k$ and $x_1=p_j$.

Let us write $\mathcal{D}=\mathcal{D}_s+\mathcal{D}_l$, where

$$\mathcal{D}_s:=s_1S_1+\sum_{i=1}^{g-1}(d_iD_i+t_iT_i+r_iR_i+l_iL_i)\quad\text{and}\quad\mathcal{D}_l:=\mathcal{D}-\mathcal{D}_s.$$

Considering the local coefficients at X_i and Y_i, we find $r_i=l_i+1$ and $d_i=t_i+l_i$. Applying the index formula in (8) we obtain

$$\begin{aligned}(12)\quad 0&=e(\mathcal{D})+\mu_{\boldsymbol{x}}(\mathcal{D})+\mu_{\boldsymbol{y}}(\mathcal{D})+\mathfrak{b}(\mathcal{D}).\mathfrak{c}(\mathcal{D})-\tfrac{1}{2}g\\ &=\big(e(\mathcal{D}_l)+\big(-\tfrac12+\tfrac14+\tfrac14\big)+\tfrac14(g-2)\big)+\mu_{\boldsymbol{x}}(\mathcal{D})+\mu_{\boldsymbol{y}}(\mathcal{D})+\mathfrak{b}(\mathcal{D}).\mathfrak{c}(\mathcal{D})-\tfrac12 g\\ &\geq\mu_{\boldsymbol{x}}(\mathcal{D})+\mu_{\boldsymbol{y}}(\mathcal{D})+\mathfrak{b}(\mathcal{D}).\mathfrak{c}(\mathcal{D})-\tfrac14(g+2).\end{aligned}$$

The 1–chains $\mathfrak{b}(\mathcal{D})$ and $\mathfrak{c}(\mathcal{D})$ may be written as

$$\mathfrak{b}(\mathcal{D})=\sum_{i=1}^{g}\mathfrak{b}_i(\mathcal{D})\quad\text{and}\quad\mathfrak{c}(\mathcal{D})=\sum_{i=1}^{g}\mathfrak{c}_i(\mathcal{D})$$

as before. Note that $\mathfrak{b}_1(\mathcal{D})$ is the arc on β_1 from X_1 to p_j, while $\mathfrak{c}_1(\mathcal{D})$ is the arc from one of t_1 or t_2 to X_1. Moreover, $\mathfrak{b}_g(\mathcal{D})$ is the arc on λ_1 from Q to one of r_2 or r_3, while $\mathfrak{c}_g(\mathcal{D})$ is the arc on λ_∞ from q_i to Q. Thus,

$$(13)\quad \mathfrak{b}(\mathcal{D}).\mathfrak{c}(\mathcal{D})=\big(-\tfrac14+\tfrac14\big)+\sum_{i=2}^{g-1}\big(\big(l_i+\tfrac12\big)\big(t_i-\tfrac12\big)-l_it_i\big)=-\tfrac14(g-2)+\tfrac12\sum_{i=2}^{g-1}(t_i-l_i).$$

Let us now assume that the local coefficients around x_i are a_i, b_i, b_i+l_i+1 and a_i+l_i, while the local coefficients around y_i are c_i, e_i, e_i+t_i-1 and c_i+t_i for $i=2,\dots,g-1$. The corresponding local coefficients around x_1, y_1, x_g and y_g would be (a_1, b_1, b_1+1, a_1), (c_1, e_1, e_1, c_1+1), $(0,0,1,0)$ and $(0,0,0,1)$, respectively, for some non-negative integers a_i, b_i, c_i and e_i, $i=1,\dots,g-1$. Thus,

$$\mu_{\boldsymbol{x}}(\mathcal{D}) + \mu_{\boldsymbol{y}}(\mathcal{D}) = \tfrac{1}{2} + \tfrac{1}{2}\sum_{i=1}^{g-1}((a_i+b_i+c_i+e_i)+(l_i+t_i)) \tag{14}$$

If we combine (12), (13) and (14), we find

$$0 \geq -\tfrac{1}{2}g + \left(\tfrac{1}{2}\right) + \tfrac{1}{2}\sum_{i=1}^{g-1}(a_i+b_i+c_i+e_i+2t_i) = \tfrac{1}{2}\sum_{i=1}^{g-1}(a_i+b_i+c_i+(e_i+t_i-1)+t_i).$$

As in the proof of the theorem for $\bar{f}_\infty$, this implies that $a_i=b_i=c_i=t_i=0$, while $e_i=1$ for $i=1,\dots,g-1$. It is easy to see from here that $j=k$ and complete the proof as before. □

4.4 The maps $\theta(K)$ and $\bar{\theta}(K)$

Let $\mathbb{H}_\bullet$ denote the homology of the chain complex $C_\bullet$ for $\bullet \in \{\infty, 1, 0\}$. If we choose a representative $a \in C_0$ of a class

$$[a] \in \mathrm{Ker}((f_\infty)_*) \subset \mathbb{H}_0,$$

there exists some $b \in C_1$ such that $f_\infty(a) = d_1(b)$. Then $d_\infty(f_0(b)) = f_0(d_1(b)) = f_0(f_\infty(a)) = 0$, so $f_0(b)$ is closed and represents a class in $\mathbb{H}_\infty$. If we replace b with another element $b' = b + \Delta b$ such that $d_1(b') = f_\infty(a)$, Δb is closed (ie it represents an element in $\mathbb{H}_1$). The difference $f_0(b') - f_0(b) = f_0(\Delta b)$ is an element in $\mathrm{Im}((f_0)_*)$. Thus, the class

$$\theta([a]) = [f_0(b)] \in \mathrm{Coker}((f_0)_*)$$

is well defined. This gives a homomorphism

$$\theta = \theta(K)\colon \mathrm{Ker}((f_\infty)_*) \longrightarrow \mathrm{Coker}((f_0)_*).$$

Similarly, we define the map $\bar{\theta} = \bar{\theta}(K)\colon \mathrm{Ker}((\bar{f}_\infty)_*) \to \mathrm{Coker}((\bar{f}_0)_*)$ from

$$\bar{f}_\infty\colon C_0 \longrightarrow \mathbb{C}_1 \quad \text{and} \quad \bar{f}_0\colon C_1 \longrightarrow C_\infty.$$

Proposition 4.6 *The maps*

$$\theta(K)\colon \mathrm{Ker}(\mathfrak{f}_\infty(K)) \longrightarrow \mathrm{Coker}(\mathfrak{f}_0(K)) \quad \textit{and} \quad \bar{\theta}(K)\colon \mathrm{Ker}(\bar{\mathfrak{f}}_\infty(K)) \longrightarrow \mathrm{Coker}(\bar{\mathfrak{f}}_0(K))$$

are the inverses of the maps induced by $\mathfrak{f}_1(K)$, $\bar{\mathfrak{f}}_1(K)\colon \mathbb{H}_\infty(K) \to \mathbb{H}_0(K)$ *which sit in the exact sequences:*

$$\begin{tikzcd} \mathbb{H}_0(K) \arrow[dr, "\mathfrak{f}_\infty(K)"'] & & \mathbb{H}_\infty(K) \arrow[ll, "\mathfrak{f}_1(K)"'] \\ & \mathbb{H}_1(K) \arrow[ur, "\mathfrak{f}_0(K)"'] & \end{tikzcd} \quad \textit{and} \quad \begin{tikzcd} \mathbb{H}_0(K) \arrow[dr, "\bar{\mathfrak{f}}_\infty(K)"'] & & \mathbb{H}_\infty(K) \arrow[ll, "\bar{\mathfrak{f}}_1(K)"'] \\ & \mathbb{H}_1(K) \arrow[ur, "\bar{\mathfrak{f}}_0(K)"'] & \end{tikzcd}$$

Proof For this purpose, let us assume that the Heegaard diagram

$$(\Sigma, \boldsymbol{\alpha}, \widehat{\boldsymbol{\beta}}, \{\lambda_0, \lambda_1, \lambda_\infty\}; u, v, w, \bar{u}, \bar{v}, \bar{w})$$

is constructed from a special Heegaard diagram as before. Let $\boldsymbol{\beta}_\bullet$ for $\bullet \in \{0, 1, \infty\}$ denote the set $\boldsymbol{\beta}_\bullet = \{\beta_1^\bullet, \dots, \beta_{g-1}^\bullet, \lambda_\bullet\}$ constructed before. Let us furthermore assume that $\boldsymbol{\beta}_1'$ is a set of g simple closed curves, where the first $g-1$ of them are small Hamiltonian isotopes of the first $g-1$ curves in $\boldsymbol{\beta}_1$ (with two transverse intersection points with the corresponding simple closed curve in $\boldsymbol{\beta}_1$) while the last (g^{th}) curve is denoted by λ_1'. We assume that λ_1' is a Hamiltonian isotope of λ_1, which is very close to the juxtaposition of the curves λ_0 and λ_∞.

Consider the two Heegaard quadruples

$$H = (\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \boldsymbol{\beta}_\infty; u, v, w) \quad \text{and} \quad H' = (\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}_0, \boldsymbol{\beta}_1', \boldsymbol{\beta}_\infty; u, v, w).$$

Let us denote the triangle maps associated with the first Heegaard diagram by

$$\mathfrak{f}_0(H) = \phi(H \setminus \boldsymbol{\beta}_0)\colon C_1(K; H) \longrightarrow C_\infty(K; H),$$
$$\mathfrak{f}_\infty(H) = \phi(H \setminus \boldsymbol{\beta}_\infty)\colon C_0(K; H) \longrightarrow C_1(K; H),$$

while the triangle maps associated with the Heegaard quadruple H' are denoted by

$$\mathfrak{f}_0(H') = \phi(H' \setminus \boldsymbol{\beta}_0)\colon C_1(K; H') \longrightarrow C_\infty(K; H') = C_\infty(K; H),$$
$$\mathfrak{f}_\infty(H') = \phi(H' \setminus \boldsymbol{\beta}_\infty)\colon C_0(K; H') = C_0(K; H) \longrightarrow C_1(K; H').$$

The holomorphic triangle map $\mathfrak{f}_1(H) = \mathfrak{f}_1(H')\colon C_\infty(K; H) \to C_0(K; H)$ may be defined using the Heegaard triple $(\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}_\infty, \boldsymbol{\beta}_0; u, v, w)$. Count of the holomorphic rectangles in H and H', respectively, that avoid the punctures u, v and w gives the homomorphisms

$$\Phi_1\colon C_0(K; H) \longrightarrow C_\infty(K; H) \quad \text{and} \quad \Phi_1'\colon C_0(K; H) \longrightarrow C_\infty(K; H)$$

such that

$$d_\infty \circ \Phi_1 + \Phi_1 \circ d_0 = \mathfrak{f}_0(H) \circ \mathfrak{f}_\infty(H) \quad \text{and} \quad d_\infty \circ \Phi_1' + \Phi_1' \circ d_0 = \mathfrak{f}_0(H') \circ \mathfrak{f}_\infty(H').$$

The interesting observation is that both Φ_1 and Φ_1' vanish when the Heegaard diagram is chosen as above. The reason for the first vanishing result is that there are no positive squares connecting the four intersection points

$$\boldsymbol{x} \in \mathbb{T}_\alpha \cap \mathbb{T}_{\beta_0}, \quad \Theta_{0,1} \in \mathbb{T}_{\beta_0} \cap \mathbb{T}_{\beta_1}, \quad \Theta_{1,\infty} \in \mathbb{T}_{\beta_1} \cap \mathbb{T}_{\beta_\infty} \quad \text{and} \quad \boldsymbol{y} \in \mathbb{T}_{\beta_\infty} \cap \mathbb{T}_\alpha.$$

In fact, $n_{\bar{u}}(\square) = n_{\overline{w}}(\square) = 1$ for every square class $\square \in \pi_2^+(\boldsymbol{x}, \Theta_{0,1}, \Theta_{1,\infty}, \boldsymbol{y})$. Thus, two opposite quarters around the intersection point r_1 have zero coefficient, while one other quadrant has coefficient 1. Since r_1 is not among the intersection points in any of $\boldsymbol{x}$, $\boldsymbol{y}$, $\Theta_{0,1}$ and $\Theta_{1,\infty}$, the coefficient of the last quadrant around r_1 is -1 and the contribution of $\square$ is thus trivial. A similar argument implies that Φ_1' is zero.

For $\bullet \in \{0, 1, \infty\}$, the Heegaard triple $H_\bullet = (\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}_\bullet', \boldsymbol{\beta}_\bullet; u, v, w)$ gives

$$\imath_\bullet = \imath(H_\bullet)\colon C_\bullet(K; H') \longrightarrow C_\bullet(K; H).$$

The homomorphisms $\imath_0$ and $\imath_\infty$ are the identity maps of $C_0(K; H)$ and $C_\infty(K; H)$, respectively. The Heegaard quadruple

$$(\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}_0, \boldsymbol{\beta}_1', \boldsymbol{\beta}_1; u, v, w)$$

determines a holomorphic square map

$$\Psi_\infty\colon C_0(K; H) \longrightarrow C_1(K; H).$$

Considering different possible degenerations of a holomorphic square of Maslov index zero, one finds the relation

$$d_1 \circ \Psi_\infty + \Psi_\infty \circ d_0 = \imath_1 \circ \mathfrak{f}_\infty(H') + \mathfrak{f}_\infty(H). \tag{15}$$

Finally, one may consider the Heegaard 5–tuple

$$(\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}_0, \boldsymbol{\beta}_1', \boldsymbol{\beta}_1, \boldsymbol{\beta}_\infty; u, v, w),$$

which may be used to construct a pentagon map $Q\colon C_0(K; H) \to C_\infty(K; H)$. Considering all possible degenerations of a holomorphic pentagon of Maslov index -1, one obtains the relation

$$d_\infty \circ Q + Q \circ d_0 = \Psi_0 \circ \mathfrak{f}_\infty(H') + \mathfrak{f}_0(H) \circ \Psi_\infty, \tag{16}$$

where $\Psi_0\colon C_1(K; H') \to C_\infty(K; H)$ is the holomorphic square map associated with $(\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}_1', \boldsymbol{\beta}_1, \boldsymbol{\beta}_\infty; u, v, w)$. The reason for the above equality is that the contributing holomorphic squares in the Heegaard quadruple $(\Sigma, \boldsymbol{\beta}_0, \boldsymbol{\beta}_1', \boldsymbol{\beta}_1, \boldsymbol{\beta}_\infty; u, v, w)$ come in cancelling pairs, while there is a single contributing holomorphic triangle corresponding to each of the Heegaard triples

$$(\Sigma, \boldsymbol{\beta}_1', \boldsymbol{\beta}_1, \boldsymbol{\beta}_\infty; u, v, w) \quad \text{and} \quad (\Sigma, \boldsymbol{\beta}_0, \boldsymbol{\beta}_1', \boldsymbol{\beta}_1; u, v, w).$$

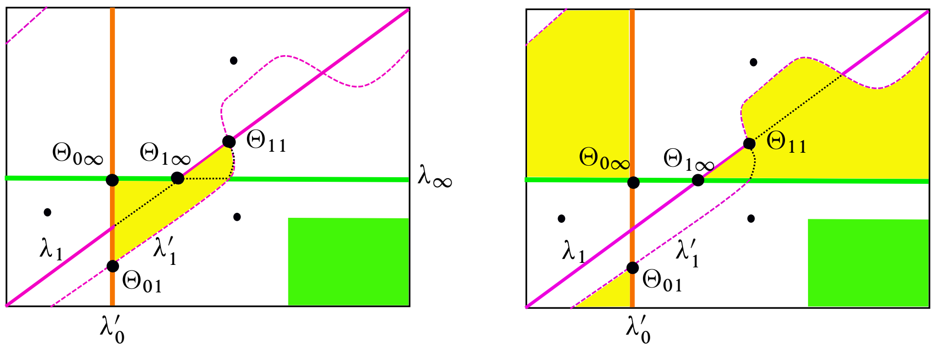


Figure 8: The domains for a cancelling pair of contributing squares connecting Θ_{01}, $\Theta_{1,1}$, $\Theta_{1,\infty}$ and $\Theta_{0,\infty}$

Figure 8 illustrates the domain for a cancelling pair of contributing squares. Moreover, the maps Φ_1 and Φ'_1, which may potentially contribute, are trivial.

Our choice of λ'_1 and the fact that the Heegaard diagram is nice imply that we have a short exact sequence

$$0 \longrightarrow C_0(K;H) \xrightarrow{\mathfrak{f}_\infty(H')} C_1(K;H') \xrightarrow{\mathfrak{f}_0(H')} C_\infty(K;H) \longrightarrow 0.$$

Correspondingly, an isomorphism $\theta'\colon \mathrm{Ker}(\mathfrak{f}_\infty(K)) \to \mathrm{Coker}(\mathfrak{f}_0(K))$ may be constructed. Choose some closed element $a \in C_0(K;H)$ and let $\mathfrak{f}_\infty(H')(a) = d'_1(b')$ for some $b' \in C_1(K;H')$. By (15),

$$\mathfrak{f}_\infty(H)(a) = (\imath_1 \circ \mathfrak{f}_\infty(H') + d_1 \circ \Psi_\infty)(a) = d_1(\imath_1(b') + \Psi_\infty(a)) =: d_1(b).$$

Using (15) and (16) we compute

$$\begin{aligned}
\mathfrak{f}_0(H)(b) &= \mathfrak{f}_0(H)(\imath_1(b') + \Psi_\infty(a)) \\
&= \mathfrak{f}_0(H')(b') + (d_\infty \circ \Psi_0 + \Psi_0 \circ d'_1)(b') + (\mathfrak{f}_0(H) \circ \Psi_\infty)(a) && \text{by (15)} \\
&= \mathfrak{f}_0(H')(b') + d_\infty \circ \Psi_0(b') + (\Psi_0 \circ \mathfrak{f}_\infty(H') + \mathfrak{f}_0(H) \circ \Psi_\infty)(a) \\
&= \mathfrak{f}_0(H')(b') + d_\infty(Q(a) + \Psi_0(b')) && \text{by (16).}
\end{aligned}$$

This means that the maps $\theta(K)$ and θ', as maps from $\mathrm{Ker}(\mathfrak{f}_\infty(K))$ to $\mathrm{Coker}(\mathfrak{f}_0(K))$, are the same. However, the map θ' is the inverse of the connecting homomorphism $\delta\colon \mathrm{Coker}(\mathfrak{f}_0(K)) \to \mathrm{Ker}(\mathfrak{f}_\infty(K))$ resulting from the short exact sequence

$$0 \longrightarrow C_0(K;H) \xrightarrow{\mathfrak{f}_\infty(H')} C_1(K;H') \xrightarrow{\mathfrak{f}_0(H')} C_\infty(K;H) \longrightarrow 0.$$

The above observations imply the claim for $\theta(K)$. The proof for the map $\bar{\theta}(K)$ is similarly reduced to showing the triviality of the holomorphic square map corresponding to the Heegaard quadruple $(\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \boldsymbol{\beta}_\infty; \bar{u}, \bar{v}, \bar{w})$.

The domain of every contributing holomorphic square corresponding to the aforementioned punctured Heegaard diagram has coefficient zero at $\bar{u}$, $\bar{v}$, $\bar{w}$ and w, and coefficient 1 at u and v. This implies that two opposite quadrants around r_3 have coefficient zero, while a third quadrant has coefficient 1. Since r_3 cannot be among the intersection points on the vertices of the square, the fourth quadrant around r_3 has coefficient -1. This contradiction gives the triviality of the holomorphic square map corresponding to $(\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \boldsymbol{\beta}_\infty; \bar{u}, \bar{v}, \bar{w})$ and completes the proof. □

5 Gluing the knot complements

5.1 Extracting a chain complex for splicing

Given two Heegaard diagrams for the complements of the knots K_1 and K_2, one may construct a Heegaard diagram for $Y(K_1, K_2)$ as follows, similar to the construction of Eftekhary [2]. Let

$$H_i = (\Sigma_i, \boldsymbol{\alpha}^i, \widehat{\boldsymbol{\beta}}^i \cup \{\mu_i, \lambda_i\})$$

denote the Heegaard diagram for K_i with Heegaard surface Σ_i, and with μ_i the meridian for K_i and λ_i a zero-framed longitude for it which cuts μ_i in a single point O_i. Then the Heegaard diagram for the three-manifold $Y = Y(K_1, K_2)$ obtained by splicing the complement of $K_1 \subset Y_1$ and the complement of $K_2 \subset Y_2$ is constructed as follows. Attach a 1–handle to $\Sigma_1 \cup \Sigma_2$, with attaching circles placed at the intersections O_1 and O_2. Use four parallel segments on this 1–handle to connect the four intersections of $\mu_1 \cup \lambda_1$ with one of the attaching circles to the four intersections of $\mu_2 \cup \lambda_2$ with the other attaching circle, so that intersection points on μ_1 are joined to the intersection points on λ_2. The union of the remaining parts from μ_1 and λ_2 with two of the four parallel line segments gives a simple closed curve on Σ, which will be denoted by $\mu_1 \# \lambda_2$. The simple closed curve $\lambda_1 \# \mu_2$ is constructed in a similar way. Let

$$\boldsymbol{\alpha} = \boldsymbol{\alpha}^1 \cup \boldsymbol{\alpha}^2 \quad \text{and} \quad \boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}^1 \cup \widehat{\boldsymbol{\beta}}^2 \cup \{\mu_1 \# \lambda_2, \lambda_1 \# \mu_2\}.$$

The resulting Heegaard diagram $H = (\Sigma, \boldsymbol{\alpha}, \boldsymbol{\beta})$ is a Heegaard diagram for the three-manifold obtained by splicing the two knot complements.

If the initial Heegaard diagrams H_i are special (see Definition 4.2) one may assume that the Heegaard diagram H will have one bad region and the rest of the regions are either bigons or rectangles. Thus, the combinatorial algorithm of Sarkar and Wang [16]

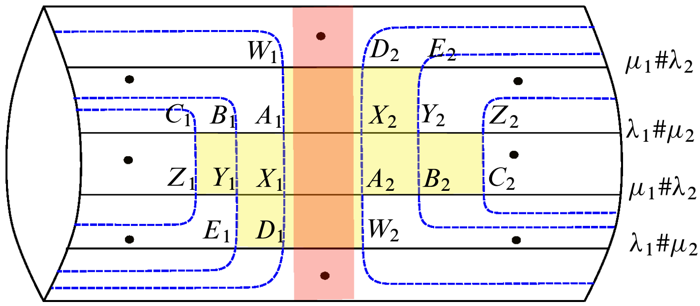


Figure 9: The cylinder illustrates a neighbourhood of the 1–handle used for attaching the two Heegaard diagrams. The union of the domains of the disks intersecting the 1–handle and contributing to the differential is shaded yellow.

may be used to compute its (hat) Heegaard Floer homology with $\mathbb{F}$ coefficients. Let z denote a marked point which is placed in the aforementioned bad region. The marked point z corresponds to the marked points $z_i \in \Sigma_i$, $i = 1, 2$. We may also choose a second marked point z_i' for the Heegaard diagram H_i which is placed next to O_i and in the quadrant opposite to the quadrant containing z_i.

Define the chain complexes M^i and L^i associated with $K_i \subset Y_i$ using the Heegaard diagrams

$$(\Sigma_i, \boldsymbol{\alpha}^i, \widehat{\boldsymbol{\beta}}^i \cup \{\mu_i\}; z_i, z_i') \quad \text{and} \quad (\Sigma_i, \boldsymbol{\alpha}^i, \widehat{\boldsymbol{\beta}}^i \cup \{\lambda_i\}; z_i, z_i'),$$

respectively. Note that the generators of the complex C associated with the Heegaard diagram H are in correspondence, either with the generators of $M = M^1 \otimes M^2$ or the generators of $L = L^1 \otimes L^2$, ie the $\mathbb{F}$–module C may be identified with $M \oplus L$. Denote the differential of M by d_M and the differential of L by d_L. The domain of every disk which contributes to the differential of C is then a rectangle or a bigon in the diagram. Such a disk may either stay in one of the Σ_i or intersect both Σ_1 and Σ_2. The disks that stay in one of the Σ_i correspond to the differentials d_M and d_L of the complexes M and L. Only a few rectangles can intersect both Σ_i and miss the marked point z (see Figure 9), while no bigons can intersect both Σ_1 and Σ_2. Because of the way the bad region (the region containing the marked point) enters the neighbourhood of the 1–handle, the rectangles which intersect both Σ_1 and Σ_2 stay in the neighbourhood of the 1–handle. The contribution of such rectangles may be described after introducing some extra notation.

The assumption on the Heegaard diagrams H_1 and H_2 from Lemma 4.1 implies that the local picture around O_i looks like the genus-4 surface illustrated on the top of Figure 3. Denote the intersection points on H_i which correspond to A, B, C, D, E, X, Y, Z and W by A_i, B_i, C_i, D_i, E_i, X_i, Y_i, Z_i and W_i, respectively.

The generators of $M^i \oplus L^i$ are the tuples $\boldsymbol{x} = (x_1, \ldots, x_{g_i})$ such that, for a permutation $\sigma\colon \{1, \ldots, g_i\} \to \{1, \ldots, g_i\}$, we have $x_j \in \alpha_{\sigma(j)} \cap \beta_j$ for $j = 1, \ldots, g_i - 1$ and $x_{g_i} \in \alpha_{\sigma(g_i)} \cap (\mu_i \cup \lambda_i)$. The complex M^i is generated by those $\boldsymbol{x}$ such that $x_{g_i} \in \mu_i$, and the complex L^i is generated by the g_i–tuples $\boldsymbol{x} = (x_1, \ldots, x_{g_i})$ with $x_{g_i} \in \lambda_i$. The homology of the complex M^i is the knot Floer homology $\widehat{\mathrm{HFK}}(K_i)$ and the homology of the complex L^i is the longitude Floer homology $\widehat{\mathrm{HFL}}(K_i)$. The homomorphisms $\Phi^i\colon M^i \to L^i$ over $\boldsymbol{x} = (x_1, \ldots, x_{g_i}) \in M^i$ are defined by

$$\Phi^i(\boldsymbol{x}) = \begin{cases} (x_1, \ldots, x_{g_i-1}, X_i) & \text{if } x_{g_i} = A_i, \\ (x_1, \ldots, x_{g_i-1}, Y_i) & \text{if } x_{g_i} = B_i, \\ (x_1, \ldots, x_{g_i-1}, Z_i) & \text{if } x_{g_i} = C_i, \\ 0 & \text{otherwise.} \end{cases}$$

The corresponding contributing triangles are $[A_i O_i X_i]$, $[B_i O_i Y_i]$ and $[C_i O_i Z_i]$. The map Φ thus corresponds to the changes $x_{g_i} \to y_{g_i}$ which are one of the following: $A_i \to X_i$, $B_i \to Y_i$ or $C_i \to Z_i$. Similarly, the homomorphisms $\Psi^i_1\colon L^i \to M^i$ correspond to the triangles $[W_i O_i A_i]$ and, over $\boldsymbol{x} = (x_1, \ldots, x_{g_i}) \in L^i$, are defined by

$$\Psi^i_1(\boldsymbol{x}) = \begin{cases} (x_1, \ldots, x_{g_i-1}, A_i) & \text{if } x_{g_i} = W_i, \\ 0 & \text{otherwise.} \end{cases}$$

Define the maps Ψ^i_2, $\Psi^i_3\colon L^i \to M^i$, where Ψ^i_2 corresponds to the changes $X_i \to D_i$ and $Y_i \to E_i$, and Ψ^i_3 corresponds to $W_i \to D_i$. Thus the triangles contributing to Ψ^i_2 are $[X_i O_i D_i]$ and $[Y_i O_i E_i]$, while the only triangle contributing to Ψ^i_3 is

$$[W_i O_i D_i] = [W_i O_i A_i] \cup [A_i O_i X_i] \cup [X_i O_i D_i].$$

The contribution of the rectangles which intersect both Σ_1 and Σ_2 to the differential of the complex $C = M \oplus L$ may thus be described by the maps

$$\Phi = \Phi^1 \otimes \Phi^2\colon L^1 \otimes L^2 \longrightarrow M^1 \otimes M^2,$$

$$\left.\begin{aligned} \Psi_1 &= \Psi^1_1 \otimes \Psi^2_2 \\ \Psi_2 &= \Psi^1_2 \otimes \Psi^2_1 \\ \Psi_3 &= \Psi^1_3 \otimes \Psi^2_3 \end{aligned}\right\}\colon M^1 \otimes M^2 \longrightarrow L^1 \otimes L^2.$$

In other words, the differential of the complex $C = M \oplus L$ is the homomorphism

$$d = d_C = \begin{pmatrix} d_M & \Phi \\ \sum_{i=1}^3 \Psi_i & d_L \end{pmatrix}.$$

Proposition 5.1 *The complexes M^i and L^i are identified with the mapping cones of $\bar{f}^i_\infty = \bar{f}_\infty(K_i)$ and $f^i_0 = f_0(K_i)$, respectively. More precisely, the $\mathbb{F}$–module M^i is isomorphic to the direct sum of $C_1(K_i)$ and $C_0(K_i)$, while L^i is isomorphic to the* *direct sum of $C_\infty(K_i)$ and $C_1(K_i)$. Moreover, the differentials d_{M^i} and d_{L^i} of M^i and L^i are identified as*

$$d_{L_i}(c_1,c_\infty)=(d_1^i(c_1),d_\infty^i(c_\infty)+f_0^i(c_1))\quad \textit{for all } (c_1,c_\infty)\in C_1(K_i)\oplus C_\infty(K_i),$$
$$d_{M_i}(c_0,c_1)=(d_0^i(c_0),d_1^i(c_1)+\bar{f}_\infty^i(c_0))\quad \textit{for all } (c_0,c_1)\in C_0(K_i)\oplus C_1(K_i).$$

Proof We sketch the proof of the claim for L^i. The corresponding claim for M^i is proved in a completely similar way. Consider the labelling of the intersection points of the α–curves with the curves $\lambda_\infty(K_i)$, $\lambda_1(K_i)$ and $\lambda(K_i)$ as in Figure 5. The intersection points with the α–curves on $\lambda_1(K_i)$ are r_1, r_2 and r_3. The intersection points with the α–curves on $\lambda_\infty(K_i)$ are $s_1,s_2,\dots,s_n,q_3,q_4,\dots,q_n$ and the intersection points with the α–curves on $\lambda(K_i)$ are $S_1,S_2,\dots,S_n,Q_3,Q_4,\dots,Q_n$ and R_1,R_2,R_3. Define the $\mathbb{F}$–module isomorphism

$$I_i\colon C_1(K_i)\oplus C_\infty(K_i)\longrightarrow L^i,\quad I_i(\boldsymbol{x}=(x_1,\dots,x_{g_i})):=(x_1,\dots,x_{g_i-1},I_i(x_{g_i})),$$

where I_i changes the letter in the labelling of an intersection point to a capital letter (so $I_i(r_j)=R_j$, $I_i(s_j)=S_j$ and $I_i(q_j)=Q_j$). Straightforward combinatorics may be used to verify $d_{L^i}(I_i(\boldsymbol{x}))=I_i(d_\infty^i(\boldsymbol{x}))$ for every generator $\boldsymbol{x}$ of $C_\infty(K_i)$ and $d_{L^i}(I_i(\boldsymbol{x}))=I_i(d_1^i(\boldsymbol{x}))+I_i(f_0^i(\boldsymbol{x}))$ for every generator $\boldsymbol{x}$ of $C_1(K_i)$. □

Under the identification of M^i with the mapping cone of $\bar{f}_\infty^i$ and the identification of L^i with the mapping cone of f_0^i, the map Φ has a simple description: it is the map that takes $C_1(K_i)$ in the mapping cone of $f_0\colon C_1(K_i)\to C_\infty(K_i)$ to the complex $C_1(K_i)$ in the mapping cone of $\bar{f}_\infty^i\colon C_0(K_i)\to C_1(K_i)$ via the identity map of $C_1(K_i)$. Furthermore, the map f_∞^i from $C_0(K_i)$ in M^i to $C_1(K_i)$ in L^i is identified with the triangle map Ψ_1^i. The induced map $\bar{f}_0^i$ from the copy of $C_1(K_i)$ in M to the copy of $C_\infty(K_i)$ in L^i is the triangle map Ψ_2^i. The map Ψ_3^i is obtained from the composition map $\bar{f}_0^i\circ f_\infty^i\colon C_0(K_i)\to C_\infty(K_i)$. Set $C_\bullet^i=C_\bullet(K_i)$. If we replace the mapping cone of $f_0^i\colon C_1^i\to C_\infty^i$ for L^i, replace the mapping cone $\bar{f}_\infty^i\colon C_0^i\to C_1^i$ for M^i, and also replace Φ^i and Ψ_j^i with the appropriate descriptions in terms of $\bar{f}_0^i$ and f_∞^i, we obtain an alternative description of the complex C.

The *cube* $\square=\square(f_\bullet^i,\bar{f}_\bullet^i\mid \bullet=0,\infty,\ i=1,2)$ associated with the knots K_1 and K_2, the corresponding complexes $C_\bullet^i$, $i=1,2$, $\bullet\in\{0,1,\infty\}$, and the maps f_0^i, $\bar{f}_0^i\colon C_1^i\to C_\infty^i$ and f_∞^i, $\bar{f}_\infty^i\colon C_0^i\to C_1^i$ is the chain complex $(\square,d_\square)$ associated with the graph of complexes represented by the following cube:

$$
\begin{tikzcd}
C^1_\infty\otimes C^2_\infty & & C^1_\infty\otimes C^2_1 \arrow[ll, "I\otimes f_0^2"'] & \\
& C^1_1\otimes C^2_\infty \arrow[ul, "f_0^1\otimes I"'] & & C^1_1\otimes C^2_1 \arrow[ul, "f_0^1\otimes I"'] \arrow[ll, "I\otimes f_0^2"'] \arrow[dd, "I"] \\
C^1_0\otimes C^2_0 \arrow[uu, "\bar f_0^1 f_\infty^1\otimes \bar f_0^2 f_\infty^2"] \arrow[rr, "\bar f^1_\infty\otimes I"] \arrow[dr, "I\otimes \bar f^2_\infty"'] & & C^1_1\otimes C^2_0 \arrow[uu, "\bar f_0^1\otimes f_\infty^2"'] \arrow[dr, "I\otimes \bar f^2_\infty"'] & \\
& C^1_0\otimes C^2_1 \arrow[uu, "f^1_\infty\otimes \bar f^2_0"] \arrow[rr, "\bar f^1_\infty\otimes I"'] & & C^1_1\otimes C^2_1
\end{tikzcd}
\tag{17}
$$

Proposition 5.2 *With the above notation fixed, the complex (C,d) is identified, as a chain complex, with the cube*

$$\big(\boxed{} = \boxed{}(f^i_\bullet, \bar f^i_\bullet \mid \bullet = 0,\infty,\ i=1,2), d_{\boxed{}}\big).$$

5.2 The linear algebra of the cubes

Let $\mathbb{H}^i_\bullet$ denote the homology of the chain complex $(C^i_\bullet, d^i_\bullet)$ for $i=1,2,\ \bullet\in\{0,1,\infty\}$. Set $\mathbb{H}_{\bullet,\star} = \mathbb{H}^1_\bullet\otimes\mathbb{H}^2_\star$ for $\bullet,\star\in\{0,1,\infty\}$. Abusing the notation, the map induced on homology by $f^i_\bullet$ will also be denoted $\mathfrak{f}^i_\bullet$ and the map induced on homology by $\bar f^i_\bullet$ will be denoted $\bar{\mathfrak{f}}^i_\bullet$.

Following the discussion of Section 5.3, we may choose appropriate decompositions $C^i_\bullet = A^i_\bullet\oplus\mathbb{H}^i_\bullet\oplus A^i_\bullet$ such that the differential $d^i_\bullet$ takes the form

$$d^i_\bullet = \begin{pmatrix} 0 & 0 & I\\ 0 & 0 & 0\\ 0 & 0 & 0\end{pmatrix}.$$

Correspondingly, we find the matrices $G(f^i_\bullet) = (f^i_\bullet)_*$ and $G(\bar f^i_\bullet) = (\bar f^i_\bullet)_*$, which will be denoted by $\mathfrak{f}^i_\bullet$ and $\bar{\mathfrak{f}}^i_\bullet$, as well as the matrices

$$M(f^i_\bullet),\quad M(\bar f^i_\bullet),\quad P(f^i_\bullet),\quad P(\bar f^i_\bullet),\quad Q(f^i_\bullet),\quad Q(\bar f^i_\bullet),\quad N(f^i_\bullet)\quad\text{and}\quad N(\bar f^i_\bullet).$$

The maps $Q(\mathfrak{f}^i_0)P(\mathfrak{f}^i_\infty)$ and $Q(\bar{\mathfrak{f}}^i_0)P(\bar{\mathfrak{f}}^i_\infty)$ from $\mathbb{H}^i_0$ to $\mathbb{H}^i_\infty$ extend the homomorphisms

$$\theta^i\colon \mathrm{Ker}(\mathfrak{f}^i_\infty)\longrightarrow \mathrm{Coker}(\mathfrak{f}^i_0)\quad\text{and}\quad \bar\theta^i\colon \mathrm{Ker}(\bar{\mathfrak{f}}^i_\infty)\longrightarrow \mathrm{Coker}(\bar{\mathfrak{f}}^i_0),$$

associated with the knot $K_i \subset Y_i$. These extensions are still denoted by θ^i and $\bar{\theta}^i$, respectively.

Lemma 2.4 implies that the homology of $(\square, d_{\square})$ is isomorphic to the homology of the chain complex $(\mathbb{H}, d_{\mathbb{H}})$ associated with the graph of chain complexes determined by the cube of Figure 2.

Proposition 5.3 *Let $(\mathbb{H}, d_{\mathbb{H}})$ denote the complex obtained from the cube $(\square, d_{\square})$ by applying Lemma 2.4. Then $(\mathbb{H}, d_{\mathbb{H}})$ is identified with the complex shown in Figure 2 provided that the maps $\theta^i = \theta(K_i)$ and $\bar{\theta}^i = \bar{\theta}(K_i)$ are given as above.*

Proof If Lemma 2.4 is applied, we obtain the same oriented graph (ie the same new edges) and the same complexes on the vertices. The directed edge from $\mathbb{H}_{0,0}$ to $\mathbb{H}_{\infty,\infty}$ is labelled by the map

$$\bar{\mathfrak{f}}_0^1\mathfrak{f}_\infty^1 \otimes \bar{\mathfrak{f}}_0^2\mathfrak{f}_\infty^2 + Q(f_0^1 \otimes I)N(f_\infty^1 \otimes \bar{f}_0^2)P(I \otimes \bar{f}_\infty^2) + Q(I \otimes f_0^2)N(\bar{f}_0^1 \otimes f_\infty^2)P(\bar{f}_\infty^1 \otimes I),$$

which is, by Lemma 2.6, equal to

$$\bar{\mathfrak{f}}_0^1\mathfrak{f}_\infty^1 \otimes \bar{\mathfrak{f}}_0^2\mathfrak{f}_\infty^2 + \theta^1 \otimes \bar{\theta}^2 + \bar{\theta}^1 \otimes \theta^2.$$

The map corresponding to the dashed edge from $\mathbb{H}_{0,0}$ to $\mathbb{H}_{1,\infty}$ is, by Lemma 2.5,

$$Q(f_\infty^1 \otimes \bar{f}_0^2)P(I \otimes \bar{f}_\infty^2) = \mathfrak{f}_\infty^1 \otimes (Q(\bar{f}_0^2)P(\bar{f}_\infty^2)) = \mathfrak{f}_\infty^1 \otimes \bar{\theta}^2.$$

The maps corresponding to the rest of dashed directed edges may be computed in a completely similar way. This completes the proof of Proposition 5.3. □

Remark 5.4 (1) Note that $Y(K_1, K_2) = Y(-K_1, -K_2)$. One may assume that $\mathfrak{f}_\bullet(-K) = \bar{\mathfrak{f}}_\bullet(K)$ and $\bar{\mathfrak{f}}_\bullet(-K) = \mathfrak{f}_\bullet(K)$, implying that $\widehat{\mathrm{HF}}(Y(-K_1, -K_2))$ is isomorphic to the homology of the complex determined by the oriented graph in Figure 2, where all barred maps change to the corresponding unbarred maps and all unbarred maps change to the corresponding barred maps.

(2) Proposition 5.3 is still weaker than Theorem 1.1, since the extensions of θ^i and $\bar{\theta}^i$ to maps from $\mathbb{H}_0^i$ to $\mathbb{H}_\infty^i$ are not arbitrary yet. In fact, without freedom in choosing these two extensions (which will be proved by the end of the current section) Theorem 1.1 stays bound to the information from the corresponding nice Heegaard diagram and has much less significance.

5.3 Simplifications of the splicing formula

We now apply Lemma 2.3 to the splicing formula of Proposition 5.3 and make some cancellations. The first cancellation comes from setting $C=\mathbb{H}$, $A=\mathbb{H}_{1,1}$ and

$$B=(\mathbb{H}_{\infty,\infty}\oplus\mathbb{H}_{1,\infty}\oplus\mathbb{H}_{\infty,1})\oplus(\mathbb{H}_{0,1}\oplus\mathbb{H}_{1,0}\oplus\mathbb{H}_{0,0})=\mathbb{E}_1\oplus\mathbb{E}_2.$$

We thus have $\widehat{\mathrm{HF}}(Y)=H_*(B,d_B)$, where

$$d_B=\begin{pmatrix}0 & \mathfrak{f}_0^1\otimes I & I\otimes\mathfrak{f}_0^2 & \theta^1\otimes\bar{\mathfrak{f}}_0^2 & \bar{\mathfrak{f}}_0^1\otimes\theta^2 & \Gamma\\ 0 & 0 & 0 & \Phi & I\otimes(\mathfrak{f}_0^2\circ\bar{\mathfrak{f}}_\infty^2) & \mathfrak{f}_\infty^1\otimes\bar{\theta}^2\\ 0 & 0 & 0 & (\mathfrak{f}_0^1\circ\bar{\mathfrak{f}}_\infty^1)\otimes I & \Psi & \bar{\theta}^1\otimes\mathfrak{f}_\infty^2\\ 0 & 0 & 0 & 0 & 0 & I\otimes\bar{\mathfrak{f}}_\infty^2\\ 0 & 0 & 0 & 0 & 0 & \bar{\mathfrak{f}}_\infty^1\otimes I\\ 0 & 0 & 0 & 0 & 0 & 0\end{pmatrix}$$

with $\Gamma=(\bar{\mathfrak{f}}_0^1\circ\mathfrak{f}_\infty^1)\otimes(\bar{\mathfrak{f}}_0^2\circ\mathfrak{f}_\infty^2)+\theta^1\otimes\bar{\theta}^2+\bar{\theta}^1\otimes\theta^1$, $\Phi=\bar{\mathfrak{f}}_\infty^1\otimes\mathfrak{f}_0^2+\mathfrak{f}_\infty^1\otimes\bar{\mathfrak{f}}_0^2$ and $\Psi=\mathfrak{f}_0^1\otimes\bar{\mathfrak{f}}_\infty^2+\bar{\mathfrak{f}}_0^1\otimes\mathfrak{f}_\infty^2$.

The dimension of the $\mathbb{F}$–vector space $H_*(B,d_B)$ only depends on the rank of the kernel and the cokernel of the matrix d_B. Define a pair of matrices M_1 and M_2 to be equivalent if $\mathrm{Ker}(M_1)\simeq\mathrm{Ker}(M_2)$ and $\mathrm{Coker}(M_1)\simeq\mathrm{Coker}(M_2)$. For a matrix M, let

$$\imath(M):=\mathrm{Ker}(M)\oplus\mathrm{Coker}(M)\quad\text{and}\quad i(M):=\mathrm{rnk}(\imath(M)).$$

If M_1 and M_2 are equivalent matrices then $\imath(M_1)\simeq\imath(M_2)$ and $i(M_1)=i(M_2)$.

We make a change of basis for $\mathbb{E}_2$ which is given by the matrix

$$\begin{pmatrix}\tau_0(K_1)\otimes\tau_1(K_2) & 0 & 0\\ 0 & \tau_1(K_1)\otimes\tau_0(K_2) & 0\\ 0 & 0 & \tau_0(K_1)\otimes\tau_0(K_2)\end{pmatrix}.$$

The matrix d_B is thus equivalent to the matrix

$$d_B'=\begin{pmatrix}0 & \mathfrak{f}_0^1\otimes I & I\otimes\mathfrak{f}_0^2 & \theta^1\tau_0^1\otimes\tau_\infty^2\mathfrak{f}_0^2 & \tau_\infty^1\mathfrak{f}_0^1\otimes\theta^2\tau_0^2 & \Gamma\\ 0 & 0 & 0 & \Phi & \tau_1^1\otimes\mathfrak{f}_0^2\tau_1^2\mathfrak{f}_\infty^2 & \mathfrak{f}_\infty^1\tau_0^1\otimes\bar{\theta}^2\tau_0^2\\ 0 & 0 & 0 & \mathfrak{f}_0^1\tau_1^1\mathfrak{f}_\infty^1\otimes\tau_1^1 & \Psi & \bar{\theta}^1\tau_0^1\otimes\mathfrak{f}_\infty^2\tau_0^2\\ 0 & 0 & 0 & 0 & 0 & I\otimes\mathfrak{f}_\infty^2\\ 0 & 0 & 0 & 0 & 0 & \mathfrak{f}_\infty^1\otimes I\\ 0 & 0 & 0 & 0 & 0 & 0\end{pmatrix}.$$

with

$$\begin{aligned}
\Phi &= \tau_1^1 \mathfrak{f}_\infty^1 \otimes \mathfrak{f}_0^2 \tau_1^2 + \mathfrak{f}_\infty^1 \tau_0^1 \otimes \tau_\infty^2 \mathfrak{f}_0^2, \\
\Psi &= \mathfrak{f}_0^1 \tau_1^1 \otimes \tau_1^2 \mathfrak{f}_\infty^2 + \tau_\infty^1 \mathfrak{f}_0^1 \otimes \mathfrak{f}_\infty^2 \tau_0^2, \\
\Gamma &= \tau_\infty^1 \mathfrak{f}_0^1 \bar{\mathfrak{f}}_\infty^1 \otimes \tau_\infty^2 \mathfrak{f}_0^2 \bar{\mathfrak{f}}_\infty^2 + \theta^1 \tau_0^1 \otimes \bar{\theta}^2 \tau_0^2 + \bar{\theta}^1 \tau_0^1 \otimes \theta^2 \tau_0^2.
\end{aligned}$$

Let us use the decompositions of (5) for K_1 and K_2 to obtain a 24×24 block decomposition of d'_B. Moreover, following the discussion at the end of Section 3.3 we may assume that, in the corresponding decompositions,

$$\theta^i = \begin{pmatrix} 0 & I \\ 0 & 0 \end{pmatrix} \quad \text{and} \quad \tau_\infty^i \bar{\theta}^i \tau_0^i = \begin{pmatrix} M^i & I \\ P^i M^i & P^i \end{pmatrix}.$$

Each entry in the above 6×6 decomposition for $d'_{\mathbb{B}}$ corresponds to a 4×4 submatrix of the aforementioned 24×24 decomposition. For instance, the $(1,4)$ entry $\theta^1 \tau_0^1 \otimes \tau_\infty^2 \mathfrak{f}_0^2$ corresponds to

$$\begin{aligned}
\begin{pmatrix} 0 & I \\ 0 & 0 \end{pmatrix} \begin{pmatrix} A_0^1 & B_0^1 \\ C_0^1 & D_0^1 \end{pmatrix} \otimes \begin{pmatrix} A_\infty^2 & B_\infty^2 \\ C_\infty^2 & D_\infty^2 \end{pmatrix} \begin{pmatrix} 0 & 0 \\ I & 0 \end{pmatrix} &= \begin{pmatrix} C_0^1 & D_0^1 \\ 0 & 0 \end{pmatrix} \otimes \begin{pmatrix} B_\infty^2 & 0 \\ D_\infty^2 & 0 \end{pmatrix} \\
&= \begin{pmatrix} C_0^1 \otimes B_\infty^2 & 0 & D_0^1 \otimes B_\infty^2 & 0 \\ C_0^1 \otimes D_\infty^2 & 0 & D_0^1 \otimes D_\infty^2 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.
\end{aligned}$$

For another instance, the $(3,5)$ entry corresponds to

$$\begin{aligned}
&\begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ A_1^1 \otimes B_1^2 & 0 & B_1^1 \otimes B_1^2 & 0 \\ A_1^1 \otimes D_1^2 & 0 & B_1^1 \otimes D_1^2 & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 & 0 & 0 \\ B_\infty^1 \otimes A_0^2 & B_\infty^1 \otimes B_0^2 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ D_\infty^1 \otimes A_0^2 & D_\infty^1 \otimes B_0^2 & 0 & 0 \end{pmatrix}. \\
&\qquad = \begin{pmatrix} 0 & 0 & 0 & 0 \\ B_\infty^1 \otimes A_0^2 & B_\infty^1 \otimes B_0^2 & 0 & 0 \\ A_1^1 \otimes B_1^2 & 0 & B_1^1 \otimes B_1^2 & 0 \\ A_1^1 \otimes D_1^2 + D_\infty^1 \otimes A_0^2 & D_\infty^1 \otimes B_0^2 & B_1^1 \otimes D_1^2 & 0 \end{pmatrix}.
\end{aligned}$$

The aforementioned 24×24 decomposition includes identity matrices as the entries determined by the following block coordinates:

$$(2,9), \quad (3,5), \quad (4,6), \quad (14,21), \quad (16,23) \quad \text{and} \quad (20,22).$$

We use the above six identity matrices for cancellation to obtain an equivalent matrix $d = \left(\begin{smallmatrix} 0 & D \\ 0 & 0 \end{smallmatrix}\right)$ over $\mathbb{B}_1 \oplus \mathbb{B}_2$, where $\mathbb{A}_{\bullet\star} = \mathbb{A}_\bullet(K_1) \otimes \mathbb{A}_\star(K_2)$ and

$$\mathbb{B}_1 = \mathbb{A}_{11} \oplus \mathbb{A}_{\infty 1} \oplus \mathbb{A}_{\infty 0} \oplus \mathbb{A}_{1\infty} \oplus \mathbb{A}_{0\infty} \oplus \mathbb{A}_{00},$$
$$\mathbb{B}_2 = \mathbb{A}_{\infty 0} \oplus \mathbb{A}_{10} \oplus \mathbb{A}_{\infty\infty} \oplus \mathbb{A}_{0\infty} \oplus \mathbb{A}_{01} \oplus \mathbb{A}_{11}.$$

Rearrange the rows and the columns of the matrix D so that D corresponds to the rows 11, 7, 8, 10, 12, 1 and the columns 19, 13, 15, 17, 18, 24 in the above 24×24 decomposition to obtain the matrix

$$\begin{pmatrix}
B_1^1\otimes B_1^2 & B_1^1\otimes A_1^2 & 0 & A_1^1\otimes B_1^2 & 0 & 0\\
0 & A_0^1\otimes B_\infty^2 & B_0^1\otimes B_\infty^2 & 0 & 0 & B_0^1\otimes(A_\infty^2+B_\infty^2P^2)\\
D_1^1\otimes B_1^2 & D_1^1\otimes A_1^2+A_0^1\otimes D_\infty^2 & B_0^1\otimes D_\infty^2 & C_1^1\otimes B_1^2 & 0 & B_0^1\otimes(C_\infty^2+D_\infty^2P^2)\\
0 & 0 & 0 & B_\infty^1\otimes A_0^2 & B_\infty^1\otimes B_0^2 & (A_\infty^1+B_\infty^1P^1)\otimes B_0^2\\
B_1^1\otimes D_1^2 & B_1^1\otimes C_1^2 & 0 & D_\infty^1\otimes A_0^2+A_1^1\otimes D_1^2 & D_\infty^1\otimes B_0^2 & (C_\infty^1+D_\infty^1P^1)\otimes B_0^2\\
0 & C_0^1\otimes B_\infty^2 & D_0^1\otimes B_\infty^2 & B_\infty^1\otimes C_0^2 & B_\infty^1\otimes D_0^2 & \Gamma
\end{pmatrix}$$

with $\Gamma = B_\infty^1 B_1^1 B_0^1 \otimes B_\infty^2 B_1^2 B_0^2 + (A_\infty^1 + B_\infty^1 P^1)\otimes D_0^2 + D_0^1 \otimes (A_\infty^2 + B_\infty^2 P^2)$.

This matrix is in turn equivalent to the matrix $\mathfrak{D} = \mathfrak{D}(K_1, K_2)$ below, which is obtained by adding $I \otimes P^2$ times the third column and $P^1 \otimes I$ times the fifth column to the last column of the above matrix:

$$\mathfrak{D} = \begin{pmatrix}
B_1^1\otimes B_1^2 & C_1^1\otimes A_1^2 & 0 & A_1^1\otimes B_1^2 & 0 & 0\\
0 & A_0^1\otimes B_\infty^2 & B_0^1\otimes B_\infty^2 & 0 & 0 & B_0^1\otimes A_\infty^2\\
D_1^1\otimes B_1^2 & D_1^1\otimes A_1^2+A_0^1\otimes D_\infty^2 & B_0^1\otimes D_\infty^2 & C_1^1\otimes B_1^2 & 0 & B_0^1\otimes C_\infty^2\\
0 & 0 & 0 & B_\infty^1\otimes A_0^2 & B_\infty^1\otimes B_0^2 & A_\infty^1\otimes B_0^2\\
B_1^1\otimes D_1^2 & B_1^1\otimes C_1^2 & 0 & D_\infty^1\otimes A_0^2+A_1^1\otimes D_1^2 & D_\infty^1\otimes B_0^2 & C_\infty^1\otimes B_0^2\\
0 & C_0^1\otimes B_\infty^2 & D_0^1\otimes B_\infty^2 & B_\infty^1\otimes C_0^2 & B_\infty^1\otimes D_0^2 & \Psi
\end{pmatrix},$$

where $\Psi = A_\infty^1 \otimes D_0^2 + D_0^1 \otimes A_\infty^2 + X^1 \otimes X^2$ and $X^i = X(K_i) = B_\infty^i B_1^i B_0^i$ for $i = 1, 2$.

Combining Proposition 5.3 with the above observations, we find:

Proposition 5.5 *Let $K_i \subset Y_i$, $i = 1, 2$, denote null-homologous knots and $Y(K_1, K_2)$ denote the three-manifold obtained by splicing the complement of K_1 with the complement of K_2. With the above definition of $\mathfrak{D}(K_1, K_2)$,*

$$\widehat{\mathrm{HF}}(Y(K_1, K_2), \mathbb{F}) \simeq \imath(\mathfrak{D}(K_1, K_2)).$$

Corollary 5.6 *The splicing formula of Proposition 5.3 is independent of the choice of extensions θ^i and $\bar\theta^i$.*

Proof The fact that the matrices P^i and M^i do not appear in the matrix $\mathfrak{D}(K_1, K_2)$ implies that the choice of the extensions θ^i, $\bar\theta^i$: $\mathbb{H}_0^i \to \mathbb{H}_\infty^i$ does not change the rank of the homology group in the splicing formula of Proposition 5.3 or Theorem 1.1. □

With the above corollary in place, the proof of Theorem 1.1 is now complete.

Definition 5.7 For a pair of knots $K_i \subset Y_i$, $i = 1, 2$, define

$$\begin{aligned}\chi(K_1, K_2) := (h_1(K_1) - h_\infty(K_1))(h_1(K_2) - h_\infty(K_2)) \\ - (h_1(K_1) - h_0(K_1))(h_1(K_2) - h_0(K_2)).\end{aligned}$$

Note that $\chi(K_1, K_2)$ is in fact the difference between the ranks of $\mathbb{B}_1 = \mathbb{B}_1(K_1, K_2)$ and $\mathbb{B}_2 = \mathbb{B}_2(K_1, K_2)$. In the corresponding $\mathbb{Z}/2\mathbb{Z}$–grading on $\mathbb{B}_1 \oplus \mathbb{B}_2$, $\chi(K_1, K_2)$ is thus the Euler characteristic of the chain complex $(\mathbb{B}_1 \oplus \mathbb{B}_2, d)$.

Corollary 5.8 *With the above notation fixed,*

$$\mathrm{rnk}\big(\widehat{\mathrm{HF}}(Y(K_1, K_2))\big) \geq |\chi(K_1, K_2)|.$$

Proof It is enough to note that

$$\chi(K_1, K_2) = \mathrm{rnk}\big(\mathrm{Ker}(\mathfrak{D}(K_1, K_2))\big) - \mathrm{rnk}\big(\mathrm{Coker}(\mathfrak{D}(K_1, K_2))\big). \qquad \square$$

Consider the matrices

$$P_L = \begin{pmatrix} I \otimes A_1^2 & 0 & 0 & 0 & I \otimes B_1^2 & 0 \\ 0 & I \otimes A_\infty^2 & I \otimes B_\infty^2 & 0 & 0 & 0 \\ 0 & I \otimes C_\infty^2 & I \otimes D_\infty^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & I \otimes A_0^2 & 0 & I \otimes B_0^2 \\ I \otimes C_1^2 & 0 & 0 & 0 & I \otimes D_1^2 & 0 \\ 0 & 0 & 0 & I \otimes C_0^2 & 0 & I \otimes D_0^2 \end{pmatrix},$$

$$P_R = \begin{pmatrix} D_1^1 \otimes I & 0 & 0 & C_1^1 \otimes I & 0 & 0 \\ 0 & A_0^1 \otimes I & B_0^1 \otimes I & 0 & 0 & 0 \\ 0 & C_0^1 \otimes I & D_0^1 \otimes I & 0 & 0 & 0 \\ B_1^1 \otimes I & 0 & 0 & A_1^1 \otimes I & 0 & 0 \\ 0 & 0 & 0 & 0 & D_\infty^1 \otimes I & C_\infty^1 \otimes I \\ 0 & 0 & 0 & 0 & B_\infty^1 \otimes I & A_\infty^1 \otimes I \end{pmatrix}.$$

Since $P_R^2 = P_L^2 = \mathrm{Id}$, both P_R and P_L are invertible and $\mathfrak{D}(K_1, K_2)$ is equivalent to $\mathfrak{D}'(K_1, K_2) = P_L \mathfrak{D}(K_1, K_2) P_R$. The matrix $\mathfrak{D}'(K_1, K_2)$ has the block presentation

$$\begin{pmatrix} D_\infty^1 B_1^1 \otimes B_1^2 A_0^2 & B_1^1 A_0^1 \otimes I & B_1^1 B_0^1 \otimes I & D_\infty^1 A_1^1 \otimes B_1^2 A_0^2 & I \otimes B_1^2 B_0^2 & 0 \\ I \otimes B_\infty^2 B_1^2 & D_1^1 A_0^1 \otimes B_\infty^2 A_1^2 & D_1^1 B_0^1 \otimes B_\infty^2 A_1^2 & 0 & B_0^1 B_\infty^1 \otimes I & B_0^1 A_\infty^1 \otimes I \\ I \otimes D_\infty^2 B_1^2 & \Psi_1 & D_1^1 B_0^1 \otimes D_\infty^2 A_1^2 & 0 & 0 & 0 \\ B_\infty^1 B_1^1 \otimes I & 0 & I \otimes B_0^2 B_\infty^2 & B_\infty^1 A_1^1 \otimes I & \Gamma_1 & \Gamma_2 \\ D_\infty^1 B_1^1 \otimes D_1^2 A_0^2 & 0 & 0 & \Psi_2 & I \otimes D_1^2 B_0^2 & 0 \\ 0 & 0 & I \otimes D_0^2 B_\infty^2 & 0 & \Gamma_3 & \Gamma_4 \end{pmatrix}$$

with

$$\begin{aligned}
\Psi_1 &= I\otimes I + D_1^1A_0^1\otimes D_\infty^2A_1^2, & \Gamma_1 &= D_0^1B_\infty^1\otimes B_0^2A_\infty^2 + X^1B_\infty^1\otimes B_0^2X^2,\\
\Psi_2 &= I\otimes I + D_\infty^1A_1^1\otimes D_1^2A_0^2, & \Gamma_2 &= D_0^1A_\infty^1\otimes B_0^2A_\infty^2 + X^1A_\infty^1\otimes B_0^2X^2,\\
& & \Gamma_3 &= D_0^1B_\infty^1\otimes D_0^2A_\infty^2 + X^1B_\infty^1\otimes D_0^2X^2,\\
& & \Gamma_4 &= I\otimes I + D_0^1A_\infty^1\otimes D_0^2A_\infty^2 + X^1A_\infty^1\otimes D_0^2X^2,
\end{aligned}$$

and is easier to use in actual computations. Note that

$$\imath(\mathfrak{D}'(K_1,K_2)) \simeq \imath(\mathfrak{D}(K_1,K_2)) \simeq \widehat{\mathrm{HF}}(Y(K_1,K_2),\mathbb{F}).$$

6 Splicing with the trefoil

6.1 The maps $\mathfrak{f}_\bullet$ and $\bar{\mathfrak{f}}_\bullet$ for the trefoils

Let us now consider the case of the right-handed trefoil, which will be denoted by R. Thus, $h_\infty(R)=h_1(R)=3$ and $h_0(R)=4$. Moreover, $y_\infty(R)=y_1(R)=1$, while $y_0(R)=2$ (see Eftekhary [3, Section 5]). Since $\mathbb{H}_\bullet(R,i)=\mathbb{F}$ for $\bullet=1,\infty$, $i=0,\pm1$, the maps $\tau_1(R)$ and $\tau_\infty(R)$ are forced and we only need to determine $\tau_0(R)$.

The decompositions of $\mathbb{H}_\infty(R)=\mathbb{H}_1(R)=\mathbb{F}^3$ according to relative Spin^c classes give

$$\mathbb{H}_1(R)=\langle a,b,c\rangle_{\mathbb{F}} \quad\text{and}\quad \mathbb{H}_\infty(R)=\langle a',b',c'\rangle_{\mathbb{F}},$$

where a, a' are generators in relative Spin^c class -1, b, b' are generators in relative Spin^c class 0 and c, c' are generators in relative Spin^c class $+1$. The homomorphisms $\mathfrak{f}_0(R)$ and $\bar{\mathfrak{f}}_0(R)$ have the following block forms in the corresponding basis:

$$\mathfrak{f}_0(R)=\begin{pmatrix}\alpha&0&0\\0&\beta&0\\0&0&\gamma\end{pmatrix} \quad\text{and}\quad \bar{\mathfrak{f}}_0(R)=\begin{pmatrix}\gamma&0&0\\0&\beta&0\\0&0&\alpha\end{pmatrix}. \tag{18}$$

From (7) we know that the ranks of $\mathfrak{f}_0(R)$ and $\bar{\mathfrak{f}}_0(R)$ are equal to 1, ie precisely one of α, β and γ is equal to 1 and the other two are zero. Moreover, the rank of $\mathfrak{f}_0(R)+\bar{\mathfrak{f}}_0(R)$ is 2, ie precisely two of $\alpha+\gamma$, $\alpha+\gamma$, 2β are non-zero. Since the coefficient ring is $\mathbb{F}$, $2b$ is automatically zero. Thus, $\alpha=1$ and $\beta=\gamma=0$, or $\gamma=1$ and $\alpha=\beta=0$.

The generator a of $\mathbb{H}_1(R)$ is not in the image of $\mathfrak{f}_\infty(R)$, since $\mathfrak{f}_\infty(R,-1)$ is trivial. Hence a is not in the kernel of $\mathfrak{f}_0(R,-1)$. Thus, from the above two possibilities the former is the case, ie in (18) we get $\alpha=1$ and $\beta=\gamma=0$.

The rank of $\bar{\mathfrak{f}}_\infty(R)$ is equal to 2 according to (7). Moreover, $\langle a,b\rangle_{\mathbb{F}}$ is already in the image of $\bar{\mathfrak{f}}_0(R)$. Thus, $\bar{\mathfrak{f}}_0(R)$ is surjective onto $\mathbb{H}_1(R,-1)\oplus\mathbb{H}_1(R,0)$. Let us use a basis a'', b'' for $\mathbb{H}_0\big(R,-\frac{1}{2}\big)$ which contains some pre-image a'' of a under $\bar{\mathfrak{f}}_\infty$ and an element b'' in the kernel of $\bar{\mathfrak{f}}_\infty$. Use the dual basis $\tau_0(b'')$, $\tau_0(a'')$ for $\mathbb{H}_0\big(R,\frac{1}{2}\big)$. The basis $\{a'',b'',\tau_0(b''),\tau_0(a'')\}$ for $\mathbb{H}_0(R)$ is thus invariant under $\tau_0=\tau_0(R)$. Correspondingly, we get

$$\bar{\mathfrak{f}}_\infty(R)=\begin{pmatrix}1&0&0&0\\0&0&x&y\\0&0&0&0\end{pmatrix}\quad\text{and}\quad \mathfrak{f}_\infty(R)=\begin{pmatrix}0&0&0&0\\y&x&0&0\\0&0&0&1\end{pmatrix}. \tag{19}$$

If $x=0$ then $y=1$, since the rank of $\mathfrak{f}_\infty(R)$ is equal to 2. The rank of $\mathfrak{f}_\infty(R)+\bar{\mathfrak{f}}_\infty(R)$ is then equal to 2; on the other hand, (7) implies that this rank is 3, a contradiction. The contradiction implies that $x=1$. Replacing a'' with $a''-yb''$, we obtain the presentation of $\mathfrak{f}_\infty(R)$ and $\bar{\mathfrak{f}}_\infty(R)$ in a new basis for $\mathbb{H}_0(R)$ (which is still invariant under the involution $\tau_0(R)$) corresponding to the values $x=1$ and $y=0$ in (19). From here, by taking into account the fact that the map $\theta(R)$ increases the Spin^c grading by $\frac{1}{2}$ while $\bar{\theta}(R)$ decreases the Spin^c grading by $\frac{1}{2}$,

$$\theta(R)=\begin{pmatrix}0&0&0&0\\1&0&0&0\\0&0&1&0\end{pmatrix}\quad\text{and}\quad \bar{\theta}(R)=\begin{pmatrix}0&1&0&0\\0&0&0&1\\0&0&0&0\end{pmatrix}. \tag{20}$$

The above computations imply that $a_0(R)=1$ while $a_1(R)=a_\infty(R)=2$. Moreover, we may take

$$\begin{aligned}
A_0(R)=D_0(R)&=\begin{pmatrix}0&0\\0&0\end{pmatrix}, & B_0(R)=C_0(R)&=\begin{pmatrix}0&1\\1&0\end{pmatrix},\\
A_1(R)=D_\infty(R)&=(0), & D_1(R)=A_\infty(R)&=\begin{pmatrix}0&0\\0&1\end{pmatrix},
\end{aligned}$$
$$C_1(R)=B_\infty(R)=B_1^T(R)=C_\infty^T(R)=\begin{pmatrix}1\\0\end{pmatrix}. \tag{21}$$

For the left-handed trefoil, a similar argument may be used for the computation, which is sketched below. The rank of $\mathfrak{f}_0(L)$ is 2 and the rank of $\mathfrak{f}_\infty(L)$ is 3. The latter implies that the rank of $\mathfrak{f}_\infty(L,1)$ is 1, the rank of $\mathfrak{f}_\infty(L,0)$ is 2 and the rank of $\mathfrak{f}_\infty(L,-1)$ is zero. Correspondingly, the ranks of $\mathfrak{f}_0(L,1)$, $\mathfrak{f}_0(L,0)$ and $\mathfrak{f}_0(L,-1)$ are equal to 0, 1 and 1, respectively. If the images of $\mathfrak{f}_\infty(L,0)$ and $\bar{\mathfrak{f}}_\infty(L,0)$ are identical, the maps $\mathfrak{f}_0(L,0)$ and $\bar{\mathfrak{f}}_0(L,0)$ are forced to be identical, since $\mathbb{H}_\infty(L,0)$ is 1–dimensional. In particular, $\mathfrak{f}_0(L,0)+\bar{\mathfrak{f}}_0(L,0)$ is trivial. Hence the rank of $\mathfrak{f}_0(L)+\bar{\mathfrak{f}}_0(L)$ is at most 2, which is in contradiction with $y_0(L)=2$. The 2–dimensional subspaces $\mathrm{Im}(\mathfrak{f}_\infty(L,0))$ and $\mathrm{Im}(\bar{\mathfrak{f}}_\infty(L,0))$ of $\mathbb{H}_1(L,0)$ are thus different. From here, their intersection is 1–dimensional and is generated by some $\tau_1(L)$–invariant element $\mathfrak{f}_\infty(b)$ with $b \in \mathbb{H}_0\big(L, -\frac{1}{2}\big)$.

Let $a \in \mathbb{H}_0\big(L, \frac{1}{2}\big)$ denote the unique non-trivial vector in the kernel of $\mathfrak{f}_\infty(L)$. Let us first assume that $b = \tau_0(a)$. Complete a to a basis (a, c) for $\mathbb{H}_0\big(L, \frac{1}{2}\big)$. Then

$$\{a, c, \tau_0(a), \tau_0(c)\}$$

is an ordered basis for $\mathbb{H}_0(L)$. Correspondingly, we obtain the basis

$$\big\{\mathfrak{f}_\infty(c), \bar{\mathfrak{f}}_\infty(c), \mathfrak{f}_\infty(\tau_0(a)), \mathfrak{f}_\infty(\tau_0(c)), \tau_1(\mathfrak{f}_\infty(c))\big\}$$

for $\mathbb{H}_1(L)$ and the matrices $\mathfrak{f}_\infty(L)$ and $\bar{\mathfrak{f}}_\infty(L)$ take the following forms, respectively:

$$\mathfrak{f}_\infty(L) = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix} \quad \text{and} \quad \bar{\mathfrak{f}}_\infty(L) = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

In particular, the matrix

$$\mathfrak{f}_\infty(L) + \bar{\mathfrak{f}}_\infty(L) = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

is a matrix of rank 3, while we should have

$$\operatorname{rnk}(\mathfrak{f}_\infty(L) + \bar{\mathfrak{f}}_\infty(L)) = \tfrac{1}{2}(h_0(L) + h_1(L) - y_\infty(L)) = 4.$$

This contradiction implies that b is different from $\tau_0(a)$, so we may take $(\tau_0(a), b)$ as a basis for $\mathbb{H}_0\big(L, \frac{1}{2}\big)$. Correspondingly, we obtain the basis

$$\big\{\tau_0(a), b, \tau_0(b), a\big\}$$

for $\mathbb{H}_0(L)$. As a basis for $\mathbb{H}_1(L, 0)$ we obtain the three vectors $\mathfrak{f}_\infty(a)$, $\bar{\mathfrak{f}}_\infty(a)$ and $\bar{\mathfrak{f}}_\infty(b)$. This basis is completed to the (ordered) basis for $\mathbb{H}_1(L)$

$$\big\{\mathfrak{f}_\infty(a), \bar{\mathfrak{f}}_\infty(b), \mathfrak{f}_\infty(b), \bar{\mathfrak{f}}_\infty(a), \tau_1(\mathfrak{f}_\infty(b))\big\}.$$

Finally, we choose the following basis for $\mathbb{H}_\infty(L)$:

$$\big\{\bar{\mathfrak{f}}_0(\mathfrak{f}_\infty(b)), \mathfrak{f}_0(\bar{\mathfrak{f}}_\infty(a)), \tau_\infty(\bar{\mathfrak{f}}_0(\mathfrak{f}_\infty(b)))\big\}.$$

In these bases, we may compute

$$\mathfrak{f}_\infty(L) = \begin{pmatrix} 1&0&0&0\\ 0&1&0&0\\ 0&0&1&0\\ 0&0&0&0 \end{pmatrix} \quad \text{and} \quad \mathfrak{f}_0(L) = \begin{pmatrix} 0&0&0&0&0\\ 0&0&0&1&0\\ 0&0&0&0&1 \end{pmatrix}.$$

Moreover, after re-ordering the elements of the above bases, we find the presentations

$$D_0(L) = A_\infty(L) = 0, \quad A_1(L) = \begin{pmatrix} 0&0\\ 0&0 \end{pmatrix}, \quad B_1(L) = C_1^T(L) = \begin{pmatrix} 1&0&0\\ 0&0&1 \end{pmatrix},$$

$$A_0(L) = \begin{pmatrix} 0&0&0\\ 0&0&1\\ 0&1&0 \end{pmatrix}, \quad D_1(L) = \begin{pmatrix} 0&0&0\\ 0&1&0\\ 0&0&0 \end{pmatrix}, \quad D_\infty(L) = \begin{pmatrix} 1&0\\ 0&0 \end{pmatrix},$$

$$B_0(L) = C_0^T(L) = \begin{pmatrix} 1\\ 0\\ 0 \end{pmatrix} \quad \text{and} \quad B_\infty(L) = C_\infty^T(L) = \begin{pmatrix} 0&1 \end{pmatrix}.$$

6.2 Splicing a knot complement with the complement of a trefoil

For a knot $K \subset Y$, let $Y(R,K)$ denote the three-manifold obtained by splicing the complement of $K \subset Y$ with the complements of the right-handed trefoil. We study the rank $r_r(K)$ of $\widehat{\mathrm{HF}}(Y(R,K))$ in this subsection. With the notation of Section 5.3, $r_r(K) = i(\mathfrak{D}'(R,K))$. Replacing the block forms of (21) in $\mathfrak{D}'(R,K)$, we find

$$\mathfrak{D}'(R,K) = \begin{pmatrix}
0 & 0 & 0 & 0 & 0 & I & 0 & B_1B_0 & 0 & 0\\
B_\infty B_1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & I\\
0 & B_\infty B_1 & 0 & 0 & B_\infty A_1 & 0 & 0 & I & 0 & 0\\
D_\infty B_1 & 0 & I & 0 & 0 & 0 & 0 & 0 & 0 & 0\\
0 & D_\infty B_1 & 0 & I & D_\infty A_1 & 0 & 0 & 0 & 0 & 0\\
I & 0 & 0 & 0 & B_0B_\infty & 0 & 0 & 0 & 0 & B_0X\\
0 & 0 & 0 & 0 & 0 & B_0B_\infty & 0 & 0 & 0 & 0\\
0 & 0 & 0 & 0 & 0 & 0 & I & D_1B_0 & 0 & 0\\
0 & 0 & 0 & 0 & D_0B_\infty & 0 & 0 & 0 & I & D_0X\\
0 & 0 & 0 & 0 & 0 & D_0B_\infty & 0 & 0 & 0 & I
\end{pmatrix},$$

where $A_\bullet = A_\bullet(K)$, $B_\bullet = B_\bullet(K)$, $C_\bullet = C_\bullet(K)$, $D_\bullet = D_\bullet(K)$ and $X = X(K)$ for $\bullet \in \{0,1,\infty\}$. Doing a series of cancellations that correspond to the identity matrices which appear as the

$$(1,6), \quad (3,8), \quad (4,3), \quad (5,4), \quad (6,1), \quad (8,7), \quad (9,9) \quad \text{and} \quad (10,10)$$

entries in the above block presentation, we obtain the equivalent matrix

$$R_r(K) := \begin{pmatrix} 0 & B_0XB_\infty \\ XB_\infty B_1 & XB_\infty A_1 + D_0XB_\infty \end{pmatrix}. \tag{22}$$

Corollary 6.1 *For a knot $K \subset Y$, let $Y(R,K)$ denote the three-manifold obtained by splicing the complement of K and the complement of the trefoil. Then*

$$\widehat{\mathrm{HF}}(Y(R,K)) = \imath(R_r(K)). \tag{23}$$

Proof The claim follows immediately from the above discussion. □

For the trefoils, our computations imply that

$$\begin{aligned} X(R)B_\infty(R) = X(L)B_\infty(L) = 0 &\implies R_r(R) = R_r(L) = 0 \\ &\implies |\widehat{\mathrm{HF}}(Y(R,R))| = 7,\ |\widehat{\mathrm{HF}}(Y(R,L))| = 9. \end{aligned}$$

The above computations agree with the computations of Hedden and Levine [5].

Corollary 6.2 *For every knot K in a homology sphere Y we have*

$$\begin{aligned} |\widehat{\mathrm{HF}}(Y(R,K))| &\geq (a_0(K) + a_1(K) + 2a_\infty(K)) - 4\min\{a_0(K), a_1(K), a_\infty(K)\} \\ &= 4\max\{h_0(K), h_1(K), h_\infty(K)\} - (h_0(K) + h_1(K) + 2h_\infty(K)). \end{aligned}$$

Moreover, if $Y(R,K)$ is a homology sphere L–space, K is trivial and Y is a homology sphere L–space.

Proof Let $M = M(K) = X(K)B_\infty(K)$ and note that

$$\begin{aligned} \mathrm{rnk}(R_r(K)) &= \mathrm{rnk}\begin{pmatrix} 0 & B_0(K)M \\ MB_1(K) & MA_1(K) + D_0(K)M \end{pmatrix} \\ &\leq \mathrm{rnk}\begin{pmatrix} MB_1(K) & MA_1(K) \end{pmatrix} + \mathrm{rnk}\begin{pmatrix} B_0(K)M \\ D_0(K)M \end{pmatrix} \\ &= 2\,\mathrm{rnk}(M) \\ &\leq 2\,\mathrm{rnk}(X(K)). \end{aligned}$$

For every knot $K \subset Y$ as above note that the rank of $X = X(K)$ is at most equal to the minimum of the sizes of the matrices $B_0(K)$, $B_1(K)$ and $B_\infty(K)$, which is

$$\min\{a_0(K), a_1(K), a_\infty(K)\}.$$

Since $R_r(K)$ is of size $h_0(K) \times h_1(K) = (a_1(K) + a_\infty(K)) \times (a_0(K) + a_\infty(K))$, this proves the first part of the corollary.

Let us assume that $\mathrm{rnk}\big(\widehat{\mathrm{HF}}(Y(R,K))\big) = 1$. From here we find

$$\begin{aligned}(a_0(K)+a_1(K)+2a_\infty(K)) - 4\min\{a_0(K), a_1(K), a_\infty(K)\}\\ = (a_0(K)+a_1(K)+2a_\infty(K)) - 4\,\mathrm{rnk}(M) = 1.\end{aligned}$$

Since $a_1(K)$ and $a_\infty(K)$ have the same parity while the parity of $a_0(K)$ is different from the parity of both $a_1(K)$ and $a_\infty(K)$, one can easily conclude that $a_0(K)-1 = a_1(K) = a_\infty(K)$. Let a denote the common value $a_1(K) = a_\infty(K)$. Then the rank of M is a and both $B_0(K)$ and $X(K)$ are invertible. We may thus assume that $A_0(K) = D_0(K) = 0$. Since

$$\mathrm{rnk}(\mathfrak{f}_\infty(K) + \bar{\mathfrak{f}}_\infty(K)) = \mathrm{rnk}\begin{pmatrix} B_1(K)A_0(K) & B_1(K)B_0(K) \\ I + D_1(K)A_0(K) & D_1(K)B_0(K)\end{pmatrix} = 2a,$$

the three-manifold Y is an L–space. Since splicing K with the trefoil is also a homology sphere L–space, we conclude that K is trivial, by [5, Theorem 1]. $\square$

Appendix: Bordered Floer homology for knot complements

The first draft of this paper appeared while the theory of bordered Floer homology was being developed. With bordered Floer homology conventions widely known to the Heegaard Floer community, the referee recommended the inclusion of an appendix which addresses the contribution of this paper within the realm of bordered Floer homology.

Let $K \subset Y$ denote a null-homologous knot inside the three-manifold Y and let $H = (\Sigma, \boldsymbol{\alpha}, \widehat{\boldsymbol{\beta}} \cup \{\lambda, \mu\}; z)$ denote a special Heegaard diagram for K, as constructed in Lemma 4.1. In particular, H is a nice Heegaard diagram for the bordered three-manifold Y_K determined by $K \subset Y$ in the sense of Lipshitz, Ozsváth and Thurston [6]. The bordered Floer complex $\widehat{\mathrm{CFD}}(Y_K)$ may then be constructed from the chain complexes $M = M(K)$ and $L = L(K)$ (which are described in Proposition 5.1 as the mapping cones of $\bar{\mathfrak{f}}_\infty(K)\colon C_0(K) \to C_1(K)$ and $\mathfrak{f}_0(K)\colon C_1(K) \to C_\infty(K)$, respectively) and the chain maps $\Phi = \Phi(K)\colon L \to M$ and $\Psi_i = \Psi_i(K)\colon M \to L$, $i = 1, 2, 3$.

More precisely and following the notation of [7, Section 4.2], the idempotents ι_0 and ι_1 and the chords ρ_1, ρ_2, ρ_3, $\rho_{12} = \rho_1\rho_2$, $\rho_{23} = \rho_2\rho_3$ and $\rho_{123} = \rho_1\rho_2\rho_3$ form an $\mathbb{F}$–basis for the differential graded algebra associated with the torus boundary:

$$\mathcal{A}(T^2, 0) = \left\langle \begin{tikzcd} \iota_0 \bullet \arrow[r, bend left, "\rho_1"] \arrow[r, bend right, "\rho_3"'] & \bullet\, \iota_1 \arrow[l, "\rho_2"'] \end{tikzcd} \right\rangle \Big/ \, (\rho_2\rho_1 = \rho_3\rho_2 = 0).$$

The module $\widehat{\mathrm{CFD}}(Y_K)$ is generated (over $\mathcal{A}(T^2,0)$) by the generators of M and L. For a generator $\boldsymbol{x}$ of L we have

$$I(\boldsymbol{x})=\iota_0 \quad\text{and}\quad \partial(\boldsymbol{x})=d_L(\boldsymbol{x})+\rho_1\Psi_1(\boldsymbol{x})+\rho_3\Psi_2(\boldsymbol{x})+\rho_{123}\Psi_3(\boldsymbol{x}), \tag{24}$$

while for a generator $\boldsymbol{y}$ of M we have

$$I(\boldsymbol{y})=\iota_1 \quad\text{and}\quad \partial(\boldsymbol{y})=d_M(\boldsymbol{y})+\rho_2\Phi(\boldsymbol{y}). \tag{25}$$

The splicing formula of (17) is then just the gluing formula for bordered Floer homology, ie [6, Theorem 1.3]. A related discussion is carried over in [6, Section 8].

Definition A.1 The chain complexes $(C_\bullet(K),d_\bullet)$, $\bullet\in\{0,1,\infty\}$, and the chain maps $f_\bullet(K)$, $\bar{f}_\bullet(K)$, $\bullet\in\{0,\infty\}$, are called *admissible data* associated with the knot K if they satisfy the following conditions:

- The homology of the complex $(C_\bullet(K),d_\bullet)$ is $\mathbb{H}_\bullet(K)$.
- The maps induced by $f_\bullet(K)$ and $\bar{f}_\bullet(K)$ in homology (under the identification of the homology of $(C_\bullet(K),d_\bullet)$ with $\mathbb{H}_\bullet(K)$) are $\mathfrak{f}_\bullet(K)$ and $\bar{\mathfrak{f}}_\bullet(K)$, respectively.
- We have $f_0(K)\circ f_\infty(K)=\bar{f}_0(K)\circ\bar{f}_\infty(K)=0$.
- The corresponding maps

$$\theta(K)\colon \mathrm{Ker}(\mathfrak{f}_\infty(K))\longrightarrow \mathrm{Coker}(\mathfrak{f}_0(K)),$$
$$\bar\theta(K)\colon \mathrm{Ker}(\bar{\mathfrak{f}}_\infty(K))\longrightarrow \mathrm{Coker}(\bar{\mathfrak{f}}_0(K))$$

  are isomorphisms and are the inverses of the maps induced by $\mathfrak{f}_1(K)$ and $\bar{\mathfrak{f}}_1(K)$, respectively.

The proof of Theorem 1.1 implies that $(C_\bullet^i,d_\bullet^i)$ and the chain maps $f_\bullet^i$, $\bar{f}_\bullet^i$ for $\bullet\in\{0,\infty\}$ and $i=1,2$ in (17) may be replaced by other admissible data corresponding to the knots K_1 and K_2. orrespondingly, the bordered Floer complex associated with any knot $K\subset Y$ may be constructed from admissible data associated with K. More precisely, we have the following proposition:

Proposition A.2 *Suppose that the chain complexes $(C_\bullet(K),d_\bullet)$, $\bullet\in\{0,1,\infty\}$, and the chain maps $f_\bullet=f_\bullet(K)$, $\bar{f}_\bullet=\bar{f}_\bullet(K)$, $\bullet\in\{0,\infty\}$, are admissible data associated with the knot $K\subset Y$ and set*

$$M(K)=C_0(K)\oplus C_1(K),\qquad L(K)=C_1(K)\oplus C_\infty(K).$$

The bordered Floer complex $\widehat{\mathrm{CFD}}(Y_K)$ may then be constructed as the left module over the differential graded algebra $\mathcal{A}(T^2,0)$ which is generated by $\iota_0.L(K)$ and $\iota_1.M(K)$, and equipped with the differential $\partial\colon \widehat{\mathrm{CFD}}(Y_K)\to\widehat{\mathrm{CFD}}(Y_K)$ defined by

$$\partial\begin{pmatrix}\boldsymbol{x}\\ \boldsymbol{y}\end{pmatrix} = \begin{cases} \begin{pmatrix} d_0(\boldsymbol{x}) \\ \bar{f}_\infty(\boldsymbol{x})+d_1(\boldsymbol{y})\end{pmatrix} + \rho_2.\begin{pmatrix}0\\ \boldsymbol{x}\end{pmatrix} & \text{if } \begin{pmatrix}\boldsymbol{x}\\ \boldsymbol{y}\end{pmatrix}\in M(K),\\ \begin{pmatrix} d_1(\boldsymbol{x}) \\ f_0(\boldsymbol{x})+d_\infty(\boldsymbol{y})\end{pmatrix} + \begin{pmatrix}\rho_1 f_\infty(\boldsymbol{x}) \\ \rho_3\bar{f}_0(\boldsymbol{y})+\rho_1\rho_2\rho_3\bar{f}_0(f_\infty(\boldsymbol{x}))\end{pmatrix} & \text{if } \begin{pmatrix}\boldsymbol{x}\\ \boldsymbol{y}\end{pmatrix}\in L(K).\end{cases} \tag{26}$$

In particular, let the $\mathbb{F}$–modules $\mathbb{A}_\bullet=\mathbb{A}_\bullet(K)$, $\bullet\in\{0,1,\infty\}$, and the matrices $A_\bullet=A_\bullet(K)$, $B_\bullet=B_\bullet(K)$, $C_\bullet=C_\bullet(K)$ and $D_\bullet=D_\bullet(K)$ be defined as in Section 3.3. Set

$$(C_0(K),d_0)=(\mathbb{A}_\infty\oplus\mathbb{A}_1,0),\quad (C_\infty(K),d_\infty)=(\mathbb{A}_1\oplus\mathbb{A}_0,0),$$
$$C_1(K)=\mathbb{A}_1\oplus\mathbb{A}_0\oplus\mathbb{A}_\infty\oplus\mathbb{A}_1 \quad\text{and}\quad d_1=\begin{pmatrix}0&0&0&0\\0&0&0&0\\0&0&0&0\\ I_{\mathbb{A}_1}&0&0&0\end{pmatrix}.$$

Correspondingly, define

$$f_\infty(K)=\begin{pmatrix}0&0\\0&0\\I&0\\0&I\end{pmatrix},\quad f_0(K)=\begin{pmatrix}I&0&0&0\\0&I&0&0\end{pmatrix}\quad\text{and}\quad \tau_1(K)=\begin{pmatrix}0&0&0&0\\0&A_1&B_1&0\\0&C_1&D_1&0\\0&0&0&0\end{pmatrix}$$

and set $\bar{f}_\infty(K)=\tau_1(K)f_\infty(K)\tau_0(K)$ and $\bar{f}_0(K)=\tau_\infty(K)f_0(K)\tau_1(K)$. The data associated with K consisting of $(C_\bullet(K),d_\bullet)$ and $f_\bullet(K)$, $\bar{f}_\bullet(K)$, $\bullet\in\{0,\infty\}$ is then admissible.

Corresponding to the above admissible data and associated with $K\subset Y$, we may construct the bordered Floer complex for K via

$$M(K)=C_0(K)\oplus C_1(K)=\mathbb{A}_\infty\oplus\mathbb{A}_1\oplus\mathbb{A}_1\oplus\mathbb{A}_0\oplus\mathbb{A}_\infty\oplus\mathbb{A}_1,$$
$$L(K)=C_1(K)\oplus C_\infty(K)=\mathbb{A}_1\oplus\mathbb{A}_0\oplus\mathbb{A}_\infty\oplus\mathbb{A}_1\oplus\mathbb{A}_1\oplus\mathbb{A}_0,$$
$$d_M=\begin{pmatrix}0&0&0&0&0&0\\0&0&0&0&0&0\\0&0&0&0&0&0\\B_1A_0&B_1B_0&0&0&0&0\\D_1A_0&D_1B_0&I&0&0&0\\0&0&0&0&0&0\end{pmatrix},\quad d_L=\begin{pmatrix}0&0&0&0&0&0\\0&0&0&0&0&0\\0&0&0&0&0&0\\I&0&0&0&0&0\\0&0&I&0&0&0\\0&0&0&I&0&0\end{pmatrix},$$

$$\Phi(K)=\begin{pmatrix}0&0&0&0&0&0\\0&0&0&0&0&0\\I&0&0&0&0&0\\0&I&0&0&0&0\\0&0&I&0&0&0\\0&0&0&I&0&0\end{pmatrix},\quad \Psi_1(K)=\begin{pmatrix}0&0&0&0&0&0\\0&0&0&0&0&0\\I&0&0&0&0&0\\0&I&0&0&0&0\\0&0&0&0&0&0\\0&0&0&0&0&0\end{pmatrix},$$

$$\Psi_2(K)=\begin{pmatrix}0&0&0&0&0&0\\0&0&0&0&0&0\\0&0&0&0&0&0\\0&0&0&0&0&0\\0&0&0&B_\infty A_1&B_\infty B_1&0\\0&0&0&D_\infty A_1&D_\infty B_1&0\end{pmatrix}\quad\text{and}\quad \Psi_3(K)=\Psi_2(K)\Phi(K)\Psi_1(K)$$

as the left module over the differential graded algebra $\mathcal{A}(T^2,0)$ generated by $\iota_0.L$ and $\iota_1.M$ and equipped with the differential $\partial\colon \widehat{\mathrm{CFD}}(Y_K)\to\widehat{\mathrm{CFD}}(Y_K)$ defined by the equations (24) and (25).

Remark A.3 Simultaneous computation of the matrices $\tau_\bullet(K)=\left(\begin{smallmatrix}A_\bullet & B_\bullet\\ C_\bullet & D_\bullet\end{smallmatrix}\right)$ is a priori quite difficult, as we observed in the case of trefoils in Section 6. This makes the above description of the bordered Floer homology hard to use even for knots $K\subset Y$ where we have complete understanding of the Heegaard Floer complex associated with K. However, it is possible to construct admissible data associated with $K\subset Y$ completely in terms of the filtered chain complex $\mathrm{CF}^\infty(Y,K;\mathbb{F})$, as will be discussed in the revision of [4].

References

[1] **A Alishahi**, **E Eftekhary**, *A refinement of sutured Floer homology*, J. Symplectic Geom. 13 (2015) 609–743

[2] **E Eftekhary**, *Longitude Floer homology and the Whitehead double*, Algebr. Geom. Topol. 5 (2005) 1389–1418 MR2171814

[3] **E Eftekhary**, *A combinatorial approach to surgery formulas in Heegaard Floer homology*, Algebr. Geom. Topol. 9 (2009) 2225–2246 MR2558310

[4] **E Eftekhary**, *Bordered Floer homology and existence of incompressible tori in homology spheres*, preprint (2015) `arXiv:1504.05329`

[5] **M Hedden**, **A S Levine**, *Splicing knot complements and bordered Floer homology*, preprint `arXiv:1210.7055v1` To appear in J. Reine Angew. Math.

[6] **R Lipshitz**, **O Ozsváth**, **D Thurston**, *Bordered Heegaard Floer homology: Invariance and pairing*, preprint (2014) `arXiv:0810.0687`

[7] **R Lipshitz**, **O Ozsváth**, **D Thurston**, *Notes on bordered Floer homology*, preprint (2014) `arXiv:1211.6791`

[8] **C Manolescu**, **P Ozsváth**, **S Sarkar**, *A combinatorial description of knot Floer homology*, Ann. of Math. 169 (2009) 633–660 MR2480614

[9] **C Manolescu**, **P Ozsváth**, **Z Szabó**, **D Thurston**, *On combinatorial link Floer homology*, Geom. Topol. 11 (2007) 2339–2412 MR2372850

[10] **P Ozsváth**, **A I Stipsicz**, *Contact surgeries and the transverse invariant in knot Floer homology*, J. Inst. Math. Jussieu 9 (2010) 601–632 MR2650809

[11] **P Ozsváth**, **Z Szabó**, *Holomorphic disks and knot invariants*, Adv. Math. 186 (2004) 58–116 MR2065507

[12] **P Ozsváth**, **Z Szabó**, *Holomorphic disks and topological invariants for closed three-manifolds*, Ann. of Math. 159 (2004) 1027–1158 MR2113019

[13] **P Ozsváth**, **Z Szabó**, *On the Heegaard Floer homology of branched double-covers*, Adv. Math. 194 (2005) 1–33 MR2141852

[14] **P S Ozsváth**, **Z Szabó**, *Knot Floer homology and integer surgeries*, Algebr. Geom. Topol. 8 (2008) 101–153 MR2377279

[15] **S Sarkar**, *Maslov index formulas for Whitney n–gons*, J. Symplectic Geom. 9 (2011) 251–270 MR2811652

[16] **S Sarkar**, **J Wang**, *An algorithm for computing some Heegaard Floer homologies*, Ann. of Math. 171 (2010) 1213–1236 MR2630063

School of Mathematics, Institute for Research in Fundamental Sciences (IPM)
PO Box 19395-5746, Tehran 19395, Iran

eaman@ipm.ir

http://math.ipm.ir/~eftekhary

Received: 4 November 2013 Revised: 19 February 2015

Algebraic & Geometric Topology 15 (2015) 3215–3252

Higher Hochschild cohomology of the Lubin–Tate ring spectrum

GEOFFROY HOREL

We construct a spectral sequence computing factorization homology of an $\mathcal{E}_d$–algebra in spectra using as an input an algebraic version of higher Hochschild homology due to Pirashvili. This induces a full computation of higher Hochschild cohomology when the algebra is étale. As an application, we compute higher Hochschild cohomology of the Lubin–Tate ring spectrum.

55P43; 16E40, 55P48

This paper is devoted to higher Hochschild cohomology. Given E an $\mathcal{E}_\infty$–ring spectrum, the Hochschild cohomology of an associative algebra A in $\mathbf{Mod}_E$ with coefficients in a bimodule M is the derived homomorphisms object in the category of A–A–bimodules with source A and target M. Higher Hochschild cohomology is the generalization of this construction when A is an $\mathcal{E}_d$–algebra instead of an associative algebra. In this case, we need to replace the notion of bimodule by the notion of operadic $\mathcal{E}_d$–module and the definition becomes

$$\mathrm{HH}_{\mathcal{E}_d}(A|E, M) = \mathbb{R}\underline{\mathrm{Hom}}_{\mathbf{Mod}_A^{\mathcal{E}_d}}(A, M),$$

where $\underline{\mathrm{Hom}}_{\mathbf{Mod}_A^{\mathcal{E}_d}}$ denotes the homomorphism object in the category of operadic $\mathcal{E}_d$–modules over A.

For practical reasons, we use a different but equivalent definition of higher Hochschild cohomology inspired by factorization homology. For A an $\mathcal{E}_d$–algebra in $\mathbf{Mod}_E$ and V a d–dimensional framed manifold, there is a spectrum $\int_V A$ called the factorization homology of A over V. This construction is functorial with respect to maps of $\mathcal{E}_d$–algebras and with respect to embeddings of framed d–manifolds. Moreover, $V \mapsto \int_V A$ is a symmetric monoidal functor. This implies that $\int_{S^{d-1}\times\mathbb{R}} A$ is an $\mathcal{E}_1$–algebra in spectra. This $\mathcal{E}_1$–algebra serves as a universal enveloping algebra for the category of operadic $\mathcal{E}_d$–modules over A. More precisely, we prove in Proposition 3.19 the identity

$$\mathrm{HH}_{\mathcal{E}_d}(A|E, M) \simeq \mathbb{R}\underline{\mathrm{Hom}}_A^{S^{d-1}\times[0,1]}(A, M),$$

Published: 12 January 2016 DOI: 10.2140/agt.2015.15.3215

where the right-hand side is an explicit construction given by a homotopy limit of a certain functor over the poset of disks on the manifold $S^{d-1}\times[0,1]$. In Corollary 3.15, we prove an equivalence

$$\mathbb{R}\underline{\mathrm{Hom}}_A^{S^{d-1}\times[0,1]}(A,M)\simeq \mathbb{R}\underline{\mathrm{Hom}}^{[0,1]}_{\int_{S^{d-1}\times(0,1)}A}(A,M),$$

where the right-hand side is a suitable generalization of the homomorphisms between left modules over an $\mathcal{E}_1$– (as opposed to associative) algebra. Thus, we reduce the computation of higher Hochschild cohomology to the computation of the derived homomorphisms between two left modules over an $\mathcal{E}_1$–algebra.

With this last description, we see that, in order to make explicit computations of higher Hochschild cohomology, the first step is to compute $\int_{S^{d-1}\times\mathbb{R}} A$ with its $\mathcal{E}_1$–structure. In Section 5, we construct a spectral sequence that computes the factorization homology of an $\mathcal{E}_d$–algebra over any framed manifold:

Proposition 5.4 *Let A be an $\mathcal{E}_d$–algebra in $\mathbf{Mod}_E$, let M be a framed d–manifold and let K be a homology theory with a $\mathbb{Z}/2$–equivariant Künneth isomorphism. There is a spectral sequence*

$$\mathrm{E}^2_{s,t}=\mathrm{HH}^M_{s,t}(K_*A)\Longrightarrow K_{s+t}\left(\int_M A\right).$$

Let us say a few words about the E^2–page. Given a commutative ring k, Pirashvili defines a functor $(X,A)\mapsto \mathrm{HH}^X(A)$, where X is a simplicial set, A is a commutative algebra in k–modules and $\mathrm{HH}^X(A)$ is a chain complex of k–modules. When $X=S^1$, this object is quasi-isomorphic to ordinary Hochschild homology. Our spectral sequence computing factorization homology is given by Pirashvili's higher Hochschild homology on the E^2–page.

In Section 6, we make an explicit computation in the case of the Lubin–Tate spectrum (also known as Morava E–theory) E_n. Using the étaleness of the algebra $(K_n)_*E_n$, we can prove that for any $\mathcal{E}_d$–structure on E_n that induces the correct multiplication on K_n–homology, the unit map

$$E_n\to\int_{S^{d-1}\times\mathbb{R}}E_n$$

is a K_n–homology equivalence. Using the fact that E_n is K_n–local, this implies the following theorem:

Proposition 6.4 *The map $\mathrm{HH}_{\mathcal{E}_d}(E_n)\to E_n$ is a weak equivalence.*

In Section 7, we prove an étale base-change theorem for étale algebras:

Theorem 7.9 *Let T be a commutative algebra in $\mathbf{Mod}_E$ that is (K–locally) étale as an $\mathcal{E}_d$–algebra. That is to say that the $\mathcal{E}_d$–version of the cotangent complex of E defined in Definition 2.7 of Francis* [6] *is (K–locally) contractible. Then, for any (K–local) $\mathcal{E}_d$–algebra A over T, the base-change map*

$$\mathrm{HH}_{\mathcal{E}_d}(A|E) \xrightarrow{\sim} \mathrm{HH}_{\mathcal{E}_d}(A|T)$$

is an equivalence.

In particular, this result combined with our computation implies that for any K_n–local $\mathcal{E}_d$–algebra A over E_n, the base-change map

$$\mathrm{HH}_{\mathcal{E}_d}(A|E_n) \to \mathrm{HH}_{\mathcal{E}_d}(A|\mathbb{S})$$

is a weak equivalence.

The full strength of the results proved in this paper is unnecessary in the case of E_n since it is known to be a commutative ring spectrum. However, we think that the method presented here could be used in other contexts, where one has to deal with $\mathcal{E}_d$–algebras that are not commutative.

Conventions

We denote by $\boldsymbol{S}$ the category of simplicial sets with its usual model structure. We use boldface letters to denote categories. We use calligraphic letters like $\mathcal{A}$ to denote operads. All our categories and operads are enriched in $\boldsymbol{S}$. Note that given a topological operad or category, we can turn it into a simplicially enriched operad or category by applying the functor Sing to each mapping space. We allow ourselves to do this operation implicitly.

We denote by $\mathbf{Mod}_E$ the simplicial category of modules over a commutative symmetric ring spectrum E. This category is symmetric monoidal for the relative tensor product over E. Moreover, it has two model structures: the positive model structure, denoted by $\mathbf{Mod}_E^+$, and the absolute model structure, denoted by $\mathbf{Mod}_E$. We refer the reader to Section 1 for more details. We often write $\boldsymbol{C}$ instead of $\mathbf{Mod}_E$ in the sections where the results do not depend a lot on the symmetric monoidal model category.

Acknowledgements

This work is part of my PhD thesis at MIT and I wish to thank Haynes Miller for his guidance. This paper benefited a lot from conversations with Clark Barwick, David Ayala, John Francis and Luis Alexandre Pereira. I also want to thank the anonymous referee for several suggestions that helped me improve the exposition.

1 Review of operads and factorization homology

We recall a few notations. We denote by **Fin** the category whose objects are the nonnegative integers and with

$$\mathbf{Fin}(m,n) = \mathbf{Set}(\{1,\dots,m\},\{1,\dots,n\}).$$

We abuse notation and write n for the finite set $\{1,\dots,n\}$.

To an operad $\mathcal{M}$ with one color, we can assign its PROP $\boldsymbol{M}$. This is a category whose set of objects coincides with the set of objects of **Fin** and with

$$\boldsymbol{M}(m,n) = \bigsqcup_{f\in\mathbf{Fin}(m,n)} \prod_{i\in n} \mathcal{M}(f^{-1}(i)).$$

Note that **Fin** is the PROP associated to the commutative operad. The construction of the associated PROP is a functor from operads to categories. In particular, the unique map $\mathcal{M}\to\mathcal{C}om$ induces a map $\boldsymbol{M}\to\mathbf{Fin}$.

An $\mathcal{M}$–algebra A in a simplicially enriched symmetric monoidal category $\boldsymbol{C}$ induces a symmetric monoidal simplicial functor $\boldsymbol{M}\to\boldsymbol{C}$ that we also denote by A.

Let E be a commutative ring in symmetric spectra. We denote by $\mathbf{Mod}_E^+$ the category of modules over E equipped with the positive model structure (constructed in Schwede [17, Theorem III.3.2] under the name projective positive stable model structure). The category $\mathbf{Mod}_E^+$ is a closed symmetric monoidal model category for the smash product over E (denoted by $-\otimes_E -$). It is also a simplicial model category. Moreover, the two structures are compatible in the sense that the tensor of simplicial sets and E–modules

$$-\otimes -\colon \boldsymbol{S}\times\mathbf{Mod}_E^+ \to \mathbf{Mod}_E^+$$

sending (X,M) to $(E\wedge\Sigma_+^\infty X)\otimes_E M$ is a Quillen left bifunctor.

There is another model structure on $\mathbf{Mod}_E$ called the absolute model structure and that we denote by $\mathbf{Mod}_E$ (its construction can also be found in [17, Thorem III.3.2]). Its weak equivalences are the same as in the positive model structure but there are more cofibrations. In particular, the important fact for us is that the unit E is cofibrant in the absolute model structure but not in the positive model structure. The model category $\mathbf{Mod}_E$ is also a closed symmetric monoidal simplicial model category. The advantage of the positive model structure is that the smash product is much better behaved. In particular, the following theorem would be false for the absolute model structure:

Theorem 1.1 *The category $\mathbf{Mod}_E^+$ is a closed symmetric monoidal cofibrantly generated simplicial model category satisfying the following properties:*

- *For any operad* $\mathcal{M}$ *in* $\boldsymbol{S}$, *the category* $\mathbf{Mod}_E^+[\mathcal{M}]$ *of* $\mathcal{M}$*–algebras in* $\mathbf{Mod}_E^+$ *has a model category structure where weak equivalences and fibrations are created by the forgetful functor* $\mathbf{Mod}_E^+[\mathcal{M}] \to (\mathbf{Mod}_E^+)^{\mathrm{Col}(\mathcal{M})}$.
- *If* $\alpha\colon \mathcal{M} \to \mathcal{N}$ *is a is a map of operads, the adjunction*
$$\alpha_!\colon \mathbf{Mod}_E^+[\mathcal{M}] \leftrightarrows \mathbf{Mod}_E^+[\mathcal{N}] :\alpha^*$$
is a Quillen adjunction. It is, moreover, a Quillen equivalence if α *is a weak equivalence.*
- *The forgetful functor* $\mathbf{Mod}_E^+[\mathcal{M}] \to (\mathbf{Mod}_E)^{\mathrm{Col}(\mathcal{M})}$ *sends cofibrant objects to cofibrant objects.*

Proof See Theorems 3.4.1 and 3.4.3 of Pavlov and Scholbach [14]. □

Remark 1.2 All the operads that we consider in this work have a finite number of colors. The only kind of weak equivalences we will have to consider are maps that induce a bijection on the set of colors and induce weak equivalences on each space of operations.

The little disk operad

There is a topological category whose objects are d–manifolds without boundary and with space of maps between M and N given by $\mathrm{Emb}(M,N)$, the topological space of smooth embeddings with the weak C^1 topology.

Definition 1.3 A *framed d–manifold* is a pair (M,σ_M) where M is a d–manifold and σ_M is a smooth section of the $\mathrm{GL}(d)$–principal bundle $\mathrm{Fr}(TM)$.

If M and N are two framed d–manifolds, we define a space of framed embeddings, denoted by $\mathrm{Emb}_f(M,N)$ as in Definition V.8.3 of Andrade [1]. We now recall this construction. First, given a diagram

$$\begin{tikzcd} & Y \arrow[d, "v"] \\ X \arrow[r, "u"'] & Z \end{tikzcd}$$

in the category of topological spaces over a fixed topological space W, we define its homotopy pullback as in [1, Chapter V.9] to be the space of triples $(y,p,z) \in X \times Z^{[0,1]} \times Y$ such that $p(0) = u(x)$, $p(1) = v(y)$ and such that the image of p in $W^{[0,1]}$ is a constant path. It can be shown that this is indeed a model for the homotopy pullback in the model category $\mathbf{Top}_{/W}$.

Definition 1.4 Let M and N be two framed d–dimensional manifolds. The *topological space of framed embeddings from M to N*, denoted by $\mathrm{Emb}_f(M,N)$, is given by the following homotopy pullback in the category of topological spaces over $\mathrm{Map}(M,N)$:

$$\begin{tikzcd}
\mathrm{Emb}_f(M,N) \arrow[r] \arrow[d] & \mathrm{Map}(M,N) \arrow[d] \\
\mathrm{Emb}(M,N) \arrow[r] & \mathrm{Map}_{\mathrm{GL}(d)}(\mathrm{Fr}(TM),\mathrm{Fr}(TN))
\end{tikzcd}$$

The right-hand side map is obtained as the composite

$$\mathrm{Map}(M,N) \to \mathrm{Map}_{\mathrm{GL}(d)}(M\times\mathrm{GL}(d), N\times\mathrm{GL}(d)) \cong \mathrm{Map}_{\mathrm{GL}(d)}(\mathrm{Fr}(TM),\mathrm{Fr}(TN)),$$

where the first map is obtained by taking the product with $\mathrm{GL}(d)$ and the second map comes from the identifications $\mathrm{Fr}(TM)\cong M\times\mathrm{GL}(d)$ and $\mathrm{Fr}(TN)\cong N\times\mathrm{GL}(d)$ induced by our choice of framing on M and N.

Andrade explains in [1, Definition V.10.1] that there are well-defined composition maps

$$\mathrm{Emb}_f(M,N)\times\mathrm{Emb}_f(N,P)\to\mathrm{Emb}_f(M,P)$$

allowing the construction of a topological category $f\mathbf{Man}_d$.

We denote by D the open disk of dimension d.

Proposition 1.5 *The evaluation at the center of the disks induces a weak equivalence*

$$\mathrm{Emb}_f(D^{\sqcup p},M)\to\mathrm{Conf}(p,M).$$

Proof See [1, Proposition V.4.5] or Proposition 6.6 of Horel [10]. □

Definition 1.6 The *little d–disk operad* $\mathcal{E}_d$ is the one-color operad whose n^{th} space is

$$\mathcal{E}_d(n)=\mathrm{Emb}_f(D^{\sqcup n},D)$$

and whose composition is induced by composition of embeddings. We denote by $\boldsymbol{E}_d$ the PROP of the operad $\mathcal{E}_d$.

Remark 1.7 This model of the little d–disk operad was introduced by Andrade [1]. Using Proposition 1.5, it is not hard to show that this definition is weakly equivalent to any other definition of the little d–disk operad.

Factorization homology

From now on, until we say otherwise, we denote by $(\boldsymbol{C}^+, \otimes, \mathbb{I})$ the symmetric monoidal category $\mathbf{Mod}_E$ with its positive model structure and by $\boldsymbol{C}$ the same category equipped with the absolute model structure. We do this partly to simplify the notations but mostly to emphasize that our arguments hold in greater generality modulo a few easy modifications.

Definition 1.8 Let A be a cofibrant object of $\boldsymbol{C}^+[\mathcal{E}_d]$. We define the *factorization homology with coefficients in* A by the coend

$$\int_M A := \mathrm{Emb}_f(-, M) \otimes_{\boldsymbol{E}_d} A.$$

This functor sends weak equivalences between cofibrant algebras to weak equivalences.

Proposition 1.9 *The functor $M \mapsto \int_M A$ is a simplicial and symmetric monoidal functor from the category $f\mathbf{Man}_d$ to the category $\boldsymbol{C}$.*

Proof See [10, Definition 7.3] and the paragraph following it. □

Let M be an object of $f\mathbf{Man}_d$. Let $\boldsymbol{D}(M)$ be the poset of subsets of M that are diffeomorphic to a disjoint union of disks. Let us choose for each object V of $\boldsymbol{D}(M)$ a framed diffeomorphism $V \cong D^{\sqcup n}$ for some uniquely determined n. Each inclusion $V \subset V'$ in $\boldsymbol{D}(M)$ induces a morphism $D^{\sqcup n} \to D^{\sqcup n'}$ in $\boldsymbol{E}_d$ by composing with the chosen parametrization. Therefore, each choice of parametrization induces a functor $\boldsymbol{D}(M) \to \boldsymbol{E}_d$. Up to homotopy this choice is unique, since the space of automorphisms of D in $\boldsymbol{E}_d$ is contractible.

In the following we assume that we have one of these functors $\delta\colon \boldsymbol{D}(M) \to \boldsymbol{E}_d$. We fix a cofibrant algebra $A\colon \boldsymbol{E}_d \to \boldsymbol{C}$.

Proposition 1.10 *There is a weak equivalence*

$$\mathrm{hocolim}_{V \in \boldsymbol{D}(M)} A(\delta V) \simeq \int_M A.$$

Proof See [10, Corollary 7.7]. □

2 Modules over $\mathcal{E}_d$–algebras

We define the notion of an S_τ–shaped module. These are modules over $\mathcal{E}_d$–algebras that are studied in detail in Horel [11].

Definition 2.1 A *d–framing* of a closed $(d-1)$–manifold S is a trivialization of the d–dimensional bundle $TS \oplus \mathbb{R}$, where $\mathbb{R}$ is a trivial line bundle.

For M a d–manifold with boundary and m a point of ∂M, we say that a vector $u \in T_m M$ is pointing inward if it is not in $T_m \partial M$ and there is a curve $\gamma\colon [0,1) \to M$ whose derivative at 0 is u.

Definition 2.2 Let S be a closed $(d-1)$–manifold. An *S–manifold* is a d–manifold with boundary M together with the data of

- a diffeomorphism $f\colon S \to \partial M$,
- a non-vanishing section ϕ of the restriction of the vector bundle TM on ∂M which is such that $\phi(m)$ is pointing inward for any m in ∂M.

Definition 2.3 Let τ be a d–framing of S. Let $i\colon T\partial M \to TM_{|\partial M}$ be the obvious inclusion. A *framed S_τ–manifold* is an S–manifold (M, f, ϕ) with the data of a framing of TM such that the composite

$$TS \oplus \mathbb{R} \xrightarrow{Tf \oplus \mathbb{R}} T(\partial M) \oplus \mathbb{R} \xrightarrow{i \oplus \phi} TM_{|\partial M}$$

sends τ to the given framing on the right-hand side.

For $E \to M$ a d–dimensional vector bundle, we denote by $\mathrm{Fr}(E)$ the $\mathrm{GL}(d)$–bundle over M whose fiber over m is the space of bases of the vector space E_m. Note that a trivialization of E is exactly the data of a section of $\mathrm{Fr}(E)$.

For (M, f, ϕ) and (M, g, ψ) two framed S_τ–manifolds, we denote by

$$\mathrm{Map}^{S_\tau}_{\mathrm{GL}(d)}(\mathrm{Fr}(TM), \mathrm{Fr}(TN))$$

the space of morphisms of $\mathrm{GL}(d)$–bundles whose underlying map $M \to N$ sends the boundary to the boundary and whose restriction to the boundary is fiberwise the identity (via the identification of both boundaries with S and of both tangent bundles with $TS \oplus \mathbb{R}$).

Definition 2.4 Let (M, f, ϕ) and (M, g, ψ) be two framed S_τ–manifolds. Let $\mathrm{Map}^S(M, N)$ be the topological space of maps between M and N that commute with the maps $f\colon S \to M$ and $g\colon S \to N$. Similarly, let $\mathrm{Emb}^S(M, N)$ be the topological space of embeddings that commute with the maps from S. The *topological space of framed embeddings from M to N*, denoted by $\mathrm{Emb}_f^{S_\tau}(M, N)$, is the following homotopy pullback taken in the category of topological spaces over $\mathrm{Map}^S(M, N)$:

$$\begin{tikzcd}
\mathrm{Emb}_f^{S_\tau}(M, N) \arrow[r] \arrow[d] & \mathrm{Map}^S(M, N) \arrow[d] \\
\mathrm{Emb}^S(M, N) \arrow[r] & \mathrm{Map}^{S_\tau}_{\mathrm{GL}(d)}(\mathrm{Fr}(TM), \mathrm{Fr}(TN))
\end{tikzcd}$$

Recall that a right module over an operad $\mathcal{M}$ is an $\boldsymbol{S}$–enriched functor $\boldsymbol{M}^{\mathrm{op}} \to \boldsymbol{S}$. We denote by $\mathbf{Mod}_{\mathcal{M}}$ the category of right modules over $\mathcal{M}$.

Definition 2.5 Let (S, τ) be a d–framed $(d-1)$–manifold. We define a right $\mathcal{E}_d$–module S_τ by the formula

$$S_\tau(n) = \mathrm{Emb}_f^{S_\tau}\big(D^{\sqcup n} \sqcup (S \times [0,1)), S \times [0,1)\big).$$

Recall, that there is a symmetric monoidal structure on $\mathbf{Mod}_{\mathcal{E}_d}$. If F and G are two objects of $\mathbf{Mod}_{\mathcal{E}_d}$, we can view them as contravariant functors on the groupoid Σ of finite sets and bijections. Then their tensor product is the left Kan extension of the functor

$$(n, m) \mapsto F(n) \times G(m)$$

along the functor $\Sigma^{\mathrm{op}} \times \Sigma^{\mathrm{op}} \to \Sigma^{\mathrm{op}}$ sending a pair of finite sets to their disjoint union.

Construction 2.6 We give S_τ the structure of an associative algebra in $\mathbf{Mod}_{\mathcal{E}_d}$. Let ϕ be an element of $S_\tau(m)$ and ψ be an element of $S_\tau(n)$. Let ψ^S be the restriction of ψ to $S \times [0,1)$. We define $\psi \square \phi$ to be the element of $S_\tau(m+n)$ whose restriction to $S \times [0,1) \sqcup D^{\sqcup m}$ is $\psi^S \circ \phi$ and whose restriction to $D^{\sqcup n}$ is $\psi_{|D^{\sqcup n}}$.

The operation

$$-\square-\colon S_\tau(n) \times S_\tau(m) \to S_\tau(n+m)$$

makes S_τ into an associative algebra in the symmetric monoidal category of right $\mathcal{E}_d$–modules.

Definition 2.7 The colored operad $S_\tau\mathcal{M}od$ has two colors a and m. Its only non-empty spaces of operations are

$$S_\tau\mathcal{M}od(\underbrace{a,\dots,a}_{n};a)=\mathcal{E}_d(n) \quad \text{and} \quad S_\tau\mathcal{M}od(\underbrace{a,\dots,a}_{n},m;m)=S_\tau(n).$$

The composition involves the operad structure on $\mathcal{E}_d$, the right $\mathcal{E}_d$–module structure on S_τ and the associative algebra structure on S_τ.

Again, $(\boldsymbol{C}^+,\otimes,\mathbb{I})$ denotes the symmetric monoidal model category $\mathbf{Mod}_E^+$ and $\boldsymbol{C}$ denotes the same category but with its absolute model structure. An algebra in $\boldsymbol{C}$ over $S_\tau\mathcal{M}od$ consists of a pair of objects (A,M) where A is an $\mathcal{E}_d$–algebra and M is equipped with an action of A of the form

$$\mathrm{Emb}_f^{S_\tau}(S\times[0,1)\sqcup D^{\sqcup n},S\times[0,1))\otimes M\otimes A^{\otimes n}\to M.$$

Definition 2.8 Let A be an $\mathcal{E}_d$–algebra in $\boldsymbol{C}$. We define the *category of S_τ–shaped modules* over A, denoted by $S_\tau\mathbf{Mod}_A$, to be the category whose objects are $S_\tau\mathcal{M}od$–algebras whose restriction to the color a is the $\mathcal{E}_d$–algebra A and whose morphisms are morphisms of $S_\tau\mathcal{M}od$–algebra inducing the identity map on A.

Remark 2.9 More generally, for any operad $\mathcal{O}$, and any right module P over $\mathcal{O}$, the above construction gives a notion of modules over $\mathcal{O}$–algebras. This construction is studied in detail in [11, Section 3].

Proposition 2.10 *Let A be an $\mathcal{E}_d$–algebra in $\boldsymbol{C}$. The coend*

$$U_A^{S_\tau}=S_\tau\otimes_{\boldsymbol{E}_d}A$$

inherits an associative algebra structure from the one on S_τ and there is an equivalence of categories between the category of left modules over $U_A^{S_\tau}$ and the category $S_\tau\mathbf{Mod}_A$.

Proof See [11, Proposition 3.9]. □

This proposition lets us put a model structure on $S_\tau\mathbf{Mod}_A$ in which the weak equivalences and fibrations are the maps that are sent to weak equivalences and fibrations by the forgetful functor $S_\tau\mathbf{Mod}_A\to\boldsymbol{C}$. Moreover, since $\boldsymbol{C}$ is a closed symmetric model category, the model category $S_\tau\mathbf{Mod}_A$ is a $\boldsymbol{C}$–enriched model category.

Example 2.11 The unit sphere inclusion $S^{d-1}\to\mathbb{R}^d$ has a trivial normal bundle. This induces a d–framing on S^{d-1}, which we denote by κ. On the other hand we have the notion of an operadic module over an $\mathcal{E}_d$–algebra A. This is an object M of $\boldsymbol{C}$ with multiplication maps

$$\mathcal{E}_d(n+1) \to \mathrm{Map}_{\boldsymbol{C}}(A^{\otimes n} \otimes M, M)$$

that are compatible with the $\mathcal{E}_d$–structure on A in a suitable way (see Definition 1.1 of Berger and Moerdijk [5]). We denote the category of such modules by $\mathbf{Mod}_A^{\mathcal{E}_d}$. The two notions are related by the following theorem:

Theorem 2.12 *For a cofibrant $\mathcal{E}_d$–algebra A, there is a Quillen equivalence*

$$S_\kappa \mathbf{Mod}_A \leftrightarrows \mathbf{Mod}_A^{\mathcal{E}_d}.$$

Moreover, the right adjoint of this equivalence commutes with the forgetful functor of both categories to $\boldsymbol{C}$.

Proof This is done in [11, Proposition 4.12]. The second claim follows from the fact that this equivalence is induced by a weak equivalence of associative algebras

$$U_A^{S_\kappa^{d-1}} \to U_A^{\mathcal{E}_d[1]},$$

where $U_A^{\mathcal{E}_d[1]}$ is the enveloping algebra of $\mathbf{Mod}_A^{\mathcal{E}_d}$ (ie it is an associative algebra such that there is an equivalence of categories $\mathbf{Mod}_{U_A^{\mathcal{E}_d[1]}} \simeq \mathbf{Mod}_A^{\mathcal{E}_d}$). □

Let S be a closed $(d-1)$–manifold and let τ be a d–framing of S. There is a map $S_\tau \to \mathrm{Emb}_f(-, S \times (0,1))$ sending an embedding $S \times [0,1) \sqcup D^{\sqcup n} \to S \times [0,1)$ to its restriction to $D^{\sqcup n}$.

Proposition 2.13 *The map $S_\tau \to \mathrm{Emb}_f(-, S \times (0,1))$ is a weak equivalence of right $\mathcal{E}_d$–modules.*

Proof This follows from [11, Proposition A.3] □

Corollary 2.14 *For a cofibrant $\mathcal{E}_d$–algebra A, there is a weak equivalence*

$$U_A^{S_\tau} \xrightarrow{\sim} \int_{S\times(0,1)} A.$$

Proof By the previous proposition, there is a weak equivalence of right $\mathcal{E}_d$–modules

$$S_\tau \xrightarrow{\sim} \mathrm{Emb}_f(-, S \times (0,1)).$$

We prove in [10, Proposition 2.8] that, for A cofibrant, the functor $-\otimes_{\boldsymbol{E}_d} A$ preserves all weak equivalences of right $\mathcal{E}_d$–modules. □

If A is an $\mathcal{E}_d$–algebra, then the object $\int_{S\times(0,1)} A$ is an $\mathcal{E}_1$–algebra. Indeed, any embedding $(0,1)^{\sqcup n} \to (0,1)$ induces an embedding $(0,1)\times S^{\sqcup n} \to (0,1)\times S$ by taking the product with S. Applying $\int_{\_} A$ to this last embedding, we get maps

$$\mathrm{Emb}^f((0,1)^{\sqcup n},(0,1)) \to \mathrm{Map}_{\boldsymbol{C}}\left(\left(\int_{S\times(0,1)} A\right)^{\otimes n}, \int_{S\times(0,1)} A\right).$$

We would like to say that the weak equivalence of the previous proposition is an equivalence of $\mathcal{E}_1$–algebras, but it is not one on the nose. However, we show in the next proposition that this is a map of S_τ–shaped modules.

Proposition 2.15 *There is an S_τ–shaped module structure on $\int_{S\times(0,1)} A$ such that the map*

$$U_A^{S_\tau} \to \int_{S\times(0,1)} A$$

is a weak equivalence of S_τ–shaped modules.

Proof Let us describe the S_τ–shaped module structure on $\int_{S\times(0,1)} A$. Let ϕ be a point in $\mathrm{Emb}_f^{S_\tau}(S\times[0,1)\sqcup D^{\sqcup n}, S\times[0,1))$. By forgetting about the boundary, ϕ defines a point in $\mathrm{Emb}_f(S\times(0,1)\sqcup D^{\sqcup n}, S\times(0,1))$ that induces a map

$$\left(\int_{S\times(0,1)} A\right)\otimes A^{\otimes n} \to \int_{S\times(0,1)} A.$$

Letting ϕ vary, this gives $\int_{S\times(0,1)} A$ the structure of an S_τ–shaped module. Moreover, the map $U_A^{S_\tau} \to \int_{S\times(0,1)} A$ is a map of S_τ–shaped modules. Since we already know that it is a weak equivalence, we are done. □

3 Higher Hochschild cohomology

In this section, we construct a geometric model for higher Hochschild cohomology. We still denote by $(\boldsymbol{C},\otimes,\mathbb{I})$ the symmetric monoidal model category $\mathbf{Mod}_E$. Our construction remains valid in other contexts (spaces, chain complexes, simplicial modules) modulo a few obvious modifications. We denote by $\underline{\mathrm{Hom}}$ the inner Hom in the category $\boldsymbol{C}$. This functor is uniquely determined by the fact that we have a natural isomorphism

$$\boldsymbol{C}(X\otimes Y, Z) \cong \boldsymbol{C}(X, \underline{\mathrm{Hom}}(Y,Z)).$$

For any associative R algebra in $\boldsymbol{C}$, the $\boldsymbol{C}$–enrichment of $\boldsymbol{C}$ induces to a $\boldsymbol{C}$–enrichment of $\mathbf{Mod}_R$. We denote by $\underline{\mathrm{Hom}}_R$ the homomorphisms object in $\mathbf{Mod}_R$.

Let A be an $\mathcal{E}_d$–algebra that we assume to be cofibrant. Our goal is to construct a functor

$$\mathbb{R}\underline{\mathrm{Hom}}_A^{S\times[0,1]}\colon S_\tau\mathbf{Mod}_A^{\mathrm{op}}\times S_\tau\mathbf{Mod}_A\to \boldsymbol{C}$$

that is weakly equivalent to $\mathbb{R}\underline{\mathrm{Hom}}_{S_\tau\mathbf{Mod}_A}(-,-):=\mathbb{R}\underline{\mathrm{Hom}}_{U_A^{S_\tau}}(-,-)$ but which is closer to the factorization homology philosophy.

For (S,τ) a d–framed $(d-1)$–manifold, we denote by $-\tau$ the d–framing on S obtained by pulling back τ along the isomorphism of the vector bundle $TS\oplus\mathbb{R}$ that is the identity on the first summand and multiplication by -1 on the second summand.

In particular, $S\times[0,1)$ is naturally an S_τ–manifold and $S\times(0,1]$ is an $S_{-\tau}$–manifold.

Definition 3.1 We denote by $\mathbf{Disk}_d^{S_\tau\sqcup S_{-\tau}}$ the topological category whose objects are the $S_\tau\sqcup S_{-\tau}$–manifolds of the form $S\times[0,1)\sqcup D^{\sqcup n}\sqcup S\times(0,1]$ with n in $\mathbb{Z}_{\geq 0}$ and whose morphisms are given by the spaces $\mathrm{Emb}_f^{S_\tau\sqcup S_{-\tau}}$.

Construction 3.2 We define a functor

$$\mathcal{F}(M,A,N)\colon (\mathbf{Disk}_d^{S_\tau\sqcup S_{-\tau}})^{\mathrm{op}}\to \boldsymbol{C}.$$

Its value on $S\times[0,1)\sqcup D^{\sqcup n}\sqcup S\times(-1,0]$ is $\underline{\mathrm{Hom}}(M\otimes A^{\otimes n},N)$.

Notice that any map in $(S_\tau\sqcup S_{-\tau})\mathbf{Mod}$ can be decomposed as a disjoint union of embeddings of the following three types:

- $S\times[0,1)\sqcup D^{\sqcup k}\to S\times[0,1)$.
- $D^{\sqcup l}\to D$ (where l is possibly zero).
- $D^{\sqcup m}\sqcup S\times(0,1]\to S\times(0,1]$.

Let ϕ be an embedding $S\times[0,1)\sqcup D^{\sqcup n}\sqcup S\times(0,1]\to S\times[0,1)\sqcup D^{\sqcup m}\sqcup S\times(0,1]$ and let

$$\phi=\phi_+\sqcup\psi_1\sqcup\cdots\sqcup\psi_r\sqcup\phi_-$$

be its decomposition with ϕ_+ of the first type, ϕ_- of the third type and ψ_i of the second type for each i. We need to extract from this data a map

$$\underline{\mathrm{Hom}}(M\otimes A^{\otimes m},N)\to\underline{\mathrm{Hom}}(M\otimes A^{\otimes n},N).$$

The action of ϕ_+ and of the ψ_i are constructed in an obvious way from the $\mathcal{E}_d$–structure of A and the S_τ–shaped module structure on M. The only non-trivial part is the action of ϕ_-. We can hence assume that $\phi=\mathrm{id}_{S\times[0,1)\sqcup D^{\sqcup p}}\sqcup\phi_-$, where ϕ_- is an embedding $D^{\sqcup n}\sqcup S\times(0,1]\to S\times(0,1]$. We want to construct

$$\underline{\mathrm{Hom}}(M\otimes A^{\otimes p},N)\to\underline{\mathrm{Hom}}(M\otimes A^{\otimes p}\otimes A^{\otimes n},N).$$

First, observe that there is a diffeomorphism $S \times [0,1) \to S \times (0,1]$ sending (s,t) to $(s, 1-t)$. This diffeomorphism sends the framing τ on $S \times [0,1)$ to the framing $-\tau$ on $S \times (0,1]$. Similarly, reflexion about the hyperplane $x_d = 0$ induces a diffeomorphism $D \to D$. Conjugating by this diffeomorphism, the embedding ϕ_- induces an embedding

$$\widetilde{\phi}_-\colon S \times [0,1) \sqcup D^{\sqcup n} \to S \times [0,1).$$

In fact, this construction induces a homeomorphism

$$\mathrm{Emb}_f^{S_{-\tau}}(S \times (0,1] \sqcup D^{\sqcup n}, S \times (0,1]) \to \mathrm{Emb}_f^{S_\tau}(S \times [0,1) \sqcup D^{\sqcup n}, S \times [0,1)).$$

Now, notice that $\underline{\mathrm{Hom}}(M \otimes A^{\otimes p}, N)$ has the structure of an S_τ–shaped A module induced from the one on N. Thus, the map $\widetilde{\phi}_-$ induces a map

$$\underline{\mathrm{Hom}}(M \otimes A^{\otimes p}, N) \otimes A^{\otimes n} \to \underline{\mathrm{Hom}}(M \otimes A^{\otimes p}, N).$$

This map is adjoint to a map

$$\underline{\mathrm{Hom}}(M \otimes A^{\otimes p}, N) \to \underline{\mathrm{Hom}}(M \otimes A^{\otimes p} \otimes A^{\otimes n}, N),$$

which we define to be the action of ϕ.

Remark 3.3 In order to be homotopically meaningful, we need a derived version of $\mathcal{F}(M, A, N)$. We claim that the homotopy type of $\mathcal{F}(M, A, N)$ only depends on the homotopy type of M, A and N as long as A is a cofibrant $\mathcal{E}_d$–algebra, M is a cofibrant object of $S_\tau\mathbf{Mod}_A$ and N is a fibrant object of $S_\tau\mathbf{Mod}_A$. Indeed, these conditions imply that

- the object M is cofibrant in $\boldsymbol{C}$, because the forgetful functor $S_\tau\mathbf{Mod}_A \to \boldsymbol{C}$ preserves cofibrations,
- A is cofibrant in $\boldsymbol{C}$,
- M is cofibrant in $\boldsymbol{C}$,
- N is fibrant in $\boldsymbol{C}$.

This implies that for all k, $\underline{\mathrm{Hom}}(M \otimes A^{\otimes k}, N) \simeq \mathbb{R}\underline{\mathrm{Hom}}(M \otimes A^{\otimes k}, N)$.

We denote by $\underline{\mathrm{hom}}$ the functor $\boldsymbol{S}^{\mathrm{op}} \times \boldsymbol{C} \to \boldsymbol{C}$ sending (X, C) to $\underline{\mathrm{Hom}}(X \otimes \mathbb{I}, C)$. Equivalently, this is the cotensor of $\boldsymbol{C}$ with $\boldsymbol{S}$ induced from the simplicial structure. For $\boldsymbol{A}$ a small simplicial category, F a functor from $\boldsymbol{A}$ to $\boldsymbol{S}$ and G a functor from $\boldsymbol{A}$ to $\boldsymbol{C}$, we denote by $\underline{\mathrm{hom}}_{\boldsymbol{A}}(F, G)$ the end

$$\int_{\boldsymbol{A}} \underline{\mathrm{hom}}(F(-), G(-)).$$

We denote by $\mathbb{R}\underline{\mathrm{hom}}_{\boldsymbol{A}}(F,G)$ the derived functor obtained by taking a cofibrant replacement of the source and a fibrant replacement of the target in the projective model structure of functors on $\boldsymbol{A}$.

Definition 3.4 We define $\mathbb{R}\underline{\mathrm{Hom}}_A^{S\times[0,1]}(M,N)$ to be the homotopy end

$$\mathbb{R}\underline{\mathrm{hom}}_{(\mathbf{Disk}_d^{S_\tau\sqcup S_{-\tau}})^{\mathrm{op}}}\big(\mathrm{Emb}_f^{S_\tau\sqcup S_{-\tau}}(-,S\times[0,1]),\mathcal{F}(QM,A,RN)\big),$$

where $QM\to M$ is a cofibrant replacement in $S_\tau\mathbf{Mod}_A$ and $N\to RN$ is a fibrant replacement.

We can now formulate the main theorem of this section.

Theorem 3.5 *There is a weak equivalence*

$$\mathbb{R}\underline{\mathrm{Hom}}_A^{S\times[0,1]}(M,N)\simeq\mathbb{R}\underline{\mathrm{Hom}}_{S_\tau\mathbf{Mod}_A}(M,N).$$

The rest of this section is devoted to the proof of this theorem. The reader willing to accept this result can safely skip the proof and move directly to the last subsection of this section.

Case of $\mathcal{E}_1$–algebras

The one-point space is a 0–manifold. This manifold has two 1–framings, which we call the negative and positive framing. By definition, a 1–framing of the point is the data of a basis of $\mathbb{R}$ as a $\mathbb{R}$–vector space. The positive framing is the one given by 1 and the negative framing is the one given by -1. Thus, by Definition 2.5, we get two right modules over $\mathcal{E}_1$. We denote by $\mathcal{R}$ the one corresponding to the negative framing and $\mathcal{L}$ the one corresponding to the positive framing.

Definition 3.6 A *left module* over an $\mathcal{E}_1$–algebra A is an object of the category $\mathcal{L}\mathbf{Mod}_A$. Similarly, a *right module* over A is an object of $\mathcal{R}\mathbf{Mod}_A$.

More explicitly, an object of $\mathcal{L}\mathbf{Mod}_A$ is an object of $\boldsymbol{C}$, M together with multiplication maps

$$A^{\otimes n}\otimes M\to M$$

for each embedding $[0,1)\sqcup(0,1)^{\sqcup n}\to[0,1)$ These maps are moreover supposed to satisfy a unitality and associativity condition.

We denote by $\mathbf{Disk}_1^{-+}$ the one-dimensional version of the category $\mathbf{Disk}^{S_\tau\sqcup S_{-\tau}}$ defined in Definition 3.1. As a particular case of Definition 3.4, given a cofibrant $\mathcal{E}_1$–algebra A and two left modules M and N, we can define $\underline{\mathrm{Hom}}_A^{[0,1]}(M,N)$ and this is given by natural transformations between contravariants functors on $\mathbf{Disk}_1^{-+}$.

Definition 3.7 The *category of non-commutative intervals*, denoted by $\mathbf{Ass}^{-+}$, is a skeleton of the category whose objects are finite sets containing $\{-,+\}$ and whose morphisms are maps of finite sets f preserving $-$ and $+$ together with the extra data of a linear ordering of each fiber which is such that $-$ (resp. $+$) is the smallest (resp. largest) element in the fiber over $-$ (resp. $+$).

Note that the functor π_0, sending a disjoint union of intervals to the set of connected components, is an equivalence of topological categories from $\mathbf{Disk}_1^{-+}$ to $\mathbf{Ass}^{-+}$. In fact, we could have defined $\mathbf{Ass}^{-+}$ as the homotopy category of $\mathbf{Disk}_1^{-+}$.

Let A be an associative algebra and M and N be left modules over it. We define $F(M,A,N)$ to be the obvious functor $(\mathbf{Ass}^{-+})^{\mathrm{op}} \to \boldsymbol{C}$ sending $\{-,1,\dots,n,+\}$ to $\underline{\mathrm{Hom}}(A^{\otimes n} \otimes M, N)$. The functoriality is defined analogously to Construction 3.2.

Recall that Δ^{op} can be described as a skeleton of the category whose objects are linearly ordered sets with at least two elements and morphisms are order-preserving morphisms that preserve the minimal and maximal element.

With this description, there is an obvious functor $\Delta^{\mathrm{op}} \to \mathbf{Ass}^{-+}$ sending a totally ordered set with minimal element $-$ and maximal element $+$ to the underlying finite set and sending an order-preserving map to the underlying map with the data of the induced linear ordering of each fiber.

Recall that given a triple (M,A,N) consisting of an associative algebra A and two left modules M and N, we can form the cobar construction $C^\bullet(M,A,N)$. It is a cosimplicial object of $\boldsymbol{C}$ whose value at $[n]$ is $\underline{\mathrm{Hom}}(A^{\otimes n} \otimes M, N)$. It is classical that if A and M are cofibrant and N is fibrant, then $C^\bullet(M,A,N)$ is Reedy fibrant and its totalization is a model for the derived Hom $\mathbb{R}\underline{\mathrm{Hom}}_{\mathbf{Mod}_A}(M,N)$.

Proposition 3.8 *Let A be an associative algebra and let M and N be left modules over it. The composition of $F(M,A,N)$ with the functor $\Delta \to (\mathbf{Ass}^{-+})^{\mathrm{op}}$ is the cobar construction $C^\bullet(M,A,N)$*

Proof This is a straightforward computation. □

We denote by $P\colon (\mathbf{Ass}^{-+})^{\mathrm{op}} \to \boldsymbol{S}$ the left Kan extension of the cosimplicial space that is levelwise a point along the map $\Delta \to (\mathbf{Ass}^{-+})^{\mathrm{op}}$. Concretely, P sends a finite set with two distinguished elements $-$ and $+$ to the set of linear orderings of that set whose smallest element is $-$ and largest element is $+$, seen as a discrete space.

Corollary 3.9 *Let A be a cofibrant associative algebra and let M and N be left modules over it. Then*

$$\mathbb{R}\underline{\mathrm{Hom}}_A(M,N) \simeq \mathbb{R}\underline{\mathrm{hom}}_{\mathbf{Ass}^{-+}}(P, F(M,A,N)).$$

Proof Assume that M is cofibrant and N is fibrant. If they are not, we take an appropriate replacement. The left-hand side is

$$\mathrm{Tot}\big([n] \to C^n(M, A, N) = \underline{\mathrm{Hom}}(M \otimes A^{\otimes n}, N)\big).$$

According to the cofibrancy/fibrancy assumption, this cosimplicial functor is Reedy fibrant, therefore the totalization coincides with the homotopy limit. Hence we have

$$\mathbb{R}\underline{\mathrm{Hom}}_A(M, N) \simeq \mathbb{R}\underline{\mathrm{hom}}_{\Delta}(*, C^\bullet(M, A, N)) \simeq \mathbb{R}\underline{\mathrm{hom}}_{\mathbf{Ass}^{-+}}(P, F(M, A, N)). \quad \square$$

Proposition 3.10 *Let A be a cofibrant associative algebra and let M and N be left modules over it. Then there is a weak equivalence*

$$\mathbb{R}\underline{\mathrm{Hom}}_A^{[0,1]}(M, N) \xrightarrow{\sim} \mathbb{R}\underline{\mathrm{Hom}}_A(M, N).$$

Proof Again, we can assume that M is cofibrant and N is fibrant. By the previous corollary, the right-hand side is the derived end

$$\mathbb{R}\underline{\mathrm{hom}}_{\mathbf{Ass}^{-+}}(P, F(M, A, N)),$$

which can be computed as the totalization of the Reedy fibrant cosimplicial object

$$C^\bullet(P, \mathbf{Ass}^{-+}, F(M, A, N)).$$

Similarly, the left-hand side is the totalization of the Reedy fibrant cosimplicial object

$$C^\bullet\big(\mathrm{Emb}^{-+}(-, [0,1]), \mathbf{Disk}^{-+}, \mathcal{F}(M, A, N)\big).$$

There is an obvious map of cosimplicial objects

$$C^\bullet\big(\mathrm{Emb}^{-+}(-, [0,1]), \mathbf{Disk}^{-+}, \mathcal{F}(M, A, N)\big) \to C^\bullet(P, \mathbf{Ass}^{-+}, F(M, A, N)),$$

which is degreewise a weak equivalence. Therefore, there is a weak equivalence between the totalizations

$$\mathbb{R}\underline{\mathrm{Hom}}_A^{[0,1]}(M, N) \xrightarrow{\sim} \mathbb{R}\underline{\mathrm{Hom}}_A(M, N). \qquad \square$$

If A is an $\mathcal{E}_1$–algebra, it can be seen as an object of $\mathcal{L}\mathbf{Mod}_A$ as follows. The map

$$A \otimes A^{\otimes n} \to A,$$

corresponding to an embedding

$$\phi\colon [0,1) \sqcup (0,1)^{\sqcup n} \to [0,1)$$

is defined to be the multiplication map $A^{\otimes n+1} \to A$ corresponding to the restriction of ϕ to its interior.

We denote by (A, A^m) the $\mathcal{LM}od$–algebra consisting of A acting on itself in the above way.

Corollary 3.11 *Let A be a cofibrant $\mathcal{E}_1$–algebra and N a left module. Then*

$$\mathbb{R}\underline{\mathrm{Hom}}_A^{[0,1]}(A^m, N) \simeq N.$$

Proof The pair (A, N) forms an algebra over $\mathcal{LM}od$. The operad $\mathcal{LM}od$ is weakly equivalent to the operad $L\mathcal{M}od$ parameterizing strictly associative algebras and left modules. This implies that we can find a pair (A', N') consisting of an associative algebra and a left module together with a weak equivalence of $\mathcal{LM}od$–algebra

$$(A, N) \xrightarrow{\sim} (A', N').$$

Using the previous proposition, we have

$$\mathbb{R}\underline{\mathrm{Hom}}_A^{[0,1]}(A^m, N) \simeq \mathbb{R}\underline{\mathrm{Hom}}_{A'}(A', N') \simeq N' \simeq N. \qquad \square$$

Let $\boldsymbol{D}([0,1])$ be the poset of open sets of $[0,1]$ that are diffeomorphic to

$$[0,1) \sqcup (0,1)^{\sqcup n} \sqcup (0,1]$$

for some n. Let us choose a functor

$$\delta\colon \boldsymbol{D}([0,1]) \to \mathbf{Disk}^{-+}$$

by picking a diffeomorphism of each object of $\boldsymbol{D}([0,1])$ with an object of $\mathbf{Disk}^{-+}$.

Proposition 3.12 *There is a weak equivalence*

$$\mathbb{R}\underline{\mathrm{Hom}}_A^{[0,1]}(M, N) \simeq \mathrm{holim}_{U \in \boldsymbol{D}([0,1])^{\mathrm{op}}} \mathcal{F}(M, A, N)(\delta U).$$

Proof We can assume that M is cofibrant and N is fibrant. First, by [10, Lemma 7.8], we have a weak equivalence

$$\mathrm{Emb}_f^{S^0}(-, [0,1]) \simeq \mathrm{hocolim}_{U \in \boldsymbol{D}([0,1])} \mathrm{Emb}_f^{S^0}(-, U).$$

It follows that there is an equivalence

$$\mathbb{R}\underline{\mathrm{Hom}}_A^{[0,1]}(M, N) \simeq \mathrm{holim}_{U \in \boldsymbol{D}([0,1])^{\mathrm{op}}} \mathbb{R}\underline{\mathrm{Hom}}_A^{\delta U}(M, N).$$

Then we notice, using the Yoneda lemma, that $U \mapsto \mathbb{R}\underline{\mathrm{Hom}}_A^{\delta U}(M, N)$ is weakly equivalent as a functor to $U \mapsto \mathcal{F}(M, A, N)(\delta U)$. $\square$

Comparison with the actual homomorphisms

In this subsection, A is a cofibrant $\mathcal{E}_d$–algebra. We will compare $\mathbb{R}\underline{\mathrm{Hom}}_A^{S\times[0,1]}(M,N)$ with $\mathbb{R}\underline{\mathrm{Hom}}_{S_\tau\mathbf{Mod}_A}(M,N)$.

Construction 3.13 Let M be an S_τ–shaped module over an $\mathcal{E}_d$–algebra A. We give M the structure of a left module over the $\mathcal{E}_1$–algebra $\int_{S\times(0,1)} A$. Let

$$(0,1)^{\sqcup n}\sqcup[0,1)\to[0,1)$$

be a framed embedding. We can take the product with S and get an embedding in $f\mathbf{Man}_d^{S_\tau}$,

$$(S\times(0,1))^{\sqcup n}\sqcup S\times[0,1)\to S\times[0,1).$$

Evaluating $\int_{\_}(M,A)$ over this embedding, we find a map

$$\Big(\int_{S\times(0,1)} A\Big)^{\otimes n}\otimes M\to M.$$

All these maps give M the structure of a left $\big(\int_{S\times(0,1)} A\big)$–module.

Proposition 3.14 *Let M and N be two S_τ–shaped modules over A. There is a weak equivalence*

$$\mathbb{R}\underline{\mathrm{Hom}}_A^{S\times[0,1]}(M,N)\simeq \mathrm{holim}_{U\in\boldsymbol{D}([0,1])^{\mathrm{op}}}\mathcal{F}\Big(M,\int_{S\times(0,1)} A,N\Big)(S\times U),$$

where M and N are given the structure of left $\big(\int_{S\times(0,1)} A\big)$–modules using the previous construction.

Proof This is a variant of Proposition 3.12. We first prove that

$$\mathbb{R}\underline{\mathrm{Hom}}_A^{S\times[0,1]}(M,N)\simeq \mathrm{holim}_{U\in\boldsymbol{D}([0,1])^{\mathrm{op}}}\mathbb{R}\underline{\mathrm{Hom}}_A^{S\times U}(M,N).$$

This follows from the equivalence

$$\mathrm{hocolim}_{U\in\boldsymbol{D}([0,1])}\mathrm{Emb}_f^{S_\tau\sqcup S_{-\tau}}(-,S\times U)\simeq \mathrm{Emb}_f^{S_\tau\sqcup S_{-\tau}}(-,S\times[0,1])$$

in the category $\mathrm{Fun}((\mathbf{Disk}^{S_\tau\sqcup S_{-\tau}})^{\mathrm{op}},\boldsymbol{S})$. Then, using the Yoneda lemma, we see that the functor

$$U\mapsto \mathbb{R}\underline{\mathrm{Hom}}_A^{S\times U}(M,N)$$

is weakly equivalent to

$$U\mapsto \mathcal{F}\Big(M,\int_{S\times(0,1)} A,N\Big)(U). \qquad \square$$

Corollary 3.15 *There is a weak equivalence*

$$\mathbb{R}\underline{\mathrm{Hom}}^{[0,1]}_{\int_{S\times(0,1)}A}(M,N)\simeq\mathbb{R}\underline{\mathrm{Hom}}_A^{S\times[0,1]}(M,N).$$

Proof Both sides are weakly equivalent to

$$\mathrm{holim}_{U\in\boldsymbol{D}([0,1])^{\mathrm{op}}}\mathcal{F}\left(M,\int_{S\times(0,1)}A,N\right)(S\times U),$$

one side by the previous proposition and the other by Proposition 3.12. □

Proof of Theorem 3.5 We fix A and a fibrant S_τ–shaped module N and we let M vary. We want to compare two contravariant functors from $S_\tau\mathbf{Mod}_A$ to $\boldsymbol{C}$. Both functors preserve weak equivalences between cofibrant objects and turn homotopy colimits into homotopy limits; therefore, it suffices to check that both functors are weakly equivalent on the generator of the category of S_τ–shaped modules. In other words, it is enough to prove that

$$\mathbb{R}\underline{\mathrm{Hom}}_A^{S\times[0,1]}(U_A^{S_\tau},N)\simeq\mathbb{R}\underline{\mathrm{Hom}}_{S_\tau\mathbf{Mod}_A}(U_A^{S_\tau},N).$$

The right-hand side of the above equation can be rewritten as $\mathbb{R}\underline{\mathrm{Hom}}_{U_A^{S_\tau}}(U_A^{S_\tau},N)$, which is trivially weakly equivalent to N.

We know from Proposition 2.15 that, as S_τ–shaped modules, there is a weak equivalence

$$U_A^{S_\tau}\to\int_{S\times(0,1)}A;$$

therefore, it is enough to prove that there is a weak equivalence

$$\mathbb{R}\underline{\mathrm{Hom}}_A^{S\times[0,1]}\left(\int_{S\times(0,1)}A,N\right)\simeq N.$$

According to Corollary 3.15, it is equivalent to prove that there is a weak equivalence

$$\mathbb{R}\underline{\mathrm{Hom}}^{[0,1]}_{\int_{S\times[0,1]}A}\left(\int_{S\times(0,1)}A,N\right)\simeq N.$$

This follows directly from Corollary 3.11. □

A generalization

We can generalize Definition 3.4. In [11, Construction 6.9], given the data of a framed bordism W between d–framed manifolds of dimension $d-1$, S_σ and T_τ, we construct a left Quillen functor

$$P_W\colon S_\sigma\mathbf{Mod}_A\to T_\tau\mathbf{Mod}_A.$$

The best way to think of this functor is as follows. Factorization homology of A over W is a $U_A^{S_\sigma}$-$U_A^{T_\tau}$–bimodule. Thus, tensoring with it induces a left Quillen functor

$$S_\sigma\mathbf{Mod}_A \to T_\tau\mathbf{Mod}_A.$$

Construction 3.16 Let W be bordism from S_σ to T_τ. Let M be an S_σ–shaped module over A and let N be a T_τ–shaped module. We can construct a functor $\mathcal{F}(M, A, N)$ as in Construction 3.2 from $(\mathbf{Disk}^{S_\sigma \sqcup T_{-\tau}})^{\mathrm{op}}$ to $\boldsymbol{C}$ that sends $S\times[0,1)\sqcup D^{\sqcup n}\sqcup T\times(0,1]$ to $\underline{\mathrm{Hom}}(A^{\otimes n}\otimes M, N)$. We define $\mathbb{R}\underline{\mathrm{Hom}}_A^W(M,N)$ to be the homotopy end

$$\mathbb{R}\underline{\mathrm{Hom}}_A^W(M,N) = \mathbb{R}\underline{\mathrm{hom}}_{(\mathbf{Disk}^{S_\sigma\sqcup T_{-\tau}})^{\mathrm{op}}}(\mathrm{Emb}_f^{S_\sigma\sqcup T_{-\tau}}(-,W), \mathcal{F}(M,A,N)).$$

This construction has the following nice interpretation:

Theorem 3.17 *Let W be a bordism from S_σ to T_τ. There is a weak equivalence*

$$\mathbb{R}\underline{\mathrm{Hom}}_A^W(M,N) \simeq \mathbb{R}\underline{\mathrm{Hom}}_A^{T\times[0,1]}(\mathbb{L}P_W(M), N).$$

Proof The proof is very analogous to the proof of Theorem 3.5. □

We can now introduce our definition of higher Hochschild cohomology.

Definition 3.18 Let A be a cofibrant $\mathcal{E}_d$–algebra in $\boldsymbol{C}$ and let M be an S_κ^{d-1}–shaped module over A. The *$\mathcal{E}_d$–Hochschild cohomology* of A with coefficients in M is defined as

$$\mathrm{HH}_{\mathcal{E}_d}(A,M) = \mathbb{R}\underline{\mathrm{Hom}}_{S_\kappa^{d-1}\mathbf{Mod}_A}(A,M).$$

We now compare this definition to a more traditional definition. Let A be a cofibrant $\mathcal{E}_d$–algebra and let M be an object of $\mathbf{Mod}_A^{\mathcal{E}_d}$. By Theorem 2.12, we can see M as an S_κ^{d-1}–shaped module over A.

Proposition 3.19 *For A a cofibrant $\mathcal{E}_d$–algebra and M an object of $\mathbf{Mod}_A^{\mathcal{E}_d}$, we have a weak equivalence*

$$\mathbb{R}\underline{\mathrm{Hom}}_{\mathbf{Mod}_A^{\mathcal{E}_d}}(A,M) \simeq \mathbb{R}\underline{\mathrm{Hom}}_{S_\kappa\mathbf{Mod}_A}(A,M).$$

Proof By Theorem 2.12, we have a Quillen equivalence

$$u_!\colon S_\kappa^{d-1}\mathbf{Mod}_A \leftrightarrows \mathbf{Mod}_A^{\mathcal{E}_d} :u^*.$$

Therefore, we have a weak equivalence $\mathbb{L}u_!u^*A \to A$ in $\mathbf{Mod}_A^{\mathcal{E}_d}$. This gives us a weak equivalence

$$\mathbb{R}\underline{\mathrm{Hom}}_{\mathbf{Mod}_A^{\mathcal{E}_d}}(A,M) \to \mathbb{R}\underline{\mathrm{Hom}}_{\mathbf{Mod}_A^{\mathcal{E}_d}}(\mathbb{L}u_!u^*A,M) \simeq \mathbb{R}\underline{\mathrm{Hom}}_{S_\kappa\mathbf{Mod}_A}(u^*A,u^*M). \quad \square$$

Thus, our definition of $\mathrm{HH}_{\mathcal{E}_d}(A,M)$ coincides with the more traditional definition that we gave in the first paragraph of the introduction. According to Theorem 3.5, we have a weak equivalence $\mathrm{HH}_{\mathcal{E}_d}(A,M) \simeq \mathbb{R}\underline{\mathrm{Hom}}_A^{S^{d-1}\times[0,1]}(A,M)$. As usual, we write $\mathrm{HH}_{\mathcal{E}_d}(A)$ for $\mathrm{HH}_{\mathcal{E}_d}(A,A)$.

Proposition 3.20 *Let $\overline{D}$ be the closed unit ball in $\mathbb{R}^d$ seen as a bordism from the empty manifold to S^{d-1}_κ. There is a weak equivalence*

$$\mathrm{HH}_{\mathcal{E}_d}(A,M) \simeq \mathbb{R}\underline{\mathrm{Hom}}_A^{\overline{D}}(\mathbb{I},M).$$

Proof $\mathbb{I}$, the unit of $\boldsymbol{C}$, is an object of $\varnothing\mathbf{Mod}_A$ (note that $\varnothing\mathbf{Mod}_A$ is equivalent to the category $\boldsymbol{C}$) and $\mathbb{L}P_{\overline{D}}(\mathbb{I})$ is weakly equivalent to A. Then it suffices to apply Theorem 3.17. □

This has the following surprising consequence:

Corollary 3.21 *The group $\mathrm{Diff}_f^{S^{d-1}}(\overline{D})$ acts on $\mathrm{HH}_{\mathcal{E}_d}(A,M)$.*

Remark 3.22 The group $\mathrm{Diff}_f^{S^{d-1}}(\overline{D})$ is weakly equivalent to the homotopy fiber of the inclusion

$$\mathrm{Diff}^{S^{d-1}}(\overline{D}) \to \mathrm{Imm}^{S^{d-1}}(\overline{D},\overline{D}),$$

where the S^{d-1} superscript means that we are restricting to the diffeomorphisms or immersions which are the identity outside on $S^{d-1}=\partial\overline{D}$. In fact, the action of $\mathrm{Diff}_f^{S^{d-1}}(\overline{D})$ factors through the inverse limit of the embedding calculus tower computing this group. Since we are in the codimension-0 case, the embedding calculus tower should not be expected to converge. Even if it does not converge, it is an interesting mathematical object. In particular, using the work of Arone and Turchin [3] and Willwacher [19, Theorem 1.2], we get an action of the Grothendieck–Teichmüller Lie algebra $\mathfrak{grt}$ on the $\mathcal{E}_2$–Hochschild cohomology of an algebra over $H\mathbb{Q}$. We hope to study this action further in future work.

4 Higher Hochschild homology

Let R be a commutative graded ring. We denote by $\mathbf{Ch}_{\geq 0}(R)$ the category of non-negatively graded chain complexes. This has a model category structure in which the weak equivalences are the quasi-isomorphisms, the cofibrations are the degreewise monomorphisms with degreewise projective cokernel and the fibrations are the epimorphisms. In particular, any object is fibrant and the cofibrant objects are the degreewise projective chain complexes.

The model category $\mathbf{Ch}_{\geq 0}(R)$ is cofibrantly generated. Thus, we have the projective model category structure on functors $\mathbf{Fin} \to \mathbf{Ch}_{\geq 0}(R)$, in which weak equivalences and fibrations are objectwise. The following definition is due to Pirashvili [15, Introduction, page 151] (see also Definition 2 of Ginot, Tradler and Zeinalian [8]).

Definition 4.1 Let A be a degreewise projective commutative algebra in $\mathbf{Ch}_{\geq 0}(R)$ and let X be a simplicial set. We denote by $\mathrm{HH}^X(A|R)$ the homotopy coend

$$\mathrm{Map}(-, X) \otimes^{\mathbb{L}}_{\mathbf{Fin}} A.$$

Remark 4.2 In practice, we can take $\mathrm{HH}^X(A|R)$ to be the realization of the simplicial object

$$\mathrm{B}_\bullet(\mathrm{Map}(-, X), \mathbf{Fin}, A).$$

This construction preserves quasi-isomorphism between degreewise projective commutative algebras. In the following, $\mathrm{HH}^X(A|R)$ will be taken to be this explicit model.

This construction also sends a weak equivalence $X \xrightarrow{\sim} Y$ to a weak equivalence

$$\mathrm{HH}^X(A|R) \xrightarrow{\sim} \mathrm{HH}^Y(A|R).$$

Proposition 4.3 *Let A be a degreewise projective commutative algebra in $\mathbf{Ch}_{\geq 0}(R)$; then the functor $X \mapsto \mathrm{HH}^X(A|R)$ lifts to a functor from $\boldsymbol{S}$ to the category of commutative algebras in $\mathbf{Ch}_{\geq 0}(R)$.*

Proof The category $\mathrm{Fun}(\mathbf{Fin}^{\mathrm{op}}, \boldsymbol{S})$ equipped with the convolution tensor product is a symmetric monoidal model category (see [13, Proposition 2.2.15]). It is easy to check that there is an isomorphism

$$\mathrm{Map}(-, X) \otimes \mathrm{Map}(-, Y) \cong \mathrm{Map}(-, X \sqcup Y).$$

Moreover, since $A\colon \mathbf{Fin} \to \mathbf{Ch}_{\geq 0}(R)$ is a commutative algebra for the convolution tensor product, the object $\mathrm{HH}^X(A|R)$ is a symmetric monoidal functor in the X variable. To conclude, it suffices to observe that any simplicial set is a commutative monoid with respect to the disjoint union in a unique way and that this structure is preserved by maps in $\boldsymbol{S}$. Therefore, $\mathrm{HH}^X(A|R)$ is a commutative algebra functorially in X. □

Proposition 4.4 *Let A be a degreewise projective commutative algebra in $\mathbf{Ch}_{\geq 0}(R)$. Let*

$$\begin{array}{ccc} X & \longrightarrow & Z \\ \downarrow & & \downarrow \\ Y & \longrightarrow & P \end{array}$$

be a homotopy pushout in the category of simplicial sets. Then there is a weak equivalence

$$\mathrm{HH}^P(A|R) \simeq \left|\mathrm{B}_\bullet\big(\mathrm{HH}^Y(A|R), \mathrm{HH}^X(A|R), \mathrm{HH}^Z(A|R)\big)\right|.$$

Proof First, notice that the maps $X \to Z$ and $X \to Y$ induce commutative algebra maps $\mathrm{HH}^X(A|R) \to \mathrm{HH}^Y(A|R)$ and $\mathrm{HH}^X(A|R) \to \mathrm{HH}^Z(A|R)$. In particular, $\mathrm{HH}^Z(A|R)$ and $\mathrm{HH}^Y(A|R)$ are modules over $\mathrm{HH}^X(A|R)$. This explains the bar construction in the statement of the proposition.

We can explicitly construct P as the realization of the simplicial space

$$[p] \mapsto Y \sqcup X^{\sqcup p} \sqcup Z,$$

where the face maps are induced by the codiagonals and the maps $X \to Y$ and $X \to Z$ and the degeneracies are induced by the maps from the empty simplicial set to X, Y and Z.

For a finite set S, and any simplicial space $U_\bullet$, there is an isomorphism

$$|U_\bullet^S| \cong |U_\bullet|^S.$$

Therefore, there is a weak equivalence of functors on **Fin**,

$$\mathrm{Map}(-, P) \simeq \left|\mathrm{B}_\bullet\big(\mathrm{Map}(-, Y), \mathrm{Map}(-, X), \mathrm{Map}(-, Z)\big)\right|,$$

where the bar construction on the right-hand side is in the category Fun(**Fin**, $\boldsymbol{S}$) with the convolution tensor product.

We can form the following bisimplicial object in $\mathbf{Ch}_{\geq 0}(R)$:

$$\mathrm{B}_\bullet\big(\mathrm{B}_\bullet(\mathrm{Map}(-, Y), \mathrm{Map}(-, X), \mathrm{Map}(-, Z)), \mathbf{Fin}, A\big).$$

By the previous observation, if we realize first with respect to the inner simplicial variable and then the outer one, we find something equivalent to $\mathrm{HH}^P(A|R)$. If we first realize with respect to the outer variable, we find

$$\mathrm{B}_\bullet\big(\mathrm{HH}^Y(A|R), \mathrm{HH}^X(A|R), \mathrm{HH}^Z(A|R)\big).$$

The two realizations are equivalent. This concludes the proof. □

Corollary 4.5 *Let A be a degreewise projective commutative algebra in $\mathbf{Ch}_{\geq 0}(R)$, then $\mathrm{HH}^{S^1}(A)$ is quasi-isomorphic to the Hochschild chains on A.*

Proof We can write S^1 as the homotopy pushout of:

$$\begin{tikzcd} S^0 \arrow[r] \arrow[d] & \mathrm{pt} \\ \mathrm{pt} & \end{tikzcd}$$

If S is a finite set $\mathrm{HH}^S(A) = A^{\otimes S}$ with the obvious commutative algebra structure. In particular, the previous theorem gives

$$\mathrm{HH}^{S^1}(A) \simeq |\mathrm{B}_\bullet(A, A \otimes A, A)|.$$

Since $A = A^{\mathrm{op}}$, the right-hand side is quasi-isomorphic to $A \otimes^{\mathbb{L}}_{A \otimes A^{\mathrm{op}}} A$. □

5 The spectral sequence

We construct a spectral sequence converging to factorization homology. Its E^2–page is identified with higher Hochschild homology. For R a $\mathbb{Z}$–graded ring, we denote by $\mathbf{GrMod}_R$ the category of $\mathbb{Z}$–graded left R–modules.

Definition 5.1 Let $\boldsymbol{I}$ be a small discrete category and let $F\colon \boldsymbol{I} \to \mathbf{GrMod}_R$ be a functor landing in the category of graded modules over R. We define *the homology of $\boldsymbol{I}$ with coefficients in F* to be the homology groups of the homotopy colimit of F seen as a functor concentrated in homological degree 0 from $\boldsymbol{I}$ to $\mathbf{Ch}_{\geq 0}(\mathbf{GrMod}_R)$. We write $\mathrm{H}^R_*(\boldsymbol{I}, F)$ for the homology of $\boldsymbol{I}$ with coefficients in F.

Note that since we consider graded modules, the chain complexes are graded chain complexes. This means that each homology group is graded. We denote by $\mathrm{H}^R_{s,t}(\boldsymbol{I}, F)$ the degree-t part of the s^{th} homology group. The index s lives in $\mathbb{Z}_{\geq 0}$ and the index t lives in $\mathbb{Z}$. There is an explicit model for this homology. We construct the simplicial object of $\mathbf{GrMod}_R$ whose p–simplices are

$$\mathrm{B}_p(R, \boldsymbol{I}, F) = \bigoplus_{i_0 \to \cdots \to i_p} F(i_0).$$

We can form the normalized chain complex associated to this simplicial object in $\mathbf{GrMod}_R$ and we get a non-negatively graded chain complex in $\mathbf{GrMod}_R$. Its homology groups are the homology groups of $\boldsymbol{I}$ with coefficients in F.

Recall that if E is an associative algebra in symmetric spectra, then $E_* = \pi_*(E)$ is an associative ring in graded abelian groups and, if M is a left E–module, then $\pi_*(M)$ is an object of $\mathbf{GrMod}_{E_*}$.

Proposition 5.2 *Let $F\colon \boldsymbol{I} \to \mathbf{Mod}_E$ be a functor from a discrete category to the category of left modules over an associative algebra in symmetric spectra E. There is a spectral sequence of E_*–modules*

$$\mathrm{E}^2_{s,t} \cong \mathrm{H}^{E_*}_{s,t}(\boldsymbol{I}, \pi_*(F)) \Longrightarrow \pi_{s+t}(\mathrm{hocolim}_{\boldsymbol{I}} F).$$

Proof The homotopy colimit can be computed by taking an objectwise cofibrant replacement of F and then the realization of the bar construction

$$\mathrm{hocolim}_{\boldsymbol{I}} F \simeq |\mathrm{B}_\bullet(*, \boldsymbol{I}, QF(-))|.$$

We can then use the standard spectral sequence associated to a simplicial object □

Now assume that E is commutative. Let A be an $\mathcal{E}_d$–algebra in $\mathbf{Mod}_E$. Let M be a framed d–manifold and let $\boldsymbol{D}(M)$ be the poset of open sets of M that are diffeomorphic to a disjoint union of copies of D. We know from Proposition 1.10 that the factorization homology of A over M can be computed as the homotopy colimit of the composition

$$\boldsymbol{D}(M) \xrightarrow{\delta} \boldsymbol{E}_d \xrightarrow{A} \mathbf{Mod}_E.$$

Hence, we are in a situation where we can apply the previous proposition. We get a spectral sequence of E_*–modules

$$\mathrm{H}^{E_*}_{s,t}(\boldsymbol{D}(M), \pi_*(A \circ \delta)) \Longrightarrow \pi_{s+t}\left(\int_M A\right).$$

We want to exploit the fact that A is a monoidal functor to obtain a more explicit model for the left-hand side in some cases.

From now on, K denotes an associative algebra in spectra whose associated homology theory has a $\mathbb{Z}/2$–equivariant Künneth isomorphism. That is, we assume that the obvious map

$$K_*(X) \otimes_{K_*} K_*(Y) \to K_*(X \wedge Y)$$

is an isomorphism of functors of the pair (X, Y) that is equivariant with respect to the obvious $\mathbb{Z}/2$–action on both sides. Examples of such ring spectra are the Eilenberg–MacLane spectra Hk for any field k and $K(n)$, the Morava K–theory of height n at odd primes.

We just smash the simplicial object computing $\mathrm{hocolim}_{\boldsymbol{D}(M)} A(\delta -)$ with K in each degree and take the associated spectral sequence. We then get a spectral sequence of $K_*(E)$–modules

$$\mathrm{H}^{K_*E}_*(\boldsymbol{D}(M), K_*(A \circ \delta)) \Longrightarrow K_*\left(\int_M A\right).$$

Now we want to identify $K_*(A \circ \delta)$ as a functor on $\boldsymbol{D}(M)$.

Proposition 5.3 *If $d=1$, $K_*(A)$ is an associative algebra in K_*E–modules. If $d>1$, $K_*(A)$ is a commutative algebra in the category of K_*E–modules.*

Proof An $\mathcal{E}_1$ algebra in $\mathbf{Mod}_E$ is in particular an associative algebra in $\mathrm{Ho}(\mathbf{Mod}_E)$ and an $\mathcal{E}_d$–algebra with $d>1$ is a commutative algebra in $\mathrm{Ho}(\mathbf{Mod}_E)$. The result then follows from the fact that the functor

$$K_*\colon \mathrm{Ho}(\mathbf{Mod}_E) \to \mathbf{GrMod}_{K_*E}$$

is symmetric monoidal. □

Now, we focus on the case where $d>1$. We have an obvious functor $\alpha\colon \boldsymbol{D}(M) \to \mathbf{Fin}$ that sends a configuration of disks on M to its set of connected components. In particular, we can consider the functor

$$\boldsymbol{D}(M) \xrightarrow{\alpha} \mathbf{Fin} \xrightarrow{K_*(A)} \mathbf{GrMod}_{K_*E},$$

where the second map is induced by the commutative algebra structure on $K_*(A)$ that we have constructed in the previous proposition. It is clear that this functor coincides with the functor obtained by applying K_* to the composite

$$\boldsymbol{D}(M) \xrightarrow{\delta} \boldsymbol{E}_d \xrightarrow{A} \mathbf{Mod}_E.$$

From this, we deduce the following proposition:

Proposition 5.4 *There is an isomorphism*

$$\mathrm{H}_*^{K_*E}(\boldsymbol{D}(M), K_*(A\circ\delta)) \cong \mathrm{HH}_*^{\mathrm{Sing}(M)}(K_*A|K_*E).$$

In particular, there is a spectral sequence

$$\mathrm{HH}_s^{\mathrm{Sing}(M)}(K_*A|K_*E)_t \Longrightarrow K_{s+t}\Big(\int_M A\Big).$$

Proof The first claim immediately implies the second.

In order to prove the first claim, we observe that we have weak equivalences

$$* \otimes^{\mathbb{L}}_{\boldsymbol{D}(M)} K_*(A\circ\delta) \simeq \mathbb{L}\alpha_! * \otimes^{\mathbb{L}}_{\mathbf{Fin}} K_*(A),$$

where $*$ denotes the constant functor with value $*$.

We have $\mathbb{L}\alpha_! * (S) = \mathrm{hocolim}_{U\in \boldsymbol{D}(M)} \mathbf{Fin}(S, \pi_0(U))$. By [11, Proposition 5.3], this contravariant functor on $\mathbf{Fin}$ coincides up to weak equivalences with $S \mapsto \mathrm{Sing}(M)^S$. □

Remark 5.5 The spectral sequence above still exists if K does not have a Künneth isomorphism as long as K_*A is flat as a K_*–module. We leave the details to the interested reader.

Multiplicative structure

Let us start with the general homotopy colimit spectral sequence.

Proposition 5.6 *Let $F\colon \boldsymbol{I} \to \mathbf{Mod}_E$ and $G\colon \boldsymbol{J} \to \mathbf{Mod}_E$ be functors. We have the equivalence*

$$\mathrm{hocolim}_{\boldsymbol{I}\times\boldsymbol{J}} F \otimes_E G \simeq (\mathrm{hocolim}_{\boldsymbol{I}} F) \otimes_E (\mathrm{hocolim}_{\boldsymbol{J}} G).$$

Proof Assume F and G are objectwise cofibrant. The right-hand side is the homotopy colimit over $\Delta^{\mathrm{op}} \times \Delta^{\mathrm{op}}$ of

$$\mathrm{B}_\bullet(*, \boldsymbol{I}, F) \otimes \mathrm{B}_\bullet(*, \boldsymbol{J}, G).$$

The diagonal of this bisimplicial object is exactly

$$\mathrm{B}_\bullet(*, \boldsymbol{I} \times \boldsymbol{J}, F \otimes_E G).$$

Since $\Delta^{\mathrm{op}} \to \Delta^{\mathrm{op}} \times \Delta^{\mathrm{op}}$ is homotopy cofinal, we are done. $\square$

We denote by $\mathrm{E}^r_{**}(\boldsymbol{I}, F)$ and $\mathrm{E}^r_{**}(\boldsymbol{J}, G)$ the spectral sequence computing the homotopy colimit of $F\colon \boldsymbol{I} \to \mathbf{Mod}_E$ and $G\colon J \to \mathbf{Mod}_E$. Then there is a pairing of spectral sequences of E_*–modules

$$\mathrm{E}^r_{**}(\boldsymbol{I}, F) \otimes_{E_*} \mathrm{E}^r_{**}(\boldsymbol{J}, G) \to \mathrm{E}^r_{**}(\boldsymbol{I} \times \boldsymbol{J}, F \otimes_E G).$$

Let us specialize to the case of factorization homology. We consider an $\mathcal{E}_d$–algebra A in $\mathbf{Mod}_E$, a homology theory with $\mathbb{Z}/2$–equivariant Künneth isomorphism K and a framed manifold M of dimension d. We denote by $\mathrm{E}^r_{**}(M, A, K)$ the spectral sequence of the previous section.

Proposition 5.7 *Let M and N be two framed d–manifolds. There is a pairing of spectral sequences*

$$\mathrm{E}^r_{**}(M, A, K) \otimes_{K_*E} \mathrm{E}^r_{**}(N, A, K) \to \mathrm{E}^r_{**}(M \sqcup N, A, K).$$

Proof We observe that $\boldsymbol{D}(M \sqcup N) \cong \boldsymbol{D}(M) \times \boldsymbol{D}(N)$ and that $A \otimes_E A$ as a functor on $\boldsymbol{D}(M) \times \boldsymbol{D}(N)$ is equivalent to A as a functor on $\boldsymbol{D}(M \sqcup N)$. Then the pairing of spectral sequences of the previous paragraph reduces exactly to the desired result. $\square$

The topological category $f\mathbf{Man}_d$ of framed d–manifolds and framed embeddings has a symmetric monoidal structure given by the disjoint union operation. This induces a symmetric monoidal structure on the ordinary category $\pi_0 f\mathbf{Man}_d$ which is the category obtained by applying π_0 to each mapping space of $f\mathbf{Man}_d$. We say that a framed d–manifold is an associative algebra up to isotopy if it has the structure of an associative algebra in $\pi_0 f\mathbf{Man}_d$. Examples of manifolds with such a structure are obtained by starting with a d–framed $(d-1)$–manifold N and then constructing the framed d–manifold $M = N \times (-1, 1)$. This manifold M has the structure of an $\mathcal{E}_1$–algebra in $f\mathbf{Man}_d$. In particular, it is an associative algebra up to isotopy.

There is a similar story in $\boldsymbol{S}$. This category has a symmetric monoidal structure with respect to the coproduct $\sqcup$. Any object has a unique commutative algebra structure given by the codiagonal $X \sqcup X \to X$. In particular, if M is an associative algebra up to isotopy, this structure reduces to the canonical multiplication on $\mathrm{Sing}(M)$.

Proposition 5.8 *Let M be a framed manifold of dimension $d \geq 2$ with the structure of an associative algebra up to isotopy. Let A be an $\mathcal{E}_d$–algebra. The spectral sequence $\mathrm{E}^r_{**}(M, A, K)$ has a commutative multiplicative structure converging to the associative algebra structure on $K_* \int_M A$. On the E^2–page, the multiplication is induced by the unique commutative algebra structure on $\mathrm{Sing}(M)$ in the category $(\boldsymbol{S}, \sqcup)$. Moreover, this structure is functorial with respect to embeddings of d–manifolds $M \to M'$ preserving the multiplication up to isotopy.*

Proof According to the previous proposition there is a multiplicative structure on the spectral sequence converging to the associative algebra structure on $K_* \int_M A$.

It is easy to see that the multiplication on the E^2–page is what is stated in the proposition. Since $\mathrm{Sing}(M)$ is commutative, the multiplication on the E^2–page is commutative. The homology of a commutative differential graded algebra is a commutative algebra, therefore the multiplication is commutative on each page.

The functoriality is clear. □

Now we want to construct an edge homomorphism. Let S be a $(d-1)$–manifold with a d–framing τ. Let ϕ be a framed embedding of $\mathbb{R}^{d-1} \times \mathbb{R}$ into $S \times \mathbb{R}$ commuting with the projection to $\mathbb{R}$. Applying factorization homology, we get a map of $\mathcal{E}_1$–algebras

$$u_\phi\colon A \cong \int_{\mathbb{R}^{d-1}\times\mathbb{R}} A \to \int_{S\times\mathbb{R}} A.$$

On the other hand, for any point x of $S \times \mathbb{R}$ we get a morphism of commutative algebras over K_*E,

$$u_x\colon K_*(A) \cong \mathrm{HH}^{\mathrm{pt}}(K_*A|K_*E) \to \mathrm{HH}^{\mathrm{Sing}(S)}(K_*A|K_*E).$$

Proposition 5.9 *For any framed embedding $\phi\colon \mathbb{R}^{d-1}\times\mathbb{R}\to S\times\mathbb{R}$, there is an edge homomorphism*

$$K_*A \to \mathrm{E}^r_{0,*}(S\times\mathbb{R}, A, K).$$

On the E^2–page it is identified with the K_*E–algebra homomorphism

$$u_{\phi(0,0)}\colon K_*(A)\to \mathrm{HH}^{\mathrm{pt}}(K_*A|K_*E)\to \mathrm{HH}^{\mathrm{Sing}(S)}(K_*A|K_*E)$$

and it converges to the K_*E–algebra homomorphism

$$K_*(u_\phi)\colon K_*A\to K_*\int_{N\times\mathbb{R}} A.$$

Proof The spectral sequence computing $K_*\int_{\mathbb{R}^{d-1}\times\mathbb{R}} A$ has its E^2–page K_*A concentrated on the 0^{th} column. For degree reasons, it degenerates. Then the result follows directly from the functoriality of the spectral sequence applied to the map ϕ. □

Note that the edge homomorphism only depends on the connected component of the image of ϕ. In the case of the sphere $S^{d-1}\times\mathbb{R}$ with the framing κ, we have a stronger result:

Lemma 5.10 *For any framed embedding $\phi\colon \mathbb{R}^{d-1}\times\mathbb{R}\to (S^{d-1}\times\mathbb{R})_\kappa$ commuting with the projection to $\mathbb{R}$, the map*

$$u_\phi\colon A\to \int_{S^{d-1}\times\mathbb{R}} A$$

has a section in the homotopy category of $\mathbf{Mod}_E$.

Proof There is an embedding

$$S^{d-1}\times\mathbb{R}\to\mathbb{R}^d$$

sending (θ,x) to $e^x\theta$. This embedding preserves the framing up to isotopy. Moreover, since $\mathrm{Emb}_f(\mathbb{R}^d,\mathbb{R}^d)$ is contractible, the composite

$$\mathbb{R}^d \xrightarrow{\phi} S^{d-1}\times\mathbb{R}\to\mathbb{R}^d$$

is isotopic to the identity. We can apply $\int_{\_} A$ to this sequence of morphisms of framed manifolds and we obtain the desired section. □

Although we will not need it, this has the following immediate corollary:

Corollary 5.11 *The image of the edge homomorphism in $\mathrm{E}^r_{**}((S^{d-1}\times\mathbb{R})_\kappa, A, K)$ consists of permanent cycles.*

Remark 5.12 Our geometric description of higher Hochschild cohomology (Definition 3.4) can be used to construct a similar spectral sequence calculating $K_*\mathrm{HH}_{\mathcal{E}_d}(A)$ whose E_2–page is a cohomological version of the higher Hochschild cohomology defined by Ginot [7]. However, this spectral sequence does not always converge.

6 Computations

Proposition 6.1 *Let A_* be a degreewise projective commutative graded algebra over a commutative graded ring R_*. Assume that A_* is a filtered colimit of étale algebras over R_*. Then, for all $d \geq 1$, the unit map*

$$A_* \to \mathrm{HH}^{S^d}(A_*|R_*)$$

is a quasi-isomorphism of commutative R_*–algebras.

Proof We proceed by induction on d. For $d=1$, $\mathrm{HH}^{S^1}(A_*|R_*)$ is quasi-isomorphic to the ordinary Hochschild homology $\mathrm{HH}(A_*|R_*)$ by Corollary 4.5. If A_* is étale, the result is well known (see for instance [18, Étale descent theorem, page 368]). If A_* is a filtered colimit of étale algebras, the result follows from the fact that Hochschild homology commutes with filtered colimits.

Now assume that $A_* \to \mathrm{HH}^{S^{d-1}}(A_*|R_*)$ is a quasi-isomorphism of commutative algebras. The sphere S^d is part of the following homotopy pushout diagram:

$$\begin{tikzcd} S^{d-1} \arrow[r] \arrow[d] & \mathrm{pt} \arrow[d] \\ \mathrm{pt} \arrow[r] & S^d \end{tikzcd}$$

Applying Proposition 4.4, we find

$$\mathrm{HH}^{S^d}(A_*|R) \simeq |\mathrm{B}_\bullet(A_*, \mathrm{HH}^{S^{d-1}}(A_*|R_*), A_*)|.$$

The quasi-isomorphism $A_* \to \mathrm{HH}^{S^{d-1}}(A_*|R_*)$ induces a degreewise quasi-isomorphism between Reedy cofibrant simplicial objects:

$$\mathrm{B}_\bullet(A_*, A_*, A_*) \to \mathrm{B}_\bullet(A_*, \mathrm{HH}^{S^{d-1}}(A_*|R_*), A_*).$$

This induces a quasi-isomorphism between their realizations,

$$A_* \simeq \mathrm{HH}^{S^d}(A_*|R_*). \qquad \square$$

Corollary 6.2 *Let A be an $\mathcal{E}_d$–algebra in $\mathbf{C}$ such that $K_*(A)$ is a filtered colimits of étale algebras over K_*; then the unit map*

$$A \to \int_{S^{d-1}\times\mathbb{R}} A$$

is a K–local equivalence.

Proof The K–homology of this map can be computed as the edge homomorphism of the spectral sequence $\mathrm{E}^2(S^{d-1}\times\mathbb{R}, A, K)$. By the previous proposition, the edge homomorphism is an isomorphism on the E^2–page. Therefore, the spectral sequence collapses at the E^2–page for degree reasons. □

Let us fix a prime p. We denote by E_n the Lubin–Tate ring spectrum of height n at p and by K_n the 2–periodic Morava K–theory of height n. Recall that

$$\begin{aligned} (E_n)_* &\cong \mathbb{W}(\mathbb{F}_{p^n})[\![u_1,\dots,u_{n-1}]\!][u^{\pm 1}], && |u_i|=0,\ |u|=2,\\ (K_n)_* &\cong \mathbb{F}_{p^n}[u^{\pm 1}] = (E_n)_*/(p,u_1,\dots,u_{n-1}). \end{aligned}$$

The spectrum E_n is known to have a unique $\mathcal{E}_1$–structure inducing the correct multiplication on homotopy groups (this is a theorem of Hopkins and Miller; see [16]) and a unique commutative structure (see [9, Corollary 7.6]). As far as we know, there is no published proof that the space of $\mathcal{E}_d$–structure for $d\geq 2$ is contractible, although evidence suggests that this is the case. The ring spectrum K_n has a $\mathbb{Z}/2$–equivariant Künneth isomorphism if p is odd. If $p=2$, the equivariance is not satisfied in general but it is true if we restrict $(K_n)_*$ to spectra whose K_n–homology is concentrated in even degree, like E_n. Our argument works at $p=2$ modulo this minor modification.

Corollary 6.3 *For any positive integer n and any $\mathcal{E}_d$–algebra structure on E_n inducing the correct multiplication on homotopy groups, the unit map*

$$E_n \to \int_{S^{d-1}\times\mathbb{R}} E_n$$

induces an isomorphism in K_n–homology.

Proof By [12, Corollary 4.10], for any such $\mathcal{E}_d$–structure on E we have

$$(K_n)_*(E_n) \cong C(\Gamma, (K_n)_*).$$

Here the right-hand side denotes the set of continuous maps $\Gamma \to (K_n)_*$, where Γ is the Morava stabilizer group with its profinite topology and $(K_n)_*$ is given the discrete topology. By definition of a profinite group, the group Γ is an inverse limit

$\Gamma = \lim_U \Gamma/U$ taken over the filtered poset of open finite index subgroups U of Γ. Thus, we have

$$C(\Gamma, (K_n)_*) = \operatorname{colim}_U C(\Gamma/U, (K_n)_*).$$

This expresses $(K_n)_* E_n$ as a filtered colimit of étale algebras over $(K_n)_*$. Using Corollary 6.2, we get the desired result. $\square$

Proposition 6.4 *With the same notations, the map $\mathrm{HH}_{\mathcal{E}_d}(E_n) \to E_n$ is an equivalence.*

Proof We have

$$\mathrm{HH}_{\mathcal{E}_d}(E_n) \simeq \mathbb{R}\underline{\mathrm{Hom}}_{\int_{S^{d-1}\times\mathbb{R}} E_n}(E_n, E_n).$$

This can be computed as the end

$$\underline{\hom}_{\mathbf{Disk}^{-+}}\left(\mathrm{Emb}^{S^0}(-, [0,1]), \mathcal{F}\left(E_n, \int_{S^{d-1}\times\mathbb{R}} E_n, E_n\right)\right).$$

The spectrum E_n is $K(n)$–local; therefore, $\underline{\mathrm{Hom}}(-, E_n)$ sends $K(n)$–equivalences to equivalences. This implies that

$$\mathcal{F}\left(E_n, \int_{S^{d-1}\times\mathbb{R}} E_n, E_n\right) \simeq \mathcal{F}(E_n, E_n, E_n).$$

Therefore, we have

$$\mathrm{HH}_{\mathcal{E}_d}(E_n) \simeq \mathbb{R}\underline{\mathrm{Hom}}_{E_n}(E_n, E_n). \qquad \square$$

We can prove a variant of the previous result. Let $E(n) = BP/(v_{n+1}, v_{n+2}, \dots)[v_n^{-1}]$ be the Johnson–Wilson spectrum and let $K(n)$ be the v_n periodic Morava K–theory with $K(n)_* = E(n)/(p, v_1, \dots, v_{n-1}) = \mathbb{F}_p[v_n^{\pm 1}]$. Let $\widehat{E}(n)$ be $L_{K(n)}E(n)$.

Proposition 6.5 *For any $\mathcal{E}_d$–algebra structure on $\widehat{E}(n)$ inducing the correct multiplication on homotopy groups, the action map*

$$\mathrm{HH}_{\mathcal{E}_d}(\widehat{E}(n)) \to \widehat{E}(n)$$

is a weak equivalence.

Proof The proof is exactly the same once we know that $K(n)_*\widehat{E}(n)$ is the commutative ring

$$K(n)_*\widehat{E}(n) = C(\Gamma, K(n)_*),$$

where Γ is again the Morava stabilizer group. $\square$

7 Étale base change for Hochschild cohomology

In this section we put the previous result in the wider context of derived algebraic geometry over $\mathcal{E}_d$–algebras. This section is inspired by Francis [6].

We let $(\boldsymbol{C}, \otimes, \mathbb{I})$ denote the category $\mathbf{Mod}_E$ but the arguments hold more generally. Note however that we need $\boldsymbol{C}$ to be stable in this section.

There is a "polar coordinate" embedding $S^{d-1} \times (0,1) \to D$ sending (θ, r) to $e^{r-1}\theta$.

Definition 7.1 Let A be an $\mathcal{E}_d$–algebra in $\boldsymbol{C}$. The cotangent complex L_A of A is defined to be the n–fold desuspension of the cofiber of the map

$$\int_{S^{d-1} \times \mathbb{R}} A \to \int_{\mathbb{R}^{\mathbb{d}}} A \cong A$$

induced by the polar coordinate embedding.

Proposition 7.2 *This coincides with the cotangent complex of A defined by Francis.*

Proof Both sides of the map commute with homotopy colimits of $\mathcal{E}_d$–algebras; therefore, it suffices to check the claim for free $\mathcal{E}_d$–algebras. Let $A = F_{\mathcal{E}_d}(V)$. Using [4, Proposition 5.5], we see that

$$\int_{S^{d-1} \times (0,1)} F_{\mathcal{E}_d}(V) \simeq \bigvee_{i \geq 0} \mathrm{Conf}(i, S^{d-1} \times (0,1)) \otimes_{\Sigma_i} V^{\otimes i}$$

and, similarly,

$$\int_D F_{\mathcal{E}_d}(V) \simeq \bigvee_{i \geq 0} \mathrm{Conf}(i, D) \otimes_{\Sigma_i} V^{\otimes i}.$$

On the other hand, it is proved in [6, Theorem 2.26] that there is a cofiber sequence

$$\int_{S^{d-1} \times (0,1)} A \to A \to L_A[n].$$

Moreover, the proof of [6, Theorem 2.26] is based on an explicit computation in the free case and an inspection of this proof shows that the first map in the above cofiber sequence coincides with the polar embedding map. □

Remark 7.3 The above definition is a bit ad hoc. Francis actually defines in [6, Definition 2.10] the cotangent complex as the object representing the $\mathcal{E}_d$–derivations. That is, we have a weak equivalence

$$\mathbb{R}\underline{\mathrm{Hom}}_{S^{d-1}_\kappa \mathbf{Mod}_A}(L_A, M) \simeq \mathbb{R}\underline{\mathrm{Hom}}_{\boldsymbol{C}[\mathcal{E}_d]/A}(A, A \oplus M) := \mathrm{Der}(A, M).$$

The fact that the two definitions coincide is [6, Theorem 2.26].

Definition 7.4 We say that an $\mathcal{E}_d$–algebra A is étale if L_A is contractible. More generally, given an object Z in $\boldsymbol{C}$, we say that A is Z–locally étale if $Z \otimes L_A$ is contractible.

We say that a map $X \to Y$ in $\boldsymbol{C}$ is a Z–local weak equivalence if the induced map $X \otimes^{\mathbb{L}} Z \to Y \otimes^{\mathbb{L}} Z$ is a weak equivalence.

An equivalent formulation of the previous definition is that A is (Z–locally) étale if the unit map $A \to \int_{S^{d-1}\times(0,1)} A$ is a (Z–local) equivalence. Indeed we have shown in Lemma 5.10 that the unit map is a section of $\int_{S^{d-1}\times(0,1)} A \to A$.

Proposition 7.5 *If A is a commutative algebra and is (Z–locally) étale as an $\mathcal{E}_d$–algebra, then it is (Z–locally) étale as an $\mathcal{E}_{d+1}$–algebra.*

Proof We have proved in [11, Theorem 5.8] that, for A a commutative algebra, $\int_M A$ is equivalent to $\mathrm{Sing}(M) \otimes A$ (ie the tensor in the category of commutative algebras in $\mathbf{Mod}_E$). Then the proof is the same as the proof of Proposition 6.1. □

Remark 7.6 More generally, using the excision property for factorization homology (see [4, Lemma 3.18]), we can prove that if A is $\mathcal{E}_{d+1}$ and is (Z–locally) étale as an $\mathcal{E}_d$–algebra, it is (Z–locally) étale as an $\mathcal{E}_{d+1}$–algebra.

Remark 7.7 If A is a commutative algebra, then A is étale as an $\mathcal{E}_2$–algebra if and only if it is formally THH–étale (ie if the map $A \to \mathrm{THH}(A)$ is an equivalence). Indeed, for commutative algebras (and in fact for $\mathcal{E}_3$–algebras), $\mathrm{THH}(A)$ coincides with $\int_{S^1\times\mathbb{R}} A$. Note that this is *not* true for $\mathcal{E}_2$–algebras, as the product framing on $S^1 \times \mathbb{R}$ is not connected to the κ–framing in the space of framings of $S^1 \times \mathbb{R}$.

Recall that an object U of $\boldsymbol{C}$ is said to be Z–local if, for all Z–local weak equivalences $X \to Y$, the induced map

$$\mathbb{R}\underline{\mathrm{Hom}}(Y, U) \to \mathbb{R}\underline{\mathrm{Hom}}(X, U)$$

is a weak equivalence in $\boldsymbol{C}$.

Lemma 7.8 *Let $u\colon R \to S$ be a map of cofibrant associative algebras in $\boldsymbol{C}$ that is a Z–local weak equivalence and let M and N be two left modules over S with N Z–local in $\boldsymbol{C}$. Then the map*

$$\mathbb{R}\underline{\mathrm{Hom}}_{\mathbf{Mod}_S}(M, N) \to \mathbb{R}\underline{\mathrm{Hom}}_{\mathbf{Mod}_R}(u^*M, u^*N)$$

is a weak equivalence.

Proof The left-hand side can be computed as the homotopy limit of the cobar construction

$$[n] \mapsto \underline{\mathrm{Hom}}(S^{\otimes n} \otimes M, N).$$

Similarly, the left-hand side can be computed as the homotopy limit of

$$[n] \mapsto \underline{\mathrm{Hom}}(R^{\otimes n} \otimes M, N).$$

Since $R \to S$ is a Z–local weak equivalence, so is $R^{\otimes n} \otimes M \to S^{\otimes n} \otimes M$ for each n. Thus, since N is Z–local, the two cosimplicial objects are weakly equivalent. This implies that they have weakly equivalent homotopy limits. □

We can now state and prove the main theorem of this section.

Theorem 7.9 *Let T be a commutative algebra in $\mathbf{C}$ that is (Z–locally) étale as an $\mathcal{E}_d$–algebra over $\mathbb{I}$. Then, for any $\mathcal{E}_d$–algebra A over T (that is Z–local as an object of $\mathbf{C}$), the base-change map*

$$\mathrm{HH}_{\mathcal{E}_d}(A) \to \mathrm{HH}_{\mathcal{E}_d}(A|T)$$

is a weak equivalence.

Proof We write $A|T$ whenever we want to emphasize the fact that we are viewing A as an $\mathcal{E}_d$–algebra over T.

By Proposition 2.11 of Francis [6], there is cofiber sequence

$$u_! L_T \to L_A \to L_{A|T},$$

where $u\colon T \to A$ is the unit map and $u_!$ is the corresponding functor

$$u_!\colon S^{d-1}_\kappa \mathbf{Mod}_T \to S^{d-1}_\kappa \mathbf{Mod}_A.$$

By hypothesis, L_T is (Z–locally) contractible; therefore, $L_A \to L_{A|T}$ is a (Z–local) equivalence. We have a base-change map of cofiber sequences:

$$\begin{tikzcd}
\Sigma^{d-1} L_A \arrow[r] \arrow[d] & \int_{S^{d-1}\times(0,1)} A \arrow[r] \arrow[d] & A \arrow[r] \arrow[d, "\mathrm{id}"] & \Sigma^d L_A \arrow[d] \\
\Sigma^{d-1} L_{A|T} \arrow[r] & \int_{S^{d-1}\times(0,1)} A|T \arrow[r] & A \arrow[r] & \Sigma^d L_{A|T}
\end{tikzcd}$$

This implies that $\int_{S^{d-1}\times(0,1)} A \to \int_{S^{d-1}\times(0,1)} A|T$ is a (Z–local) equivalence.

We can form the commutative diagram

$$\begin{tikzcd} U_A^{S_\kappa^{d-1}} \arrow[r] \arrow[d] & \int_{S^{d-1}\times(0,1)} A \arrow[d] \\ U_{A|T}^{S_\kappa^{d-1}} \arrow[r] & \int_{S^{d-1}\times(0,1)} A|T, \end{tikzcd}$$

where the horizontal maps are the maps of Corollary 2.14. These maps are weak equivalences by Corollary 2.14. Thus, the map $U_A^{S_\kappa^{d-1}} \to U_{A|T}^{S_\kappa^{d-1}}$ is a (Z–local) weak equivalence of associative algebras. The theorem follows from this fact and the previous lemma. □

Remark 7.10 The computation of Section 6 implies that $S \to E_n$ is $K(n)$–locally an étale morphism of $\mathcal{E}_d$–algebras for all d. Therefore, given a $K(n)$–local E_n–algebra A, we can compute its (higher) Hochschild cohomology over E_n or over S without affecting the result. This fact is used by Angeltveit [2, Theorem 6.9] in the case of ordinary Hochschild cohomology.

References

[1] **R Andrade**, *From manifolds to invariants of E_n–algebras*, PhD thesis, Massachusetts Institute of Technology (2010) MR2898647 Available at `http://math.stanford.edu/~randrade/thesis/`

[2] **V Angeltveit**, *Topological Hochschild homology and cohomology of A_∞ ring spectra*, Geom. Topol. 12 (2008) 987–1032 MR2403804

[3] **G Arone**, **V Turchin**, *Graph-complexes computing the rational homotopy of high dimensional analogues of spaces of long knots*, preprint (2013) `arXiv:1108.1001`

[4] **D Ayala**, **J Francis**, *Factorization homology of topological manifolds*, J. Topology (2015) online publication

[5] **C Berger**, **I Moerdijk**, *On the derived category of an algebra over an operad*, Georgian Math. J. 16 (2009) 13–28 MR2527612

[6] **J Francis**, *The tangent complex and Hochschild cohomology of $\mathcal{E}_n$–rings*, Compos. Math. 149 (2013) 430–480 MR3040746

[7] **G Ginot**, *Higher order Hochschild cohomology*, C. R. Math. Acad. Sci. Paris 346 (2008) 5–10 MR2383113

[8] **G Ginot**, **T Tradler**, **M Zeinalian**, *Higher Hochschild homology, topological chiral homology and factorization algebras*, Comm. Math. Phys. 326 (2014) 635–686 MR3173402

[9] **P G Goerss**, **M J Hopkins**, *Moduli spaces of commutative ring spectra*, from: "Structured ring spectra", (A Baker, B Richter, editors), London Math. Soc. Lecture Note Ser. 315, Cambridge Univ. Press (2004) 151–200 MR2125040

[10] **G Horel**, *Factorization homology and calculus à la Kontsevich Soibelman*, preprint (2015) arXiv:1307.0322

[11] **G Horel**, *Operads, modules and topological field theories*, preprint (2015) arXiv:1405.5409

[12] **M Hovey**, *Operations and co-operations in Morava E–theory*, Homology Homotopy Appl. 6 (2004) 201–236 MR2076002

[13] **S B Isaacson**, *Cubical homotopy theory and monoidal model categories*, PhD thesis, Harvard (2009) MR2713397 Available at http://search.proquest.com/docview/304891860

[14] **D Pavlov**, **J Scholbach**, *Symmetric operads in abstract symmetric spectra*, preprint (2014) arXiv:1410.5699

[15] **T Pirashvili**, *Hodge decomposition for higher order Hochschild homology*, Ann. Sci. École Norm. Sup. 33 (2000) 151–179 MR1755114

[16] **C Rezk**, *Notes on the Hopkins–Miller theorem*, from: "Homotopy theory via algebraic geometry and group representations", (M Mahowald, S Priddy, editors), Contemp. Math. 220, Amer. Math. Soc. (1998) 313–366 MR1642902

[17] **S Schwede**, *An untitled book project about symmetric spectra*, preprint (2007) Available at http://www.math.uni-bonn.de/people/schwede/SymSpec.pdf

[18] **C A Weibel**, **S C Geller**, *Étale descent for Hochschild and cyclic homology*, Comment. Math. Helv. 66 (1991) 368–388 MR1120653

[19] **T Willwacher**, *M Kontsevich's graph complex and the Grothendieck–Teichmüller Lie algebra*, Invent. Math. 200 (2015) 671–760 MR3348138

Mathematisches Institut, Universität Münster
Einsteinstrasse 62, D-48149 Münster, Germany

geoffroy.horel@gmail.com

http://geoffroy.horel.org

Received: 17 March 2014 Revised: 24 March 2015

msp

Algebraic & Geometric Topology 15 (2015) 3253–3303

Fixed-point free circle actions on 4–manifolds

WEIMIN CHEN

This paper is concerned with fixed-point free $\mathbb{S}^1$–actions (smooth or locally linear) on orientable 4–manifolds. We show that the fundamental group plays a predominant role in the equivariant classification of such 4–manifolds. In particular, it is shown that for any finitely presented group with infinite center there are at most finitely many distinct smooth (resp. topological) 4–manifolds which support a fixed-point free smooth (resp. locally linear) $\mathbb{S}^1$–action and realize the given group as the fundamental group. A similar statement holds for the number of equivalence classes of fixed-point free $\mathbb{S}^1$–actions under some further conditions on the fundamental group. The connection between the classification of the $\mathbb{S}^1$–manifolds and the fundamental group is given by a certain decomposition, called a *fiber-sum decomposition*, of the $\mathbb{S}^1$–manifolds. More concretely, each fiber-sum decomposition naturally gives rise to a Z–splitting of the fundamental group. There are two technical results in this paper which play a central role in our considerations. One states that the Z–splitting is a canonical JSJ decomposition of the fundamental group in the sense of Rips and Sela. Another asserts that if the fundamental group has infinite center, then the homotopy class of principal orbits of any fixed-point free $\mathbb{S}^1$–action on the 4–manifold must be infinite, unless the 4–manifold is the mapping torus of a periodic diffeomorphism of some elliptic 3–manifold.

57S15; 57M07, 57M50

1 Introduction

Locally linear $\mathbb{S}^1$–actions on oriented 4–manifolds were classified by Fintushel up to orientation-preserving equivariant homeomorphisms (for smooth $\mathbb{S}^1$–actions the classification is up to orientation-preserving equivariant diffeomorphisms); see [16; 17; 18]. One associates to each locally linear $\mathbb{S}^1$–action a legally weighted 3–manifold, which is the orbit space decorated with certain orbit-type data and a characteristic class of the $\mathbb{S}^1$–action. The equivariant classification of the $\mathbb{S}^1$–four-manifolds is then given by the isomorphism classes of the corresponding legally weighted 3–manifolds.

An important technique for studying locally linear $\mathbb{S}^1$–actions on 4–manifolds is a replacement trick due to Pao [35]. Pao's trick allows one to trade a certain weighted

Published: 12 January 2016 DOI: 10.2140/agt.2015.15.3253

circle in a legally weighted 3–manifold for a pair of fixed points, or to have the weighted circle deleted and a 3–ball removed from the legally weighted 3–manifold. (In particular, Pao's replacement trick applies only to locally linear $\mathbb{S}^1$–actions with a nonempty fixed-point set.) This procedure has the effect of replacing the given $\mathbb{S}^1$–action by another (nonequivalent) $\mathbb{S}^1$–action on the same 4–manifold. Besides the construction of locally linear, nonlinear $\mathbb{S}^1$–actions on $\mathbb{S}^4$ in the original paper [35], the following are some of the further implications of Pao's trick when combined with the classification results of Fintushel in [16; 17; 18]:

- If a 4–manifold X admits a locally linear (resp. smooth) $\mathbb{S}^1$–action with a pair of fixed points or a fixed 2–sphere, then X admits infinitely many nonequivalent locally linear (resp. smooth) $\mathbb{S}^1$–actions; see [35]. (There are many examples of such 4–manifolds, including a large class of simply connected 4–manifolds.)
- Modulo the 3–dimensional Poincaré conjecture (which is now resolved [36]), a simply connected, smooth $\mathbb{S}^1$–four-manifold is diffeomorphic to a connected sum of $\mathbb{S}^4$, $\pm\mathbb{CP}^2$, or $\mathbb{S}^2\times\mathbb{S}^2$; see [18], compare also [48].
- If an oriented 4–manifold with $b_2^+\geq 1$ admits a locally linear (resp. smooth) $\mathbb{S}^1$–action having at least one fixed point, then it contains a topologically (resp. smoothly) embedded, essential 2–sphere of nonnegative self-intersection; see Baldridge [3, Theorem 2.1].[1] In particular, the Hurwitz map $\pi_2\to H_2$ has infinite image. Baldridge's theorem gives a useful obstruction for the existence of $\mathbb{S}^1$–actions with fixed points, particularly for the smooth case as such a smoothly embedded 2–sphere constrains the Seiberg–Witten invariants of the 4–manifold; see [19].

In this paper we study fixed-point free $\mathbb{S}^1$–actions on orientable 4–manifolds, either smooth or locally linear depending on which category (ie smooth or topological) we work in. The arguments are valid for both categories; for simplicity, we shall work mainly in the smooth category. Our results indicate that the equivariant classification of fixed-point free $\mathbb{S}^1$–actions, where there is a lack of Pao's replacement trick, is sharply different from that of $\mathbb{S}^1$–actions with fixed points. In particular, we show that under reasonable assumptions the fundamental group plays a predominant role in the equivariant classification of 4–manifolds with a fixed-point free $\mathbb{S}^1$–action. We showcase this phenomenon with the following two theorems.

[1] Baldridge [3] works in the smooth category, but the arguments are valid in the locally linear category as well.

Theorem 1.1 *Let X be an orientable 4–manifold such that*

(i) *the center of $\pi_1(X)$ is infinite cyclic,*

(ii) *$\pi_1(X)$ is single-ended and is not isomorphic to the fundamental group of a Klein bottle, and*

(iii) *any canonical JSJ decomposition of $\pi_1(X)$ contains a vertex subgroup which is not isomorphic to an HNN extension of a finite cyclic group.*

Then there exists a constant $C > 0$, depending only on $\pi_1(X)$, such that the number of equivalence classes of fixed-point free $\mathbb{S}^1$–actions (smooth or locally linear) on X is bounded by C.

Theorem 1.2 *Let G be a finitely presented group with infinite center. There exists a constant $C > 0$, depending only on G, such that the number of diffeomorphism classes (resp. homeomorphism classes) of orientable 4–manifolds admitting a fixed-point free, smooth (resp. locally linear) $\mathbb{S}^1$–action, whose fundamental group is isomorphic to G, is bounded by C.*

Our approach to equivariant classification (resp. classification) of fixed-point free $\mathbb{S}^1$–four-manifolds differs from the traditional approach of legally weighted 3–manifolds (see Fintushel [16; 17; 18]) where, in our method, geometric group theory plays a prominent role. The central notion in our approach is a certain decomposition of the $\mathbb{S}^1$–manifolds which is called a *fiber-sum decomposition*; see Definition 1.3. Such a decomposition gives rise to a Z–splitting of the fundamental group of the manifold, and the central result of this paper states that the Z–splitting is a canonical JSJ decomposition of the fundamental group in the sense of Rips and Sela [38, Theorem 1.5]. We also point out that the methods of this paper are essentially different from those in Hillman [25], where homotopy/homeomorphism classifications of $\mathbb{S}^1$–bundles over certain 3–manifolds are given. In particular, the diffeomorphism classification result in Theorem 1.2 is not accessible by the surgery-theoretic techniques employed in Hillman [25].

The orbit map of a fixed-point free $\mathbb{S}^1$–action on an orientable 4–manifold defines a Seifert-type $\mathbb{S}^1$–fibration of the 4–manifold, giving the orbit space a structure of a closed, orientable 3–dimensional orbifold whose singular set consists of a disjoint union of embedded circles, called *singular circles*. (Equivalently, the 4–manifold is the total space of a principal $\mathbb{S}^1$–bundle over the 3–orbifold.) With this understood, the building blocks of a fiber-sum decomposition are oriented fixed-point free $\mathbb{S}^1$–four-manifolds whose corresponding orbit space is an irreducible 3–orbifold. We shall call such $\mathbb{S}^1$–four-manifolds *irreducible*. Note that the orientation of the 4–manifold determines an orientation of the base 3–orbifold, as the fibers of the Seifert–type $\mathbb{S}^1$–fibration are canonically oriented.

Definition 1.3 (Fiber-sum decomposition) Let X be a smooth orientable 4–manifold. Suppose we are given a finite set of smooth *oriented* 4–manifolds X_i, $i \in I$, with the following significance.

(i) For each $i \in I$, there is a fixed-point free $\mathbb{S}^1$–action on X_i with orbit map $\pi_i\colon X_i \to Y_i$ where Y_i is irreducible.

(ii) There is a finite set J such that, for each $j \in J$, there exists a pair of distinct points $y_{j,1}$, $y_{j,2} \in \bigsqcup_{i\in I} Y_i$ which have the same multiplicity if singular.

(iii) Let $F_{j,1}$ and $F_{j,2}$ be the fibers of $\bigsqcup_i \pi_i\colon \bigsqcup_i X_i \to \bigsqcup_i Y_i$ over $y_{j,1}$ and $y_{j,2}$, respectively. For each $j \in J$, there is an orientation-reversing but fiber-wise orientation-preserving, fiber-preserving diffeomorphism $\phi_j\colon \partial Nd(F_{j,1}) \to \partial Nd(F_{j,2})$.

(iv) For any $i \in I$, $j \in J$, if Y_i contains exactly one of the points $y_{j,1}$, $y_{j,2}$, say $y_{j,1} \in Y_i$, then the homotopy class of the fiber $F_{j,1}$ generates a *proper* subgroup of $\pi_1(X_i)$.

With the above understood, we say that X admits a *fiber-sum decomposition* if there exists a diffeomorphism between X and the oriented 4–manifold

$$\bigsqcup_{i\in I} X_i \setminus \bigsqcup_{j\in J} (Nd(F_{j,1}) \sqcup Nd(F_{j,2}))/\sim \bigsqcup_{j\in J} \phi_j,$$

and, given such a diffeomorphism, we say that X is *fiber-sum-decomposed* into X_i along N_j, where each $N_j \cong \mathbb{S}^1 \times \mathbb{S}^2$ is the image of $\partial Nd(F_{j,1})$ (or equivalently, $\partial Nd(F_{j,2})$) in X. Furthermore, the irreducible $\mathbb{S}^1$–four-manifolds X_i are called the *factors* of the fiber-sum decomposition.

Remarks The isotopy classification of diffeomorphisms of $\mathbb{S}^1 \times \mathbb{S}^2$ is given by $\pi_0(O(2)\times O(3)\times \Omega O(3))$; see Hatcher [23]. In particular, there are two distinct isotopy classes of homologically trivial diffeomorphisms because of the factor $\pi_0(\Omega O(3)) = \pi_1 \mathrm{SO}(3) = \mathbb{Z}_2$. However, the isotopy class of the diffeomorphism $\phi_j\colon \partial Nd(F_{j,1}) \to \partial Nd(F_{j,2})$ is uniquely determined because of the requirement that it be fiber-preserving.

It turns out that the class of 4–manifolds which admit a fiber-sum decomposition are precisely the smooth, fixed-point free $\mathbb{S}^1$–four-manifolds whose fundamental group has infinite center. In order to understand this, we recall that a fixed-point free $\mathbb{S}^1$–action is called *injective* (and so is the corresponding $\mathbb{S}^1$–four-manifold) if the homotopy class of the principal orbits has infinite order. With this understood, note that in Definition 1.3 each 3–orbifold Y_i is irreducible. It follows easily that the $\mathbb{S}^1$–action on each X_i must be injective. Moreover, it is clear that the $\mathbb{S}^1$–actions on X_i descend to a fixed-point free $\mathbb{S}^1$–action on X, which is also injective. On the other hand, given any injective $\mathbb{S}^1$–action, the orbit space (as a 3–orbifold) admits a certain kind of spherical decompositions which are called *reduced* (see Lemma 2.4 for details), and any such spherical decomposition naturally gives rise to a fiber-sum decomposition of the 4–manifold (for more details see the proof of Theorem 1.4).

In summary, a 4–manifold admits a fiber-sum decomposition if and only if it admits an injective fixed-point free $\mathbb{S}^1$–action. Note that the homotopy class of the principal orbits of the $\mathbb{S}^1$–action lies in the center of the fundamental group of the 4–manifold. In particular, π_1 of an injective fixed-point free $\mathbb{S}^1$–four-manifold has infinite center. The converse is given in the following theorem.

Theorem 1.4 *Let X be a smooth (resp. locally linear), fixed-point free $\mathbb{S}^1$–four-manifold whose fundamental group has infinite center. Then the $\mathbb{S}^1$–action must be injective unless X is diffeomorphic (resp. homeomorphic) to the mapping torus of a periodic diffeomorphism of some elliptic 3–manifold. As a consequence, any smooth, fixed-point free $\mathbb{S}^1$–four-manifold whose π_1 has infinite center admits a fiber-sum decomposition.*

Note that in the case where X is diffeomorphic to the mapping torus of a periodic diffeomorphism of some elliptic 3–manifold, X admits another fixed-point free $\mathbb{S}^1$–action which is injective. So in any event, the 4–manifold admits a fiber-sum decomposition.

We remark that the fundamental group of a smooth, fixed-point free $\mathbb{S}^1$–four-manifold with nontrivial Seiberg–Witten invariant must have infinite center; see [11; 12].

With the preceding understood, the main theme of this paper is to recover the fiber-sum decompositions of an injective $\mathbb{S}^1$–four-manifold from its fundamental group. The main results are summarized in Theorems 1.5 and 1.6 below.

In order to describe the results, observe that given any fiber-sum decomposition of X into factors X_i along N_j, there is an associated finite graph of groups where the vertex groups and edge groups are given by $\pi_1(X_i)$ and $\pi_1(N_j)$, respectively, such that $\pi_1(X)$ is isomorphic to the fundamental group of the graph of groups. Such a presentation of $\pi_1(X)$ is called a *Z–splitting* as each edge group $\pi_1(N_j)$ is infinite cyclic. An in-depth study of Z–splittings of single-ended finitely generated groups was given in [38] by Rips and Sela; in particular, they showed the existence of certain "universal" Z–splittings for each single-ended finitely presented group, which are called *canonical JSJ decompositions*.

Theorem 1.5, which is the main technical result of this paper, asserts that the Z–splitting associated to a fiber-sum-decomposition is a canonical JSJ decomposition of the fundamental group.

Theorem 1.5 *Let X and X' be smooth 4–manifolds which are fiber-sum-decomposed into X_i along N_j and $X'_{i'}$ along $N'_{j'}$, respectively. Suppose $\pi_1(X)$ and $\pi_1(X')$ are single-ended and are not isomorphic to the fundamental group of a 2–torus or a Klein bottle. Then the following hold.*

(1) *The Z–splitting of $\pi_1(X)$ associated to the given fiber-sum-decomposition of X is a canonical JSJ decomposition.*[2]

(2) *Assume further that the submanifolds N_j and $N'_{j'}$ are null-homologous in X and X', respectively, and let $\alpha\colon \pi_1(X)\to\pi_1(X')$ be any isomorphism. Then after modifying the embeddings of N_j and $N'_{j'}$ by fiber-preserving isotopies if necessary, $\alpha\colon \pi_1(X)\to\pi_1(X')$ may be enhanced to an isomorphism between the Z–splittings of $\pi_1(X)$ and $\pi_1(X')$ associated to the new fiber-sum decompositions of X and X', respectively.*

Remarks (1) Canonical JSJ decompositions are not unique as Z–splittings. Nevertheless, Theorem 1.5(1) implies that the number of factors X_i, the number of submanifolds N_j, as well as the conjugacy classes of subgroups $\pi_1(X_i)$ and $\pi_1(N_j)$, depend only on $\pi_1(X)$; see Proposition 3.5 for details. We shall also point out that in the course of the proof of Theorem 1.5, the group $\pi_1(X)$ is shown to have the property that it admits no hyperbolic-hyperbolic elementary Z–splittings; see Lemma 3.1.

(2) The stronger uniqueness in Theorem 1.5(2) corresponds to the uniqueness of canonical JSJ decompositions up to a sequence of *slidings, conjugations, and conjugations of boundary monomorphisms*. Such uniqueness has been established for torsion-free (Gromov) hyperbolic groups (see Sela [43, Theorem 1.7]), but remains open in general for single-ended finitely presented groups; see [38, page 106].

(3) The assumption that the submanifolds N_j are null-homologous in X is equivalent to the assumption that the underlying graph of the associated Z–splitting of $\pi_1(X)$ is a tree. By Theorem 1.5(1), this assumption depends only on the group $\pi_1(X)$.

[2]There is an annoying collapse of terminology here as a canonical JSJ decomposition of $\pi_1(X)$ corresponds *not* to the JSJ decomposition of the base 3–orbifold, but to a reduced spherical decomposition of the 3–orbifold; see Lemma 2.4.

(4) It is worth pointing out that the considerations in this paper provide an almost ideal setting for the need for developing the algebraic theory of Rips and Sela on Z–splittings of single-ended finitely presented groups [38].

The next theorem, Theorem 1.6, is concerned with the building blocks of fiber-sum decompositions. In particular, it is shown that in most of the cases the diffeomorphism class of an irreducible $\mathbb{S}^1$–four-manifold is determined by the fundamental group. To state the result, we remark that a finitely generated group with infinite center is either single-ended or double-ended; see Lemma 4.1.

Theorem 1.6 *Let X and X' be irreducible $\mathbb{S}^1$–four-manifolds, and let $\alpha\colon \pi_1(X) \to \pi_1(X')$ be any isomorphism.*

(1) *If $\pi_1(X)$ and $\pi_1(X')$ are single-ended, then there exists a diffeomorphism $\phi\colon X \to X'$ such that $\phi_* = \alpha\colon \pi_1(X) \to \pi_1(X')$.*

(2) *If $\pi_1(X)$ and $\pi_1(X')$ are double-ended, then X and X' are each the mapping-torus of a periodic diffeomorphism of an elliptic 3–manifold. Moreover, there exists a diffeomorphism $\phi\colon X \to X'$ such that $\phi_* = \alpha\colon \pi_1(X) \to \pi_1(X')$, if the elliptic 3–manifold is not a lens space.*

Finally, the cases which are not covered in Theorems 1.5 and 1.6, ie when $\pi_1(X)$ is isomorphic to the fundamental group of a 2–torus or a Klein bottle, are handled separately. In particular, we direct the reader's attention to two classification theorems of fixed-point free $\mathbb{S}^1$–four-manifolds. One is concerned with the situation where the center of π_1 is of rank greater than 1; the other is about the situation where π_1 is isomorphic to the π_1 of a Klein bottle. See Theorems 4.3 and 6.2 for more details.

With the preceding understood, Theorem 1.1 follows readily from Theorems 1.5 and 1.6. Theorem 1.2 also follows from Theorems 1.5 and 1.6 with the additional help of Theorems 1.4, 4.3 and 6.2.

Having reviewed the main theorems, we now give a few remarks about the technical aspect of this paper. Our arguments rely heavily on the recent advances in 3–dimensional topology, particularly those centered around the resolution of Thurston's Geometrization Conjecture (henceforth referred to as the *Geometrization theorem*; see [4; 36]; see also [14]). For instance, Lemma 5.2(which asserts that an orientable 3–orbifold is Seifert fibered if π_1^{orb} has infinite center, and furthermore, if $\pi_2^{\mathrm{orb}} \neq 0$, it is the mapping torus of a periodic diffeomorphism of a 2–orbifold with finite π_1^{orb}) played a key role in the proofs of several theorems of this paper. The proof of this lemma involves several particular forms of the Geometrization theorem, which include the earlier work of

Meeks and Scott [32] on finite group actions on Seifert 3–manifolds, the resolution of the Seifert fiber space conjecture due to Gabai [20] (and independently Casson and Jungreis [9]), as well as the more recent Orbifold theorem of Boileau, Leeb and Porti [4] and the resolution of Poincaré conjecture; see [36]. On the other hand, as we mentioned earlier, this paper also draws considerably from geometric group theory, particularly the work of Rips and Sela on Z–splittings of single-ended finitely presented groups; see [38].

Before ending the introduction, we point out a corollary of Theorem 1.4 which is of independent interest.

Corollary 1.7 *Let X be a 4–manifold whose fundamental group has infinite center. If X admits a locally linear, fixed-point free $\mathbb{S}^1$–action, then there are no embedded 2–spheres with odd self-intersection in X. In particular, X is minimal.*

We end the introduction with the following questions, which are naturally suggested by the results of this paper (see [44; 45; 46; 34; 47] for some relevant problems and results in dimension three).

Question 1.8 Let X be an oriented, smooth, fixed-point free $\mathbb{S}^1$–four-manifold whose fundamental group has infinite center.

(1) Is the diffeomorphism type of X determined by its homeomorphism type?

(2) Can one express the Seiberg–Witten invariant of X in terms of topological invariants of the manifold?

The organization of the rest of the paper is as follows. In Section 2, we first review some basic definitions and facts about 2–orbifolds and 3–orbifolds, and then we prove several preliminary lemmas which will be used in later sections. Section 3 is devoted to the proof of Theorem 1.5; it begins with a brief review of the Bass–Serre theory of groups acting on trees (in particular, the definition of graph of groups and its fundamental group), as well as a review on the relevant part of the work of Rips and Sela in [38] concerning Z–splittings of single-ended finitely presented groups. The proof of Theorem 1.6 is given in Section 4, and so is the classification of fixed-point free $\mathbb{S}^1$–four-manifolds whose π_1 has a center of rank greater than 1. Section 5 is devoted to Theorem 1.4; in particular, we prove the key lemma, Lemma 5.2, in this section. Corollary 1.7 asserting minimality of injective $\mathbb{S}^1$–four-manifolds is proven here as well. Section 6 contains the proofs of Theorems 1.1 and 1.2, as well as the classification of fixed-point free $\mathbb{S}^1$–four-manifolds whose π_1 is isomorphic to the π_1 of a Klein bottle.

Throughout this paper, we shall adopt the following notation: the center of a group G is denoted by $z(G)$.

2 Recollections and preliminary lemmas

For the reader's convenience, we shall begin by giving a brief review on the relevant definitions and basic facts about 2–orbifolds and 3–orbifolds; for more details, see eg [41; 5]. Recall first that an orbifold (not necessarily orientable) is called *good* if it is the quotient of a manifold by a properly discontinuous action of a discrete group; otherwise it is call *bad*. It is called *very good* if it is the quotient of a manifold by a finite group action. All orbifolds are assumed to be connected and closed (ie compact without boundary) unless mentioned otherwise.

An orientable 2–orbifold is given by a closed orientable surface as the underlying space, with isolated singular points where the local groups are cyclic, generated by a rotation. For a nonorientable 2–orbifold, if the underlying space has a nonempty boundary, the singular set will also contain the boundary of the underlying space, which is a polygon with local groups being either a reflection through a line in $\mathbb{R}^2$ or a dihedral group D_{2n} generated by two reflections through lines making an angle π/n. With this understood, a *teardrop* is a 2–sphere with one singular point. A *spindle* is a 2–sphere with two singular points of different multiplicities (ie the orders of the local groups). A *football* is a 2–sphere with two singular points of the same multiplicity. A *turnover* is a 2–sphere with three singular points. Except for a teardrop or a spindle, all orientable 2–orbifolds are very good. An orientable 2–orbifold is called *spherical* (resp. *toric*, resp. *hyperbolic*) if it is the quotient of a 2–sphere (resp. 2–torus, resp. closed surface of genus >1) by a finite group. A 2–orbifold is spherical if and only if it is either a nonsingular sphere, a football, or a turnover with multiplicities $(2,2,n)$, $(2,3,3)$, $(2,3,4)$, or $(2,3,5)$. The turnovers correspond to the quotient of 2–sphere by the action of a dihedral group D_{2n} or one of the platonic groups T_{12}, O_{24}, or I_{60}.

All 2–suborbifolds in a 3–orbifold are assumed to be orientable. There is a special class of 3–orbifolds which are important for the considerations in this paper; these are the 3–orbifolds which do not contain any bad 2–suborbifolds. It is a consequence of the Geometrization theorem (see [4; 31]) that if a 3–orbifold does not contain any bad 2–suborbifolds, then it must be very good, ie it is the quotient of a 3–manifold by a finite group action. For simplicity, we shall call such a 3–orbifold *good*.

An orientable 3–orbifold (with or without boundary) is called *spherical* (resp. *discal*) if it is the quotient of the 3–sphere (resp. 3–ball) by a finite isometry group. A good 3–orbifold is called *irreducible* if every spherical 2–suborbifold bounds a discal 3–orbifold. An irreducible 3–orbifold is called *atoroidal* if it contains no essential toric 2–suborbifold. A 3–orbifold (not necessarily orientable) is called *Seifert fibered* if it is the total space of an orbifold bundle over a 2–orbifold (not necessarily orientable) with generic fiber a circle or a mirrored interval. (A mirrored interval is the quotient of a circle by an orientation-reversing involution.) It is easily seen that a generic fiber of an orientable Seifert fibered 3–orbifold must be a circle. Moreover, if the base 2–orbifold is orientable, then the singular set of the Seifert fibered 3–orbifold must consist of a union of fibers.

The rest of this section is occupied by a number of preliminary lemmas. The following lemma about the center of an amalgamated product or an HNN extension is well known to the experts. However, for the sake of completeness, we include a statement and a proof of the lemma here.

Lemma 2.1 (1) *If $A \neq C \neq B$, then the center of $A *_C B$ is contained in C.*

(2) *Let $C \subset A$ be a subgroup and $\alpha\colon C \to A$ be an injective homomorphism, and let $A *_C \alpha$ denote the corresponding HNN extension. Suppose $x \in z(A *_C \alpha)$. Then either $x \in C$, or x is nontorsion, $C = A = \alpha(C)$, and $A *_C \alpha$ is isomorphic to $A *_C \alpha'$ for some $\alpha'\colon A \to A$ which is of finite order.*

Proof For a proof of part (1), see Magnus, Karrass and Solitar [30], Corollary 4.5, page 211. We shall give a proof for part (2) here. An element of $A *_C \alpha$ can be uniquely represented by a reduced word; see eg Scott and Wall [42]. Lemma 2.1 is a direct consequence of this fact.

More concretely, recall that the group $A *_C \alpha$ is generated by elements of A and a letter t with additional relations $tct^{-1} = \alpha(c)$ for all $c \in C$. We let T and T_α be the sets of some fixed choices of representatives of the right cosets of C and $\alpha(C)$ in A, respectively. Then a reduced word in $A *_C \alpha$ takes the form

$$a_1 t^{\epsilon_1} a_2 t^{\epsilon_2} \cdots a_n t^{\epsilon_n} a_{n+1},$$

where $\epsilon_i = \pm 1$, $a_i \in T$ if $\epsilon_i = +1$, $a_i \in T_\alpha$ if $\epsilon_i = -1$, and furthermore, $a_i \neq 1$ if $\epsilon_{i-1} \neq \epsilon_i$, and a_{n+1} is allowed to be an arbitrary element of A.

Let $x = a_1 t^{\epsilon_1} a_2 t^{\epsilon_2} \cdots a_n t^{\epsilon_n} a_{n+1}$ be an element of the center (here $n = 0$ represents the case where $x \in A$). If $n = 0$, then by $tx = xt$ it is clear that $x = a_{n+1} \in C$ which obeys $\alpha(x) = x$. Suppose $n > 0$. If $a_1 \neq 1$, then the uniqueness of representation by reduced words implies that $tx \neq xt$, which is a contradiction. If $a_1 = 1$, then $t^{-\epsilon_1} x = x t^{-\epsilon_1}$ implies that $a_2 = 1$. Iterating this process, we see that $x = t^l a_{n+1}$ for some $0 \neq l \in \mathbb{Z}$. It follows from $t^{-1}x = xt^{-1}$ that $a_{n+1} = t a_{n+1} t^{-1}$, which implies that $a_{n+1} \in C$ and $\alpha(a_{n+1}) = a_{n+1}$. Furthermore, the commutativity of t and a_{n+1} also implies that $x = t^l a_{n+1}$ is nontorsion. To see $C = A = \alpha(C)$, note that if there is an $a \in T$ or T_α such that $a \neq 1$, then one has $ax \neq xa$ which is a contradiction. This implies that $C = A = \alpha(C)$. Now for any $c \in C$,

$$t^l a_{n+1} \alpha^l(c) = x\alpha^l(c) = \alpha^l(c)x = \alpha^l(c) t^l a_{n+1} = t^l c a_{n+1},$$

which implies that $a_{n+1}\alpha^l(c) = ca_{n+1}$ for any $c \in C = A$. Let $\alpha'\colon A \to A$ be defined by $\alpha'(c) = a_{n+1}\alpha(c)a_{n+1}^{-1}$. Then it follows from $\alpha(a_{n+1}) = a_{n+1}$ that

$$(\alpha')^l(c) = a_{n+1}\alpha^l(c)a_{n+1}^{-1} = c, \quad \forall c \in A.$$

Now note that $A *_C \alpha$ is isomorphic to $A *_C \alpha'$ where α' has finite order l. This completes the proof of the lemma. □

For our purposes in this paper, it is important to understand the center of the fundamental group of a 2–orbifold or a 3–orbifold.

Lemma 2.2 *Let Σ be a 2–orbifold (not necessarily orientable) such that $z(\pi_1^{\mathrm{orb}}(\Sigma))$ is nontrivial. Then the following statements hold true.*

(a) *If Σ is orientable, then it is either a football, a spindle with non-coprime multiplicities, a turnover with multiplicities $(2,2,2)$, or a nonsingular torus.*

(b) *If Σ is nonorientable, then its orientable double cover $\tilde{\Sigma}$ must lie in the following list: a nonsingular sphere, a teardrop, a spindle, a football, a turnover with multiplicities $(2,2,2)$, or a nonsingular torus. Moreover, $z(\pi_1^{\mathrm{orb}}(\Sigma))$ is torsion-free if and only if $\tilde{\Sigma}$ is a nonsingular torus.*

Proof Suppose Σ is orientable. If Σ is bad, then it must be a spindle with non-coprime multiplicities because this is the only case where $\pi_1^{\mathrm{orb}}(\Sigma)$ is nontrivial. Assume Σ is good. If Σ is spherical, then it must be a football or a turnover with multiplicities $(2,2,2)$, because the other groups, ie D_{2n} with $n \neq 2$, T_{12}, O_{24}, I_{60}, all have trivial center. If Σ is toric, then it must be a nonsingular torus because the fundamental group of a toric turnover is centerless. Finally, Σ can not be hyperbolic because a cocompact Fuchsian group has trivial center.

Suppose Σ is nonorientable, and let $\tilde{\Sigma}$ be the orientable 2–orbifold which doubly covers Σ. Note that $\mathbb{Z}_2$ acts on $\tilde{\Sigma}$ via deck transformations. We shall discuss the proof in two cases: (i) the deck transformations are free, (ii) the deck transformations are not free.

In case (i), the underlying space $|\Sigma|$ is a nonorientable, closed surface. We can decompose $|\Sigma|$ as the union of $\mathbb{RP}^2 \setminus D^2$ and an orientable surface with one boundary component along their boundaries. Correspondingly, we have a decomposition of Σ as the union of (nonsingular) $\mathbb{RP}^2 \setminus D^2$ and an orientable 2–orbifold Σ' with one boundary component. It follows that $z(\pi_1^{\mathrm{orb}}(\Sigma))$ being nontrivial forces $\pi_1^{\mathrm{orb}}(\Sigma')$ to be finite (see Lemma 2.1(1) and Lemma 2.2(a)), so that Σ' must be either a (nonsingular) D^2 or $D^2/\mathbb{Z}_m$ with $m > 1$. This shows that the double cover $\tilde{\Sigma}$ is either a (nonsingular) sphere or a football.

In case (ii), if $z(\pi_1^{\mathrm{orb}}(\widetilde{\Sigma}))$ is nontrivial, then we are done by part (a). Moreover, if $\widetilde{\Sigma}$ is a nonsingular torus, the fixed-point set of the deck transformation consists of a union of circles. Since the deck transformation is orientation-reversing, the Lefschetz fixed-point theorem implies that the action on $H_1(\widetilde{\Sigma};\mathbb{R})$ must have eigenvalues $+1$ and -1. It follows then that $z(\pi_1^{\mathrm{orb}}(\Sigma))=\mathbb{Z}$ in this case. If $z(\pi_1^{\mathrm{orb}}(\widetilde{\Sigma}))$ is trivial, then $z(\pi_1^{\mathrm{orb}}(\Sigma))=\mathbb{Z}_2$ and acts on $\widetilde{\Sigma}$ via deck transformations. Let p be a fixed-point of the deck transformation. Since $\pi_1^{\mathrm{orb}}(\widetilde{\Sigma})\to\pi_1(|\widetilde{\Sigma}|)$ is surjective, the induced action of $z(\pi_1^{\mathrm{orb}}(\Sigma))=\mathbb{Z}_2$ on $\pi_1(|\widetilde{\Sigma}|,p)$ must be trivial. This implies that the Lefschetz number of the action of $z(\pi_1^{\mathrm{orb}}(\Sigma))=\mathbb{Z}_2$ on $|\widetilde{\Sigma}|$ equals -2 times the genus of $|\widetilde{\Sigma}|$. The Lefschetz fixed-point theorem then implies that $|\widetilde{\Sigma}|$ has genus zero. If $\widetilde{\Sigma}$ is bad, then clearly we are done. If $\widetilde{\Sigma}$ is good, then it is the quotient of an orientable closed surface Σ' by a finite group. Note that $z(\pi_1^{\mathrm{orb}}(\Sigma))=\mathbb{Z}_2$ also acts on Σ' via deck transformations which are orientation-reversing. The same argument as above shows that Σ' must have genus zero. In other words, $\widetilde{\Sigma}$ is spherical. It follows easily that it must be either a (nonsingular) sphere, a football or a turnover with multiplicities $(2,2,2)$. (In fact $\widetilde{\Sigma}$ is a sphere because we assume $z(\pi_1^{\mathrm{orb}}(\widetilde{\Sigma}))$ is trivial.) Hence the lemma. □

Lemma 2.3 *Let Y denote an irreducible 3–orbifold with infinite $\pi_1^{\mathrm{orb}}(Y)$. Then $z(\pi_1^{\mathrm{orb}}(Y))$ is torsion-free.*

Proof By the JSJ-decomposition theorem for 3–orbifolds (see [5, Theorem 3.3]), there is a finite collection (possibly empty) of disjoint, essential toric 2–suborbifolds Σ_j, $j=1,2,\dots,m$, which split Y into 3–suborbifolds Y_i, $i=1,2,\dots,n$, such that each Y_i is either Seifert fibered or atoroidal. This presents $\pi_1^{\mathrm{orb}}(Y)$ as the fundamental group of a finite graph of groups, where the vertex groups are $\pi_1^{\mathrm{orb}}(Y_i)$ and the edge groups are $\pi_1^{\mathrm{orb}}(\Sigma_j)$. If $\{\Sigma_j\}$ is not empty, then the torsion part of $z(\pi_1^{\mathrm{orb}}(Y))$ must lie in the edge groups $z(\pi_1^{\mathrm{orb}}(\Sigma_j))$; see Lemma 2.1. By Lemma 2.2(a), $z(\pi_1^{\mathrm{orb}}(\Sigma_j))$ is torsion-free, which implies that $z(\pi_1^{\mathrm{orb}}(Y))$ is torsion-free when $\{\Sigma_j\}$ is not empty.

Suppose $\{\Sigma_j\}$ is empty. Then Y is either Seifert fibered or atoroidal. Assume Y is Seifert fibered first, and let $\pi\colon Y\to B$ be a Seifert fibration. There is an induced exact sequence (see [5, Proposition 2.12])

$$1\to C\to\pi_1^{\mathrm{orb}}(Y)\xrightarrow{\pi_*}\pi_1^{\mathrm{orb}}(B)\to 1,$$

where C is cyclic or dihedral (either finite or infinite). In addition, C is finite if and only if $\pi_1^{\mathrm{orb}}(Y)$ is finite. In the present case Y only has a 1–dimensional singular set, so that a generic fiber of π must be a circle. Consequently, C is cyclic in the above exact sequence. Since $\pi_1^{\mathrm{orb}}(Y)$ is infinite, we have $C=\mathbb{Z}$. On the other hand,

$C = \pi_1(\mathbb{S}^1)/\mathrm{Image}\,\delta$, where $\delta\colon \pi_2^{\mathrm{orb}}(B) \to \pi_1(\mathbb{S}^1)$ is the connecting homomorphism in the exact sequence of homotopy groups associated to the Seifert fibration $\pi\colon Y \to B$. (For the definition of homotopy groups of orbifolds and the exact sequence associated to an orbifold fibration, see [21; 22; 10].) As C is infinite, δ must be the zero map, and consequently $\pi_*\colon \pi_2^{\mathrm{orb}}(Y) \to \pi_2^{\mathrm{orb}}(B)$ is surjective. By the assumption that Y is irreducible, its universal cover $\widetilde{Y}$ is also irreducible; see [5, Theorem 3.23]. Consequently, we have $\pi_2^{\mathrm{orb}}(Y) = \pi_2(\widetilde{Y}) = 0$ which implies that $\pi_2^{\mathrm{orb}}(B) = 0$. Now observing that a bad 2–orbifold and a spherical 2–orbifold must have nontrivial π_2^{orb}, we conclude, by Lemma 2.2, that $z(\pi_1^{\mathrm{orb}}(B))$ must be torsion-free. It follows easily that $z(\pi_1^{\mathrm{orb}}(Y))$ is torsion-free in this case.

It remains to consider the case where Y is atoroidal. If Y is nonsingular (ie a 3–manifold), then $\pi_1^{\mathrm{orb}}(Y) = \pi_1(Y)$ is torsion-free; hence $z(\pi_1^{\mathrm{orb}}(Y))$ must be torsion-free. If Y is an honest orbifold, then by the Orbifold theorem of Boileau, Leeb and Porti (see [4, Corollary 1.2]), Y is geometric. In fact, we will need the following more precise statement: Y has a metric of constant curvature or is Seifert fibered. It is clear that, since $\pi_1^{\mathrm{orb}}(Y)$ is infinite, we only need to discuss the following two cases: (i) Y is hyperbolic, (ii) Y is Euclidean.

Suppose Y is hyperbolic. Then there is a hyperbolic 3–manifold Y' and a finite group of isometries G such that $Y = Y'/G$. Now suppose $z(\pi_1^{\mathrm{orb}}(Y))$ is not torsion-free, and let $g \in z(\pi_1^{\mathrm{orb}}(Y))$ be a torsion element. Then since $\pi_1(Y')$ is torsion-free, g may be regarded as an element of G, and it acts on Y' via deck transformations. Moreover, g must have a fixed point, say $p \in Y'$. This gives rise to an automorphism g_* of $\pi_1(Y', p)$, which is trivial because $g \in z(\pi_1^{\mathrm{orb}}(Y))$. By Mostow Rigidity, $g\colon Y' \to Y'$ is trivial, which is a contradiction.

Suppose Y is Euclidean. By the Bieberbach theorem (see [41, page 443]), Y is finitely covered by T^3 with deck transformation group G. Let $x \in z(\pi_1^{\mathrm{orb}}(Y))$ be a torsion element. Then x may be regarded as an element of G and acts on T^3 via deck transformations. Furthermore, x must have a fixed point, say $p \in T^3$. Since x is central, the induced automorphism $x_*\colon \pi_1(T^3, p) \to \pi_1(T^3, p)$ must be trivial. It follows that x is trivial, which is a contradiction.

This completes the proof of the lemma. □

Given any good 3–orbifold Y which is not irreducible, one can cut Y open along a finite system of spherical 2–suborbifolds into pieces which are irreducible. More precisely, by the spherical decomposition theorem (see [5, Theorem 3.2]), there is a finite, nonempty collection of disjoint spherical 2–suborbifolds $\{\Sigma_j\}$ such that each component Y_i of $Y \setminus \{\Sigma_j\}$ becomes an irreducible 3–orbifold after capping-off the boundary spherical 2–suborbifolds by the corresponding discal 3–orbifolds.

For the purpose in this paper, a slightly improved version of the above statement is needed. More concretely, given any system of spherical 2–suborbifolds $\{\Sigma_j\}$ of Y, let $\{Y_i\}$ be the set of components of $Y \setminus \{\Sigma_j\}$. We say that Σ_j is separating (resp. nonseparating) in Y_i if Σ_j is a boundary component (resp. a nonseparating spherical 2–suborbifold) of the closure of Y_i in Y. (Note that Σ_j can be a nonseparating spherical 2–suborbifold of Y but is separating in Y_i.) With this understood, we say that the corresponding spherical decomposition of Y is *reduced* if for any Σ_j, Y_i such that Σ_j is separating in Y_i, $\pi_1^{\mathrm{orb}}(\Sigma_j)$ is a proper subgroup of $\pi_1^{\mathrm{orb}}(Y_i)$ under the inclusion of Σ_j in the closure of Y_i in Y.

Lemma 2.4 *For any good, reducible 3–orbifold Y, there exists a reduced spherical decomposition of Y into irreducible 3–orbifolds.*

Proof Given any spherical decomposition of Y into irreducible pieces, which always exists (see [5, Theorem 3.2]), we can modify it into a reduced spherical decomposition as follows. Let $\{\Sigma_j\}$ be the corresponding system of spherical 2–suborbifolds and let $\{Y_i\}$ be the set of components of $Y \setminus \{\Sigma_j\}$. Suppose for some i, j, Σ_j is separating in Y_i and $\pi_1^{\mathrm{orb}}(\Sigma_j) = \pi_1^{\mathrm{orb}}(Y_i)$. Let $Y_k \in \{Y_i\}$ be the other component whose closure in Y also contains Σ_j as a boundary component. Then observe that the 3–orbifold obtained from capping-off $Y_k \cup \Sigma_j \cup Y_i$ is the same as that obtained from capping-off Y_k. This is because, by the Geometrization theorem, the 3–orbifold obtained from capping-off the boundary components of Y_i other than Σ_j is a discal 3–orbifold with boundary Σ_j. Consequently, if we remove Σ_j from $\{\Sigma_j\}$, the corresponding spherical decomposition still splits Y into irreducible pieces. Continuing this process, we arrive at a reduced spherical decomposition in finitely many steps. Hence the lemma. $\square$

We remark that given any spherical decomposition of a good 3–orbifold Y, with $\{\Sigma_j\}$ being the system of spherical 2–suborbifolds and $\{Y_i\}$ being the set of components of $Y \setminus \{\Sigma_j\}$, one has a corresponding finite graph of groups whose vertex groups and edge groups are given by $\{\pi_1^{\mathrm{orb}}(Y_i)\}$ and $\{\pi_1^{\mathrm{orb}}(\Sigma_j)\}$ respectively, such that $\pi_1^{\mathrm{orb}}(Y)$ is naturally isomorphic to the fundamental group of the graph of groups. When the spherical decomposition is reduced, the corresponding graph of groups is also reduced in the sense that an edge group is always a proper subgroup of the vertex groups as long as the end points of the edge are distinct vertices. Given any finite graph of groups, one can always modify it into a reduced one without changing the isomorphism class of the fundamental groups by collapsing a number of edges. Lemma 2.4 is simply a manifestation of this principle in the geometric setting of spherical decomposition of 3–orbifolds. When there are no nonseparating spherical 2–suborbifolds, the existence and uniqueness of reduced spherical decompositions were proven in [37] (called *efficient splittings* therein).

Next we give a classification of certain orientation-preserving finite group actions on $\mathbb{S}^1 \times \mathbb{S}^2$. The case where the actions are free or have only isolated exceptional orbits was discussed in Meeks and Scott [32, Theorem 8.4]. Our discussion relies on the Equivariant sphere theorem of Meeks and Yau (see [33]) and Geometrization of finite group actions on $\mathbb{S}^3$ (compare also Dinkelbach and Leeb [14] via equivariant Ricci flow).

In order to state the result, we shall fix the following convention and notations. We orient $\mathbb{S}^3$ as the boundary of the unit ball in $\mathbb{C}^2$ and consider certain orientation-preserving $\mathbb{Z}_{2m}$–actions on $\mathbb{S}^3$. When m is even, there is only one such action up to a change of generators of $\mathbb{Z}_{2m}$. When m is odd, there are two nonequivalent such actions, and we shall denote the quotient orbifolds by $\mathbb{RP}^3_m$ and $\widetilde{\mathbb{RP}^3}_m$, respectively. More concretely, we fix a generator t of $\mathbb{Z}_{2m}$, and let

$$\mathbb{RP}^3_m = \mathbb{S}^3/\mathbb{Z}_{2m}, \qquad \text{where } t\cdot(z_1,z_2) = \left(-z_1, \exp\left(\frac{\pi i}{m}\right)z_2\right),$$

and

$$\widetilde{\mathbb{RP}^3}_m = \mathbb{S}^3/\mathbb{Z}_{2m}, \qquad \text{where } t\cdot(z_1,z_2) = \left(-z_1, \exp\left(\frac{(m+1)\pi i}{m}\right)z_2\right), \quad m \text{ is odd.}$$

Note that when $m > 1$, these actions can be characterized by the fact that the whole group has no fixed points but the index 2 subgroup fixes an unknotted circle. Moreover, the difference between $\mathbb{RP}^3_m$ and $\widetilde{\mathbb{RP}^3}_m$ is that the singular set of $\widetilde{\mathbb{RP}^3}_m$ has two components, of multiplicities 2 and m, respectively, while the singular set of $\mathbb{RP}^3_m$ has only one component, of multiplicity m.

Lemma 2.5 *Let G be a finite group that acts on $\mathbb{S}^1 \times \mathbb{S}^2$ preserving the orientation.*

- *Suppose the action of G is homologically trivial. Then $\mathbb{S}^1 \times \mathbb{S}^2/G$ is the mapping torus of a periodic diffeomorphism of some spherical 2–orbifold.*
- *Suppose G is cyclic and is generated by t which is homologically nontrivial. Then the quotient orbifold $\mathbb{S}^1 \times \mathbb{S}^2/G$ is diffeomorphic to one of the following:*

$$\mathbb{RP}^3_m \#_m \mathbb{RP}^3_m, \quad \mathbb{RP}^3_m \#_m \widetilde{\mathbb{RP}^3}_m, \quad \text{or} \quad \widetilde{\mathbb{RP}^3}_m \#_m \widetilde{\mathbb{RP}^3}_m,$$

where $\#_m$ denotes the connected sum of orbifolds over a point of multiplicity m, such that a generator of the π_1^{orb} of $\mathbb{S}^2/\mathbb{Z}_m$ has the same image on both sides of the connected sum.

Proof First of all, by the Equivariant sphere theorem of Meeks and Yau (see [33, page 480]), there exists a finite set of embedded 2–spheres $\{\Sigma_i\}$ of $\mathbb{S}^1 \times \mathbb{S}^2$ that is G–invariant and generates the π_2 as a π_1–module. Since $\pi_2(\mathbb{S}^1 \times \mathbb{S}^2)$ has rank 1, we may assume G acts on the set of spheres $\{\Sigma_i\}$ transitively. It follows easily from the Geometrization theorem that when cutting $\mathbb{S}^1\times\mathbb{S}^2$ open along the Σ_i's, each component Y_j of $\mathbb{S}^1\times\mathbb{S}^2\setminus\{\Sigma_i\}$ is a 3–manifold diffeomorphic to the product of $\mathbb{S}^2$ with an interval.

For convenience of the argument, we shall consider the following finite graph Γ, where the vertices correspond to the components Y_j and the edges to the embedded spheres Σ_i, and Σ_i is incident to Y_j if and only if Σ_i is contained in the closure of Y_j. Clearly Γ is homeomorphic to a circle, and there is an induced simplicial action of G on Γ. We denote by G_0 the subgroup of G which acts trivially on Γ.

Suppose G_0 is nontrivial. We pick an embedded sphere Σ_i and cut $\mathbb{S}^1\times\mathbb{S}^2$ open along Σ_i. Because Σ_i is G_0–invariant, we can close up $\mathbb{S}^1\times\mathbb{S}^2\setminus\Sigma_i$ and obtain a G_0–action on $\mathbb{S}^3$. By the Geometrization theorem, the action of G_0 is given by an isometry, which implies that the original G_0–action on $\mathbb{S}^1\times\mathbb{S}^2$ is a product action that is trivial on the $\mathbb{S}^1$–factor. Note that we are done if $G=G_0$.

Assume $G\neq G_0$ and consider the action of G. In the case where G acts homologically trivially, G/G_0 acts effectively on Γ by rotations. This implies that $\mathbb{S}^1\times\mathbb{S}^2/G$ is the mapping torus of the 2–orbifold $\mathbb{S}^2/G_0$ for some periodic diffeomorphism of $\mathbb{S}^2/G_0$ which generates G/G_0. The lemma follows easily in this case.

Suppose G is generated by t which is homologically nontrivial. Then the induced action of t on Γ must be given by a reflection, and G_0 is an index 2 subgroup. Furthermore, G_0 is cyclic in this case and the action of G_0 on $\mathbb{S}^2$ is given by rotations. The order of t is even, say $2m$, and there are two possibilities for the induced action of t on the graph Γ: (i) t has an invariant edge, (ii) t fixes two vertices.

In case (i), t leaves an embedded sphere Σ_i invariant (which is the only one because by assumption G acts transitively on the set of spheres $\{\Sigma_i\}$). The induced action of t on Σ_i is orientation-reversing, and there are two nonequivalent actions when m is odd. More concretely, if we identify Σ_i with the unit sphere in $\mathbb{R}^3=\mathbb{R}\times\mathbb{C}$, then the actions are given by

$$t\cdot(x,z)=\left(-x,\exp\left(\frac{\pi i}{m}\right)z\right),\qquad\text{where } (x,z)\in\mathbb{R}\times\mathbb{C},$$

and

$$t\cdot(x,z)=\left(-x,\exp\left(\frac{(m+1)\pi i}{m}\right)z\right),\qquad\text{where } (x,z)\in\mathbb{R}\times\mathbb{C},\ m \text{ is odd.}$$

It follows easily that the quotient of a t–invariant regular neighborhood of Σ_i is diffeomorphic to either $\mathbb{RP}^3_m$ or $\widetilde{\mathbb{RP}^3}_m$ with a ball centered at a singular point of multiplicity m removed. Moreover, the complement of the t–invariant regular neighborhood is a

3–manifold Y_j that is diffeomorphic to the product of $\mathbb{S}^2$ with an interval. The action of t on Y_j can be naturally extended to an t–action on $\mathbb{S}^3$ by capping-off the boundary of Y_j, which, by the Geometrization theorem, is equivalent to an isometry. Note that when $m > 1$, t^2 has a 1–dimensional fixed-point set. It follows easily that $Y_j/\langle t\rangle$ is also diffeomorphic to either $\mathbb{RP}^3_m$ or $\widetilde{\mathbb{RP}^3}_m$ with a ball centered at a singular point of multiplicity m removed, and $\mathbb{S}^1\times\mathbb{S}^2/G$ is diffeomorphic to either $\mathbb{RP}^3_m \#_m \mathbb{RP}^3_m$, or $\mathbb{RP}^3_m \#_m \widetilde{\mathbb{RP}^3}_m$, or $\widetilde{\mathbb{RP}^3}_m \#_m \widetilde{\mathbb{RP}^3}_m$ as claimed.

In case (ii) where t fixes two vertices of the graph Γ, the set $\{\Sigma_i\}$ has two elements Σ_1 and Σ_2, and $\mathbb{S}^1\times\mathbb{S}^2\setminus\{\Sigma_i\}$ has two components Y_1 and Y_2, such that Y_1 and Y_2 are t–invariant and t switches Σ_1 and Σ_2. Similarly, the t–actions on Y_1 and Y_2 can be extended to a t–action on $\mathbb{S}^3$ by capping-off the boundary, and by the Geometrization theorem, the quotients of Y_1 and Y_2 by t are diffeomorphic to either $\mathbb{RP}^3_m$ or $\widetilde{\mathbb{RP}^3}_m$, and the lemma follows in this case too. □

We end with a lemma concerning existence of Seifert-type T^2–fibrations on a 4–manifold.

Lemma 2.6 *Let $\pi\colon X\to Y$ be a principal $\mathbb{S}^1$–bundle over an orientable 3–orbifold where Y is Seifert fibered. If the homotopy class of a regular fiber of the Seifert fibration on Y lies in the image of $z(\pi_1(X))$ under $\pi_*\colon \pi_1(X)\to\pi_1^{\mathrm{orb}}(Y)$, then $\pi\colon X\to Y$ may be extended to a principal T^2–bundle over a 2–orbifold.*

Proof Let pr: $Y\to B$ be the Seifert fibration on Y where B is a 2–orbifold. (We note that B must be orientable because the class of a regular fiber of pr lies in the center $z(\pi_1^{\mathrm{orb}}(Y))$.) Then the composition of π with pr, $\Pi\colon X\to B$, defines X as a T^2–bundle over B. We shall prove that Π is principal, which is equivalent to the condition that Π has a trivial monodromy representation.

To see that the monodromy representation of Π is trivial, we consider an arbitrary loop γ in B lying in the complement of the singular set. Pick a base point $b_0\in\gamma$, and a base point $x_0\in\Pi^{-1}(b_0)$. Choose a section γ' of Π over γ through x_0, and a loop δ containing x_0 in X that is a section of π over the fiber of pr at b_0. Let h be the fiber of π containing x_0. With this understood, the monodromy representation of Π is trivial if and only if the classes of h, δ, and γ' in $\pi_1(X)$ commute. But this is clear because the classes of both h and δ lie in the center $z(\pi_1(X))$. Hence the lemma. □

3 Fiber-sum decomposition and fundamental group

This section contains three subsections. Section 3.1 is devoted to a review of Bass–Serre theory and Rips–Zela theory, and it also contains a proof of Lemma 3.1 and Lemma 3.2.

Section 3.2 is occupied by a proof of Theorem 1.5(1), as given through Lemma 3.3, Lemma 3.4, and Proposition 3.5. The last subsection, Section 3.3, contains the proof for Theorem 1.5(2).

3.1 Some recollections in geometric group theory

We begin with a brief review of the Bass–Serre theory of groups acting on trees, see eg [13; 42] for more details.

Let Γ be a connected, nonempty graph, with the set of vertices and edges denoted by $V\Gamma$ and $E\Gamma$, respectively, and the incidence functions denoted by $\iota, \tau\colon E\Gamma \to V\Gamma$. Recall that a group of graphs, denoted by G_Γ, consists of the following data: each $v \in V\Gamma$ and $e \in E\Gamma$ is assigned with a group $G(v)$ and $G(e)$, respectively, and for each $e \in E\Gamma$ there is a pair of boundary monomorphisms $\alpha\colon G(e) \to G(\iota e)$ and $\omega\colon G(e) \to G(\tau e)$.

Let Γ_0 be a maximal tree in Γ. The fundamental group of G_Γ with respect to Γ_0, denoted by $\pi(G_\Gamma, \Gamma_0)$, is the group given by the following presentation:

- Generating set: $\{t_e \mid e \in E\Gamma\} \cup \bigcup_{v \in V\Gamma} G(v)$.
- Relations: the relations for $G(v)$, $\forall v \in V\Gamma$; $t_e^{-1}\alpha(g)t_e = \omega(g)$, $\forall g \in G(e)$, $\forall e \in E\Gamma$; and $t_e = 1$, $\forall e \in E\Gamma_0 = E\Gamma \cap \Gamma_0$.

It is known that the isomorphism class of $\pi(G_\Gamma, \Gamma_0)$ is independent of Γ_0, and it is called the *fundamental group* of the graph of groups G_Γ.

Given any graph of groups G_Γ, there is a canonically constructed tree T, called the *Bass–Serre tree*, together with a canonical action of the fundamental group of G_Γ. Moreover, the graph of groups G_Γ can be recovered from the action of its fundamental group on the Bass–Serre tree in a canonical way, which we describe below.

Let G be a group acting on a tree T without inversion, ie the action sends vertices to vertices and edges to edges such that every edge invariant under the action is being fixed. Let Γ be the quotient graph, and $p\colon T \to \Gamma$ be the quotient map. Let $T' \subset T$ be a subset and $T_0 \subset T'$ be a subtree of T. We call T' a *fundamental G–transversal* in T with subtree T_0 if (i) $p\colon T' \to \Gamma$ is bijective, and (ii) $p\colon T_0 \to \Gamma$ is onto a maximal tree in Γ. It is known that such a pair (T', T_0) always exists. Note that by (i), one can give a canonical graph structure to T' as follows: $VT' = VT \cap T'$, $ET' = ET \cap T'$, and the incidence functions $\bar\iota, \bar\tau\colon ET' \to VT'$ are defined by the equations

$$p(\bar\iota e) = p(\iota e), \quad p(\bar\tau e) = p(\tau e), \quad \forall e \in ET'.$$

(Here ι, τ are the incidence functions of T.) Note that by (ii), T_0 is a maximal tree in T' with respect to this graph structure, and $\bar{\iota}e = \iota e$, $\bar{\tau}e = \tau e$ for any $e \in ET_0$.

Now, given any fundamental G–transversal T' with subtree T_0, one can canonically construct a graph of groups G_Γ as follows, where Γ and T' are identified as graphs. For any $v \in VT'$, we assign to it the group $G(v) = G_v = \{g \in G \mid gv = v\}$, and for any $e \in ET'$, we assign to it the group $G(e) = G_e = \{g \in G \mid ge = e\}$. The boundary monomorphisms $\alpha\colon G(e) \to G(\bar{\iota}e)$, $\omega\colon G(e) \to G(\bar{\tau}e)$ are defined as follows. For any $e \in ET'$, pick $g_e, h_e \in G$ such that $g_e\bar{\iota}e = \iota e$, $h_e\bar{\tau}e = \tau e$, where for any $e \in ET_0$, $g_e = h_e = 1$. Then for any $g \in G(e)$, define $\alpha(g) = g_e^{-1}gg_e$ and $\omega(g) = h_e^{-1}gh_e$ (note that $G(e) \subset G(\iota e)$, $G(e) \subset G(\tau e)$).

There is an obvious homomorphism $\phi\colon \pi(G_\Gamma, T_0) \to G$ which sends t_e to $g_e^{-1}h_e \in G$. The fundamental theorem of the Bass–Serre theory asserts that ϕ is an isomorphism. Moreover, when T is the Bass–Serre tree of a graph of groups G_Γ and G is the fundamental group of G_Γ with the canonical action on T, the graph of groups G_Γ can be recovered in the above manner.

Next we review the Rips–Sela theory; see [38] for more details. Given any group G, a *Z–splitting* of G is a presentation of G as the fundamental group of a finite graph of groups where all the edge groups are infinite cyclic. *Elementary Z–splittings* are Z–splittings for which the graph of groups contains only one edge, ie an amalgamated product or an HNN extension. Given a Z–splitting of G and an elementary Z–splitting of a vertex group of the Z–splitting that is compatible with the boundary monomorphisms, there is a naturally defined new Z–splitting of G, which is called an *elementary refinement*, where the new graph of groups is obtained by replacing the vertex in the original graph by the corresponding one edge graph. A *refinement* of a Z–splitting is the result of a sequence of elementary refinements. The inverse operation of a refinement is called a *collapse*.

The fundamental result in the Rips–Sela theory concerns the existence of certain universal Z–splittings of a single-ended finitely presented group, called *canonical JSJ decompositions*, from which all other Z–splittings of the group can be derived in a certain organized way (involving refinement or collapse). The starting point of this work is an analysis of the interactions between two distinct elementary Z–splittings. To be more concrete, let $G = A_i *_{C_i} B_i$ (or $A_i *_{C_i}$) be two given elementary Z–splittings, where C_i is generated by c_i, for $i = 1, 2$. The element c_2 is called *elliptic* with respect to the first splitting if it is contained in a conjugate of A_1 or B_1, and *hyperbolic* otherwise, and similarly for c_1 with respect to the second splitting. With this understood, one of the basic result in the Rips–Sela theory (see [38, Theorem 2.1]) asserts that if G is freely indecomposable, then c_1 and c_2 are simultaneously elliptic or simultaneously hyperbolic.

The bulk of the Rips–Sela theory is devoted to the analysis of hyperbolic-hyperbolic splittings. Our first observation is that, for a group G with infinite $z(G)$, hyperbolic-hyperbolic splittings seldom occur, which greatly simplifies the situation.

Lemma 3.1 *Let G be a single-ended group with infinite $z(G)$, and suppose G is not isomorphic to the fundamental group of a 2–torus or Klein bottle. Then*

(i) *the center $z(G)$ is contained in the edge groups of every reduced Z–splitting of G, and*

(ii) *there are no hyperbolic-hyperbolic elementary Z–splittings of G.*

Proof We shall first prove part (i) of the lemma, where it suffices to consider only the case of elementary Z–splittings. Let $G = A *_C B$ or $A *_C$ be an elementary Z–splitting, where $A \neq C \neq B$. By Lemma 2.1, if the splitting is an amalgamated product, then C contains $z(G)$. If the splitting is an HNN extension and C does not contain $z(G)$, then $A = C = \langle c\rangle$ which is infinite cyclic, and G is isomorphic to $A *_A \alpha$ for a finite order automorphism α of A. Clearly α is either the identity or $\alpha\colon c \mapsto c^{-1}$, which implies that G is isomorphic to the fundamental group of a 2–torus or Klein bottle. Hence part (i) of the lemma.

As for part (ii), suppose to the contrary, there is a pair of hyperbolic-hyperbolic elementary Z–splittings $G = A_i *_{C_i} B_i$ (or $A_i *_{C_i}$), $i = 1, 2$, where C_i is generated by c_i. We first note that the hyperbolicity implies that the splittings are reduced. Then by part (i), there are integers $m, n > 0$ such that $c_1^m, c_2^n \in z(G)$, so that c_1^m and c_2^n commute. With this understood, Theorem 3.6 in Rips and Sela [38] implies that G is isomorphic to the fundamental group of either a 2–torus, or a Klein bottle, or a Euclidean 2–branched projective plane, or a Euclidean 4–branched sphere (an explicit presentation of these groups are given in Proposition 3.3 of [38], page 63). The case of 2–torus or Klein bottle is excluded by the assumptions of the lemma, and the rest of the cases are excluded by the fact that G has infinite center; see Lemma 2.2. (Note that in Theorem 3.6 of [38], there is the assumption that G is a freely indecomposable group which does not split over $\mathbb{Z}_2$. By Stallings' End theorem (see eg [42, Theorem 6.1]), G satisfies this assumption because of being single-ended.) Hence the lemma. □

We remark that hyperbolic-hyperbolic splittings do occur. For example, let G be the fundamental group of a Klein bottle. Then $G = A *_A \alpha$, where $A = \langle c\rangle$ is infinite cyclic and $\alpha\colon c \mapsto c^{-1}$, and $G = A *_C A$, where C is the index 2 subgroup of the infinite cyclic group A, are a pair of hyperbolic-hyperbolic splittings of G.

Let G be a single-ended group with infinite $z(G)$, which is not isomorphic to the fundamental group of a 2–torus or Klein bottle. Let T be the Bass–Serre tree of a reduced Z–splitting of G, and let V be the subset of the set of vertices VT which consists of v such that the isotropy subgroup G_v fixes a vertex $v' \neq v$. The subset V is clearly G–invariant, which gives rise to a G–invariant partition $(V, VT \setminus V)$ of VT. The following lemma is concerned with the structure of V.

Lemma 3.2 *There exists a collection of infinite cyclic subgroups G_i of G, $i \in I$, which has the following significance.*

- *For each $i \in I$, let V_i be the subset of V consisting of v such that $G_v = G_i$, and let $H_i \equiv \{t \in G \mid tgt^{-1} = g, \forall g \in G_i\}$ be the centralizer of G_i. Then H_i acts transitively on V_i.*
- *For each $i \in I$, let $\{g_j \mid j \in J(i)\}$ be a fixed choice of representatives of the right cosets of H_i in G, where the right coset H_i is represented by $g_j = 1$. Then $\{g_j(V_i) \mid j \in J(i), i \in I\}$ forms a partition of V.*

Proof Let $v \in V$ be any element, and let $v' \neq v$ be fixed under G_v. Since T is a tree, there exists a unique reduced path γ in T which connects v and v'. Because G_v fixes both v and v', and because γ is unique, G_v must also fix γ. In particular, if e is the edge in γ which is incident to v, then it follows easily that $G_v = G_e$, which implies that G_v is infinite cyclic.

Let v_1 be the other vertex in γ to which e is incident. Since the Z–splitting is reduced, v_1 must lie in the same orbit of v under the action of G. In other words, there is a $t \in G$ such that $t \cdot v = v_1$. Suppose $G_v = G_e$ is generated by c. Then $G_{v_1} = tG_vt^{-1}$ is generated by $c_1 \equiv tct^{-1}$. Furthermore, $c \in G_e \subset G_{v_1}$, so that $c = c_1^n$ for some $n \in \mathbb{Z}$. On the other hand, by Lemma 3.1(i), there exists a nonzero $m \in \mathbb{Z}$ such that $c^m \in z(G)$. Consequently,

$$c_1^m = (tct^{-1})^m = tc^mt^{-1} = c^m = c_1^{nm},$$

which implies $n = 1$. With $c = c_1 = tct^{-1}$, it follows that t lies in the centralizer of G_v, and moreover, $G_{v_1} = G_v$. Repeating this argument to v_1, we see that there is a t' lying in the centralizer of G_v, such that $t' \cdot v = v'$ and $G_{v'} = G_v$. Now if we let $V(v)$ be the subset of V consisting of elements whose isotropy subgroup equals G_v, and let $H(v)$ be the centralizer of G_v, then $H(v)$ acts transitively on $V(v)$.

The above analysis shows that the following relation $\sim$ on V is an equivalence relation: $v' \sim v$ if and only if G_v fixes v'. The equivalence relation gives rise to a partition of V. It is clear that one can choose a subset $\{V_i \mid i \in I\}$ of equivalence classes such that this partition can be described as $\{g_j(V_i) \mid j \in J(i), i \in I\}$, where G_i is the isotropy subgroup of the vertices in V_i, and g_j, $j \in J(i)$, is some fixed representative of the right coset of the centralizer H_i of G_i in G, with $g_j = 1$ for the right coset of H_i. This completes the proof of the lemma. □

3.2 Proof of Theorem 1.5(1)

By assumption, X is fiber-sum-decomposed into X_i along N_j. This gives rise to a Z–splitting of $\pi_1(X)$ which will be denoted by Λ, with vertex groups and edge groups given by $\pi_1(X_i)$ and $\pi_1(N_j)$, respectively. Note that Definition 1.3 (iv) implies that the Z–splitting Λ is reduced. Furthermore, we shall point out that by Lemma 3.1(i), $z(\pi_1(X))$ is contained in every edge group of Λ. On the other hand, recall that the fiber-sum decomposition of X gives rise to a canonical injective $\mathbb{S}^1$–action on X. We denote the orbit map by $\pi\colon X \to Y$, where we shall point out that Y is naturally a good orbifold, ie it does not contain any bad 2–suborbifolds. Let Σ_j be the spherical 2–suborbifold of Y over which N_j is Seifert fibered under π. Then it follows easily that the decomposition of Y in Y_i along Σ_j is a reduced spherical decomposition, where Y_i is the irreducible 3–orbifold in the orbit map $\pi_i\colon X_i \to Y_i$ that comes with the fiber-sum decomposition of X; see Definition 1.3.

Let Λ_{JSJ} be a canonical JSJ decomposition of $\pi_1(X)$ as constructed in [38]. We will show that Λ_{JSJ} and Λ are equivalent as canonical JSJ decompositions of $\pi_1(X)$ as described in [38]. To this end, we consider the Bass–Serre trees T_{JSJ} and T of Λ_{JSJ} and Λ, respectively, each equipped with the canonical action of $\pi_1(X)$. As for notations, recall that for any vertex v or edge e of T_{JSJ} or T, the corresponding isotropy subgroups of $\pi_1(X)$ are denoted by G_v or G_e, respectively.

Lemma 3.3 *For any $w \in VT$, G_w fixes a vertex of T_{JSJ}.*

Proof We consider the induced action of G_w on the Bass–Serre tree T_{JSJ}, and for any vertex v and edge e of T_{JSJ}, we denote by G'_v and G'_e the isotropy subgroups of the G_w–action at v and e, respectively. By Theorem 4.12 in [13], there are the following three possibilities.

(a) G_w fixes a vertex of T_{JSJ}.

(b) There is a reduced infinite path, $v_0, e_1^{\epsilon_1}, v_1, e_2^{\epsilon_2}, \dots$, in T_{JSJ} such that
$$G'_{v_0} \subset G'_{v_1} \subset \cdots, \quad G_w = \bigcup_{n\ge 0} G'_{v_n} = \bigcup_{n \ge 1} G'_{e_n},$$
and for all $n \ge 1$, $G_w \ne G'_{e_n}$.

(c) Some element of G_w translates some edge e of T_{JSJ}, and for $C \equiv G'_e$, either $G_w = B *_C D$ with $B \ne C \ne D$, or $G_w = B *_C$.

It remains to show that neither (b) nor (c) can occur. First, applying Lemma 3.1(i) to the $\pi_1(X)$–action on T_{JSJ}, we see that $z(\pi_1(X))$ fixes every edge of T_{JSJ}. Secondly, note that there is a factor X_i such that G_w is conjugate to the subgroup $\pi_1(X_i)$ in $\pi_1(X)$. Finally, if h denotes the homotopy class of a regular fiber of $\pi\colon X \to Y$, then $h \in z(\pi_1(X)) \cap G_w$, so that $h \in G'_e$ for every edge e of T_{JSJ}.

With the preceding understood, we consider case (b) first. In this case, we have

$$\pi_1^{\mathrm{orb}}(Y_i) \cong \pi_1(X_i)/\langle h\rangle \cong G_w/\langle h\rangle = \bigcup_{n\geq 1} G'_{e_n}/\langle h\rangle = \bigcup_{n\geq 1} F_n,$$

where F_n is a finite group, $F_n \subset F_{n+1}$, and $G_w/\langle h\rangle \neq F_n$ for all $n \geq 1$. Clearly, $\pi_1^{\mathrm{orb}}(Y_i)$ can not be finite. To rule out the case where $\pi_1^{\mathrm{orb}}(Y_i)$ is infinite, we note that $\pi_1^{\mathrm{orb}}(Y_i)$ has a finite index torsion-free subgroup H by the Geometrization theorem; see [4; 31]. Let $\widetilde{H}$ be the corresponding subgroup of $G_w/\langle h\rangle$ under $\pi_1^{\mathrm{orb}}(Y_i) \cong G_w/\langle h\rangle$. Then $\widetilde{H} = \bigcup_{n\geq 1} F_n \cap \widetilde{H} = \bigcup_{n\geq 1} \varnothing = \varnothing$, which is a contradiction. Hence case (b) is excluded.

For case (c), we set $C' = C/\langle h\rangle$, $B' = B/\langle h\rangle$, and $D' = D/\langle h\rangle$; then

$$G_w/\langle h\rangle = B' *_{C'} D', \text{ with } B' \neq C' \neq D', \text{ or } G_w/\langle h\rangle = B' *_{C'}.$$

Since C' is a finite group, $G_w/\langle h\rangle$ has more than one end by Stallings' End theorem; see eg [42, Theorem 6.1]. However, since Y_i is irreducible, the number of ends of $\pi_1^{\mathrm{orb}}(Y_i)$ is at most 1, which is a contradiction to $\pi_1^{\mathrm{orb}}(Y_i) \cong G_w/\langle h\rangle$. This rules out case (c), and the lemma is proved. □

Lemma 3.4 *There exists a $\pi_1(X)$–equivariant bijection $\phi\colon VT \to VT_{JSJ}$. In particular, for any $w \in VT$, $G_w = G_{\phi(w)}$.*

Proof First, we let V (resp. V_{JSJ}) be the subset of VT (resp. VT_{JSJ}) described in Lemma 3.2, and let G_i, V_i, H_i, g_j, $j \in J(i)$, $i \in I$, be as defined in Lemma 3.2 for VT.

Given any $w \in VT$, G_w fixes a vertex $v \in VT_{JSJ}$ by Lemma 3.3. On the other hand, since $\pi_1(X)$ has no hyperbolic-hyperbolic splittings (see Lemma 3.1(ii)), it follows from the construction of canonical JSJ decompositions in [38] that the action of G_v on T must also fix a vertex, say w'. One has the obvious inclusion relations $G_w \subset G_v \subset G_{w'}$. By Lemma 3.2, one always has $G_w = G_{w'}$, so that $G_v = G_w$ must hold. We will discuss according to cases (i) $w \in VT \setminus V$ and (ii) $w \in V$.

In case (i), $w' = w$. We claim that $v \in VT_{JSJ} \setminus V_{JSJ}$; in particular, v is uniquely determined by w. To see this, suppose there is a $v_1 \neq v$ such that $G_{v_1} = G_v$. Then by Lemma 3.2 there is a t lying in the centralizer of G_v such that $v_1 = t \cdot v$. In particular, t is not in $G_v = G_w$. This implies that $t \cdot w \neq w$, but $G_{t\cdot w} = G_w$, which is a contradiction to the assumption that $w \in VT \setminus V$. With this understood, we define

ϕ from $VT \setminus V$ to $VT_{JSJ} \setminus V_{JSJ}$ by setting $\phi(w) = v$. It follows easily that ϕ is a $\pi_1(X)$–equivariant bijection between $TV \setminus V$ and $VT_{JSJ} \setminus V_{JSJ}$. (The surjectivity part uses the fact that for any vertex $v \in VT_{JSJ}$, the action of G_v on T fixes a vertex. This is a consequence of Lemma 3.1(ii) by the construction of JSJ decompositions in [38].)

In case (ii) where $w \in V$, v also lies in V_{JSJ} by a similar argument. We shall define $\phi\colon V \to V_{JSJ}$ as follows. Let $V_{i,JSJ}$ be the subset of V_{JSJ} consisting of vertices whose isotropy subgroups are given by G_i. Then for any fixed choice of $w_i \in V_i$, $v_i \in V_{i,JSJ}$, there is an H_i–equivariant bijection $\phi\colon V_i \to V_{i,JSJ}$ sending w_i to v_i. Using the elements g_j, $j \in J(i)$, we can uniquely extend ϕ to a $\pi_1(X)$–equivariant bijection from $\bigcup_{j\in J(i)} g_j(V_i)$ to $\bigcup_{j\in J(i)} g_j(V_{i,JSJ})$, which defines ϕ from V to V_{JSJ}. This completes the proof of the lemma. □

According to Rips and Sela [38, Theorem 7.1], canonical JSJ decompositions of a single-ended, finitely presented group G are determined up to the following equivalence relation: the Bass–Serre trees are G–homotopy equivalent relative to the set of vertices. With this understood, Theorem 1.5(1) follows from part (1) of the following proposition. In (2)–(4) we list some consequences of (1) which will be used later in the proofs of Theorem 1.5(2), Theorem 1.1, and Theorem 1.2.

Proposition 3.5 (1) *There exist subdivisions T' and T'_{JSJ} of T and T_{JSJ}, respectively, and $\pi_1(X)$–equivariant simplicial maps $h_1\colon T' \to T_{JSJ}$ and $h_2\colon T'_{JSJ} \to T$ extending ϕ and ϕ^{-1}, respectively (ϕ as in Lemma 3.4), such that $h_2 \circ h_1$ and $h_1 \circ h_2$ are $\pi_1(X)$–homotopic, relative to the set of vertices, to the corresponding identity maps.*

(2) *There exists a bijection $\widehat{\phi}\colon V\Lambda \to V\Lambda_{JSJ}$, such that for any factor X_i of the fiber-sum decomposition of X, $\pi_1(X_i)$ is conjugate in $\pi_1(X)$ to the vertex group at the vertex $\widehat{\phi}(X_i)$ of Λ_{JSJ}. In particular, the number of factors X_i and the conjugacy classes of subgroups $\pi_1(X_i)$ depend only on $\pi_1(X)$.*

(3) *The cardinality of $\{N_j\}$ depends only on $\pi_1(X)$.*

(4) *For any N_j, there is an edge e_j of the graph of Λ_{JSJ} such that $\pi_1(N_j)$ is conjugate in $\pi_1(X)$ to the edge group at e_j, and vice versa. In particular, the set of conjugacy classes of subgroups $\pi_1(N_j)$ depends only on $\pi_1(X)$.*

Proof Fixing a choice of ϕ in Lemma 3.4, we shall define the subdivision T' of T and the simplicial map $h_1\colon T' \to T$ as follows. For any edge $e \in ET$, there is a unique reduced path in T_{JSJ} which starts from $\phi(\iota e)$ and ends at $\phi(\tau e)$. There is a unique subdivision of e such that ϕ can be extended to a simplicial map over

e. Doing this to every edge of T, we obtain the subdivision T' and the simplicial map h_1. The whole construction is clearly $\pi_1(X)$–equivariant because ϕ is $\pi_1(X)$–equivariant and reduced paths with fixed ends in a tree are unique. The subdivision T'_{JSJ} and the simplicial map h_2 are constructed similarly with ϕ replaced by ϕ^{-1}. One can further subdivide T' (still denoted by T' for simplicity) so that h_1 can be regarded as a simplicial map to the subdivision T'_{JSJ} of T_{JSJ}. With this understood, $h_2 \circ h_1\colon T' \to T$ is $\pi_1(X)$–homotopic to the identity map relative to the set of vertices VT because (i) it is identity on VT, and (ii) T is a tree. The statement about $h_1 \circ h_2$ follows similarly. This finishes the proof of part (1).

Part (2) is a direct consequence of Lemma 3.4. For part (3), recall that the set of edges of Λ is identified with the set $\{N_j\}$. With this understood, observe that the underlying graphs of Λ and Λ_{JSJ}, which are given by $T/\pi_1(X)$ and $T_{JSJ}/\pi_1(X)$, respectively, are homotopy equivalent, so that they have the same Euler characteristics. This shows that the Euler characteristic of Λ, ie the number of vertices minus the number of edges of Λ, depends only on $\pi_1(X)$. It follows that the cardinality of $\{N_j\}$ depends only on $\pi_1(X)$.

Finally, we give a proof for part (4). For any N_j, we choose an edge e of T whose $\pi_1(X)$–orbit corresponds to N_j. As we have shown in the proof of part (1), $h_2 \circ h_1(e)$ is a path in T which has the same initial and terminal points as e. Since T is a tree, the loop formed by $h_2 \circ h_1(e)$ and e^{-1} must be reduced, which implies that e lies in the image of $h_2 \circ h_1(e)$. Let e' be an edge of T_{JSJ} lying in the path $h_1(e)$ such that e is contained in the path $h_2(e')$. Then by the construction of h_1, h_2 in part (1), we have $G_e \subset G_{e'} \subset G_e$, which implies that $G_e = G_{e'}$. We name e_j to be the edge of Λ_{JSJ} which corresponds to the $\pi_1(X)$–orbit of e'. Then it follows that $\pi_1(N_j)$ is conjugate to the edge group of Λ_{JSJ} at e_j. Part (4) follows easily. This completes the proof of Proposition 3.5. □

3.3 Proof of Theorem 1.5(2)

Before turning to the proof of Theorem 1.5(2), we first give a geometric interpretation of the conjugacy classes of subgroups $\pi_1(N_j)$ in $\pi_1(X)$. We begin by observing that the submanifolds N_j fall into two different types as follows. Let Γ be the subgroup of $\pi_1(X)$ generated by the homotopy class of a regular fiber of $\pi\colon X \to Y$. Then N_j falls into two cases according to (i) $\Gamma = \pi_1(N_j)$, or (ii) Γ is a proper subgroup of $\pi_1(N_j)$. It is clear that case (i) corresponds to the case where Σ_j is an ordinary 2–sphere.

With the preceding understood, we have the following lemma.

Lemma 3.6 (1) *Suppose* Γ *is a proper subgroup of* $\pi_1(N_j)$ *for some* j. *Then for any* N_k, *if* $g^{-1}\pi_1(N_j)g \subset \pi_1(N_k)$ *for some* $g \in \pi_1(X)$, *then* $g^{-1}\pi_1(N_j)g = \pi_1(N_k)$. *In particular, if* $\pi_1(N_j) = z(\pi_1(X))$, *then* $\pi_1(N_k) = z(\pi_1(X))$ *for any* k.

(2) *Let* N_j *and* N_k *be given which are over* Σ_j *and* Σ_k, *respectively. Suppose there are components* γ_j *and* γ_k *of the singular set of* Y *such that* $\Sigma_j \cap \gamma_j \neq \varnothing$, $\Sigma_k \cap \gamma_k \neq \varnothing$, *and suppose that* $\pi_1(N_j)$ *and* $\pi_1(N_k)$ *are conjugate in* $\pi_1(X)$. *Then* γ_j *and* γ_k *are equivalent in the following sense: either* $\gamma_j = \gamma_k$, *or there are components of the singular set of* Y, $\gamma_0, \gamma_1, \dots, \gamma_N$, *and spherical* 2*–suborbifolds* $\Sigma_1, \dots, \Sigma_N \in \{\Sigma_j\}$, *such that*

$$\gamma_{\alpha-1} \cap \Sigma_\alpha \cap \gamma_\alpha \neq \varnothing, \quad \alpha = 1, 2, \dots, N.$$

Proof For part (1), let N_j and N_k be Seifert fibered over Σ_j and Σ_k, respectively, under $\pi\colon X \to Y$. Since Γ is a proper subgroup of $\pi_1(N_j)$ and $g^{-1}\pi_1(N_j)g \subset \pi_1(N_k)$ for some $g \in \pi_1(X)$, Γ is also a proper subgroup of $\pi_1(N_k)$. Consequently, there are components γ_j and γ_k of the singular set of Y such that $\Sigma_j \cap \gamma_j \neq \varnothing$ and $\Sigma_k \cap \gamma_k \neq \varnothing$. If $\gamma_j = \gamma_k$, one clearly has $g^{-1}\pi_1(N_j)g = \pi_1(N_k)$ as claimed.

Suppose $\gamma_j \neq \gamma_k$. We denote by Y_0 the 3–orbifold obtained from Y by removing a regular neighborhood of all singular circles of Y except γ_k. Note that Y_0 is a good 3–orbifold as Y is good. We let $\widehat{Y}_0$ be a 3–manifold cover of Y_0. We shall apply the Equivariant loop theorem (see eg [5, Theorem 3.19]) to $\widehat{Y}_0$ as follows. Denote by F a component of $\partial\widehat{Y}_0$ which contains the preimage of a meridian of γ_j. Then observe that the assumption $g^{-1}\pi_1(N_j)g \subset \pi_1(N_k)$ for some $g \in \pi_1(X)$ implies that F is not π_1–injective. Hence, by the Equivariant loop theorem, there is an equivariant compression 2–disc $\widehat{D}$ in $\widehat{Y}_0$ with $\partial\widehat{D} \subset F$. The group action on $\widehat{D}$ contains exactly one fixed point, which implies that the image of $\widehat{D}$ under the covering map $\widehat{Y}_0 \to Y_0$ is an embedded 2–disc D in $|Y_0|$ intersecting γ_k at exactly one point. Furthermore, it follows easily that ∂D must be a meridian of γ_j. Closing up D in $|Y|$, we obtain an embedded 2–sphere Σ, which intersects each of γ_j and γ_k at exactly one point and intersects no other singular circles. Since Y contains no bad 2–suborbifolds, it follows that γ_j and γ_k must have the same multiplicity, which implies that $g^{-1}\pi_1(N_j)g = \pi_1(N_k)$ as claimed. If $\pi_1(N_j) = z(\pi_1(X))$, then $\pi_1(N_j) \subset \pi_1(N_k)$ for any k by Lemma 3.1(i), which implies that $\pi_1(N_k) = z(\pi_1(X))$ for any k. This finishes off the proof of part (1).

Next we prove part (2). The idea is to show that up to replacing one or both of γ_j and γ_k by some singular circles that are equivalent in the sense described in part (2) of the lemma, the embedded 2–sphere Σ which we constructed in the previous paragraph can be modified so that it lies in the complement of the spherical 2–suborbifolds $\{\Sigma_j\}$. To this end, we first perturb Σ so that it intersects each element of $\{\Sigma_j\}$ transversely and the intersection occurs in the complement of the singular set of Y. Now we fix our attention on a $\Sigma' \in \{\Sigma_j\}$ such that $\Sigma \cap \Sigma' \neq \varnothing$. Let $l \in \Sigma \cap \Sigma'$ be a circle (if there is any) which bounds a disc $D \subset \Sigma'$ such that (i) D contains no singular points, and (ii) D contains no intersection points with Σ. Let D_1 and D_2 be the two discs into which l divides Σ. Then both $D_1 \cup D$, $D_2 \cup D$ are embedded 2–spheres in Y. Since Y contains no bad 2–suborbifolds, it follows easily that exactly one of D_1 and D_2, say D_1, contains no singular points. With this understood, we shall modify Σ by replacing D_1 with D and slightly perturbing it by an isotopy so that the new surface does not intersect Σ' in a neighborhood of D. In order to keep the notation simple, we shall still denote the resulting embedded 2–sphere by Σ. It is easily seen that the above procedure has the effect of removing the component l from $\Sigma \cap \Sigma'$, and moreover, it does not create new intersection points of Σ with any element of $\{\Sigma_j\}$. By repeating this procedure, we may assume now that the intersection of Σ with any element $\Sigma' \in \{\Sigma_j\}$ is either empty, or it consists of a union of circles each of which divides Σ' into two discs, each containing exactly one singular point.

One can further reduce the number of components of $\Sigma \cap \Sigma'$ to at most one. To see this, let l and l' be a pair of components of $\Sigma \cap \Sigma'$ such that l and l' bound an annulus $A' \subset \Sigma'$ and l bounds a disc $D' \subset \Sigma'$ where A' and D' do not contain any intersection points with Σ (note that if the number of components of $\Sigma \cap \Sigma'$ is greater than 1, such a pair always exists). Then the annulus $A \subset \Sigma$ bounded by l and l' does not contain any singular points, because otherwise, either l or l', say l, will bound a disc $D \subset \Sigma$ containing no singular points, and furthermore, D and a disc in Σ' bounded by l form an embedded 2–sphere in Y containing exactly one singular point, contradicting the fact that Y is pseudogood. With this understood, we modify Σ by replacing the annulus A with A', and as before, after applying a small isotopy the pair of components l and l' are removed and no new intersection points are created. By repeating this procedure, we may assume that for each $\Sigma' \in \{\Sigma_j\}$, the intersection $\Sigma \cap \Sigma'$ consists of at most one component.

Now we are at the final stage of modifying Σ. Let l be a circle of intersection of Σ with a $\Sigma' \in \{\Sigma_j\}$ such that l bounds a disc $D \subset \Sigma$ which does not intersect with any other elements of $\{\Sigma_j\}$. (Such l always exists, or Σ lies in the complement of $\{\Sigma_j\}$.) Let $D' \subset \Sigma'$ be a disc bounded by l. Then $D \cup D'$ is an embedded 2–sphere which can be perturbed so that it lies in the complement of $\{\Sigma_j\}$. Call it $\widehat{\Sigma}$, and suppose that $\widehat{\Sigma}$ lies in Y_i, which is an irreducible 3–orbifold. Furthermore, without loss of generality we assume D contains a singular point in γ_j, and we denote by γ_j' the singular circle which intersects with D'. We claim that γ_j and γ_j' are equivalent in the sense described in part (2) of the lemma. To see this, note that $\widehat{\Sigma}$ bounds a discal 3–orbifold in Y_i by the irreducibility of Y_i. In particular, there is an arc γ lying in the singular set of Y_i which connects the two singular points on $\widehat{\Sigma}$. If γ does not intersect any elements of $\{\Sigma_j\}$, then $\gamma_j = \gamma_j'$; hence they are equivalent. Suppose $\Sigma_1, \dots, \Sigma_N$ are the elements of $\{\Sigma_j\}$ that intersect with γ. Then there are subarcs $I_1, \dots, I_N$ of γ, where I_α is contained in the discal 3–orbifold in Y_i bounded by Σ_α, $1 \le \alpha \le N$. Clearly there are singular circles $\gamma_0, \gamma_1, \dots, \gamma_N$ such that the end points of I_α lie in $\gamma_{\alpha-1}$ and γ_α, respectively. It follows easily that γ_j and γ_j' are equivalent through $\gamma_0, \dots, \gamma_N$ and $\Sigma_1, \dots, \Sigma_N$. With this understood, we replace γ_j by γ_j', and we modify Σ by replacing D by D'. The new embedded 2–sphere can be perturbed slightly so that it does not intersect Σ' and no new intersection points with elements of $\{\Sigma_j\}$ were created. Furthermore, it intersects with each of the singular circles γ_k and γ_j' in exactly one point and contains no other singular points. By repeating this procedure, we obtain an embedded 2–sphere, which is still denoted by Σ, such that (i) Σ is in the complement of the elements of $\{\Sigma_j\}$, and (ii) Σ contains exactly two singular points lying on some singular components $\widehat{\gamma}_j$ and $\widehat{\gamma}_k$, which are equivalent to γ_j and γ_k, respectively. As we have shown earlier, $\widehat{\gamma}_j$ and $\widehat{\gamma}_k$ are equivalent, which implies that γ_j and γ_k are equivalent. This finishes the proof of the lemma. □

In summary, the conjugacy classes of subgroups $\pi_1(N_j)$ (which are the conjugacy classes of the edge groups of Λ) can be classified as follows: (i) there is a distinguished conjugacy class, ie the class of those $\pi_1(N_j) = \Gamma$, and this conjugacy class can be characterized by the fact that the corresponding Σ_j are ordinary 2–spheres; (ii) for any other conjugacy class where $\pi_1(N_j)$ contains Γ as a proper subgroup, there is an associated equivalence class of singular circles as described in Lemma 3.6(2), which is characterized by the fact that $\pi_1(N_j)$ belongs to the conjugacy class if and only if the corresponding Σ_j intersects with a singular circle belonging to the equivalence class. With this understood, we shall show in the next lemma that, by modifying the embeddings of N_j via fiber-preserving isotopies (with respect to $\pi\colon X \to Y$) if necessary, one can bring the underlying graph of the Z–splitting Λ into a certain normal form. We should point out that modifying the embeddings of N_j via fiber-preserving isotopies does not change the conjugacy classes of the edge groups of the Z–splitting.

Lemma 3.7 *For any given vertex v of Λ, and any conjugacy class of edge groups of Λ that are contained in the vertex group $G(v)$ up to conjugacy, one can modify the embeddings of those N_j via fiber-preserving isotopies, where $\pi_1(N_j)$ belongs to the given conjugacy class of edge groups, such that the Z–splitting of $\pi_1(X)$ associated to the new fiber-sum decomposition of X has the following property: for any edge e, if $G(e)$ belongs to the given conjugacy class of edge groups, then e is incident to v.*

Proof First of all, we observe that modifying the embeddings of N_j via fiber-preserving isotopies corresponds to moving one of the points $y_{j,1}$ or $y_{j,2}$ (see Definition 1.3) via isotopies, and moreover, for any Y_i, the edge which corresponds to N_j is incident to the vertex corresponding to X_i if and only if one of the points $y_{j,1}$ or $y_{j,2}$ lies in Y_i.

Now with the vertex v and the conjugacy class of edge groups given as in the lemma, we denote by X_0 the irreducible $\mathbb{S}^1$–four-manifold corresponding to v, and denote by Y_0 the corresponding irreducible 3–orbifold. We first note that the case where the given conjugacy class of edge groups is the distinguished one, ie where $\pi_1(N_j) = \Gamma$, is trivial, because in this case Σ_j is an ordinary 2–sphere and hence the points $y_{j,1}$ and $y_{j,2}$ are both lying in the complement of the singular set. For any other conjugacy class of edge groups, there is an associated equivalence class of singular circles as described in Lemma 3.6(2). Since the edge groups belonging to the given conjugacy class are contained in the vertex group $G(v) = \pi_1(X_0)$ up to conjugacy, there must be a singular circle belonging to the equivalence class which has nonempty intersection with the irreducible 3–orbifold Y_0. We pick one such singular circle and denote it by γ_0, and we set $I_0 \equiv Y_0 \cap \gamma_0 \neq \varnothing$. Now consider any N_j such that $\pi_1(N_j)$ belongs to the given conjugacy class of edge groups and $\Sigma_j \cap \gamma_0 \neq \varnothing$. There are two possibilities: (i) Σ_j intersects γ_0 at two points; (ii) Σ_j intersects γ_0 at only one point. Consider case (i) first. If we cut Y open along Σ_j and then fill in the 3–discal neighborhoods of $y_{j,1}$ and $y_{j,2}$, the singular circle γ_0 is turned into two components, one of which, denoted by γ', contains I_0. Without loss of generality, assume $y_{j,1}$ is contained in γ'. Then by moving $y_{j,1}$ along γ' via isotopy if necessary, we may arrange such that $y_{j,1} \in I_0$. Now consider case (ii). Let γ_1 be the singular circle which contains the other singular point on Σ_j. Then when we cut Y open along Σ_j and fill in the 3–discal neighborhoods of $y_{j,1}$ and $y_{j,2}$, the two components γ_0 and γ_1 are turned into one component, denoted by γ'. In this case, one can always arrange so that $y_{j,1} \in I_0$, by moving $y_{j,1}$ via isotopy along γ'. Note that after moving $y_{j,1}$ via isotopy and then performing the connected sum operation to get back to Y, the singular circles γ_0 and γ_1 are turned into γ_0' and γ_1', respectively, both of which have nonempty intersection with Y_0. With this last property understood, observe that we can now perform the operation described above to any N_j such that $\Sigma_j \cap \gamma_1' \neq \varnothing$. The lemma follows by an induction process. □

We remark that applying Lemma 3.7 to a Z–splitting Λ does not change the sets $V\Lambda$ and $E\Lambda$; it only changes the incident function. From the construction of Bass–Serre trees (see [13]), it follows particularly that neither the action of $\pi_1(X)$ on the vertex set of the Bass–Serre tree T of Λ changes, nor does the $\pi_1(X)$–equivariant bijection ϕ in Lemma 3.4.

Proof of Theorem 1.5(2) First of all, we shall reformulate the problem as follows. We denote the group $\pi_1(X')$ by G and identify $\pi_1(X)$ with G via the given isomorphism $\alpha\colon \pi_1(X) \to \pi_1(X')$. With this understood, let Λ and Λ' be the Z–splittings of G associated to the given fiber-sum decompositions of X and X', respectively. We shall prove that after modifying the embeddings of N_j and N_j' via fiber-preserving isotopies if necessary, Λ and Λ' may be arranged to be isomorphic as Z–splittings of G. Note that the assumption that N_j and N_j' are null-homologous is equivalent to that the underlying graphs of Λ and Λ' are trees. We shall denote by T and T' the Bass–Serre trees of Λ and Λ', respectively. By Lemma 3.4, there exists a G–equivariant bijection ϕ from VT onto VT', which induces a bijection $\hat{\phi}\colon V\Lambda \to V\Lambda'$ and a family of isomorphisms of vertex groups $\rho_v\colon G(v) \to G(v')$ given by conjugation by elements of G, where $v \in V\Lambda$ and $v' = \hat{\phi}(v) \in V\Lambda'$.

First consider the special case where $\pi_1(N_j) = z(G) = \pi_1(N_j')$ for all N_j and N_j'. We fix a vertex $v \in V\Lambda$ and let $v' = \hat{\phi}(v) \in V\Lambda'$ be the corresponding vertex. Then we apply Lemma 3.7 to Λ and Λ' so that for the resulting new Z–splittings, which are still denoted by Λ and Λ' for simplicity, every edge $e \in E\Lambda$ and $e' \in E\Lambda'$ is incident to v and v', respectively. With this understood, there is an isomorphism of the underlying graphs of Λ and Λ', extending $\hat{\phi}\colon V\Lambda \to V\Lambda'$. Since by assumption all the edge groups of Λ and Λ' are given by the center $z(G)$, it follows easily that the family of isomorphisms ρ_v can be extended to an isomorphism of the Z–splittings Λ and Λ'. This finishes the proof for the special case where $\pi_1(N_j) = z(G) = \pi_1(N_j')$ for all N_j and N_j'.

Suppose $\pi_1(N_j) = z(G)$ for all N_j does not hold. Then by Lemma 3.6(1), the condition that π_1 of a regular fiber of $\pi\colon X \to Y$ is a proper subgroup of $\pi_1(N_j)$ for some N_j is equivalent to the more convenient condition that $\pi_1(N_j) \neq z(G)$, as the latter is formulated without reference to $\pi\colon X \to Y$. On the other hand, by Proposition 3.5(4), $\pi_1(N_j') = z(G)$ for all N_j' also does not hold. Accordingly, one can divide the set of edges $E\Lambda$ (resp. $E\Lambda'$) into two groups by the following rules:

(I) $e \in E\Lambda$ (resp. $e' \in E\Lambda'$) belongs to (I) if and only if $G(e) \neq z(G)$ (resp. $G(e') \neq z(G)$).

(II) $e \in E\Lambda$ (resp. $e' \in E\Lambda'$) belongs to (II) if and only if $G(e) = z(G)$ (resp. $G(e') = z(G)$).

Pick a vertex $v \in V\Lambda$, and without loss of generality, assume that there is an edge e belonging to (I) such that $G(e)$ is conjugate to a subgroup of $G(v)$. We denote the set of such edges by E_v. Then by Lemma 3.7, we can assume that any $e \in E_v$ is incident to v. Furthermore, we can assume (again with the help of Lemma 3.7) that any $e \in E\Lambda$ belonging to (II) is not incident to v by the fact that $E_v \neq \varnothing$. With this understood, we denote by Γ_v the minimal subgraph containing v and E_v and by G_{Γ_v} the corresponding subgraph of groups supported by Γ_v. Finally, we let $v' = \hat{\phi}(v) \in V\Lambda'$ be the corresponding vertex in the Z–splitting Λ'. We make the same arrangement as above for the vertex v' with the corresponding notations in which v is replaced by v'.

Our next goal is to construct an isomorphism between the subgraphs of groups G_{Γ_v} and $G_{\Gamma_{v'}}$, extending the given isomorphism $\rho_v\colon G(v) \to G(v')$. To this end, we pick a fundamental G–transversal for G_{Γ_v} as follows. Let $\tilde{v}$ be a vertex of the Bass–Serre tree T whose G–orbit is v. For each $e \in E_v$, we choose an edge $\tilde{e} \in ET$ incident to $\tilde{v}$, whose G–orbit is e. We let $\Gamma_{\tilde{v}}$ be the minimal subgraph of T containing $\tilde{v}$ and $\tilde{e}$, $\forall e \in E_v$. Then it is clear that $\Gamma_{\tilde{v}}$ is a fundamental G–transversal for G_{Γ_v}. With this understood, we shall construct a fundamental G–transversal for $G_{\Gamma_{v'}}$ as follows.

We set $\tilde{v}' = \phi(\tilde{v})$, where $\phi\colon VT \to VT'$ is the G–equivariant bijection coming from Lemma 3.4, which induces $\hat{\phi}\colon V\Lambda \to V\Lambda'$. For any edge $\tilde{e} \in \Gamma_{\tilde{v}}$, we denote by $\widetilde{w}$ the vertex other than $\tilde{v}$ to which $\tilde{e}$ is incident, and set $\widetilde{w}' = \phi(\widetilde{w})$ correspondingly. Then as in the proof of Proposition 3.5, there exists a unique reduced path in T' connecting $\tilde{v}'$ to $\widetilde{w}'$,

$$v_0 = \tilde{v}', \quad e_1^{\epsilon_1}, \quad v_1, \quad e_2^{\epsilon_2}, \quad \dots, \quad e_n^{\epsilon_n}, \quad v_n = \widetilde{w}',$$

such that $G_{\tilde{e}} \subset G_{e_i}$ for all i and that there exists a j with $G_{e_j} = G_{\tilde{e}}$. Let $\hat{e}_i \in E\Lambda'$ be the G–orbit of e_i. Then since the edge $e \in E_v$ belongs to (I), it follows that $\hat{e}_j \in E\Lambda'$ also belongs to (I) because $G_{e_j} = G_{\tilde{e}}$. Now with $G_{e_j} = G_{\tilde{e}} \subset G_{e_i}$, it follows from Lemma 3.6(1) that $G_{e_j} = G_{e_i}$ for all i, which implies that the edge groups $G(\hat{e}_i)$ belong to the same conjugacy class in G. It follows that the vertices v_k, where k is even, must be in the same G–orbit, and that n must be odd. In particular, v_{n-1} and $v_0 = \tilde{v}'$ are in the same G–orbit. We fix a choice of $g_{\tilde{e}} \in G$ such that $g_{\tilde{e}} v_{n-1} = \tilde{v}'$, set $\tilde{e}' = e_n$, and let $w \in V\Lambda$ and $w' \in V\Lambda'$ be the G–orbit of $\widetilde{w}$ and $\widetilde{w}'$, respectively. Then the G–orbit $e' \in E\Lambda'$ of $\tilde{e}'$ is incident to the vertices v' and w'. It follows that e', w' are part of the subgraph $\Gamma_{v'}$, and $v \mapsto v'$, $e \mapsto e'$ and $w \mapsto w'$ define an isomorphism between Γ_v and $\Gamma_{v'}$.

Suppose $\rho_v\colon G(v) \to G(v')$ is given by $h \mapsto g_{\tilde{v}} h g_{\tilde{v}}^{-1}$ for some $g_{\tilde{v}} \in G$, where $h \in G_{\tilde{v}}$. Then the subset $\{g_{\tilde{v}}\tilde{v}', g_{\tilde{v}}g_{\tilde{e}}\tilde{e}', g_{\tilde{v}}g_{\tilde{e}}\widetilde{w}' \mid e \in E_v, w \in \Gamma_v\}$ is a fundamental G–transversal for $G_{\Gamma_{v'}}$. Moreover, there is an isomorphism $\{\rho_v, \rho_e, \rho_w \mid e \in E_v, w \in \Gamma_v\}$ between the subgraphs of groups G_{Γ_v} and $G_{\Gamma_{v'}}$, extending the given isomorphism $\rho_v\colon G(v) \to G(v')$, where $\rho_e\colon G_{\tilde{e}} \to G_{g_{\tilde{v}}g_{\tilde{e}}\tilde{e}'}$ and $\rho_w\colon G_{\widetilde{w}} \to G_{g_{\tilde{v}}g_{\tilde{e}}\widetilde{w}'}$ are given by conjugation of $g_{\tilde{v}}g_{\tilde{e}} \in G$.

Finally, by repeating the above construction, we obtain a disjoint union of subgraphs of groups G_{Γ_k} of the Z–splitting Λ, a disjoint union of subgraphs of groups $G_{\Gamma'_k}$ of the

Z–splitting Λ', and a collection of isomorphisms $\rho_k\colon G_{\Gamma_k} \to G_{\Gamma'_k}$, such that for any edges $e \in E\Lambda \setminus \{\Gamma_k\}$ and $e' \in E\Lambda' \setminus \{\Gamma'_k\}$, $G(e) = z(G) = G(e')$. It follows easily that the isomorphisms ρ_k can be uniquely extended to an isomorphism of Z–splittings between Λ and Λ'. This finishes the proof of Theorem 1.5(2). □

4 Irreducible $\mathbb{S}^1$–four-manifolds

This section is devoted to a proof of Theorem 1.6. The proof involves a smooth classification of fixed-point free, smooth $\mathbb{S}^1$–four-manifolds whose π_1 has a center of rank greater than 1 (see Theorem 4.3), which is given at the end of the section.

The following lemma shows that a finitely generated group with infinite center is either single-ended or double-ended.

Lemma 4.1 *Let G be a finitely generated group with infinite $z(G)$ and suppose G is not single-ended. Then G is isomorphic to $A *_A \alpha$, where A is a finite group. In particular, G is double-ended.*

Proof Let $e(G)$ denote the number of ends of G. Then $e(G) \geq 1$ because G is infinite. On the other hand, by Stallings' End theorem (see eg Scott and Wall [42]), if $e(G) \geq 2$, then G splits over a finite subgroup, ie either $G = A *_C B$ with $A \neq C \neq B$, or $G = A *_C \alpha$, where in both cases C is a finite group. By Lemma 2.1, the assumption that $z(G)$ is infinite implies that the first case can not occur, and in the second case, $C = A = \alpha(C)$. In particular, A is a finite group. □

Lemma 4.2 *Let $\pi\colon X \to Y$ be the orbit map of an injective $\mathbb{S}^1$–action. Then $\pi_1(X)$ is double-ended if and only if $\pi_1^{\mathrm{orb}}(Y)$ is finite.*

Proof It suffices to show that if $\pi_1(X)$ is double-ended, then $\pi_1^{\mathrm{orb}}(Y)$ is finite; the other direction is trivial; see eg Scott and Wall [42]. To see this, note that $\pi_1(X) = A *_A \alpha$ for a finite group A by Lemma 4.1, where we recall that $A *_A \alpha$ is generated by elements of A and a letter t with additional relations $tat^{-1} = \alpha(a)$, $a \in A$. If we let H be the cyclic subgroup generated by t, then H has finite index in $\pi_1(X)$. On the other hand, if we let Γ be the subgroup generated by the homotopy class of a regular fiber of π, then $\Gamma \cap H$ has finite index in H because α is of finite order. Consequently $\Gamma \cap H$ has finite index in $\pi_1(X)$. This implies that the index of Γ in $\pi_1(X)$ is also finite, which means exactly that $\pi_1^{\mathrm{orb}}(Y)$ is finite. Hence the lemma. □

Proof of Theorem 1.6 Part (1) The proof for this part is based on the rigidity of *injective Seifert fibered space construction*, which we shall briefly review first; see Lee and Raymond [28] for more details. Suppose we are given a group π together with a short exact sequence $1 \to \Gamma \to \pi \to Q \to 1$, where $\Gamma = \mathbb{Z}^k$. Let W be a simply connected smooth manifold and consider the trivial principal $\mathbb{R}^k$–bundle $\mathbb{R}^k \times W$ over W. Let ψ be a smooth, free and properly discontinuous action of π on $\mathbb{R}^k \times W$ via bundle morphisms, such that the restriction $\psi|_\Gamma$ is given by translations via an embedding $\epsilon\colon \Gamma = \mathbb{Z}^k \to \mathbb{R}^k$ as a uniform lattice. Such an action ψ induces a smooth action of Q on W, which is denoted by ρ. The quotient space $E \equiv \mathbb{R}^k \times W/\psi(\pi)$ is a Seifert fibered space over the orbifold $W/\rho(Q)$, with regular fiber $T^k = \mathbb{R}^k/\epsilon(\Gamma)$ which is a k–dimensional torus. Conversely, a Seifert fibered space with a regular fiber T^k must arise from such a construction if the inclusion of a regular fiber induces an injective map on π_1 (such Seifert fibered spaces are called injective). In this case the short exact sequence $1 \to \Gamma \to \pi \to Q \to 1$ is part of the homotopy exact sequence associated to the corresponding fibration, with π being the π_1 of the Seifert fibered space, $\Gamma = \mathbb{Z}^k$ being the π_1 of a regular fiber, and Q being the π_1^{orb} of the base orbifold.

Given two such actions ψ_1 and ψ_2 of π, with induced embeddings $\epsilon_1, \epsilon_2\colon \Gamma \to \mathbb{R}^k$ and induced actions ρ_1 and ρ_2 of Q on W, the aforementioned rigidity theorem asserts that if ρ_1 and ρ_2 are conjugate by a diffeomorphism $h\colon W \to W$, then ψ_1 and ψ_2 are conjugate by (λ, g, h), where $\lambda \in C^\infty(W, \mathbb{R}^k)$, $g \in \mathrm{GL}(k, \mathbb{R})$, and

$$(\lambda, g, h) \cdot (v, w) = (g(v) + \lambda(h(w)), h(w)), \;\; (v, w) \in \mathbb{R}^k \times W.$$

Note that in particular, the corresponding Seifert fibered spaces $E_1 = \mathbb{R}^k \times W/\psi_1(\pi)$ and $E_2 = \mathbb{R}^k \times W/\psi_2(\pi)$ are diffeomorphic via a fiber-preserving diffeomorphism induced by (λ, g, h); see [28, page 381].

Now let E_1 and E_2 be two injective Seifert fibered spaces and let $\alpha\colon \pi_1(E_1) \to \pi_1(E_2)$ be an isomorphism. Furthermore, we assume that the universal covers of E_1 and E_2 are diffeomorphic, say given by $\mathbb{R}^k \times W$, and that the isomorphism $\alpha\colon \pi_1(E_1) \to \pi_1(E_2)$ respects the homotopy exact sequences associated to the corresponding fibrations on E_1 and E_2. Note that the latter is always true when there is a certain uniqueness of the short exact sequence $1 \to \Gamma \to \pi \to Q \to 1$, eg when $\Gamma = z(\pi)$. With this understood, we denote the group $\pi_1(E_2)$ by π and identify $\pi_1(E)$ with π via α. Then E_1 and E_2 may be regarded as arising from the injective Seifert fibered space construction for some actions ψ_1 and ψ_2 of π on $\mathbb{R}^k \times W$. Let ρ_1 and ρ_2 be the induced actions of Q on W. Then the rigidity theorem mentioned above implies that there is a fiber-preserving diffeomorphism $\phi\colon E_1 \to E_2$ such that $\phi_* = \alpha\colon \pi_1(E_1) \to \pi_1(E_2)$ if ρ_1 and ρ_2 are conjugate by a diffeomorphism $h\colon W \to W$. (Roughly speaking, the above rigidity theorem allows us to show that if the diffeomorphism classification of the base orbifolds are determined by the fundamental groups, then so are the fiber-preserving diffeomorphism classification of the corresponding Seifert fibered spaces.)

With the preceding understood, we shall now give a proof for part (1). Consider first the case where rank $z(\pi_1(X)) > 1$. A smooth classification of such fixed-point free, smooth $\mathbb{S}^1$–four-manifolds is given in Theorem 4.3, which shows that it suffices to consider the case where rank $z(\pi_1(X)) = 2$ and $\pi_2(X) = 0$. Moreover, it also shows that, in this case, X and X' arise from the above injective Seifert fibered space construction with $k = 2$ and $W = \mathbb{R}^2$. (Note that the uniqueness of the short exact sequence follows from the fact that $\Gamma = z(\pi)$; see Lemma 2.2(a).) With this understood, the existence of $\phi\colon X \to X'$ with $\phi_* = \alpha$ follows from the fact that for orientable 2–orbifolds with infinite fundamental group, any isomorphism of π_1^{orb} may be realized by a diffeomorphism of the 2–orbifolds; eg see [29].

It remains to consider the case where rank $z(\pi_1(X)) = 1$. In this case, X is an injective Seifert fibered space over a 3–orbifold Y with regular fiber $\mathbb{S}^1$, where Y is an irreducible 3–orbifold with infinite fundamental group. As Y is good, the Geometrization theorem implies that $Y = \widetilde{Y}/G$ for some aspherical 3–manifold $\widetilde{Y}$; see [31; 4]. (Note that G may be trivial here.) Furthermore, by the Geometrization theorem, $\widetilde{Y}$ admits a geometric decomposition; see eg Kleiner and Lott [27]. In particular, $\widetilde{Y}$ is either Haken, or Seifert fibered, or hyperbolic, and the universal cover of $\widetilde{Y}$ is diffeomorphic to $\mathbb{R}^3$. With this understood, we see that X arises from the injective Seifert fibered space construction with $k = 1$ and $W = \mathbb{R}^3$. (Note that the condition $\Gamma = z(\pi)$ is satisfied (see Lemma 2.3), which gives the required uniqueness for the short exact sequence $1 \to \Gamma \to \pi \to Q \to 1$.) It remains to show that for irreducible 3–orbifolds with infinite fundamental group, any isomorphism of π_1^{orb} may be realized by a diffeomorphism of the 3–orbifolds. This was verified by McCullough and Miller (see the proof of Corollary 5.3 in [31]) when $\widetilde{Y}$ is either Haken or Seifert fibered. For the remaining case, the 3–orbifolds are hyperbolic, and in this case, Mostow Rigidity implies that any isomorphism of π_1^{orb} may be realized by an isometry of the 3–orbifolds. This finishes off the proof for part (1).

Part (2): Let $\pi\colon X \to Y$ be the orbit map of the $\mathbb{S}^1$–action on X. By Lemma 4.2, this is the case precisely when Y has finite fundamental group. By the Geometrization theorem, Y is a spherical 3–orbifold, ie there is a finite subgroup G of SO(4) such that $Y = \mathbb{S}^3/G$. Note that the Euler class of $\pi\colon X \to Y$ is torsion, so that there is a 3–manifold $\widehat{Y}$ and a periodic diffeomorphism f such that $Y = \widehat{Y}/\langle f\rangle$ and X is the mapping torus of f. Moreover, by the Geometrization theorem, $\widehat{Y}$ is an elliptic 3–manifold. Similar conclusions hold for X'; ie X' is the mapping torus of a periodic diffeomorphism f' of an elliptic 3–manifold $\widehat{Y}'$.

Note that the mapping torus descriptions of X and X' imply that $\pi_1(X)$ and $\pi_1(X')$ are given by HNN extensions $\pi_1(\widehat{Y}) *_{\pi_1(\widehat{Y})} f_*$ and $\pi_1(\widehat{Y}') *_{\pi_1(\widehat{Y}')} f'_*$, respectively. This gives rise to short exact sequences

$$1 \to \pi_1(\widehat{Y}) \xrightarrow{i} \pi_1(X) \xrightarrow{p} \mathbb{Z} \to 1 \quad \text{and} \quad 1 \to \pi_1(\widehat{Y}') \xrightarrow{i'} \pi_1(X') \xrightarrow{p'} \mathbb{Z} \to 1.$$

With this understood, given any isomorphism $\alpha\colon \pi_1(X) \to \pi_1(X')$, we observe that the homomorphism $p' \circ \alpha \circ i\colon \pi_1(\widehat{Y}) \to \mathbb{Z}$ is trivial because $\pi_1(\widehat{Y})$ is finite. This implies that $\alpha \circ i\colon \pi_1(\widehat{Y}) \to \pi_1(X')$ lies in the image of $i'\colon \pi_1(\widehat{Y}') \to \pi_1(X')$. It follows easily from this consideration that $\alpha\colon \pi_1(X) \to \pi_1(X')$ induces an isomorphism $\widehat{\alpha}\colon \pi_1(\widehat{Y}) \to \pi_1(\widehat{Y}')$ such that $f'_* = \widehat{\alpha} \circ f_* \circ \widehat{\alpha}^{-1}$ as an element of $\mathrm{Out}\,(\pi_1(\widehat{Y}'))$. Suppose $\widehat{\alpha}$ can be realized by a diffeomorphism $h\colon \widehat{Y} \to \widehat{Y}'$, eg when $\widehat{Y}$ and $\widehat{Y}'$ are not lens spaces. Identifying $\widehat{Y}$ with $\widehat{Y}'$ via h, X may be regarded as the mapping torus of the periodic diffeomorphism $g = h \circ f \circ h^{-1}\colon \widehat{Y}' \to \widehat{Y}'$. Now observe that $g_* = f'_*$ as an element of $\mathrm{Out}\,(\pi_1(\widehat{Y}'))$, which implies that g and f' are homotopic, hence isotopic; see [1; 39; 7; 26; 40; 6]. The existence of $\phi\colon X \to X'$ with $\phi_* = \alpha$ follows easily from these considerations. This finishes the proof of part (2). $\square$

We end this section with the smooth classification theorem alluded to earlier. The proof of the theorem employs a key lemma, Lemma 5.2, whose proof will be given in the next section.

Theorem 4.3 *Suppose that X is a fixed-point free, smooth $\mathbb{S}^1$–four-manifold with rank $z(\pi_1(X)) > 1$. Then X belongs to one of the following cases:*

1. *If rank $z(\pi_1(X)) > 2$, then X is diffeomorphic to the 4–torus T^4.*
2. *If rank $z(\pi_1(X)) = 2$ and $\pi_2(X) \neq 0$, then X is diffeomorphic to $T^2 \times \mathbb{S}^2$.*
3. *If rank $z(\pi_1(X)) = 2$ and $\pi_2(X) = 0$, then X is diffeomorphic to $\mathbb{S}^1 \times N^3/G$, where N^3 is an irreducible Seifert 3–manifold with infinite fundamental group, and G is a finite cyclic group acting on $\mathbb{S}^1 \times N^3$ preserving the product structure and orientation on each factor, and the Seifert fibration on N^3.*

Proof Let $\pi\colon X \to Y$ be the orbit map of the $\mathbb{S}^1$–action. Note that $\pi_*\colon \pi_1(X) \to \pi_1^{\mathrm{orb}}(Y)$ is surjective, so that $\pi_*(z(\pi_1(X))$ is contained in $z(\pi^{\mathrm{orb}}(Y))$. It follows easily from rank $z(\pi_1(X)) > 1$ that $z(\pi^{\mathrm{orb}}(Y))$ is infinite. By Lemma 5.2, Y is Seifert fibered, and furthermore, by Lemma 2.6, $\pi\colon X \to Y$ extends to a principal T^2–bundle over a 2–orbifold B, which will be denoted by $\Pi\colon X \to B$. We remark that B is an orientable, closed 2–orbifold.

We begin by describing a decomposition of the principal T^2–bundle into a pair of principal $\mathbb{S}^1$–bundles over B. More concretely, given any basis (e_1, e_2) of $\pi_1(T^2)$, we let $\theta_i\colon T^2\to\mathbb{S}^1$, $i=1,2$, be the projections to the first and the second factor of the decomposition $T^2=\mathbb{S}^1\times\mathbb{S}^1$ that is determined by the basis (e_1,e_2). This gives rise to a pair of principal $\mathbb{S}^1$–bundles over B, denoted by V_1 and V_2, which are induced by θ_1 and θ_2, respectively. Note that one can recover the principal T^2–bundle $\Pi\colon X\to B$ as the pull-back bundle of $V_1\times V_2\to B\times B$ via the diagonal map $B\to B\times B$. Moreover, with a change of basis, one can always arrange V_1 to have vanishing Euler number. Indeed, under the change of basis

$$e_1=ae_1'+ce_2',\quad e_2=be_1'+de_2',$$

where $ad-bc=1$, the corresponding principal $\mathbb{S}^1$–bundles V_1' and V_2' associated to the basis (e_1',e_2') have Euler numbers

$$e(V_1')=a\cdot e(V_1)+b\cdot e(V_2),\quad e(V_2')=c\cdot e(V_1)+d\cdot e(V_2).$$

If both of $e(V_1)$ and $e(V_2)$ are nonzero, one can choose a unique pair of integers (up to a sign), (a,b), such that $e(V_1')=0$. Note that, up to a sign, $e(V_2')$ is independent of the choices of c and d. This said, we shall assume in what follows that $e(V_1)=0$.

With these preparations, we now consider case (1) where rank $z(\pi_1(X))>2$. It is clear that $z(\pi_1^{\mathrm{orb}}(B))$ is nontrivial and infinite. By Lemma 2.2(a), B must be a nonsingular torus. As $e(V_1)=0$ and B is nonsingular, V_1 is trivial, which implies that $X=\mathbb{S}^1\times V_2$. Finally, the assumption that rank $z(\pi_1(X))>2$ implies that V_2 must also be trivial. Hence X is diffeomorphic to the 4–torus T^4.

Consider case (2) where rank $z(\pi_1(X))=2$ and $\pi_2(X)\neq 0$. Note that X is a principal $\mathbb{S}^1$–bundle over V_2, which is the pull-back of the principal $\mathbb{S}^1$–bundle $V_1\to B$ via the map $V_2\to B$. The homotopy exact sequence associated to the fibration $X\to V_2$ (see Haefliger [22]) implies that $z(\pi_1^{\mathrm{orb}}(V_2))$ is infinite and $\pi_2^{\mathrm{orb}}(V_2)\neq 0$. By Lemma 5.2, V_2 is the mapping torus of a periodic diffeomorphism of a 2–orbifold Σ where $\pi_1^{\mathrm{orb}}(\Sigma)$ is finite. Now observe that $e(V_1)=0$ implies that Σ must be either $\mathbb{S}^2$ or a football. It follows easily that X is diffeomorphic to $T^2\times\mathbb{S}^2$, which finishes the proof for case (2).

For case (3) where rank $z(\pi_1(X))=2$ and $\pi_2(X)=0$, we first observe that $\pi_1^{\mathrm{orb}}(B)$ is infinite, and therefore B is good. Let $B=\widetilde{B}/\Gamma$, where $\widetilde{B}$ is a closed orientable surface and Γ is a finite group acting on $\widetilde{B}$. We let $\widetilde{X}$, $\widetilde{V}_1$, and $\widetilde{V}_2$ be the pull-backs of $X\to B$, $V_1\to B$, and $V_2\to B$ to $\widetilde{B}$ via $\widetilde{B}\to B=\widetilde{B}/\Gamma$. Then Γ acts freely on $\widetilde{X}$, giving $X=\widetilde{X}/\Gamma$ and $\widetilde{V}_1=\mathbb{S}^1\times\widetilde{B}$. Let Γ_1 be the subgroup of Γ which acts trivially on the $\mathbb{S}^1$–factor in $\widetilde{V}_1=\mathbb{S}^1\times\widetilde{B}$. Then Γ_1 acts freely on $\widetilde{V}_2$. Denote by N^3 the quotient $\widetilde{V}_2/\Gamma_1$, which is clearly an irreducible Seifert 3–manifold with infinite fundamental group. With this understood, note that $\widetilde{X}/\Gamma_1=\mathbb{S}^1\times N^3$, so that if we set $G = \Gamma/\Gamma_1$, then $X = \mathbb{S}^1 \times N^3/G$ where the action of G preserves the product structure and the orientation of each factor, as well as the Seifert fibration on N^3, as claimed. This finishes the proof of Theorem 4.3. □

Remark 4.4 Theorem 2.1 in [12] asserts that if a 4–manifold with $b_2^+ \geq 1$ has nontrivial Seiberg–Witten invariant, then the homotopy class of the principal orbits of any smooth, fixed-point free $\mathbb{S}^1$–action on the manifold must be of infinite order; in particular, the center of the fundamental group must be infinite. As a corollary of Theorem 4.3(2), the converse of the above statement is not true. More concretely, consider a ruled surface X which is a nontrivial $\mathbb{S}^2$–bundle over T^2. Note that X satisfies $b_2^+ \geq 1$, it has nontrivial Seiberg–Witten invariant, and $z(\pi_1(X))$ is infinite. However, by Theorem 4.3(2), X does not admit any smooth, fixed-point free $\mathbb{S}^1$–action. It is also interesting to note that a double cover of X, which is diffeomorphic to $\mathbb{S}^2 \times T^2$, admits a smooth, fixed-point free $\mathbb{S}^1$–action. We remark that for a closed aspherical manifold, such a correlation between the existence of circle actions and the nontriviality of the center of the fundamental group is part of a conjectured rigidity of aspherical manifolds going back to work of Borel. See [8] for some recent progress and more detailed discussions.

5 Injectivity of $\mathbb{S}^1$–actions when π_1 has infinite center

The main purpose of this section is to show that a smooth fixed-point free $\mathbb{S}^1$–four-manifold whose fundamental group has infinite center is injective, and hence admits a fiber-sum decomposition. A key role is played by Lemma 5.2, whose proof requires the use of the Geometrization theorem in various forms.

We begin with the following observation.

Lemma 5.1 *Let Y be a 3–orbifold with a singular set consisting of a union of circles. Then there is a good 3–orbifold Y_0 such that Y and Y_0 have the same underlying space, and $\pi_1^{\mathrm{orb}}(Y_0) = \pi_1^{\mathrm{orb}}(Y)$.*

Proof Denote by $|Y|$ the underlying 3–manifold of Y and by ΣY the singular set of Y, consisting of components $\gamma_1, \dots, \gamma_n$. Then $\pi_1^{\mathrm{orb}}(Y)$ admits the following presentation

$$\pi_1^{\mathrm{orb}}(Y) = \pi_1(|Y| \setminus \Sigma Y)/N.$$

Here N is the normal subgroup generated by the elements $\mu_{\gamma_i}^{m_i}$, $i = 1, 2, \dots, n$, where μ_{γ_i} is the meridian around γ_i and m_i is the multiplicity of γ_i; see [5, Proposition 2.7].

With this understood, for any bad 2–suborbifold C in Y, one has the following two possibilities:

(i) There is exactly one γ_i such that $C \cap \gamma_i \neq \varnothing$.

(ii) There are γ_i and γ_j, with $i \neq j$ and $m_i \neq m_j$, such that $C \cap \gamma_i \neq \varnothing$ and $C \cap \gamma_j \neq \varnothing$.

In case (i), the existence of such a C implies that $\mu_{\gamma_i} = 1$ in $\pi_1(|Y| \setminus \Sigma Y)$, hence $\pi_1^{\mathrm{orb}}(Y)$ is unchanged after removing γ_i from ΣY. In the resulting 3–orbifold, C is no longer a bad 2–suborbifold.

In case (ii), let $m = \gcd(m_i, m_j)$. We change Y to a new 3–orbifold by replacing the multiplicities of γ_i, γ_j with m. (In case of $m = 1$, this simply means that γ_i, γ_j are both removed from ΣY.) Note that the existence of C implies that the normal subgroup generated by $\mu_{\gamma_i}^{m_i}$ and $\mu_{\gamma_j}^{m_j}$ is the same as that generated by $\mu_{\gamma_i}^{m}$ and $\mu_{\gamma_j}^{m}$. It follows that $\pi_1^{\mathrm{orb}}(Y)$ remains unchanged in this process. Since there are only finitely many singular circles and during the process either the number of singular circles is decreased or the multiplicity of a singular circle is decreased, this process must terminate in finitely many steps. At the end, we obtain a good 3–orbifold Y_0 such that $|Y_0| = |Y|$ and $\pi_1^{\mathrm{orb}}(Y_0) = \pi_1^{\mathrm{orb}}(Y)$. Hence the lemma. □

A more conceptual view which was suggested by the referee goes as follows: introducing a notion of complexity for 3–orbifolds, say by the sum of the multiplicities of the singular circles, then the orbifold Y_0 in Lemma 5.1 is characterized as the one with the minimal complexity among the 3–orbifolds which have the same underlying space and the same fundamental group of the orbifold Y.

In the following lemma, for the definition of $\pi_2^{\mathrm{orb}}(Y)$ we refer to [21; 22; 10].

Lemma 5.2 *Let Y be an orientable 3–orbifold, not necessarily good, with a singular set consisting of a union of circles. If $z(\pi_1^{\mathrm{orb}}(Y))$ is infinite, then Y is Seifert fibered. Moreover, if $\pi_2^{\mathrm{orb}}(Y) \neq 0$, then Y is the mapping torus of a periodic diffeomorphism of a 2–orbifold with finite fundamental group.*

Proof Let Y_0 be the good 3–orbifold associated to Y from Lemma 5.1, which is clearly orientable. Then there is an orientable 3–manifold Y' equipped with a finite group action of G, such that $Y_0 = Y'/G$; see [4; 31]. Since $\pi_1^{\mathrm{orb}}(Y_0) = \pi_1^{\mathrm{orb}}(Y)$, $z(\pi_1^{\mathrm{orb}}(Y_0))$ is also infinite, and consequently, $z(\pi_1(Y'))$, which contains $\pi_1(Y') \cap z(\pi_1^{\mathrm{orb}}(Y_0))$, is infinite. As an abelian subgroup of a 3–manifold group, $z(\pi_1(Y'))$ must contain an infinite cyclic subgroup H (see [24, Theorem 9.14]), which is clearly normal in $\pi_1(Y')$.

Consider first the case where $\pi_2(Y') = 0$. By work of Gabai (see [20] and, independently, Casson and Jungreis [9]), Y' is Seifert fibered, with H being generated by a regular fiber of the Seifert fibration. Since $H \subset z(\pi_1^{\mathrm{orb}}(Y_0))$, it must be invariant under the action of G. By a theorem of Meeks and Scott (see [32, Theorem 2.2]), G preserves the Seifert fibration on Y', which implies that Y_0 is Seifert fibered. Since we assume $\pi_2(Y')=0$, Y_0 does not contain any essential spherical 2–suborbifold. From the proof of Lemma 5.1, we see that Y contains no bad 2–suborbifold, and in this case, $Y=Y_0$. This proves that Y is Seifert fibered. Note that in this case,

$$\pi_2^{\mathrm{orb}}(Y)=\pi_2^{\mathrm{orb}}(Y_0)=\pi_2(Y')=0.$$

Suppose $\pi_2(Y')\neq 0$. Since $z(\pi_1(Y'))$ is nontrivial, Y' must be prime (here we use Lemma 2.1 and the resolution of the Poincaré conjecture [36]), and consequently, $Y'=\mathbb{S}^1\times\mathbb{S}^2$. Note that G must act on $Y'=\mathbb{S}^1\times\mathbb{S}^2$ homologically trivially because the fundamental group of $Y_0=Y'/G$ is infinite. By Lemma 2.5, $Y_0=Y'/G$ is the mapping torus of a periodic diffeomorphism of some spherical 2–orbifold; in particular, Y_0 is Seifert fibered. If Y is good, then $Y=Y_0$, and the lemma follows in this case. Note that in this case,

$$\pi_2^{\mathrm{orb}}(Y)=\pi_2^{\mathrm{orb}}(Y_0)=\pi_2(Y')\neq 0.$$

It remains to consider the case where Y is not good. Recall that in the proof of Lemma 5.1, Y_0 is obtained from Y by performing a sequence of operations where, in each, either a singular circle is removed or its multiplicity is decreased. Since Y_0 is the mapping torus of a periodic diffeomorphism f of some spherical 2–orbifold Σ, it follows easily that Σ is either $\mathbb{S}^2$ or a football. Moreover, if Σ is a football, f must be isotopic to the identity map, and therefore Y_0 is diffeomorphic to $\mathbb{S}^1\times\Sigma$. It follows readily that Y is the product of $\mathbb{S}^1$ with a bad 2–orbifold B. Note that in this case,

$$\pi_2^{\mathrm{orb}}(Y)=\pi_2^{\mathrm{orb}}(B)\neq 0,$$

since a bad 2–orbifold has nontrivial π_2^{orb}.

Suppose $\Sigma=\mathbb{S}^2$, and therefore $Y_0=\mathbb{S}^1\times\mathbb{S}^2$. Note that Y can have at most two singular circles. Assume first that Y has only one singular circle, which is denoted by γ. It suffices to show that $(|Y|,\gamma)$ and $(\mathbb{S}^1\times\mathbb{S}^2,\mathbb{S}^1\times\{\mathrm{pt}\})$ are diffeomorphic. To see this, let $W=Y\setminus Nd(\gamma)$ and let μ denote a meridian of γ. Then $\pi_1^{\mathrm{orb}}(Y)=\pi_1(W)/\langle\mu^m\rangle$ where m denotes the multiplicity of γ. Since μ bounds a disc in W, and $\pi_1^{\mathrm{orb}}(Y)=\pi_1(Y_0)=\mathbb{Z}$, it follows that $\pi_1(W)=\mathbb{Z}$. Cutting W open along the disc bounded by μ, we obtain a 3–manifold W_0 with $\partial W_0=\mathbb{S}^2$ and $\pi_1(W_0)$ trivial. By the Geometrization theorem, W_0 is a 3–ball, which implies easily that $(|Y|,\gamma)$ is diffeomorphic to $(\mathbb{S}^1\times\mathbb{S}^2,\mathbb{S}^1\times\{\mathrm{pt}\})$. This shows that Y is the product of $\mathbb{S}^1$ with a teardrop. Note that $\pi_2^{\mathrm{orb}}(Y)\neq 0$ as we argued before.

Finally, suppose Y has two components, denoted by γ_1, γ_2, which have multiplicities m_1, m_2 respectively. From the construction of Y_0 in Lemma 5.1, it follows easily that m_1, m_2 are relatively prime. With this understood, it suffices to show that $(|Y|, \gamma_1, \gamma_2)$ is diffeomorphic to $(\mathbb{S}^1 \times \mathbb{S}^2, \mathbb{S}^1 \times \{\text{pt}\}, \mathbb{S}^1 \times \{\text{pt}\})$. First, as we argued in the previous case, $(|Y|, \gamma_1)$ is diffeomorphic to $(\mathbb{S}^1 \times \mathbb{S}^2, \mathbb{S}^1 \times \{\text{pt}\})$, so that if we let $W = Y \setminus Nd(\gamma_1)$, then $|W| = \mathbb{S}^1 \times D^2$. It remains to show that $(|W|, \gamma_2)$ is diffeomorphic to $(\mathbb{S}^1 \times D^2, \mathbb{S}^1 \times \{\text{pt}\})$. To see this, note that the meridians μ_1 and μ_2 of γ_1 and γ_2, respectively, bound an annulus in $W \setminus Nd(\gamma_2)$. Consequently,

$$\mathbb{Z} = \pi_1^{\text{orb}}(Y) = \pi_1(W \setminus Nd(\gamma_2))/\langle \mu_1^{m_1}, \mu_2^{m_2} \rangle = \pi_1(W \setminus Nd(\gamma_2))/\langle \mu_2 \rangle,$$

which implies that the following sequence is short exact:

$$1 \to \mathbb{Z}_{m_2} \to \pi_1^{\text{orb}}(W) = \pi_1(W \setminus Nd(\gamma_2))/\langle \mu_2^{m_2} \rangle \to \mathbb{Z} \to 1.$$

Now if we cut W open along a copy of $\{\text{pt}\} \times D^2$ in $|W| = \mathbb{S}^1 \times D^2$, we obtain a 3–orbifold W_0 with $\partial W_0 = \mathbb{S}^2/\mathbb{Z}_{m_2}$. Moreover, it follows from the above short exact sequence that $\pi_1^{\text{orb}}(W_0) = \mathbb{Z}_{m_2}$. Then the Geometrization theorem implies that W_0 is discal, from which it follows that $(|Y|, \gamma_1, \gamma_2)$ is diffeomorphic to $(\mathbb{S}^1 \times \mathbb{S}^2, \mathbb{S}^1 \times \{\text{pt}\}, \mathbb{S}^1 \times \{\text{pt}\})$, and consequently, Y is the product of $\mathbb{S}^1$ with a bad 2–orbifold. Moreover, $\pi_2^{\text{orb}}(Y) \neq 0$. This finishes the proof of the lemma. □

Proof of Theorem 1.4 Let $\pi\colon X \to Y$ be the orbit map of the fixed-point free $\mathbb{S}^1$–action. Suppose the $\mathbb{S}^1$–action is not injective. Then the homotopy class of a regular fiber of π is finite, and since $z(\pi_1(X))$ is infinite, the image of $z(\pi_1(X))$ under $\pi_*\colon \pi_1(X) \to \pi_1^{\text{orb}}(Y)$, clearly contained in $z(\pi_1^{\text{orb}}(Y))$, must also be infinite. By Lemma 5.2, either Y is irreducible, or Y is the mapping torus of a periodic diffeomorphism of a 2–orbifold with finite fundamental group. Since we assume that the homotopy class of a regular fiber of π is finite, Y can not be irreducible. Then it follows easily that X is the mapping torus of a periodic diffeomorphism of some elliptic 3–manifold.

To see that X admits a fiber-sum decomposition, it suffices to consider the case where the $\mathbb{S}^1$–action is injective. We note first that the fact that the homotopy class of a regular fiber of π has infinite order implies that the orbit space Y of the $\mathbb{S}^1$–action does not contain any bad 2–suborbifolds. In other words, Y must be good. By Lemma 2.4, Y admits a reduced spherical decomposition. More precisely, there is a system of finitely many spherical 2–suborbifolds $\Sigma_j \subset Y$ such that, after capping off the boundary of each component of $Y \setminus \bigcup_j \Sigma_j$, one obtains a collection of 3–orbifolds Y_i where each Y_i is irreducible. Furthermore, each Σ_j must be either an ordinary 2–sphere or a football, and the preimage $N_j \equiv \pi^{-1}(\Sigma_j)$ must be diffeomorphic to $\mathbb{S}^1 \times \mathbb{S}^2$, because the homotopy class of a regular fiber of π has infinite order. Finally, observe that the restriction of π on each N_j may be uniquely extended to a Seifert-type $\mathbb{S}^1$–fibration on $\mathbb{S}^1 \times B^3$ so that, correspondingly, we obtain the irreducible $\mathbb{S}^1$–four-manifolds X_i and the orbit maps $\pi_i\colon X_i \to Y_i$. It follows easily that X is fiber-sum-decomposed into X_i along N_j. We remark that the requirement that the spherical decomposition of Y be reduced ensures that Definition 1.3(iv) is satisfied. This finishes off the proof of Theorem 1.4. □

Proof of Corollary 1.7 By Theorem 1.4, it suffices to consider the case where the $\mathbb{S}^1$–action is injective. Let $\pi\colon X \to Y$ be the corresponding orbit map. We observe that Y does not contain any bad 2–suborbifolds, hence there exist a 3–manifold $\widetilde{Y}$ and a finite group G such that $Y = \widetilde{Y}/G$; see [4; 31]. On the other hand, by the homotopy exact sequence associated to $\pi\colon X \to Y$ (see Haefliger [22]), it follows easily that $\pi_*\colon \pi_2(X) \to \pi_2^{\mathrm{orb}}(Y)$ is an isomorphism. Let $\tilde{\pi}\colon \widetilde{X} \to \widetilde{Y}$ be the pull-back fibration via the projection $\widetilde{Y} \to Y$. Then $\widetilde{X}$ is a finite regular cover of X. It suffices to show that there exist no embedded 2–spheres with odd self-intersection in $\widetilde{X}$.

Suppose to the contrary, there is an embedded 2–sphere C in $\widetilde{X}$ with $C^2 \equiv 1 \pmod 2$. Consider the projection of C into $\widetilde{Y}$ under $\tilde{\pi}$. Clearly $[C] \in \pi_2(\widetilde{X})$ is nonzero. On the other hand, $\pi_*\colon \pi_2(X) \to \pi_2^{\mathrm{orb}}(Y)$ is an isomorphism, so that $\tilde{\pi}_*\colon \pi_2(\widetilde{X}) \to \pi_2(\widetilde{Y})$ is also an isomorphism. Consequently, $\tilde{\pi}|_C\colon \mathbb{S}^2 \to \widetilde{Y}$ is homotopically nontrivial. By the Sphere theorem (see [24, Theorem 4.11]), there is an embedded 2–sphere Σ in a neighborhood of $\tilde{\pi}(C)$, whose class is clearly homologous to $\tilde{\pi}_*[C]$. Observe that the Euler class of $\tilde{\pi}\colon \widetilde{X} \to \widetilde{Y}$ evaluates to 0 on Σ. This is because the pull-back of the Euler class of $\tilde{\pi}$ to $\widetilde{X}$ is zero so that the Euler class of $\tilde{\pi}$ evaluates trivially on the class of $\tilde{\pi}(C)$. This implies that the restriction of $\tilde{\pi}$ to Σ is trivial, and in particular, Σ has a section Σ' in $\widetilde{X}$. Consequently, we obtain an equation of homology classes

$$C = \Sigma' + \sum_i T_i,$$

where $T_i = \tilde{\pi}^{-1}(\gamma_i)$ for some loops $\gamma_i \subset \widetilde{Y}$; see [2, Theorem 9]. Since Σ' and all T_i have self-intersection 0, this implies $C^2 \equiv 0 \pmod 2$, which is a contradiction. This finishes the proof of Corollary 1.7. □

6 Theorems 1.1 and 1.2

This section is devoted to the proofs of Theorems 1.1 and 1.2. We remark that while Theorem 1.1 follows readily from Theorems 1.5 and 1.6, the proof of Theorem 1.2 requires some additional care in the case when each irreducible $\mathbb{S}^1$–four-manifold in the fiber-sum decomposition is a mapping torus of a periodic diffeomorphism of a lens space. Furthermore, the case when the fundamental group of the 4–manifold is isomorphic to the fundamental group of a Klein bottle needs to be dealt with separately. In all these considerations, the following lemma describing certain isotopies of periodic diffeomorphisms of $\mathbb{S}^3$ or a lens space plays a key role.

Let $Y = \mathbb{S}^3/G$ where G is a cyclic subgroup of SO(4) of order n given by

$$\lambda \cdot (z_1, z_2) = (\lambda^p z_1, \lambda^q z_2),$$

where $\lambda = \exp(2\pi i/n)$ is a n^{th} root of unity and $\gcd(n, p, q) = 1$. Set $u = \gcd(n, p)$, $v = \gcd(n, q)$. Then $\gcd(u, v) = 1$ so that uv is a divisor of $n = |\pi_1^{\text{orb}}(Y)|$, and Y has at most two singular circles of multiplicities u and v, given by $z_2 = 0$ and $z_1 = 0$, respectively.

Suppose H is a subgroup of G of order $\hat{n}$ generated by $\lambda^{n/\hat{n}}$, which acts freely on $\mathbb{S}^3$. Note that this condition is equivalent to $\gcd(\hat{n}, p) = 1$ and $\gcd(\hat{n}, q) = 1$; in particular, $\hat{n}, u, v$ are pairwise coprime so that $\hat{n} \le n/uv$. We set $\hat{Y} = \mathbb{S}^3/H$, which is either $\mathbb{S}^3$ or a lens space. With this understood, let $f\colon \hat{Y} \to \hat{Y}$ be a periodic diffeomorphism such that $Y = \hat{Y}/\langle f \rangle$.

Lemma 6.1 *For any singular circle γ of Y, say the one defined by $z_2 = 0$ which has multiplicity u, we let $\hat{\gamma}$ be the preimage of γ in $\hat{Y}$. Then there exist a periodic diffeomorphism $f'\colon \hat{Y} \to \hat{Y}$ and an isotopy $f_t\colon \hat{Y} \to \hat{Y}$ between f and f', such that:*

- *The restriction of f_t on $\hat{\gamma}$ is independent of t (in particular, $f = f'$ on $\hat{\gamma}$).*
- *f' is free on $\hat{\gamma}$ so that the image of $\hat{\gamma}$ in $Y' = \hat{Y}/\langle f' \rangle$ is not a singular circle.*
- *When $\hat{Y} = \mathbb{S}^3$, one can arrange f' such that Y' is the lens space $L(n/u, 1)$.*

Proof We first consider the case where $\hat{n} > 1$. Set $p' = p/u$, let u' be the unique integer satisfying $uu' \equiv 1 \pmod{\hat{n}}$ and $0 < u' < \hat{n}$, and consider the following action of a cyclic subgroup $G' \subset \mathrm{SO}(4)$ of order $n' = n/u$, given by

$$\delta \cdot (z_1, z_2) = (\delta^{p'} z_1, \delta^{qu'} z_2),$$

where $\delta = \exp(2\pi i/n')$ is an n'^{th} root of unity. Note that since $\lambda^{n/\hat{n}} \cdot (z_1, z_2) = \delta^{n'u/\hat{n}} \cdot (z_1, z_2)$, $H = \langle \lambda^{n/\hat{n}} \rangle = \langle \delta^{n'/\hat{n}} \rangle$ is also a subgroup of G'.

There is a k with $\gcd(n, k) = 1$ such that $f\colon \hat{Y} \to \hat{Y}$ is represented by the H–equivariant map $F\colon (z_1, z_2) \mapsto \lambda^k \cdot (z_1, z_2)$. We shall consider the H–equivariant map $F'\colon (z_1, z_2) \mapsto \delta^k \cdot (z_1, z_2)$, which has the following properties: (i) $F = F'$ on $\{(z_1, 0) \mid |z_1| = 1\}$, (ii) there is an H–equivariant isotopy F_t between F and F' which is constant in t on $\{(z_1, 0) \mid |z_1| = 1\}$. For instance, $F_t\colon (z_1, z_2) \mapsto (\delta^{kp'} z_1, \theta_t z_2)$, where $\theta_t = \exp(2tkqu'\pi i/n' + 2(1-t)kq\pi i/n)$, $0 \le t \le 1$. Let f', f_t be the descendant of F', F_t to $\widehat{Y}$ respectively. Then clearly f_t is an isotopy between f and f' that is constant on $\hat{\gamma} = \{(z_1, 0) \mid |z_1| = 1\}/H$, and f' is free on $\hat{\gamma}$ so that the image of $\hat{\gamma}$ in $Y' = \widehat{Y}/\langle f' \rangle$ is not a singular circle. This finishes the proof for the case where $\hat{n} > 1$.

Now suppose $\hat{n} = 1$, which means that H is trivial. Then instead, we consider the following action of a cyclic subgroup $G' \subset \mathrm{SO}(4)$ of order $n' = n/u$, given by

$$\delta \cdot (z_1, z_2) = (\delta^{p'} z_1, \delta^{p'} z_2).$$

The rest of the argument is the same, with $H \subset G'$ trivially true. Note that in this case, $Y' = \mathbb{S}^3/\langle f' \rangle = \mathbb{S}^3/G' = L(n/u, 1)$. This finishes the proof of Lemma 6.1. □

As an immediate corollary of Lemma 6.1, we obtain the following classification of fixed-point free smooth $\mathbb{S}^1$–four-manifolds whose fundamental group is isomorphic to the fundamental group of a Klein bottle.

Theorem 6.2 *Let X be a fixed-point free smooth $\mathbb{S}^1$–four-manifold such that $\pi_1(X)$ is isomorphic to the fundamental group of a Klein bottle. Then X is diffeomorphic to the quotient of $T^2 \times \mathbb{S}^2$ by the involution τ, where*

$$\tau\colon (x, y, z) \mapsto (-x, \bar{y}, -z) \quad \text{for } x, y \in \mathbb{S}^1 \subset \mathbb{C} \text{ and } z \in \mathbb{S}^2 \subset \mathbb{R}^3.$$

Proof As $\pi_1(X)$ is isomorphic to the fundamental group of a Klein bottle, it has the following presentation: $\pi_1(X) = \{c, t \mid tct^{-1} = c^{-1}\}$. Clearly the center $z(\pi_1(X))$ is the infinite cyclic subgroup generated by t^2. By Theorem 1.4, the $\mathbb{S}^1$–action is injective. We let $\pi\colon X \to Y$ be the corresponding orbit map. Let $m > 0$ be the multiplicity of the homotopy class of a regular fiber of π in $z(\pi_1(X))$. Then

$$\pi_1^{\mathrm{orb}}(Y) = \{c, t \mid tct^{-1} = c^{-1}, t^{2m} = 1\}.$$

Let $\widehat{Y}$ be the regular covering of Y corresponding to the infinite normal cyclic subgroup generated by c. Since $\widehat{Y}$ is good and its fundamental group is torsion-free, $\widehat{Y}$ must be a 3–manifold, and clearly, $\widehat{Y} = \mathbb{S}^1 \times \mathbb{S}^2$. The corresponding group of deck transformations on $\widehat{Y}$ is cyclic of order $2m$ and is generated by t, which sends $c \in \pi_1(\widehat{Y})$ to $-c$. By Lemma 2.5, Y is diffeomorphic to either $\mathbb{RP}^3_m \#_m \mathbb{RP}^3_m$, $\mathbb{RP}^3_m \#_m \widetilde{\mathbb{RP}^3}_m$, or $\widetilde{\mathbb{RP}^3}_m \#_m \widetilde{\mathbb{RP}^3}_m$. Consequently, X is fiber-sum-decomposed into X_1 and X_2 along N, with $\pi_i\colon X_i \to Y_i$, $i = 1, 2$, where each of Y_1 and Y_2 is either $\mathbb{RP}^3_m$ or $\widetilde{\mathbb{RP}^3}_m$, and $\pi\colon N \to \Sigma$ where Σ intersects the singular circle of multiplicity m in Y.

There are $\widehat{Y}_i$ and periodic diffeomorphisms f_i such that $Y_i = \widehat{Y}_i/\langle f_i \rangle$ and X_i is the mapping torus of f_i, where $i = 1, 2$. We apply Lemma 6.1 to Y_i, $\widehat{Y}_i$, and f_i, with γ

being the singular circle of multiplicity m. We claim that in either case, ie $Y_i = \mathbb{RP}^3_m$ or $\widetilde{\mathbb{RP}^3}_m$, $\widehat{Y}_i$ must be $\mathbb{S}^3$, ie $\hat{n} = 1$. For the case where $Y_i = \widetilde{\mathbb{RP}^3}_m$, it follows from the fact that $\widetilde{\mathbb{RP}^3}_m$ has two singular circles with multiplicities 2 and m, respectively, so that $\hat{n} \leq n/uv = 2m/2m = 1$. For the case where $Y_i = \mathbb{RP}^3_m$, a similar argument shows that $\hat{n} \leq 2$. Continuing using the notations in Lemma 6.1, we have, in this case, $p = m$, $q = 1$, $n' = 2$, and f_i' is given by multiplication by δ. If $\hat{n} = 2$ and, therefore, $\widehat{Y}_i = \mathbb{RP}^3$, f_i' is the identity map on $\widehat{Y}_i$. Consequently, as the mapping torus of f_i', X_i is diffeomorphic to $\mathbb{S}^1 \times \widehat{Y}_i$, and $\pi_1(X)$ contains a torsion subgroup of $\mathbb{Z}_2$ coming from $\pi_1(\widehat{Y}_i)$. But this contradicts the fact that $\pi_1(X)$ is isomorphic to the π_1 of a Klein bottle. Hence $\widehat{Y}_i = \mathbb{S}^3$ in both cases. We conclude by observing that each X_i is the mapping torus of the antipodal map on $\mathbb{S}^3$. We denote by $\pi_i'\colon X_i \to \mathbb{RP}^3$ the corresponding Seifert-type $\mathbb{S}^1$–fibration.

Finally, by the property in Lemma 6.1 that the restriction of f_t on $\hat{\gamma}$ is independent of t, it is easily seen that the Seifert-type $\mathbb{S}^1$–fibrations $\pi_i\colon X_i \to Y_i$ and $\pi_i'\colon X_i \to \mathbb{RP}^3$ are identical on the mapping torus of $f = f'\colon \hat{\gamma} \to \hat{\gamma}$. It follows easily that X is also fiber-sum-decomposed into X_1 and X_2 along N, with $\pi_i'\colon X_i \to \mathbb{RP}^3$ on each factor X_i. Theorem 6.2 follows easily. □

Theorem 1.1 follows immediately from the following theorem.

Theorem 6.3 *Let G be a finitely presented group such that:*

(i) *rank $z(G) = 1$.*

(ii) *G is single-ended and is not isomorphic to the π_1 of a Klein bottle.*

(iii) *Any canonical JSJ decomposition of G contains a vertex subgroup which is not isomorphic to an HNN extension of a finite cyclic group.*

Let S_G be the set of equivariant diffeomorphism classes of orientable, fixed-point free, smooth $\mathbb{S}^1$–four-manifolds X such that $\pi_1(X) = G$. Then there exists a constant $C > 0$, depending only on G, such that $\#S_G < C$.

Proof Let X be an orientable, fixed-point free, smooth $\mathbb{S}^1$–four-manifold such that $\pi_1(X) = G$. Since G is single-ended, it follows easily from Theorem 1.4 that any fixed-point free $\mathbb{S}^1$–action on X must be injective. Thus, any fixed-point free $\mathbb{S}^1$–action on X is associated with a canonical fiber-sum decomposition. Suppose X is decomposed into factors X_i along N_j. For convenience we shall fix an orientation of X, which is the one induced from the fiber-sum decomposition. Then the following data completely determine the oriented equivariant diffeomorphism class of X:

(i) The isomorphism class of the underlying graph of Λ.

(ii) For each pair of i, j such that $N_j \subset X_i$, the fiber-preserving isotopy class of embeddings of N_j in X_i for each fixed oriented, fiber-preserving diffeomorphism class of X_i.

(iii) For each i, the oriented, fiber-preserving diffeomorphism class of X_i.

These data are subject to the following constraints: the cardinalities of $\{X_i\}$ and $\{N_j\}$ and the conjugacy classes of subgroups $\pi_1(X_i)$ and $\pi_1(N_j)$ in G are determined by G; see Proposition 3.5. With this understood, our aim is to show that the number of objects in each of (i), (ii), and (iii) is bounded by a constant depending only on G.

The number of objects in (i) is clearly bounded by a constant depending only on G, since the cardinalities of $\{X_i\}$ and $\{N_j\}$ are fixed by G. For the objects in (ii) and (iii), where an index i is being fixed, we shall discuss separately according to the following three cases, (a) rank $z(\pi_1(X_i)) > 1$, (b) $\pi_1(X_i)$ is single-ended with rank $z(\pi_1(X_i)) = 1$, (c) $\pi_1(X_i)$ is double-ended.

Note that the number of objects in (ii) is bounded by the number of singular circles of Y_i plus one, so we need to show that, for each i, the number of singular circles of Y_i is bounded by a constant depending only on G. With this understood, consider case (a) where X_i is a Seifert-type T^2–fibration over a 2–orbifold B_i with infinite π_1^{orb}. As shown in the proof of Theorem 1.6(1), B_i is uniquely determined by $\pi_1(X_i)$, hence by G. On the other hand, Y_i is Seifert fibered over B_i, so that the number of singular circles of Y_i is bounded by the number of singular points of B_i, which depends only on G. In case (b), Y_i is uniquely determined by $\pi_1(X_i)$ as shown in the proof of Theorem 1.6(1), hence the number of singular circles of Y_i depends only on G. In case (c), $\pi_1^{\mathrm{orb}}(Y_i)$ is finite. The Geometrization theorem implies that Y_i is spherical. Since the singular set of Y_i consists of a union of embedded circles, the work of Dunbar in [15] shows that $Y_i = \mathbb{S}^3/G_i$, where G_i is a subgroup of SO(4) which preserves a Hopf fibration. It follows easily that the number of singular components of Y_i is universally bounded (say by 4). This shows that the number of objects in (ii) is bounded by a constant depending only on G.

Finally, we examine the boundedness of the number of objects in (iii). In case (a), the diffeomorphism class of X_i is uniquely determined by $\pi_1(X_i)$ (see Theorem 1.6(1)); however, the Seifert-type $\mathbb{S}^1$–fibration $\pi_i\colon X_i \to Y_i$ has infinitely many choices, one for each primitive element of $z(\pi_1(X_i))$. With this understood, note that, by assumption, $z(G)$ has rank 1, so there is only one possible choice for the regular fiber class of π_i in $z(\pi_1(X_i))$. This shows that $\pi_i\colon X_i \to Y_i$ is uniquely determined by G in this case. In case (b), both X_i and π_i are uniquely determined by $\pi_1(X_i)$, as shown in the proof of Theorem 1.6(1), and hence are also determined by G.

Lastly, we consider case (c). By Theorem 1.6(2), X_i is the mapping torus of a periodic diffeomorphism $f_i\colon \widehat{Y}_i \to \widehat{Y}_i$ of an elliptic 3–manifold. It follows from the proof of Theorem 1.6(2) that the number of diffeomorphism classes of X_i is bounded by a constant depending only on $\pi_1(X_i)$. In order to bound the number of fiber-preserving diffeomorphism classes, we shall employ the rigidity theorem of injective Seifert fibered space constructions as in the proof of Theorem 1.6(1), with $k=1$ and $W=\mathbb{S}^3$. With this understood, it is clear that it suffices to show that the number of possible short exact sequences $1 \to \Gamma \to \pi \to Q \to 1$ involved in the argument is bounded by a constant depending only on G. Equivalently, we will show that the multiplicity of the homotopy class of a regular fiber of π_i in $z(\pi_1(X_i))$ is bounded by a constant depending only on G.

Denote by h the homotopy class of a regular fiber. Since the conjugacy classes of the subgroups $\pi_1(X_i)$ in G depend only on G, it follows easily that it suffices to bound the multiplicity of h in $z(\pi_1(X))$. With this understood, we observe that since for each j, $z(\pi_1(X)) \subset \pi_1(N_j)$, the multiplicity of h in $z(\pi_1(X))$ is bounded by the multiplicity of h in $\pi_1(N_j)$ for every j, which equals 1 if Σ_j is an ordinary 2–sphere, and equals the multiplicity of the singular circle of Y that Σ_j intersects otherwise. In particular, if one of the Σ_j is an ordinary 2–sphere, or one of the Y_i has infinite fundamental group, we are done for (iii). (Note that, since G is single-ended, there is at least one N_j if case (c) is valid.)

Suppose $\pi_1^{\mathrm{orb}}(Y_i)$ is finite for each i and Σ_j is a football for each j. Again, since the singular set of Y_i consists of a union of embedded circles, the work of Dunbar in [15] shows that $Y_i = \mathbb{S}^3/G_i$ for a finite subgroup G_i of SO(4) that preserves a Hopf fibration. It follows that X_i is the mapping torus of a periodic diffeomorphism $f_i\colon \widehat{Y}_i \to \widehat{Y}_i$, where $\widehat{Y}_i$ has a Seifert fibration induced from the Hopf fibration and f_i preserves the Seifert fibration on $\widehat{Y}_i$. By the assumption (iii), there is a Y_i such that $\pi_1(\widehat{Y}_i)$ is nonabelian. With the following lemma (Lemma 6.4), we finish the proof by observing that $\pi_1(\widehat{Y}_i)$ is completely determined by $\pi_1(X_i)$, which depends only on G. □

Lemma 6.4 *Let $\widehat{Y}$ be an elliptic 3–manifold with nonabelian fundamental group, and let $\pi\colon \widehat{Y} \to B$ be the unique Seifert fibration on $\widehat{Y}$. Suppose $f\colon \widehat{Y} \to \widehat{Y}$ is an orientation-preserving periodic diffeomorphism that preserves π. Then the multiplicity of any singular circle of the 3–orbifold $Y = \widehat{Y}/\langle f\rangle$ is bounded by a constant depending only on the multiplicities of the singular points of B.*

Proof For any singular circle γ in Y, the multiplicity of γ equals the order of its isotropy subgroup. Let f_γ be a generator of the isotropy subgroup, which is given by f^k for some k. Since $f\colon \widehat{Y} \to \widehat{Y}$ preserves $\pi\colon \widehat{Y} \to B$, so does f_γ, and there is an induced periodic diffeomorphism $\bar{f}_\gamma\colon B \to B$ of the 2–orbifold B.

Since $\pi_1(\widehat{Y})$ is nonabelian, B is a turnover with multiplicities $(2,2,n)$, $(2,3,3)$, $(2,3,4)$, or $(2,3,5)$. We shall discuss according to the following cases: (i) $\bar{f}_\gamma$ is trivial, (ii) $\bar{f}_\gamma$ is nontrivial.

Suppose $\bar{f}_\gamma$ is trivial. Then f_γ acts as a rotation on each fiber of $\pi\colon \widehat{Y} \to B$. It follows easily that γ must be an exceptional fiber of π, and the order of f_γ is a divisor of the multiplicity of the singular point $\pi(\gamma) \in B$.

Suppose $\bar{f}_\gamma$ is nontrivial. Then there are two possibilities: (a) $\bar{f}_\gamma$ is orientation-preserving, (b) $\bar{f}_\gamma$ is orientation-reversing. In case (a), the order of $\bar{f}_\gamma$ is either 2 or 3, and $\bar{f}_\gamma$ has two isolated fixed-points. Moreover, γ must be the fiber over one of the fixed-points of $\bar{f}_\gamma$. It follows easily that the multiplicity of γ equals the order of $\bar{f}_\gamma$, which is at most 3. In case (b), $\bar{f}_\gamma$ must be a reflection over a great circle in B because $\bar{f}_\gamma$ has a nonempty fixed-point (which contains $\pi(\gamma)$, for instance). Since f is orientation-preserving, f_γ must be a reflection on the fibers over the great circle fixed under $\bar{f}_\gamma$. It follows that the multiplicity of γ equals 2 in this case. □

Proof of Theorem 1.2 This result follows from Theorem 6.3 except in the following cases:

(a) rank $z(G) > 1$.

(b) G is double-ended.

(c) G is isomorphic to the π_1 of a Klein bottle.

(d) None of the above is true, and moreover, every vertex subgroup of a canonical JSJ decomposition of G is an HNN extension of a finite cyclic group.

Cases (a), (c) are settled with the help of Theorems 4.3 and 6.2. Case (b) is settled by Theorem 1.4, Lemma 4.2, and Theorem 1.6. (Note that in case (b) where G is double-ended, we appeal to Theorem 1.6(2), where we observe that when X is a mapping torus of a periodic diffeomorphism of a lens space, the number of possible lens spaces is bounded by a constant depending only on $\pi_1(X)$.)

For case (d), we shall continue with the proof of Theorem 6.3, where we are left with the situation that $\pi_1(\widehat{Y}_i)$ is finite cyclic for each i and Σ_j is a football for each j. Recall that $Y_i = \widehat{Y}_i/\langle f_i\rangle$ for some periodic diffeomorphism f_i. Moreover, there is a Seifert fibration $\mathrm{pr}_i\colon \widehat{Y}_i \to B_i$ which is induced from a Hopf fibration and is preserved under f_i.

We shall analyze the multiplicities of the singular circles in Y_i. To this end, let γ be a singular circle and f_γ be a generator of its isotropy subgroup. Denote by $\bar{f}_\gamma\colon B_i \to B_i$ the induced map. If $\bar{f}_\gamma$ is orientation-reversing, then as we showed in the proof of Lemma 6.4, the multiplicity of γ is 2. If $\bar{f}_\gamma$ is orientation-preserving and switches the two singular points of B_i, then the multiplicity of γ is also 2, as we argued in the proof of Lemma 6.4. In the remaining cases where $\bar{f}_\gamma$ is either trivial or fixes the two singular points of B_i, or B_i has no singular points at all, the multiplicity of γ may not be bounded by a constant depending only on G, and we need to deal with it differently.

Note that in either of the remaining cases, $Y_i = \mathbb{S}^3/G_i$ for a finite subgroup G_i of $\mathrm{SO}(4)$, which is given by

$$\lambda \cdot (z_1, z_2) = (\lambda^{p_i} z_1, \lambda^{q_i} z_2),$$

where $\lambda = \exp(2\pi i/n_i)$ is a n_i^{th} root of unity and $\gcd(n_i, p_i, q_i) = 1$. Set

$$u_i = \gcd(n_i, p_i) \quad \text{and} \quad v_i = \gcd(n_i, q_i).$$

Then Y_i has at most two singular circles of multiplicities u_i and v_i, respectively. Furthermore, if Σ_j intersects the singular circle of multiplicity u_i (resp. v_i), the index of $\pi_1(N_j)$ in $\pi_1(X_i)$ is n_i/u_i (resp. n_i/v_i). Consequently, if both singular circles of Y_i are intersected by Σ_j for some j, then $u_i \le n_i/v_i$ and $v_i \le n_i/u_i$ are both bounded by a constant depending only on G; see Proposition 3.5. Clearly, we are done for (iii) in the proof of Theorem 6.3 if there exists a Y_i for which such a situation occurs.

We are left to examine the case where, for each i, there is exactly one singular circle of Y_i which is intersected by Σ_j for some j. In this case, we shall apply Lemma 6.1 and change the Seifert-type $\mathbb{S}^1$–fibrations $\pi_i\colon X_i \to Y_i$ in the fiber-sum decomposition of X to $\pi_i'\colon X_i \to Y_i'$. Note that with the new fibrations π_i', each N_j is fibered over an ordinary 2–sphere. Consequently, up to suitable modifications of the Seifert-type $\mathbb{S}^1$–fibrations, the number of objects in (iii) in the proof of Theorem 6.3 is bounded by a constant depending only on G, from which Theorem 1.2 follows. □

Acknowledgments The author is in debt to Slawomir Kwasik for many helpful conversations throughout the course of the work. This work was partly done during the author's visit at the Institute for Advanced Study, and he wishes to thank the institute for its hospitality and financial support. Finally, the author thanks the referee for a very careful reading of the article and for the many suggestions which have improved considerably the exposition of the paper. The author is partially supported by NSF grant DMS-1065784.

References

[1] **K Asano**, *Homeomorphisms of prism manifolds*, Yokohama Math. J. 26 (1978) 19–25 MR524878

[2] **S J Baldridge**, *Seiberg–Witten invariants, orbifolds, and circle actions*, Trans. Amer. Math. Soc. 355 (2003) 1669–1697 MR1946410

[3] **S Baldridge**, *Seiberg–Witten vanishing theorem for S^1–manifolds with fixed points*, Pacific J. Math. 217 (2004) 1–10 MR2105762

[4] **M Boileau**, **B Leeb**, **J Porti**, *Geometrization of* 3*–dimensional orbifolds*, Ann. of Math. 162 (2005) 195–290 MR2178962

[5] **M Boileau**, **S Maillot**, **J Porti**, *Three-dimensional orbifolds and their geometric structures*, Panoramas et Synthèses 15, Soc. Math. France, Paris (2003) MR2060653

[6] **M Boileau**, **J-P Otal**, *Scindements de Heegaard et groupe des homéotopies des petites variétés de Seifert*, Invent. Math. 106 (1991) 85–107 MR1123375

[7] **F Bonahon**, *Difféotopies des espaces lenticulaires*, Topology 22 (1983) 305–314 MR710104

[8] **S Cappell**, **S Weinberger**, **M Yan**, *Closed aspherical manifolds with center*, J. Topol. 6 (2013) 1009–1018 MR3145148

[9] **A Casson**, **D Jungreis**, *Convergence groups and Seifert fibered* 3*–manifolds*, Invent. Math. 118 (1994) 441–456 MR1296353

[10] **W Chen**, *On a notion of maps between orbifolds, II: Homotopy and CW–complex*, Commun. Contemp. Math. 8 (2006) 763–821 MR2274941

[11] **W Chen**, *Seiberg–Witten invariants of* 3*–orbifolds and non-Kähler surfaces*, J. Gökova Geom. Topol. GGT 6 (2012) 1–27 MR3011643

[12] **W Chen**, *Hurwitz-type bound, knot surgery, and smooth $\mathbb{S}^1$–four-manifolds*, Math. Z. 276 (2014) 267–279 MR3150204

[13] **W Dicks**, **M J Dunwoody**, *Groups acting on graphs*, Cambridge Studies in Advanced Mathematics 17, Cambridge Univ. Press (1989) MR1001965

[14] **J Dinkelbach**, **B Leeb**, *Equivariant Ricci flow with surgery and applications to finite group actions on geometric 3–manifolds*, Geom. Topol. 13 (2009) 1129–1173 MR2491658

[15] **W D Dunbar**, *Nonfibering spherical* 3*–orbifolds*, Trans. Amer. Math. Soc. 341 (1994) 121–142 MR1118824

[16] **R Fintushel**, *Locally smooth circle actions on homotopy* 4*–spheres*, Duke Math. J. 43 (1976) 63–70 MR0394716

[17] **R Fintushel**, *Circle actions on simply connected* 4*–manifolds*, Trans. Amer. Math. Soc. 230 (1977) 147–171 MR0458456

[18] **R Fintushel**, *Classification of circle actions on 4–manifolds*, Trans. Amer. Math. Soc. 242 (1978) 377–390 MR496815

[19] **R Fintushel**, **R J Stern**, *Immersed spheres in 4–manifolds and the immersed Thom conjecture*, Turkish J. Math. 19 (1995) 145–157 MR1349567

[20] **D Gabai**, *Convergence groups are Fuchsian groups*, Ann. of Math. 136 (1992) 447–510 MR1189862

[21] **A Haefliger**, *Homotopy and integrability*, from: "Manifolds", (N H Kuiper, editor), Lecture Notes in Mathematics 197, Springer, Berlin (1971) 133–163 MR0285027

[22] **A Haefliger**, *Groupoïdes d'holonomie et classifiants*, from: "Transversal structure of foliations", 116 (1984) 70–97 MR755163

[23] **A Hatcher**, *On the diffeomorphism group of $S^1 \times S^2$*, Proc. Amer. Math. Soc. 83 (1981) 427–430 MR624946

[24] **J Hempel**, *3–manifolds*, Ann. of Math. Studies 86, Princeton Univ. Press (1976) MR0415619

[25] **J A Hillman**, *Four-manifolds, geometries and knots*, Geometry & Topology Monographs 5 (2002) MR1943724

[26] **C Hodgson**, **J H Rubinstein**, *Involutions and isotopies of lens spaces*, from: "Knot theory and manifolds", (D Rolfsen, editor), Lecture Notes in Math. 1144, Springer, Berlin (1985) 60–96 MR823282

[27] **B Kleiner**, **J Lott**, *Notes on Perelman's papers* `arXiv:math.DG/0605667`

[28] **K B Lee**, **F Raymond**, *The role of Seifert fiber spaces in transformation groups*, from: "Group actions on manifolds", (R Schultz, editor), Contemp. Math. 36, Amer. Math. Soc. (1985) 367–425 MR780974

[29] **A M Macbeath**, *Geometrical realization of isomorphisms between plane groups*, Bull. Amer. Math. Soc. 71 (1965) 629–630 MR0177110

[30] **W Magnus**, **A Karrass**, **D Solitar**, *Combinatorial group theory: presentations of groups in terms of generators and relations*, Interscience, New York (1966) MR0207802

[31] **D McCullough**, **A Miller**, *Manifold covers of 3–orbifolds with geometric pieces*, Topology Appl. 31 (1989) 169–185 MR994409

[32] **W H Meeks, III**, **P Scott**, *Finite group actions on 3–manifolds*, Invent. Math. 86 (1986) 287–346 MR856847

[33] **W H Meeks, III**, **S T Yau**, *Topology of three-dimensional manifolds and the embedding problems in minimal surface theory*, Ann. of Math. 112 (1980) 441–484 MR595203

[34] **G Meng**, **C H Taubes**, *$\underline{\mathrm{SW}}$ = Milnor torsion*, Math. Res. Lett. 3 (1996) 661–674 MR1418579

[35] **P S Pao**, *Nonlinear circle actions on the 4–sphere and twisting spun knots*, Topology 17 (1978) 291–296 MR508892

[36] **G Perelman**, *The entropy formula for the Ricci flow and its geometric applications* arXiv:math/0211159

[37] **C Petronio**, *Spherical splitting of 3–orbifolds*, Math. Proc. Cambridge Philos. Soc. 142 (2007) 269–287 MR2314601

[38] **E Rips**, **Z Sela**, *Cyclic splittings of finitely presented groups and the canonical JSJ decomposition*, Ann. of Math. 146 (1997) 53–109 MR1469317

[39] **J H Rubinstein**, *On 3–manifolds that have finite fundamental group and contain Klein bottles*, Trans. Amer. Math. Soc. 251 (1979) 129–137 MR531972

[40] **J H Rubinstein**, **J S Birman**, *One-sided Heegaard splittings and homeotopy groups of some 3–manifolds*, Proc. London Math. Soc. 49 (1984) 517–536 MR759302

[41] **P Scott**, *The geometries of 3–manifolds*, Bull. London Math. Soc. 15 (1983) 401–487 MR705527

[42] **P Scott**, **T Wall**, *Topological methods in group theory*, from: "Homological group theory", (C T C Wall, editor), London Math. Soc. Lecture Note Ser. 36, Cambridge Univ. Press (1979) 137–203 MR564422

[43] **Z Sela**, *Structure and rigidity in (Gromov) hyperbolic groups and discrete groups in rank 1 Lie groups, II*, Geom. Funct. Anal. 7 (1997) 561–593 MR1466338

[44] **G A Swarup**, *On a theorem of C B Thomas*, J. London Math. Soc. 8 (1974) 13–21 MR0341474

[45] **C B Thomas**, *The oriented homotopy type of compact 3–manifolds*, Proc. London Math. Soc. 19 (1969) 31–44 MR0248838

[46] **V G Turaev**, *Towards the topological classification of geometric 3–manifolds*, from: "Topology and geometry — Rohlin Seminar", (O Y Viro, editor), Lecture Notes in Math. 1346, Springer, Berlin (1988) 291–323 MR970081

[47] **V Turaev**, *A combinatorial formulation for the Seiberg–Witten invariants of 3–manifolds*, Math. Res. Lett. 5 (1998) 583–598 MR1666856

[48] **T Yoshida**, *Simply connected smooth 4–manifolds which admit nontrivial smooth S^1 actions*, Math. J. Okayama Univ. 20 (1978) 25–40 MR0494173

Department of Mathematics and Statistics, University of Massachusetts
Amherst, MA 01003, USA

wchen@math.umass.edu

Received: 26 March 2014 Revised: 10 January 2015

Algebraic & Geometric Topology 15 (2015) 3305–3322

Exactly fourteen intrinsically knotted graphs have 21 edges

MINJUNG LEE
HYOUNGJUN KIM
HWA JEONG LEE
SEUNGSANG OH

Johnson, Kidwell, and Michael showed that intrinsically knotted graphs have at least 21 edges. Also it is known that K_7 and the thirteen graphs obtained from K_7 by ∇Y moves are intrinsically knotted graphs with 21 edges. We prove that these 14 graphs are the only intrinsically knotted graphs with 21 edges.

57M25, 57M27

1 Introduction

Throughout the article we will take an embedded graph to mean a graph embedded in R^3. We call a graph G *intrinsically knotted* if every embedding of the graph contains a knotted cycle. Conway and Gordon [2] showed that K_7, the complete graph with seven vertices, is an intrinsically knotted graph. A graph H is *minor* of another graph G if it can be obtained from G by contracting or deleting some edges. An intrinsically knotted graph is *minor minimal intrinsically knotted* provided no proper minor is intrinsically knotted. Robertson and Seymour [9] proved that there are only finite minor minimal intrinsically knotted graphs, but finding the complete set of them is still an open problem. However, it is well known that K_7 and the thirteen graphs obtained from this graph by ∇Y moves are minor minimal intrinsically knotted; see Conway and Gordon [2], and Kohara and Suzuki [6].

A ∇Y *move* is an exchanging operation that removes all edges of a triangle abc and inserts a new vertex v and three edges va, vb and vc as in Figure 1. Its reverse operation is called a $Y\nabla$ *move*. Since ∇Y moves preserve intrinsic knottedness (see Motwani, Raghunathan, and Saran [7]), we will only consider triangle-free graphs in the article.

From the work of Johnson, Kidwell, and Michael [5], it follows that any intrinsically knotted graph consists at least 21 edges. Here is the main theorem.

Published: 12 January 2016 DOI: 10.2140/agt.2015.15.3305

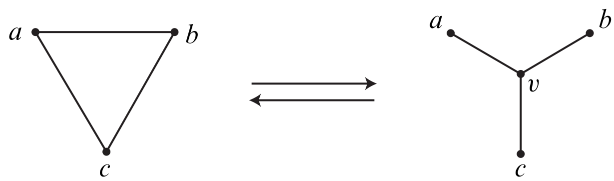


Figure 1: ∇Y and $Y\nabla$ moves

Theorem 1 *The only triangle-free intrinsically knotted graphs with exactly* 21 *edges are* H_{12} *and* C_{14}. (H_{12} *and* C_{14} *were described by Kohara and Suzuki in [6].*)

Kohara and Suzuki [6] found fourteen intrinsically knotted graphs. Goldberg, Mattman, and Naimi [3] constructed twenty graphs derived from H_{12} and C_{14} by $Y\nabla$ moves as in Figure 2, and they showed that these six graphs, N_9, N_{10}, N_{11}, N'_{10}, N'_{11}, and N'_{12}, are not intrinsically knotted. This fact was proved by Hanaki, Nikkuni, Taniyama, and Yamazaki [4] independently. Theorem 1 guarantees that all intrinsically knotted graphs with 21 edges can be obtained from H_{12} and C_{14} by $Y\nabla$ moves. Thus, we have the following theorem.

Theorem 2 *The only intrinsically knotted graphs with exactly* 21 *edges are* K_7 *and the thirteen graphs obtained from* K_7 *by* ∇Y *moves.*

This theorem gives us the complete set of fourteen minor minimal intrinsically knotted graphs with 21 edges.

Acknowledgments

The corresponding author (Seungsang Oh) was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) number NRF-2014R1A2A1A11050999.

The third author was supported by the National Research Foundation of Korea Grant funded by the Korean Government (NRF-2010-0024630).

2 Terminology

From now on let $G=(V,E)$ denote a triangle-free graph with 21 edges. Here V and E denote the sets of all vertices and edges of G, respectively. For any two distinct vertices a and b, let $\widehat{G}_{a,b}=(\widehat{V}_{a,b},\widehat{E}_{a,b})$ denote the graph obtained from G by deleting two vertices a and b, and then contracting an edge incident to a vertex of degree 1 or

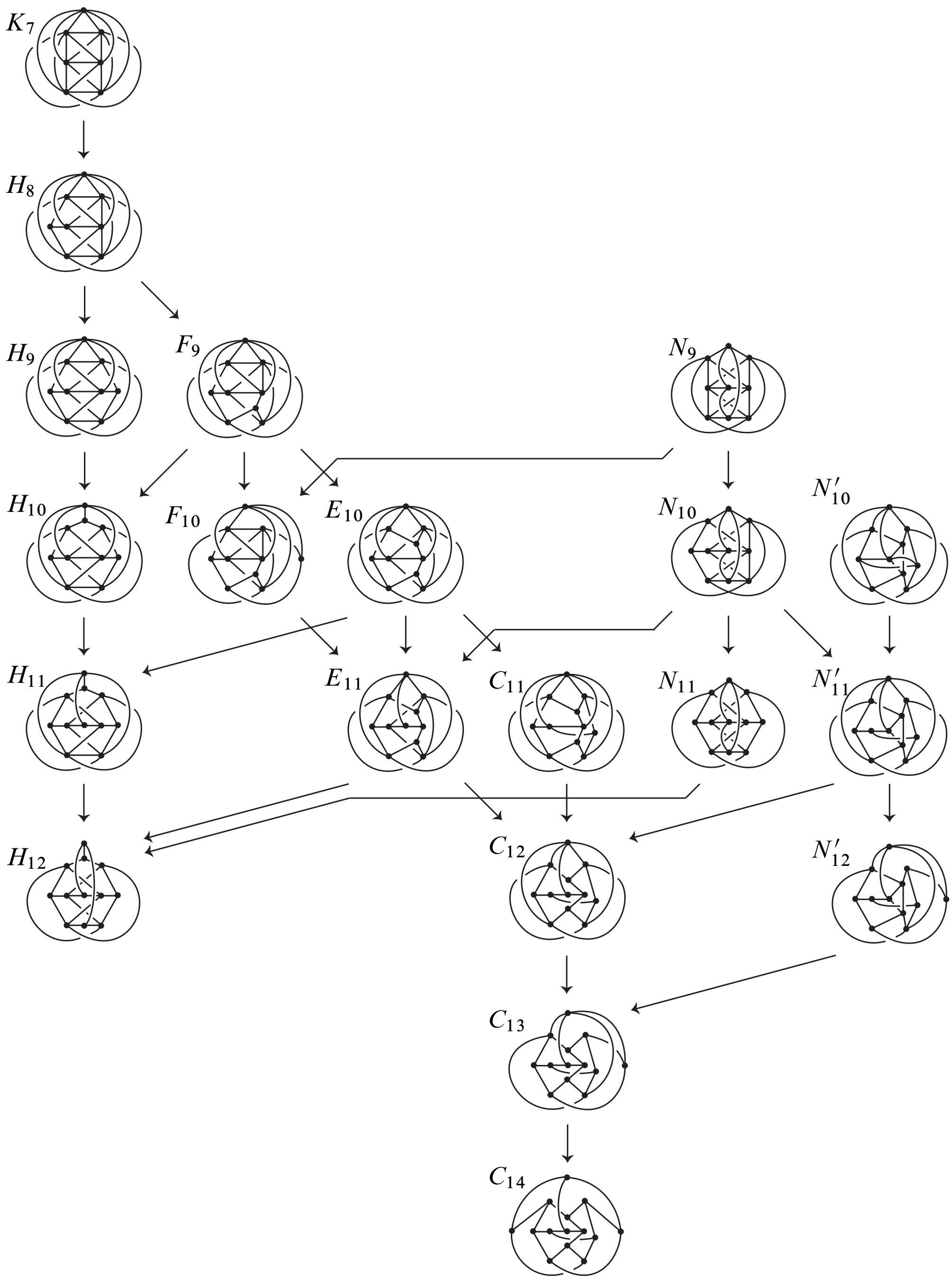


Figure 2: The graph K_7 and 19 more related graphs, where each arrow represents a ∇Y move

2 repeatedly until no vertices of degree 1 or 2 exist. Removing vertices means deleting interiors of all edges incident to these vertices as well as the resulting isolated vertices.

In a graph, the distance between two vertices a and b is the number of edges in the shortest path connecting them and is denoted by $\mathrm{dist}(a,b)$. The degree of a vertex a is denoted by $\deg(a)$. To count the number of edges of $\widehat{G}_{a,b}$, we introduce some notation.

- $E(a)$ is the set of edges which are incident to a.
- $V(a) = \{c \in V \mid \mathrm{dist}(a,c) = 1\}$.
- $V_n(a) = \{c \in V \mid \mathrm{dist}(a,c) = 1,\ \deg(c) = n\}$.
- $V_n(a,b) = V_n(a) \cap V_n(b)$.
- $V_Y(a,b) = \{c \in V \mid \exists\, d \in V_3(a,b) \text{ such that } c \in V_3(d) \setminus \{a,b\}\}$.

First consider the graph $G \setminus \{a,b\}$ for some distinct vertices a and b. In this graph each vertex of $V_3(a,b)$ has degree 1, and each vertex of $V_3(a)$, $V_3(b)$ (not in $V_3(a,b)$), and $V_4(a,b)$ has degree 2. To derive $\widehat{G}_{a,b}$, we first delete all edges incident to a and b from G, and then we also delete the remaining edges incident to $V_3(a,b)$, and finally we contract one edge of the remaining pair of edges incident to each vertex of $V_3(a)$, $V_3(b)$ (not in $V_3(a,b)$), $V_4(a,b)$, and $V_Y(a,b)$ as dotted lines in Figure 3(a). Thus, we have the following equation counting the number of edges of $\widehat{G}_{a,b}$ which is called a *count equation*:

$$|\widehat{E}_{a,b}| = 21 - |E(a) \cup E(b)| - (|V_3(a)| + |V_3(b)| - |V_3(a,b)| + |V_4(a,b)| + |V_Y(a,b)|).$$

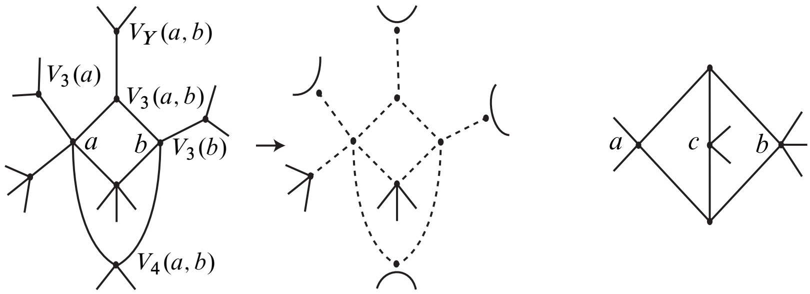
For short, $NE(a,b) = |E(a) \cup E(b)|$ and $NV_3(a,b) = |V_3(a)| + |V_3(b)| - |V_3(a,b)|$. If a and b are adjacent vertices (ie $\mathrm{dist}(a,b) = 1$), then all of $V_3(a,b)$, $V_4(a,b)$, and $V_Y(a,b)$ are empty because G is triangle-free. Note that this manner of deriving $\widehat{G}_{a,b}$ must be handled in a slightly different way when there is a vertex c in V such that more than one vertex of $V(c)$ are contained in $V_3(a,b)$ as in Figure 3(b). In this case, we usually delete or contract more edges incident to c, even though c is not in $V_Y(a,b)$.

A graph is *n–apex* if one can remove n vertices from the graph to obtain a planar graph. The following lemma gives an important condition for a graph to be not intrinsically knotted.

Lemma 3 [1; 8] *If G is 2–apex, then G is not intrinsically knotted.*

The following two lemmas play an important role for G to be 2–apex.

Lemma 4 *If $|\widehat{E}_{a,b}| \leq 8$, then $\widehat{G}_{a,b}$ is a planar graph. Thus, G is not intrinsically knotted.*



(a) (b)

Figure 3: Deriving $\widehat{G}_{a,b}$

Lemma 5 *If $|\widehat{E}_{a,b}| = 9$, then $\widehat{G}_{a,b}$ is either a planar graph or homeomorphic to $K(3,3)$. Furthermore, if $\widehat{G}_{a,b}$ is not homeomorphic to $K(3,3)$, then G is not intrinsically knotted.*

The graph $K(3,3)$ is a bipartite graph where each part has three vertices and each vertex is adjacent to every vertex in the opposite part, and so it is a triangle-free graph and every vertex has degree 3.

To prove Theorem 1, we will show that any triangle-free graph with 21 edges is eventually either a 2–apex or homeomorphic to one of H_{12} or C_{14}. Since intrinsically knotted graphs have at least 21 edges [5], it is sufficient to consider simple and connected graphs having no vertex of degree 1 or 2. Our process is constructing all possible such triangle-free graph G with 21 edges, deleting two suitable vertices a and b of G, and then counting the number of edges of $\widehat{G}_{a,b}$. If $\widehat{G}_{a,b}$ has 9 edges or less, we can use Lemma 4 or Lemma 5 in order to show that G is not intrinsically knotted. In the event that $\widehat{G}_{a,b}$ is not planar, we will show that G is homeomorphic to H_{12} or C_{14}.

Before describing the proof of Theorem 1, we introduce more notation. Since G is triangle-free, for any vertex a of G, no two vertices in $V(a)$ are adjacent. This means that $E(b)$ and $E(c)$ do not contain an edge in common for any two distinct vertices b and c in $V(a)$. We set:

- $E^2(a) = \bigcup_{b \in V(a)} E(b)$.
- $E \setminus E^2(a) = \{e_1(a), \ldots, e_{21-n}(a)\}$ if $|E^2(a)| = n < 21$.

$e_i(a)$ is called an *extra edge*, and the two endpoints of the edge are denoted as $x_i(a)$ and $y_i(a)$, where $\deg(x_i(a)) \geq \deg(y_i(a))$.

In order to visualize G, we perform the following steps. First choose a vertex a with the maximal degree among all vertices and draw $E^2(a)$. If $|E^2(a)| < 21$, draw $E \setminus E^2(a)$ apart from $E^2(a)$ as in Figure 4(a). Then all vertices of degree 1 of $E^2(a)$ and $E \setminus E^2(a)$ are merged into some vertices of degree at least 3 without adding new edges as in Figure 4(b). Let $\overline{V}(a)$ denote the set of all such vertices, and let $[\overline{V}(a)]$ denote a sequence of the degrees of vertices in $\overline{V}(a)$ as follows:

- $\overline{V}(a) = V \setminus (V(a) \cup \{a\}) = \{\overline{v}_1(a), \dots, \overline{v}_m(a)\}$ with $\deg(\overline{v}_i(a)) \geq \deg(\overline{v}_{i+1}(a))$.
- $[\overline{V}(a)] = [\deg(\overline{v}_1(a)), \dots, \deg(\overline{v}_m(a))]$.
- $\|[\overline{V}(a)]\| = \deg(\overline{v}_1(a)) + \cdots + \deg(\overline{v}_m(a))$.

The graph in Figure 4(b) is an example satisfying $\deg(a) = 5$, $|V_3(a)| = 1$, $|E^2(a)| = 19$, and $[\overline{V}(a)] = [4, 4, 4, 3, 3]$.

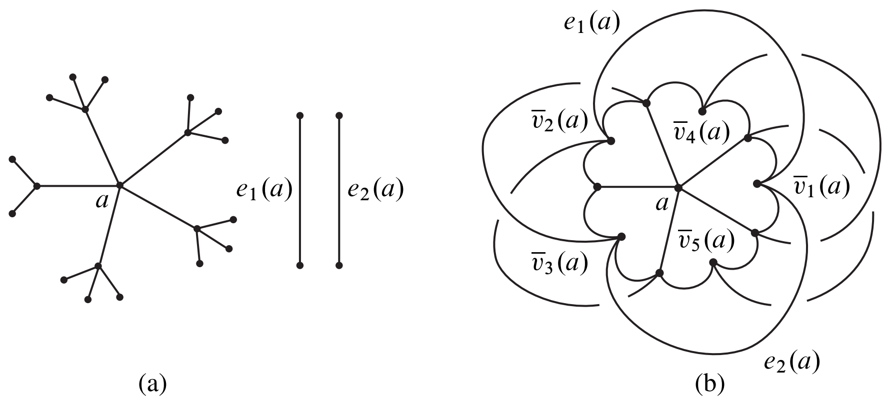


Figure 4: Visualization of G

The remaining three sections of the article are devoted to the proof of Theorem 1. From now on, a denotes one of vertices with maximal degree in G. The proof is divided into three parts according to the degree of a. In Section 3 we show that any graph G with $\deg(a) \geq 5$ cannot be intrinsically knotted. In Section 4 we show that an intrinsically knotted graph with $\deg(a) = 4$ is exactly H_{12}. Finally, in Section 5 we show that any intrinsically knotted graph, all of whose vertices have degree 3, is always C_{14}.

3 deg(a) ≥ 5

In this section we will show that for some $a', b' \in V$ either $|\widehat{E}_{a',b'}| \leq 8$ or $|\widehat{E}_{a',b'}| = 9$, but that $\widehat{G}_{a',b'}$ is not homeomorphic to $K(3,3)$ by showing that it contains a vertex of degree more than 3 or a triangle (or sometimes a bigon). Then, as a conclusion, G is not intrinsically knotted by Lemmas 4 and 5. Recall that G has 21 edges, every vertex has degree at least 3, and a has the maximal degree among them.

3.1 Case $\deg(a) \geq 6$ or $\deg(a) = 5$ with $|V_3(a)| \geq 4$

If $\deg(a) \geq 6$, then $|V_3(a)| \geq 3$. Let c be any vertex in $V_3(a)$. Choose a vertex b which has the maximal degree among $V(c) \setminus \{a\}$. Then $|E(b)| + |V_Y(a,b)| \geq 4$, since $|V_Y(a,b)| \geq 1$ when $\deg(b) = 3$. Note that $|V_3(b)| \geq |V_3(a,b)|$. By the count equation, $|\widehat{E}_{a,b}| \leq 8$ in $\widehat{G}_{a,b}$.

Suppose that $\deg(a) = 5$ and $|V_3(a)| \geq 4$. The proof is similar to the previous paragraph.

3.2 Case $\deg(a) = 5$ and $|V_3(a)| = 3$

Let b and c be two vertices of $V(a) \setminus V_3(a)$. First, suppose that both of them have degree 5. Then $NE(a,b) = 9$ and $|V_3(a)| = 3$, so $|\widehat{E}_{a,b}| \leq 9$. Furthermore, the vertex c has degree 4 in $\widehat{G}_{a,b}$, so it follows that $\widehat{G}_{a,b}$ is not homeomorphic to $K(3,3)$. Thus, G is not intrinsically knotted by Lemma 5.

Now assume that one of them, say b, has degree 4. If $V(b) \setminus \{a\}$ consists of three vertices, all of which are of degree 3, then $NE(a,b) = 8$ and $NV_3(a,b) = 6$, so $|\widehat{E}_{a,b}| \leq 7$. If not, let d be a vertex of $V(b)$ which has degree at least 4. Then $NE(a,d) \geq 9$, $|V_3(a)| = 3$, and $|V_4(a,d)| \geq 1$, because $V_4(a,d) \ni b$. This implies that $|\widehat{E}_{a,d}| \leq 8$.

3.3 Case $\deg(a) = 5$ and $|V_3(a)| = 0$

First, suppose that $V(a)$ contains a vertex of degree 5, say c. Since G has 21 edges, the other four vertices of $V(a)$ have degree 4. By the previous cases, it is sufficient to suppose that $|V_3(c)| \leq 2$. So $V(c) \setminus \{a\}$ has at least two vertices, say b and d, of degree 4 or 5. Since $|E^2(a)| = 21$ and G is triangle-free, all edges of $E(b)$ must be incident to different vertices of $V(a)$, so $|V_4(a,b)| \geq 3$. This implies that $|\widehat{E}_{a,b}| \leq 9$. Since $\widehat{G}_{a,b}$ has the vertex d of degree at least 4, it follows that $\widehat{G}_{a,b}$ is not homeomorphic to $K(3,3)$.

Now, assume that all vertices of $V(a)$ have degree 4, giving $|E^2(a)| = 20$. Let $e_1(a)$ be the extra edge and recall that two endpoints of $e_1(a)$ are $x_1(a)$ and $y_1(a)$ with $\deg(x_1(a)) \geq \deg(y_1(a))$. Since G is triangle-free, all edges of $E(x_1(a)) \cup E(y_1(a))$ except $e_1(a)$ must be incident to different vertices of $V(a)$. Thus the degrees of $x_1(a)$ and $y_1(a)$ must be either 4 and 3, or 3 and 3, respectively. If $\deg(x_1(a)) = 4$, then $|V_4(a, x_1(a))| = 3$ and $|V_3(x_1(a))| = 1$, so $|\widehat{E}_{a,x_1(a)}| = 8$. If not, $[\overline{V}(a)]$ is either $[5,3,3,3,3]$ or $[4,4,3,3,3]$, because $|[\overline{V}(a)]| = 17$. Thus $\overline{v}_1(a)$ has degree 5 or 4 and differs from $x_1(a)$ and $y_1(a)$, so $|V_4(a, \overline{v}_1(a))| \geq 4$. Therefore, $|\widehat{E}_{a,\overline{v}_1(a)}| \leq 8$.

3.4 Case deg(a) = 5 and $|V_3(a)| = 1$

In this case, $V(a)$ contains four vertices of degree 4 or 5. Let n be the number of such vertices of degree 4, and so we have $4-n$ vertices of degree 5, where $n = 2, 3, 4$. This implies that $|E^2(a)| = 21 + (2-n)$, and $n-2$ extra edges exist. If $\overline{V}(a)$ contains a vertex $\overline{v}_1(a)$ of degree 5, then five edges of $E(\overline{v}_1(a))$ are extra edges or incident to different vertices in $V(a)$. For any of the above n, at least two among these edges are incident to vertices of degree 4 in $V(a)$. Then $NE(a, \overline{v}_1(a)) = 10$, $|V_3(a)| = 1$, and $|V_4(a, \overline{v}_1(a))| \geq 2$, implying $|\widehat{E}_{a,\overline{v}_1(a)}| \leq 8$.

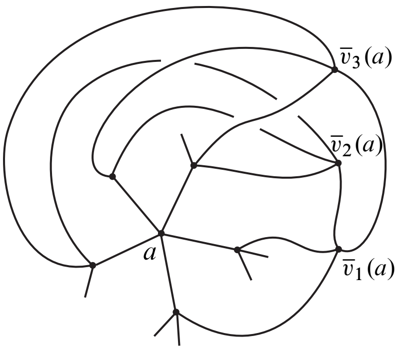
Now, suppose that $\overline{V}(a)$ contains vertices of degree 3 or 4 only. If $n = 2$, $\|[\overline{V}(a)]\| = 16$, and so $[\overline{V}(a)]$ is either $[4,4,4,4]$ or $[4,3,3,3,3]$. For any vertex b in $V_5(a)$, four edges of $E(b)$ must be incident to different vertices of $\overline{V}(a)$. Indeed, these four edges are incident to four vertices of degree 4, or at least three edges among them are incident to vertices of degree 3 in $\overline{V}(a)$. This means that the vertex b has degree 5 with either $V_3(b) = 0$ or $V_3(b) \geq 3$. Both cases are dealt with in previous cases 3.3, 3.1, and 3.2.

If $n = 3$, $\|[\overline{V}(a)]\| = 17$, and so $[\overline{V}(a)] = [4,4,3,3,3]$. Let $V_5(a) = \{b\}$. To avoid the case 3.2, four edges of $E(b)$ must be incident to two vertices of degree 4 and two vertices of degree 3 in $\overline{V}(a)$, which are $\overline{v}_1(a)$, $\overline{v}_2(a)$, $\overline{v}_3(a)$, and $\overline{v}_4(a)$. Then there is a vertex c of $V_4(a)$ such that at most one edge of $E(c)$ is incident to $\overline{v}_3(a)$ and $\overline{v}_4(a)$, ie two edges of $E(c)$ are incident to $\overline{v}_1(a)$, $\overline{v}_2(a)$, or $\overline{v}_5(a)$. This implies that $NE(b,c) = 9$ and $NV_3(b,c) + |V_4(b,c)| \geq 4$, implying $|\widehat{E}_{b,c}| \leq 8$.

Finally, if $n = 4$, $\|[\overline{V}(a)]\| = 18$, and so $[\overline{V}(a)]$ is either $[4,4,4,3,3]$ or $[3,3,3,3,3,3]$. Recall that two extra edges exist. In the former case let $\{\overline{v}_1(a), \overline{v}_2(a), \overline{v}_3(a)\}$ be the three vertices of degree 4 in $\overline{V}(a)$. For each $i = 1,2,3$, if more than two edges of $E(\overline{v}_i(a))$ are incident to $V_4(a)$, then $NE(a, \overline{v}_i(a)) = 9$, $|V_3(a)| = 1$, and $|V_4(a, \overline{v}_i(a))| \geq 3$, implying $|\widehat{E}_{a,\overline{v}_i(a)}| \leq 8$. So, each of at least two edges of $E(\overline{v}_i(a))$ must be either incident to the unique vertex of $V_3(a)$ or an extra edge. Since G is triangle-free, one of three vertices, say $\overline{v}_1(a)$, has the property that $E(\overline{v}_1(a))$ contains both extra edges, and $V(\overline{v}_1(a))$ and $V(\overline{v}_i(a))$ for each $i = 2,3$ cannot share a vertex in $V(a)$. This implies that $V(\overline{v}_2(a))$ and $V(\overline{v}_3(a))$ coincide as in Figure 5(a). Then $NE(\overline{v}_2(a), \overline{v}_3(a)) = 8$, and either $|V_4(\overline{v}_2(a), \overline{v}_3(a))| = 4$ or $|V_4(\overline{v}_2(a), \overline{v}_3(a))| = 3$ and $|V_3(\overline{v}_2(a))| = 1$. Thus, $|\widehat{E}_{\overline{v}_2(a),\overline{v}_3(a)}| \leq 9$. In $\widehat{G}_{\overline{v}_2(a),\overline{v}_3(a)}$ the vertex a still has degree 4 or 5 so that $\widehat{G}_{\overline{v}_2(a),\overline{v}_3(a)}$ is not homeomorphic to $K(3,3)$.

In the latter case, let $V_4(a) = \{b_1, b_2, b_3, b_4\}$. We claim that for some $i, j = 1,2,3,4$, $|V_3(b_i, b_j)| \leq 1$. Suppose not; that is, $|V_3(b_i, b_j)| \geq 2$ for all combinations of i and j. By some combinatorics we can derive that all 12 edges of $E(b_1) \cup E(b_2) \cup E(b_3) \cup E(b_4) \setminus E(a)$ are incident to only four vertices of $\overline{V}(a)$ as in Figure 5(b).

This means that two extra edges must be incident to the remaining two vertices of $\overline{V}(a)$ at both endpoints. But a bigon is not allowed. Therefore, without loss of generality, $|V_3(b_1, b_2)| \leq 1$. Then $NE(b_1, b_2) = 8$ and $NV_3(b_1, b_2) \geq 5$, implying $|\widehat{E}_{b_1,b_2}| \leq 8$.



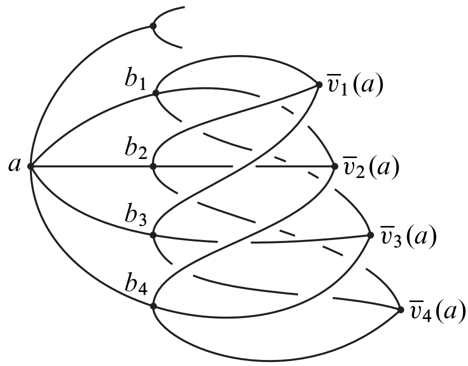


Figure 5: [4, 4, 4, 3, 3] and [3, 3, 3, 3, 3, 3] cases

3.5 Case deg$(a) = 5$ and $|V_3(a)| = 2$

If $V(a)$ contains a vertex of degree 5, say b, then the previous four cases guarantee that we only consider that $|V_3(b)| = 2$, so $NV_3(a,b) = 4$, which implies $|\widehat{E}_{a,b}| = 8$. Therefore we assume that $V(a)$ contains three vertices of degree 4. In this case three extra edges exist. Since $\|[\overline{V}(a)]\| = 19$, $[\overline{V}(a)]$ is one of [5, 5, 5, 4], [5, 5, 3, 3, 3], [5, 4, 4, 3, 3], [4, 4, 4, 4, 3], or [4, 3, 3, 3, 3, 3].

If, for some vertex $\overline{v}_i(a)$ with degree 5, one edge of $E(\overline{v}_i(a))$ is incident to $V_4(a)$, then $NE(a, \overline{v}_i(a)) = 10$, $|V_3(a)| = 2$, and $|V_4(a, \overline{v}_i(a))| \geq 1$, implying $|\widehat{E}_{a,\overline{v}_i(a)}| \leq 8$. Thus, three edges of $E(\overline{v}_i(a))$ are extra edges and the remaining two edges are incident to $V_3(a)$. In the first two cases, [5, 5, 5, 4] and [5, 5, 3, 3, 3], both $E(\overline{v}_1(a))$ and $E(\overline{v}_2(a))$ share three extra edges, but G does not have a bigon. In the third case, [5, 4, 4, 3, 3], $E(\overline{v}_1(a))$ contains three extra edges and one of these extra edges must be incident to $\overline{v}_4(a)$ or $\overline{v}_5(a)$, both of which have degree 3. Then $NE(a, \overline{v}_1(a)) = 10$ and $NV_3(a, \overline{v}_1(a)) \geq 3$, implying $|\widehat{E}_{a,\overline{v}_1(a)}| \leq 8$.

If, for some vertex $\overline{v}_i(a)$ with degree 4, two edges of $E(\overline{v}_i(a))$ are incident to $V_4(a)$, then $NE(a, \overline{v}_i(a)) = 9$, $|V_3(a)| = 2$, and $|V_4(a, \overline{v}_i(a))| \geq 2$, implying $|\widehat{E}_{a,\overline{v}_i(a)}| \leq 8$. Thus, at most one edge of $E(\overline{v}_i(a))$ is incident to $V_4(a)$. In the fourth case, [4, 4, 4, 4, 3], at least twelve among sixteen edges incident to four vertices of degree 4 in $\overline{V}(a)$ are not incident to $V_4(a)$. This is impossible because there are only two vertices in $V_3(a)$ and three extra edges. In the last case, [4, 3, 3, 3, 3, 3], since only one edge of $E(\overline{v}_1(a))$ is possibly incident to $V_4(a)$, there is a vertex b in $V_4(a)$ such that three edges of $E(b)$ are incident to vertices of degree 3 in $\overline{V}(a)$. Then $NE(a,b)=8$ and $NV_3(a,b)\geq 5$, implying $|\widehat{E}_{a,b}|\leq 8$.

4 deg(a) = 4

Since $|V|=|V_4|+|V_3|$ and $4|V_4|+3|V_3|=2|E|$, the pair $(|V_4|,|V_3|)$ has three choices: $(3,10)$, $(6,6)$, and $(9,2)$. Here, V_n denotes the set of vertices of degree n. As in the preceding section, we will show that for some $a',b'\in V$ either $|\widehat{E}_{a',b'}|\leq 8$ or $|\widehat{E}_{a',b'}|=9$, but $\widehat{G}_{a',b'}$ is not homeomorphic to $K(3,3)$, implying that G is not intrinsically knotted. But one exception occurs so that G can possibly be H_{12} when $(|V_4|,|V_3|)=(6,6)$.

4.1 Case $(|V_4|,|V_3|)=(3,10)$

First suppose that V_4 has a vertex a such that all four vertices of $V(a)$ have degree 3. Let b_1 and b_2 be the other vertices of V_4. For each $i=1,2$, $NE(a,b_i)=8$. If there is a vertex of $V_3(b_i)$ which is not contained in $V(a)$, then $NV_3(a,b_i)\geq 5$, implying $|\widehat{E}_{a,b_i}|\leq 8$. Thus each vertex of $V(b_1)$ is the vertex b_2 or contained in $V(a)$, and similarly for b_2. This implies that the number of vertices of V_3 which have distance 1 or 2 from the vertex a is at most 6. Take a vertex c of V_3 with distance at least 3 from a. Since each vertex of $V(c)$ is neither b_1 nor b_2, it has degree 3. Thus $NE(a,c)=7$ and $NV_3(a,c)\geq 7$, implying $|\widehat{E}_{a,c}|\leq 7$.

Now, we only need to consider the case that each vertex of V_4 is adjacent to at least one vertex of degree 4. Then, without loss of generality, we have vertices a, b and c of V_4 such that $V(b)$ contains a and c. If $V_3(a)$ and $V_3(c)$ do not coincide, then $|V_4(a,c)|=1$ and $NV_3(a,c)\geq 4$, implying $|\widehat{E}_{a,c}|\leq 8$. If $V_3(a)$ and $V_3(c)$ coincide and $|V_Y(a,c)|\geq 2$, then $|V_4(a,c)|=1$ and $NV_3(a,c)=3$, implying $|\widehat{E}_{a,c}|\leq 7$. If not, for the unique vertex d of $V_Y(a,c)$, $V_3(a)=V_3(c)=V(d)$. Then, for a vertex b' of $V_3(b)$, $V_3(b')$ is disjoint from $V_3(a)$. Thus $NE(a,b')=7$, $NV_3(a,b')=5$, and $|V_4(a,b')|=1$, implying $|\widehat{E}_{a,b'}|\leq 8$.

4.2 Case $(|V_4|,|V_3|)=(6,6)$

Consider the subgraph H of G consisting of all edges whose both end vertices have degree 4. Since G has six vertices of degree 3 and the same number of vertices of degree 4, H is not empty set.

Claim 1 *If H has a vertex of degree 1, then G is not intrinsically knotted.*

Proof Suppose that H has a vertex a of degree 1. Let b be the unique vertex of degree 4 in $V(a)$. If $|V_3(b)|=3$, then $NE(a,b)=7$ and $NV_3(a,b)=6$, implying $|\widehat{E}_{a,b}|\le 8$. Thus, there is another vertex c of $V_4(b)$, and so we let $V(c)=\{b,d_1,d_2,d_3\}$.

First, assume that $|V_3(c)|=0$. So the two vertices of $V(b)\setminus\{a,c\}$ must have degree 3, because the six vertices a,b,c,d_1,d_2, and d_3 in V_4 are all different. Thus $NE(a,b)=7$ and $NV_3(a,b)=5$, so $|\widehat{E}_{a,b}|\le 9$. Since $\widehat{G}_{a,b}$ has another vertex d_1 of degree 4, it follows that $\widehat{G}_{a,b}$ is not homeomorphic to $K(3,3)$.

Second, assume that $|V_3(c)|=1$, say $d_1\in V_3(c)$. If d_1 is not one of the vertices in $V(a)$, then $NE(a,c)=8$ and $NV_3(a,c)+|V_4(a,c)|=5$, implying $|\widehat{E}_{a,c}|\le 8$. So we may assume that d_1 is in $V(a)$ and let $V(d_1)=\{a,c,v_1\}$. If v_1 has degree 3, then $NV_3(a,c)+|V_4(a,c)|=4$ and $V_Y(a,c)=\{v_1\}$, implying $|\widehat{E}_{a,c}|\le 8$. Otherwise v_1 has degree 4 and it is different from d_2 and d_3. For any $i=2,3$, each vertex of $V(d_i)\setminus\{c\}$ either has degree 3 or is v_1. Thus $NE(d_2,d_3)=8$ and $NV_3(d_2,d_3)+|V_4(d_2,d_3)|\ge 4$, implying $|\widehat{E}_{d_2,d_3}|\le 9$. But $\widehat{G}_{d_2,d_3}$ has a triangle containing vertices a, b and d_1. See Figure 6(a).

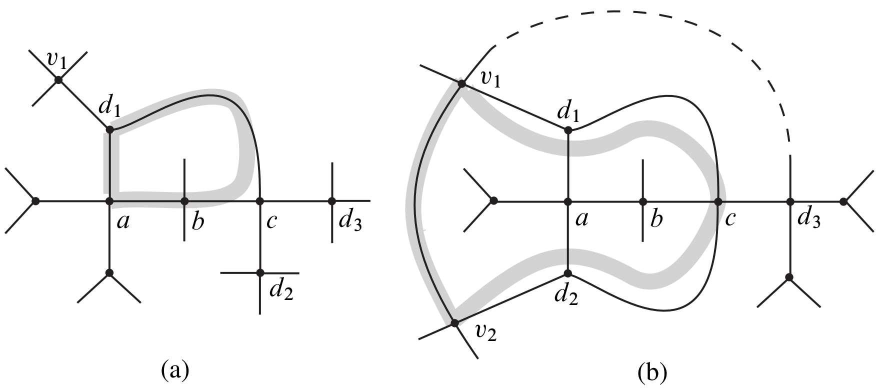


Figure 6: Some nonintrinsically knotted cases

Last, assume that $|V_3(c)|\ge 2$ and let d_1 and d_2 be two such vertices. As in the previous case, we may say that d_1 and d_2 are in $V(a)$, and $V(d_i)=\{a,c,v_i\}$ for $i=1,2$ where v_i has degree 4. When $v_1=v_2$, $|V_3(a)|=3$, $|V_4(a,c)|=1$, and v_1 has degree 2 when we construct $\widehat{G}_{a,c}$, implying $|\widehat{E}_{a,c}|\le 8$. When $\mathrm{dist}(v_1,v_2)\ge 2$, three cases occur as follows: $|V_3(v_1)|\ge 3$, $|V_3(v_2)|\ge 3$, or for both $i=1,2$ $|V_3(v_i)|=2$ and $V_4(v_i)=V_4\setminus\{a,c,v_1,v_2\}$. All three cases satisfy that $NV_3(v_1,v_2)+|V_4(v_1,v_2)|\ge 4$, implying $|\widehat{E}_{v_1,v_2}|\le 9$. But $\widehat{G}_{v_1,v_2}$ has a bigon containing vertices a and c. Finally, when $\mathrm{dist}(v_1,v_2)=1$, two cases occur as follows. If d_3 has degree 3, then by the same reason as before we may say that d_3 is also in $V(a)$, and $V(d_3)=\{a,c,v_3\}$ where v_3 has degree 4. By the previous argument any pair of v_1, v_2 and v_3 has distance 1. This implies that G contains a triangle. If d_3 has degree 4, then $|V_3(d_3)| \geq 2$, because at most one vertex of $V(d_3)$ can be v_1 or v_2. Thus, $NV_3(a, d_3) \geq 4$, implying $|\widehat{E}_{a,d_3}| \leq 9$. But $\widehat{G}_{a,d_3}$ has a triangle containing vertices c, v_1 and v_2. See Figure 6(b). □

Claim 2 *If H is not a cycle with 6 edges, then G is not intrinsically knotted.*

Proof By Claim 1, if H is not a cycle with 6 edges, then H contains a cycle with 4 or 5 edges. First assume that H contains a cycle with 5 edges. Let $\{a_1, \dots, a_5\}$ be the set of five vertices of the cycle appearing in clockwise order. If the remaining vertex b of V_4 is contained in some $V(a_i)$, say $i = 1$, then b must have distance 1 from one of a_3 and a_4, say a_3, by Claim 1. See Figure 7. If $V_3(a_2) \neq V_3(b)$, $NV_3(a_2, b) + |V_4(a_2, b)| \geq 5$, implying $|\widehat{E}_{a_2,b}| \leq 8$. Otherwise, $V_3(a_2) = V_3(b)$. Let c_1 and c_3 be the vertices of $V_3(a_1)$ and $V_3(a_3)$, respectively. If $c_1 = c_3$, we still have $|\widehat{E}_{a_2,b}| \leq 9$ and $\widehat{G}_{a_2,b}$ has a triangle containing vertices a_5, a_4 and $c_1 = c_3$. If $c_1 \neq c_3$, then $|\widehat{E}_{a_1,a_3}| \leq 9$ and $\widehat{G}_{a_1,a_3}$ has a bigon as in the figure.

If b is not contained in $V(a_i)$ for any $i = 1, \dots, 5$, then $|V_3(a_i)| = 2$. If there is a pair of vertices a_i and a_{i+2} (or a_{i-3} if $i = 4, 5$) such that $V_3(a_i)$ and $V_3(a_{i+2})$ are disjoint, then $NV_3(a_i, a_{i+2}) + |V_4(a_i, a_{i+2})| = 5$, implying $|\widehat{E}_{a_i,a_{i+2}}| \leq 8$. Otherwise, for any pair of vertices a_i and a_{i+2} (or a_{i-3} if $i = 4, 5$), $V_3(a_i)$ and $V_3(a_{i+2})$ share vertices. Then they must share only one vertex as in Figure 7(b). Since there is only one extra vertex b of degree 4, for some pair of vertices a_i and a_{i+2}, $NV_3(a_i, a_{i+2}) + |V_4(a_i, a_{i+2})| = 4$ and $V_Y(a_i, a_{i+2}) \geq 1$, implying $|\widehat{E}_{a_i,a_{i+2}}| \leq 8$.

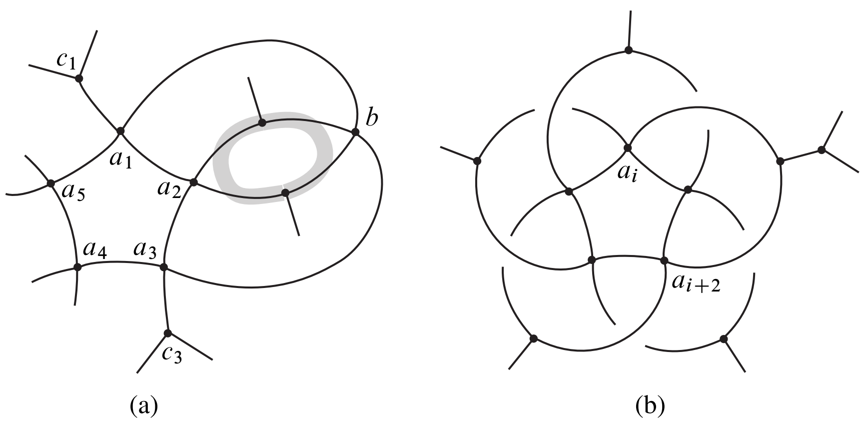


Figure 7: Cycle with 5 edges

Now, assume that H contains a cycle with 4 edges. Let $\{a_1, \dots, a_4\}$ be the set of four vertices of the cycle appearing in clockwise order. If $V(a_1)$ and $V(a_3)$ (or similarly for $V(a_2)$ and $V(a_4)$) share only two vertices, a_2 and a_4, then the remaining two vertices of V_4 must be contained in $V(a_1) \cup V(a_3)$. Otherwise, since $V(a_1) \cup V(a_3)$ has four more vertices other than a_2 and a_4, $NV_3(a_1, a_3) \geq 3$ and $|V_4(a_1, a_3)| = 2$, implying $|\widehat{E}_{a_1,a_3}| \leq 8$. By Claim 1, the two vertices have distance 1, so H contains a cycle with 5 edges which was dealt in the previous case. If $V(a_1)$ and $V(a_3)$ (or similarly for $V(a_2)$ and $V(a_4)$) share exactly three vertices, a_2, a_4 and b, then let c_1 and c_3 be the remaining vertices of $V(a_1)$ and $V(a_3)$, respectively. If both c_1 and c_3 have degree 3, then $NV_3(a_1, a_3) + |V_4(a_1, a_3)| \geq 5$. If both have degree 4, then H contains a cycle with 5 edges as in the previous case. Finally, if only c_1 (or similarly c_3) has degree 4, then, by Claim 1, $V(c_1)$ contains another vertex, say d, of V_4, and also d must have distance 1 from one of a_2 and a_4, say a_4, as in Figure 8(a). So $NV_3(a_4, c_1) + |V_4(a_4, c_1)| \geq 4$, implying $|\widehat{E}_{a_4,c_1}| \leq 9$, and $\widehat{G}_{a_4,c_1}$ has a triangle containing vertices a_2, a_3, and b. Now we may assume that $V(a_1) = V(a_3)$ and $V(a_2) = V(a_4)$. Then $NV_3(a_1, a_3) + |V_4(a_1, a_3)| = 4$, implying $|\widehat{E}_{a_1,a_3}| \leq 9$, and so $\widehat{G}_{a_1,a_3}$ has a bigon as in Figure 8(b). □

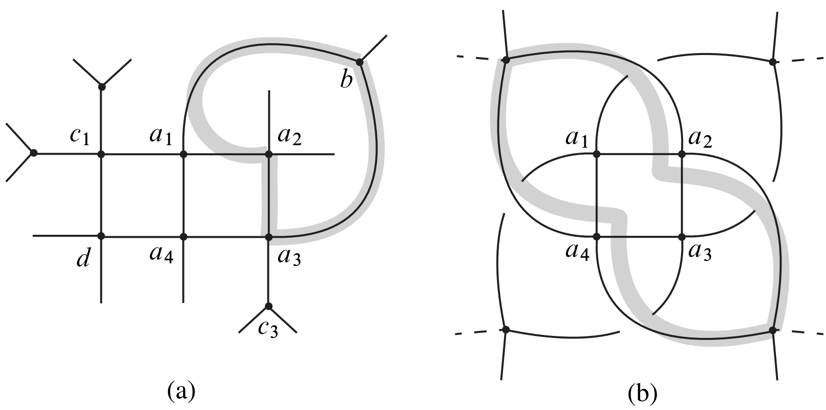


Figure 8: Cycle with 4 edges

By Claim 2, H is exactly a cycle with 6 edges. Let $\{a_1, \ldots, a_6\}$ be the set of six vertices of the cycle with a_i adjacent to a_{i+1} for $i = 1, \ldots, 5$, and a_6 adjacent to a_1. First, suppose that there is not a vertex b in V_3 such that $V(b) = \{a_1, a_3, a_5\}$. If $V_3(a_1)$ and $V_3(a_3)$ are disjoint, then $NV_3(a_1, a_3) + |V_4(a_1, a_3)| = 5$. If $V_3(a_1)$ and $V_3(a_3)$ share exactly one vertex c, then the vertex of $V(c) \setminus \{a_1, a_3\}$ is not a_5, so it should be one of $V_Y(a_1, a_3)$. Thus $NV_3(a_1, a_3) + |V_4(a_1, a_3)| + |V_Y(a_1, a_3)| = 5$. If $V_3(a_1)$ and $V_3(a_3)$ are same, then $NV_3(a_1, a_5) + |V_4(a_1, a_5)| = 5$, because $V_3(a_1)$ and $V_3(a_5)$ are disjoint. All three cases guarantee that G is not intrinsically knotted. Therefore we may assume that there are two vertices b_1 and b_2 so that $V(b_1) = \{a_1, a_3, a_5\}$ and $V(b_2) = \{a_2, a_4, a_6\}$. See Figure 9(a).

Suppose that there is a vertex c, with $c \neq b_1$, so that $V(c)$ contains a_1 and a_3. Let d_2 and d_5 be the vertices of $V_3(a_2)$ and $V_3(a_5)$, other than b_1 and b_2, respectively. If $d_2 \neq d_5$, then $NV_3(a_2, a_5) = 4$. If $d_2 = d_5$, then $NV_3(a_2, a_5) = 3$ and $V_Y(a_2, a_5)$ is not empty. Both cases provide $|\widehat{E}_{a_2,a_5}| \leq 9$, and $\widehat{G}_{a_2,a_5}$ has a triangle containing vertices a_1, a_3, and c. Therefore we may assume in general that for any vertex c, except b_1 and b_2, $V(c)$ does not contain both a_i and a_{i+2} for any $i = 1, 2, 3, 4$, and both a_i and a_{i-4} for any $i = 5, 6$.

Now we conclude $E \setminus \{E^2(b_1) \cup E^2(b_2)\}$ consists of three extra edges. Note that each vertex of these edges has degree 3, and there are four more vertices of degree 3 besides b_1 and b_2. These two facts guarantee that these extra edges must be connected as a tree. This tree can be of two types; either all three edges are incident to one vertex d, or two edges are incident to different endpoints of the other edge e, respectively. In both cases, any two edges adjoined to the tree at the same vertex at the end must be also incident to a_i and a_{i+3}, respectively, for some $i = 1, 2, 3$. Therefore, G is one of three graphs as in Figure 9(b)–(c), depending on the type of the tree. The graph G in Figure 9(b) is H_{12}, which is intrinsically knotted. But the two graphs in Figure 9(c) are not intrinsically knotted because, for some i, $|\widehat{E}_{a_i,a_{i+2}}| \leq 9$, and $\widehat{G}_{a_i,a_{i+2}}$ has a triangle.

4.3 Case $(|V_4|, |V_3|) = (9, 2)$

Let b_1 and b_2 be the vertices of V_3. Since $|V_3| = 2$, there are at least three vertices, a_1, a_2, and a_3, in V_4 such that all vertices of each $V(a_i)$ have degree 4. If $\mathrm{dist}(a_1, a_2) = 1$, then $V(a_1) \cup V(a_2)$ consists of 8 vertices of V_4, and so let c be the ninth vertex. Let d be any vertex among $V(a_1) \cup V(a_2) \setminus \{a_1, a_2\}$ which is not contained in $V(c)$. We assume that d is in $V(a_1)$. Then $V(d)$ should be contained in $V(a_2) \cup \{b_1, b_2\}$. This implies that $NE(a_2, d) = 8$ and $|V_3(d)| + |V_4(a_2, d)| \geq 4$, implying $|\widehat{E}_{a_2,d}| \leq 9$. Since c has degree 4 in $\widehat{G}_{a_2,d}$, it follows that $\widehat{G}_{a_2,d}$ is not homeomorphic to $K(3, 3)$. We have the same result for any choices of pairs among a_1, a_2, and a_3.

Now assume that the distance between any pair among a_1, a_2, and a_3 is at least 2. We separate into several cases according to the number $|V_4(a_1, a_2)|$. If $V_4(a_1, a_2) = \varnothing$ (ie $\mathrm{dist}(a_1, a_2) > 2$), then $|V_4| \geq 10$, a contradiction. If $V_4(a_1, a_2) = \{d\}$, then $V_4 = V(a_1) \cup V(a_2) \cup \{a_1, a_2\}$. This implies that $a_3 \in V(a_1) \cup V(a_2)$, so $\mathrm{dist}(a_1, a_3) = 1$ or $\mathrm{dist}(a_2, a_3) = 1$, both of which were dealt with in the previous case. If $V_4(a_1, a_2) = \{d_1, d_2\}$, then $V(d_1) \cup V(d_2) \setminus \{a_1, a_2\}$ is contained in $\{a_3, b_1, b_2\}$. This implies that each $V(d_i) \setminus \{a_1, a_2\}$ is a set of two vertices among $\{a_3, b_1, b_2\}$, so that $|V_3(d_1, d_2)| + |V_4(d_1, d_2)| \geq 4$, implying $|\widehat{E}_{d_1,d_2}| \leq 9$. Since at least two of four vertices in $V(a_1) \cup V(a_2) \setminus \{d_1, d_2\}$ still have degree 4 in $\widehat{G}_{d_1,d_2}$, it follows

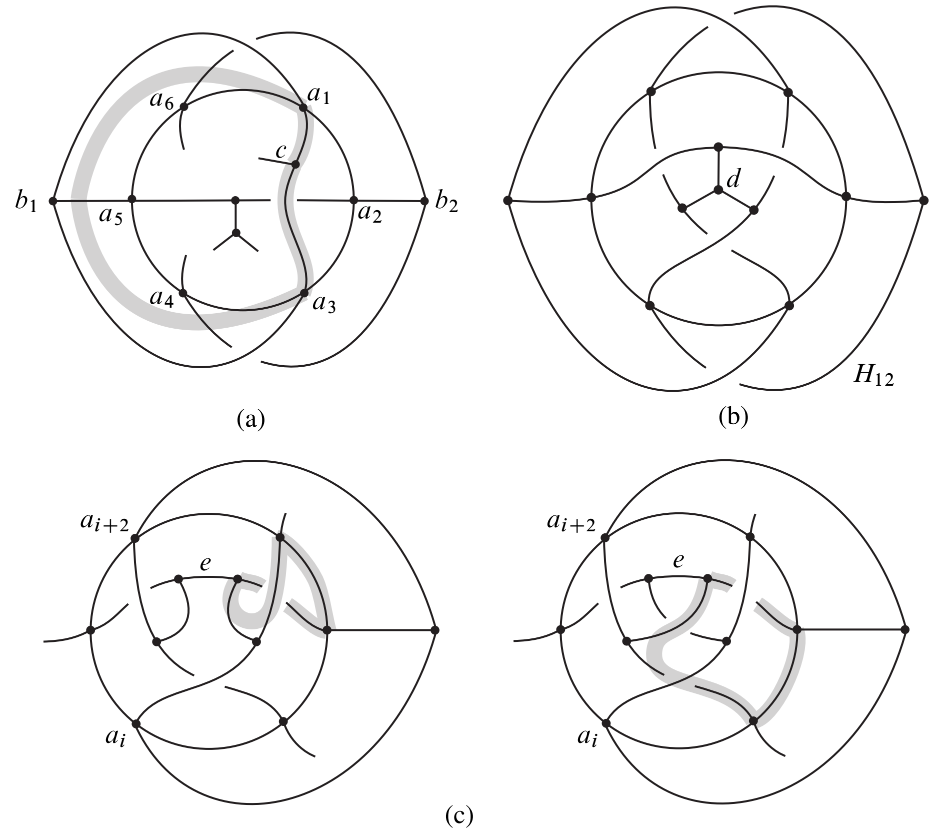


Figure 9: Constructing H_{12}

that $\widehat{G}_{d_1,d_2}$ is not homeomorphic to $K(3,3)$. If $V_4(a_1,a_2) = \{d_1,d_2,d_3\}$, then $V(d_1) \cup V(d_2) \cup V(d_3) \setminus \{a_1,a_2\}$ is contained in $\{a_3,a_4,b_1,b_2\}$, where a_3 and a_4 are the remaining two vertices of degree 4 other than $V(a_1) \cup V(a_2) \cup \{a_1,a_2\}$. Thus each $V(d_i) \setminus \{a_1,a_2\}$ is the set of two vertices among $\{a_3,a_4,b_1,b_2\}$. This implies that $|V_3(d_i,d_j)| + |V_4(d_i,d_j)| \geq 4$ for some $i, j = 1,2,3$, implying $|\widehat{E}_{d_i,d_j}| \leq 9$. Since at least one of three vertices $V(a_1) \cup V(a_2) \setminus \{d_i,d_j\}$ still has degree 4 in $\widehat{G}_{d_i,d_j}$, it follows that $\widehat{G}_{d_i,d_j}$ is not homeomorphic to $K(3,3)$. Finally, if $|V_4(a_1,a_2)| = 4$, then $|\widehat{E}_{a_1,a_2}| \leq 9$. Since $\widehat{G}_{a_1,a_2}$ still has the remaining three vertices of degree 4, it follows that $\widehat{G}_{a_1,a_2}$ is not homeomorphic to $K(3,3)$.

5 deg(a) = 3

Since we are working on the graph with 21 edges and every vertex has degree 3, there are exactly 14 vertices. First, suppose that there exists a pair of vertices a and b with

$\operatorname{dist}(a,b) \geq 4$. Then $E^2(a)$ and $E^2(b)$ can share vertices, but they do not share edges in common. Since $|E^2(a) \cup E^2(b)| = 18$ and $|V(a) \cup V(b) \cup \{a,b\}| = 8$, the 18 endpoints of $E^2(a)$, $E^2(b)$, and three extra edges which are $E \setminus \{E^2(a) \cup E^2(b)\}$, meet at six vertices. If any two edges of $E^2(a) \setminus E(a)$ (and similarly for b) are incident to one vertex c of these six vertices, take the unique vertex d of $V(a)$ which is not an endpoint of these two edges. Then $NE(b,d) = 6$ and $NV_3(b,d) = 6$, implying $|\widehat{E}_{b,d}| = 9$. But $\widehat{G}_{b,d}$ has a triangle containing c and the two vertices of $V(a) \setminus \{d\}$, so it follows that $\widehat{G}_{b,d}$ is not homeomorphic to $K(3,3)$. If not, each of these six vertices is a common endpoint of one edge of $E^2(a)$, one edge of $E^2(b)$, and one extra edge. Now, take an extra edge e and let b_1 and b_2 be the two vertices of $V(b)$ which have distance 1 from the endpoints of e. Let b_3 be the remaining vertex of $V(b)$. Then $NE(b_1,b_2) = 6$, $NV_3(b_1,b_2) = 5$, and $V_Y(b_1,b_2) = \{b_3\}$, implying $|\widehat{E}_{b_1,b_2}| = 9$. But $\widehat{G}_{b_1,b_2}$ has a triangle containing a and two vertices of $V(a)$, so it follows that $\widehat{G}_{b_1,b_2}$ is not homeomorphic to $K(3,3)$. See Figure 10(a).

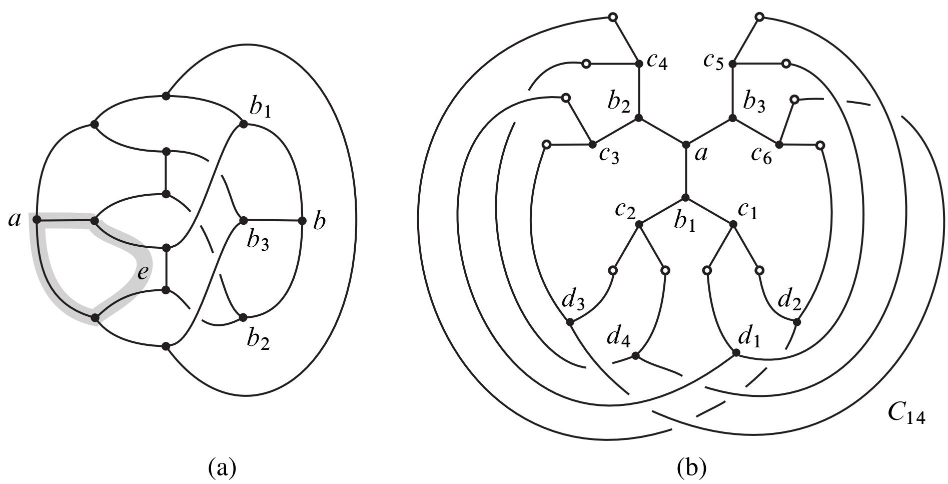


Figure 10: Constructing C_{14}

Therefore, we assume that the distance between any pair of vertices cannot exceed 3. Now we construct the intrinsically knotted graph G satisfying these conditions. Take a vertex a and let $V(a) = \{b_1, b_2, b_3\}$ and $V(b_i) = \{a, c_{2i-1}, c_{2i}\}$ for $i = 1,2,3$. As in Figure 10(b), the graph $E(a) \cup E(c_1) \cup \cdots \cup E(c_6)$ consists of 21 edges and 22 vertices. We show this is the only way to draw the graph with 21 edges such that all vertices have distance at most 3 from a and 10 vertices $a, b_1, b_2, b_3, c_1, \ldots, c_5$, and c_6 have degree 3. Now we join 12 white dots in Figure 10(b) into 4 groups indicating the remaining 4 vertices by d_1, d_2, d_3 and d_4. Thus each $V(d_j)$, $j = 1,2,3,4$, has three vertices among $c_1, \ldots, c_6$. Since the distance between any c_i and $c_{i'}$ cannot exceed 3, the following two properties must be satisfied. The first property is that $V(d_j)$ contains exactly one vertex from each group $\{c_{2i-1}, c_{2i}\}$ for $i = 1, 2, 3$. For example, if $V(d_1) = \{c_1, c_2, c_3\}$ (ie two vertices from the group $\{c_1, c_2\}$), then we can connect c_1 to at most two vertices among $\{c_4, c_5, c_6\}$ through some $E(d_j)$. This means that the distance between c_1 and one among $\{c_4, c_5, c_6\}$ exceeds 3. The second property is that different $V(d_j)$ and $V(d_{j'})$ share at most one vertex. For example, if they share two vertices c_1 and c_3, then $\text{dist}(c_1, c_4) = 4$. From these two properties, without loss of generality, we may say that

$$V(d_1) = \{c_1, c_3, c_5\}, \quad V(d_2) = \{c_1, c_4, c_6\},$$
$$V(d_3) = \{c_2, c_3, c_6\}, \quad V(d_4) = \{c_2, c_4, c_5\}$$

as drawn in Figure 10(b). This graph is exactly C_{14}.

References

[1] **P Blain**, **G Bowlin**, **T Fleming**, **J Foisy**, **J Hendricks**, **J Lacombe**, *Some results on intrinsically knotted graphs*, J. Knot Theory Ramifications 16 (2007) 749–760 MR2341314

[2] **J H Conway**, **C M Gordon**, *Knots and links in spatial graphs*, J. Graph Theory 7 (1983) 445–453 MR722061

[3] **N Goldberg**, **T W Mattman**, **R Naimi**, *Many, many more intrinsically knotted graphs*, Algebr. Geom. Topol. 14 (2014) 1801–1823 MR3212585

[4] **R Hanaki**, **R Nikkuni**, **K Taniyama**, **A Yamazaki**, *On intrinsically knotted or completely 3–linked graphs*, Pacific J. Math. 252 (2011) 407–425 MR2860431

[5] **B Johnson**, **M E Kidwell**, **T S Michael**, *Intrinsically knotted graphs have at least 21 edges*, J. Knot Theory Ramifications 19 (2010) 1423–1429 MR2746195

[6] **T Kohara**, **S Suzuki**, *Some remarks on knots and links in spatial graphs*, from: "Knots 90", (A Kawauchi, editor), de Gruyter, Berlin (1992) 435–445 MR1177440

[7] **R Motwani**, **A Raghunathan**, **H Saran**, *Constructive results from graph minors: Linkless embeddings*, from: "Proc. 29th annual symposium on foundations of computer science", IEEE, Piscataway, NJ (1988) 398–409

[8] **M Ozawa**, **Y Tsutsumi**, *Primitive spatial graphs and graph minors*, Rev. Mat. Complut. 20 (2007) 391–406 MR2351115

[9] **N Robertson**, **P D Seymour**, *Graph minors, XX: Wagner's conjecture*, J. Combin. Theory Ser. B 92 (2004) 325–357 MR2099147

ML,HK,SO: *Department of Mathematics, Korea University*
Anam-dong, Sungbuk-ku, Seoul 136-701, South Korea

HJL: *Department of Mathematical Sciences, KAIST*
291 Daehak-ro, Yuseong-gu, Daejeon 305-701, South Korea

mjmlj@korea.ac.kr, kimhjun@korea.ac.kr, hjwith@kaist.ac.kr,
seungsang@korea.ac.kr

Received: 9 June 2014 Revised: 13 March 2015

Algebraic & Geometric Topology 15 (2015) 3323–3353

Equivalence classes of augmentations and Morse complex sequences of Legendrian knots

MICHAEL B HENRY
DAN RUTHERFORD

Let L be a Legendrian knot in $\mathbb{R}^3$ with the standard contact structure. In earlier work of Henry, a map was constructed from equivalence classes of Morse complex sequences for L, which are combinatorial objects motivated by generating families, to homotopy classes of augmentations of the Legendrian contact homology algebra of L. Moreover, this map was shown to be a surjection. We show that this correspondence is, in fact, a bijection. As a corollary, homotopic augmentations determine the same graded normal ruling of L and have isomorphic linearized contact homology groups. A second corollary states that the count of equivalence classes of Morse complex sequences of a Legendrian knot is a Legendrian isotopy invariant.

57R17; 57M25, 53D40

1 Introduction

The symplectic techniques of holomorphic curves and generating families provide two effective classes of invariants of Legendrian knots in standard contact $\mathbb{R}^3$. The holomorphic curve approach, which in this low-dimensional setting takes on a combinatorial flavor, can be used to define a differential graded algebra (DGA). The DGA is known alternatively as the Legendrian contact homology DGA or the Chekanov–Eliashberg DGA and was originally defined by Chekanov [1] and Eliashberg, Givental and Hofer [5]. Generating families of Legendrian submanifolds in 1–jet spaces, including $\mathbb{R}^3$, have also been used to produce homological Legendrian invariants; see, for instance, Jordan and Traynor [13], Sabloff and Traynor [18] and Traynor [19; 20]. In addition to distinguishing Legendrian isotopy classes of knots, both the holomorphic and generating family invariants carry useful information about Lagrangian cobordisms, see Ekholm, Honda and Kálmán [4] and Sabloff and Traynor [18].

For Legendrian knots in $\mathbb{R}^3$, several close connections have been discovered between holomorphic curve and generating family invariants, although many questions remain. For example, the existence of a linear at infinity generating family for a Legendrian knot is known to be equivalent to the existence of a certain DGA morphism, called an

Published: 12 January 2016 DOI: 10.2140/agt.2015.15.3323

augmentation, from the Chekanov–Eliashberg DGA to its ground ring; see Pushkar' and Chekanov [3], Fuchs [7], Fuchs and Ishkhanov [8], Fuchs and Rutherford [9] and Sabloff [17]. However, it is unknown if this statement can be strengthened to a bijective correspondence between appropriate equivalence classes of generating families and augmentations. In this article, we approach this question using a discrete analog of a generating family called a Morse complex sequence, abbreviated MCS. MCSs have proven to be more tractable for explicit construction and computation; see, for example, Henry [10] and Henry and Rutherford [11; 12]. Section 2.2 sketches the connection between generating families and Morse complex sequences; a more complete description can be found in [11].

The concept of a Morse complex sequence originally appeared in unpublished work of Petya Pushkar, and first appears in print in the work of the first author [10] where MCSs are studied in connection with augmentations. In [10], a surjective map is defined from MCSs of L to augmentations of the Chekanov–Eliashberg DGA of L. Moreover, equivalent MCSs are mapped to homotopic augmentations. In the present article, we complement the results of [10] by showing in Lemma 3.1 that two MCSs mapped to homotopic augmentations must, in fact, be equivalent as MCSs. Combined with [10] this gives the following.

Theorem 1.1 *For any Legendrian knot $L \subset \mathbb{R}^3$ with generic front diagram, there is a bijection between equivalence classes of Morse complex sequences for L and homotopy classes of augmentations of the Chekanov–Eliashberg DGA of L.*

As a consequence, the number of MCS equivalence classes is a Legendrian isotopy invariant; see Corollary 4.1. The less immediate Corollary 4.2 combines Theorem 1.1 with previous work of the authors from [11] to deduce that homotopic augmentations must have isomorphic linearized homology groups. The set of linearized homology groups is a Legendrian isotopy invariant. Corollary 4.2 allows for a refinement of this invariant by considering multiplicities.

The remainder of the article is organized as follows. Section 2 recalls background concerning augmentations and Morse complex sequences. Section 3 contains the proof of Theorem 1.1 and Section 4 includes three corollaries.

Acknowledgments We gratefully acknowledge The Royal Academy of Belgium where conversations between the authors on this project began in August 2013 at a workshop on Legendrian submanifolds, holomorphic curves, and generating families. We also thank the workshop organizer Frédéric Bourgeois. In addition, we thank the American Institute of Mathematics for supporting a SQuaRE research group on augmentations and related topics, and we thank our fellow SQuaRE participants Dmitry Fuchs, Paul Melvin, Josh Sabloff, and Lisa Traynor.

2 Background

A *Legendrian knot* in the standard contact structure on $\mathbb{R}^3$ is a smooth knot $L\colon S^1 \to \mathbb{R}^3$ satisfying $L'(t) \in \ker(dz - y\,dx)$ for all $t \in S^1$. A smooth one-parameter family L_t, $0 \le t \le 1$, of Legendrian knots is a *Legendrian isotopy* between L_0 and L_1. The *front diagram* of L is the projection of L to the xz–plane. Every Legendrian knot is Legendrian isotopic, by an arbitrarily small Legendrian isotopy, to a Legendrian knot whose front diagram is embedded except at transverse self-intersections, called *crossings*, and semi-cubical cusps such that, in addition, all of these exceptional points have distinct x–coordinates. A Legendrian knot with such a front diagram is said to have a σ*–generic* front diagram; see, for example, the front diagram in Figure 1. In a neighborhood of an x value that is not the x–coordinate of a crossing or cusp, the front diagram looks like a collection of non-intersecting line segments commonly called the *strands* of D at x. Orient L. The *rotation number* $r(L)$ is $(d-u)/2$, where d (resp. u) is the number of cusps at which the orientation travels downward (resp. upward) with respect to the z–axis.

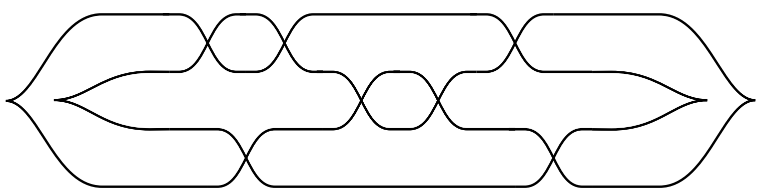


Figure 1: A σ–generic front diagram of a Legendrian knot with rotation number 0

2.1 Chekanov–Eliashberg algebra

Fix a Legendrian knot L with σ–generic front diagram D and rotation number 0. A *Maslov potential* is a map $\mu\colon L \to \mathbb{Z}$ that is constant except at cusp points of L where the Maslov potential of the lower strand of the cusp is one less than the upper strand. Let $A(D)$ be the $\mathbb{Z}/2\mathbb{Z}$ vector space generated by the labels $Q = \{q_1, \dots, q_n\}$ assigned to the crossings and right cusps of D. A generator $q \in Q$ is assigned a *grading* $|q|$, also called a *degree*, so that $|q|$ is 1 if q is a right cusp and, otherwise, $|q|$ is $\mu(T) - \mu(B)$ where T and B are the strands of D crossing at q and T has smaller slope. The graded algebra $\mathcal{A}(D)$ is the unital tensor algebra $TA(D)$. The *Chekanov–Eliashberg algebra*, written $(\mathcal{A}(D), \partial)$, is the algebra $\mathcal{A}(D)$ along with a degree -1 differential $\partial\colon \mathcal{A}(D) \to \mathcal{A}(D)$ that, in the case of the front diagram description from [15], is defined by counting certain admissible maps of the two-disk D^2 into the xz–plane.

Definition 2.1 below defines only those admissible maps needed in this article; we refer the reader to [15] for a complete definition of ∂.

An *augmentation* is an algebra homomorphism $\epsilon\colon \mathcal{A}(D) \to \mathbb{Z}/2\mathbb{Z}$ satisfying $\epsilon \circ \partial = 0$, $\epsilon(1) = 1$, and $\epsilon(q) = 1$ only if $|q| = 0$. The set $\mathrm{Aug}(D)$ is the set of all augmentations of $(\mathcal{A}(D), \partial)$. We say a crossing q is *augmented* by ϵ if $\epsilon(q) = 1$. An augmentation can be thought of as a morphism between the differential graded algebra $(\mathcal{A}(D), \partial)$ and the differential graded algebra $(\mathbb{Z}/2\mathbb{Z}, \partial')$ whose only non-zero element is in degree 0 and where $\partial' = 0$. From this perspective, there is a natural algebraic equivalence relation on $\mathrm{Aug}(D)$. Given ϵ and ϵ' in $\mathrm{Aug}(D)$, a *chain homotopy* from ϵ to ϵ' is a degree 1 linear map $H\colon (\mathcal{A}(D), \partial) \to (\mathbb{Z}/2\mathbb{Z}, \partial')$ satisfying $\epsilon - \epsilon' = \partial' \circ H + H \circ \partial$ and $H(ab) = H(a)\epsilon'(b) + (-1)^{|a|}\epsilon(a)H(b)$ for all $a, b \in \mathcal{A}(D)$. Since we are working over $\mathbb{Z}/2\mathbb{Z}$ and $\partial' = 0$, these conditions simplify to

$$\epsilon - \epsilon' = H \circ \partial \quad \text{and} \quad H(ab) = H(a)\epsilon'(b) + \epsilon(a)H(b). \tag{1}$$

By [14, Lemma 2.18], a chain homotopy H is determined by the values it takes on the degree -1 crossings of D.

We say augmentations ϵ and ϵ' are *homotopic* and write $\epsilon \simeq \epsilon'$ if there exists a chain homotopy from ϵ to ϵ'. As the notation implies and as is proven in [6], chain homotopy provides an equivalence relation on the set $\mathrm{Aug}(D)$. We let $\mathrm{Aug}^{\mathrm{ch}}(D)$ be $\mathrm{Aug}(D)/\simeq$. By [10, Proposition 4.5], the count of homotopy classes of augmentations is a Legendrian isotopy invariant.

Suppose ϵ and ϵ' are augmentations in $\mathrm{Aug}(D)$ and there exists a chain homotopy H from ϵ to ϵ'. Suppose q is a degree 0 crossing and $\langle \partial q, \prod_{i=1}^m q_{k_i} \rangle$ is 1, where $\langle \partial q, \prod_{i=1}^m q_{k_i} \rangle$ is the coefficient of $\prod_{i=1}^m q_{k_i}$ in ∂q. Then, by Equation (1),

$$\begin{aligned}
(\epsilon - \epsilon')(q) = H \circ \partial(q) &= H\Big(\prod_{i=1}^m q_{k_i} + \cdots\Big) \\
&= H\Big(\prod_{i=1}^m q_{k_i}\Big) + H(\cdots) \\
&= \sum_{j=1}^m \Big[\Big(\prod_{i=1}^{j-1} \epsilon(q_{k_i})\Big) H(q_{k_j}) \Big(\prod_{i=j+1}^m \epsilon'(q_{k_i})\Big)\Big] + H(\cdots).
\end{aligned}$$

At most one term in the sum

$$\sum_{j=1}^m \Big[\Big(\prod_{i=1}^{j-1} \epsilon(q_{k_i})\Big) H(q_{k_j}) \Big(\prod_{i=j+1}^m \epsilon'(q_{k_i})\Big)\Big]$$

may be non-zero, since ϵ and ϵ' are non-zero only on generators of degree 0 and H is non-zero only on generators of degree -1. Note that, for a fixed $j \in \{1, \dots, m\}$, the term

$$\Big(\prod_{i=1}^{j-1} \epsilon(q_{k_i})\Big) H(q_{k_j}) \Big(\prod_{i=j+1}^{m} \epsilon'(q_{k_i})\Big)$$

is non-zero if and only if $H(q_{k_j}) = 1$ holds and for $1 \le i \le j-1$ (resp. $j+1 \le i \le m$), the crossing q_{k_i} is augmented by ϵ (resp. ϵ').

The monomials $\prod_{i=1}^m q_{k_i}$ appearing in $\partial(q)$ correspond to certain mappings of the two-disk D^2 into the xz–plane that are immersions except for allowable exceptions along ∂D^2. Only monomials containing generators of degree 0 or -1 are relevant for our purposes. Therefore, we present only the description of such disks in the following definitions. Note that this restriction allows us to rule out some additional behaviors of ∂D^2 near right cusps that appear in [15] and lead to monomials that contain generators of degree 1.

Let D^2 be the disk of radius 1 centered at the origin in $\mathbb{R}^2$. Choose m points from $\partial D^2 \setminus \{(1,0)\}$. Label the chosen points $\{b_1, \dots, b_m\}$ counter-clockwise with b_1 the first point counter-clockwise from $(1,0)$.

Definition 2.1 In terms of the notation above, a $(0,-1)$*–admissible disk* is a continuous map from D^2 into the xz–plane that maps ∂D^2 to the front diagram D and is a smooth orientation preserving immersion when restricted to the interior of D^2 satisfying the following conditions:

(1) The mapping takes $(1,0)$ to a degree 0 crossing q and the image of f in a neighborhood of $(1,0)$ looks as in Figure 2(a). We say the $(0,-1)$–admissible disk *originates at* q.

(2) For exactly one $1 \le j \le m$, $f(b_j)$ is a degree -1 crossing q_{k_j} and the image of f in a neighborhood of b_j looks as in Figure 2(d) or (e).

(3) For all $i \neq j$, $f(b_i)$ is a degree 0 crossing q_{k_i} and the image of f in a neighborhood of b_i looks as in Figure 2(d) or (e).

(4) Along ∂D^2 the mapping is smooth except at $\{b_1, \dots, b_m\} \cup \{(1,0)\}$ as described in (1)–(3) and at points in $\partial D^2 \setminus (\{b_1, \dots, b_m\} \cup \{(1,0)\})$ where the image of f looks like either Figure 2(b) or (c).

We say the $(0,-1)$–admissible disk has *convex corners* at $q_{k_1}, \dots, q_{k_m}$. The $(0,-1)$–admissible disk is assigned the monomial $\prod_{i=1}^m q_{k_i}$. We say a $(0,-1)$–admissible disk is an (ϵ, ϵ', H)*–admissible disk* if, for some $1 \le j \le m$, $H(q_{k_j}) = 1$ holds and

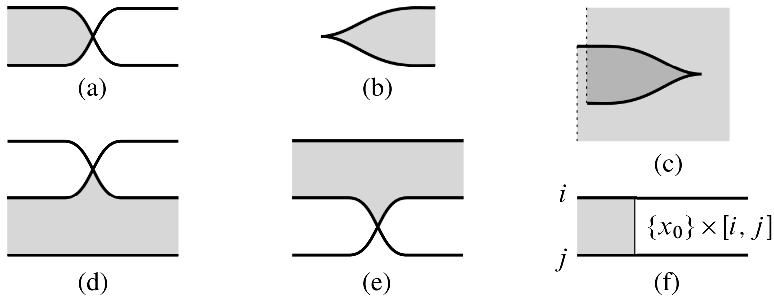


Figure 2: The possible singularities of the disk in Definition 2.1 and the half-disks in Definitions 3.3 and 3.2. The crossings in (d) and (e) are called convex corners. Near a boundary point that maps to a right cusp the image of a disk overlaps itself as indicated in (c) by the darkly shaded region.

for $1 \le i \le j-1$ (resp. $j+1 \le i \le m$), the crossing q_{k_i} is augmented by ϵ (resp. ϵ'); see Figure 3.

Henceforth, we consider admissible disks up to orientation preserving reparametrization of the domain (fixing $\{b_1, \dots, b_m\} \cup \{(1,0)\}$), and all counts of disks are up to this equivalence relation.

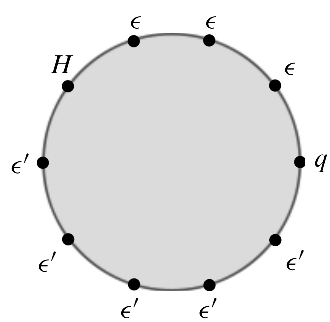


Figure 3: The domain of an (ϵ, ϵ', H)–admissible disk with labels indicating marked points mapped to crossings augmented by ϵ and ϵ' and the marked point mapped to the crossing satisfying $H(q_{k_j}) = 1$

The restrictions on the types of non-smooth points of an $(0,-1)$–admissible disk imply that q is the right-most point of the disk. From [15, Section 2], when a single q_{k_j} has degree -1 while q and all of the remaining q_{k_i} have degree 0, $\langle \partial q, \prod_{i=1}^m q_{k_i} \rangle = 1$ holds if and only if there are an odd number of $(0,-1)$–admissible disks originating at q and with monomial $\prod_{i=1}^m q_{k_i}$. Proposition 2.2 follows directly from the discussion above.

Proposition 2.2 *Suppose D is a σ–generic front diagram of a Legendrian knot and ϵ and ϵ' are augmentations in* $\mathrm{Aug}(D)$*. If q is a degree 0 crossing and H is a chain homotopy from ϵ to ϵ', then ϵ and ϵ' differ at q if and only if there are an odd number of (ϵ, ϵ', H)–admissible disks originating at q.*

2.2 Morse complex sequences

We briefly sketch the connection between generating families and Morse complex sequences and refer the reader to [11] for more details. A one-parameter family of smooth functions $f_x\colon \mathbb{R}^N \to \mathbb{R}$, parametrized by $x \in \mathbb{R}$, is a generating family for a Legendrian knot L with front diagram D if

$$D = \left\{(x,z) : z = f_x(e) \text{ for some } e \in \mathbb{R}^N \text{ satisfying } \frac{\partial f_x}{\partial e}(e) = 0\right\}.$$

With an appropriately chosen metric, a generic $x \in \mathbb{R}$ determines a Morse chain complex (C_x, d_x) on $\mathbb{R}^N$ and, as x varies, the evolution of the Morse complexes of f_x are well-understood; a cusp of D corresponds to the creation or elimination of a canceling pair of critical points and a crossing corresponds to two critical points exchanging critical values. As x varies, it is also possible for a fiberwise gradient flowline to momentarily flow between two critical points of the same index. Such an occurrence is called a handleslide and it determines an explicit chain isomorphism between successive Morse complexes. In summary, a generating family and choice of metric determine a one-parameter family of Morse chain complexes and the relationship between successive chain complexes is determined by the crossings and cusps of D and the handleslides. A Morse complex sequence on D is a finite sequence of chain complexes (C_m, d_m) and vertical marks on D that are meant to correspond to the Morse chain complexes and handleslides of a generating family and choice of metric. In addition, varying the choice of metric motivates an equivalence relation on MCSs.

Fix a Legendrian knot L with σ–generic front diagram D, rotation number 0, and Maslov potential μ. Theorem 1.1 proves that a certain surjective map in [10] from equivalence classes of Morse complex sequences to $\mathrm{Aug}^{\mathrm{ch}}(D)$ is, in fact, a bijection. We will use the definition of a Morse complex sequence given in [11]. This definition differs slightly from the definition in [10], but both definitions determine the same set of objects on L.

A *handleslide* on D is a vertical line segment disjoint from all crossings and cusps and with endpoints on strands of D that have the same Maslov potential.

Definition 2.3 A *Morse complex sequence* on a σ–generic front diagram D is the triple $\mathcal{C} = (\{(C_m, d_m)\}, \{x_m\}, H)$ satisfying:

(1) H is a set of handleslides on D.

(2) The real values $x_1 < x_2 < \cdots < x_M$ are x–coordinates distinct from the x–coordinates of crossings and cusps of D and handleslides of H. For each $1 \le m < M$, the set $\{(x,z) : x_m \le x \le x_{m+1}\}$ contains a single crossing, cusp or handleslide. The set $\{(x,z) : -\infty < x \le x_1\}$ contains the left-most left cusp and the set $\{(x,z) : x_M \le x < \infty\}$ contains the right-most right cusp.

(3) For each $1 \le m \le M$, the points of intersection of the vertical line $\{x_m\} \times \mathbb{R}$ and D are labeled $e_1, e_2, \dots, e_{s_m}$ from top to bottom. The vector space C_m is the $\mathbb{Z}$–graded $\mathbb{Z}/2\mathbb{Z}$ vector space generated by $e_1, e_2, \dots, e_{s_m}$, where the degree of each generator is the value of the Maslov potential on the corresponding strand of D, $|e_i| = \mu(e_i)$. The map $d_m \colon C_m \to C_m$ is a degree -1 differential that is triangular in the sense that

$$d_m e_i = \sum_{i<j} c_{ij} e_j, \quad c_{ij} \in \mathbb{Z}/2\mathbb{Z}.$$

(4) The coefficients $\langle d_1 e_1, e_2 \rangle$ and $\langle d_M e_1, e_2 \rangle$ are both 1. Suppose $1 \le m < M$ and let T be the tangle $D \cap \{(x,z) : x_m \le x \le x_{m+1}\}$. If T contains a left (resp. right) cusp between strands k and $k+1$, then $\langle d_{m+1} e_k, e_{k+1} \rangle$ is 1 (resp. $\langle d_m e_k, e_{k+1} \rangle$ is 1). If T contains a crossing between strands k and $k+1$, then $\langle d_m e_k, e_{k+1} \rangle$ is 0.

(5) For $1 \le m < M$, the crossing, cusp, or handleslide mark in the tangle $T = D \cap \{(x,z) : x_m \le x \le x_{m+1}\}$ determines an algebraic relationship between the chain complexes (C_m, d_m) and (C_{m+1}, d_{m+1}) as follows:

(a) **Crossing** If the crossing is between strands k and $k+1$, then the map $\phi \colon (C_m, d_m) \to (C_{m+1}, d_{m+1})$ defined by

$$\phi(e_i) = \begin{cases} e_i & \text{if } i \notin \{k, k+1\}, \\ e_{k+1} & \text{if } i = k, \\ e_k & \text{if } i = k+1 \end{cases}$$

is an isomorphism of chain complexes.

(b) **Right cusp** If the right cusp is between strands k and $k+1$, then the linear map

$$\phi(e_i) = \begin{cases} [e_i] & \text{if } i < k, \\ [e_{i+2}] & \text{if } i \ge k \end{cases}$$

is an isomorphism of chain complexes from (C_{m+1}, d_{m+1}) to the quotient of (C_m, d_m) by the acyclic subcomplex generated by $\{e_k, d_m e_k\}$.

(c) **Left cusp** The case of a left cusp is the same as the case of a right cusp, though the roles of (C_m, d_m) and (C_{m+1}, d_{m+1}) are reversed.

(d) **Handleslide** If the handleslide mark has endpoints on strands k and l with $k < l$, then the map $h_{k,l}\colon (C_m, d_m) \to (C_{m+1}, d_{m+1})$ defined by

$$h_{k,l}(e_i) = \begin{cases} e_i & \text{if } i \neq k, \\ e_k + e_l & \text{if } i = k \end{cases}$$

is an isomorphism of chain complexes.

The set $\mathrm{MCS}(D)$ is the set of all Morse complex sequences on D.

Remark 2.4 Morse complex sequences may be defined over more general coefficient rings than $\mathbb{Z}/2\mathbb{Z}$; see [12]. We restrict attention to $\mathbb{Z}/2\mathbb{Z}$ coefficients as this is also done in [10].

Definition 2.5 An MCS $\mathcal{C} = (\{(C_m, d_m)\}, \{x_m\}, H)$ in $\mathrm{MCS}(D)$ has *simple left cusps* if, for each tangle $T = \{(x,z) : x_m \leq x \leq x_{m+1}\}$ containing a left cusp between strands k and $k+1$, the chain complex (C_{m+1}, d_{m+1}) satisfies $\langle d_{m+1} e_k, e_i \rangle = \langle d_{m+1} e_{k+1}, e_i \rangle = 0$ for all $k+1 < i$ and $\langle d_{m+1} e_j, e_{k+1} \rangle = \langle d_{m+1} e_j, e_k \rangle = 0$ for all $j < k$.

The subset $\mathrm{MCS}_b(D) \subset \mathrm{MCS}(D)$ denotes the set of MCSs with simple left cusps. We use the letter b to be consistent with the notation of [10], where a left cusp is also called a "birth". This language is meant to draw a connection to the creation of a canceling pair of critical points, often called a birth, in a one-parameter family of Morse functions on a manifold.

Given an MCS $\mathcal{C} = (\{(C_m, d_m)\}, \{x_m\}, H)$ with simple left cusps, the chain complexes $\{(C_m, d_m)\}$ are uniquely determined by the crossings and cusps of D, the handleslides H, and requirements (5) (a)–(d) of Definition 2.3. Consequently, $\mathcal{C}$ may be represented visually by placing the handleslide marks H on the front diagram D; see Figure 4.

In [10] an equivalence relation on the set $\mathrm{MCS}(D)$ is defined that is motivated by a corresponding equivalence for generating families; see also [11]. Here we denote the set of equivalence classes of this relation by $\widehat{\mathrm{MCS}}(D) = \mathrm{MCS}(D)/\simeq$. We recall a version of this equivalence relation that applies to the more restricted set of MCSs with simple left cusps, $\mathrm{MCS}_b(D)$. We denote equivalence classes with respect to this relation by $\widehat{\mathrm{MCS}}_b(D)$. By [10, Proposition 3.17], the map from $\widehat{\mathrm{MCS}}_b(D)$ to $\widehat{\mathrm{MCS}}(D)$ induced by the inclusion $\mathrm{MCS}_b(D) \subset \mathrm{MCS}(D)$ is a bijection. Therefore, to prove Theorem 1.1, we need only consider MCSs in $\mathrm{MCS}_b(D)$ and MCS classes in $\widehat{\mathrm{MCS}}_b(D)$.

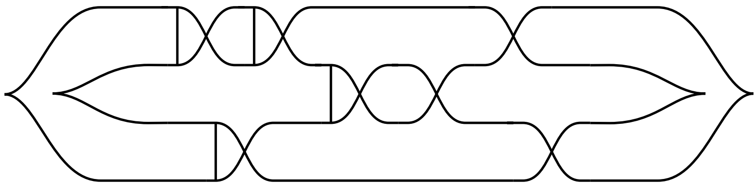


Figure 4: An MCS with simple left cusps. This MCS is also in A–form.

The equivalence relation on $\mathrm{MCS}_b(D)$ is generated by the *MCS moves* pictured in Figures 5 and 6. The numbering indicated will be used throughout this article. Additional moves result from reflecting each of the two figures in (3), (7), (9), (10) and (12) of Figure 5 about a horizontal axis and reflecting each of the two figures in (4), (9), (11) and (12) of Figure 5 about a vertical axis. The handleslide modification that results from reflecting Figure 5 (10) about a vertical axis is *not* an MCS move for MCSs with simple left cusps. (The absence of this reflected move is the only difference between the definitions of the equivalence relations on $\mathrm{MCS}_b(D)$ and $\mathrm{MCS}(D)$ discussed in the previous paragraph.) MCS move (13) requires explanation. Suppose $\mathcal{C} = (\{(C_m, d_m)\}, \{x_m\}, H)$ is an MCS on D and suppose there exist x_m and $1 \leq k < l \leq s_m$ such that $\mu(e_k) = \mu(e_l) - 1$. Then MCS move (13) introduces the collection of handleslides K defined as follows. The handleslides in K are of two types. First, if $i < k$ and $\langle d_m e_i, e_k \rangle = 1$ holds, then K contains a handleslide with endpoints on i and l. Second, if $l < j$ and $\langle d_m e_l, e_j \rangle = 1$ holds, then K contains a handleslide with endpoints on k and j.

By [10, Proposition 3.8], modifying the handleslide set of an MCS in $\mathrm{MCS}_b(D)$ as in one of the cases in Figures 5 and 6 results in another MCS in $\mathrm{MCS}_b(D)$. Therefore, the notion of equivalence in the following definition is well-defined. In addition, if an MCS move is applied to an MCS, then only those chain complexes near the location of the MCS move are affected. In other words, the MCS moves are local in the sense that they change both the handleslides and chain complexes of an MCS only in a local neighborhood.

Definition 2.6 Two MCSs $\mathcal{C}$ and $\mathcal{C}'$ in $\mathrm{MCS}_b(D)$ are *equivalent*, written $\mathcal{C} \simeq \mathcal{C}'$, if there exists a sequence $\mathcal{C}_1, \mathcal{C}_2, \dots, \mathcal{C}_s$ in $\mathrm{MCS}_b(D)$ so that $\mathcal{C} = \mathcal{C}_1$, $\mathcal{C}' = \mathcal{C}_s$, and, for all $1 \leq i < s$, the set of handleslide marks of $\mathcal{C}_i$ and $\mathcal{C}_{i+1}$ differ by exactly one MCS move. The set $\widehat{\mathrm{MCS}}_b(D)$ is the set $\mathrm{MCS}_b(D)/\simeq$.

MCSs of the following type have a standard form that makes their relationship with augmentations particularly simple to describe.

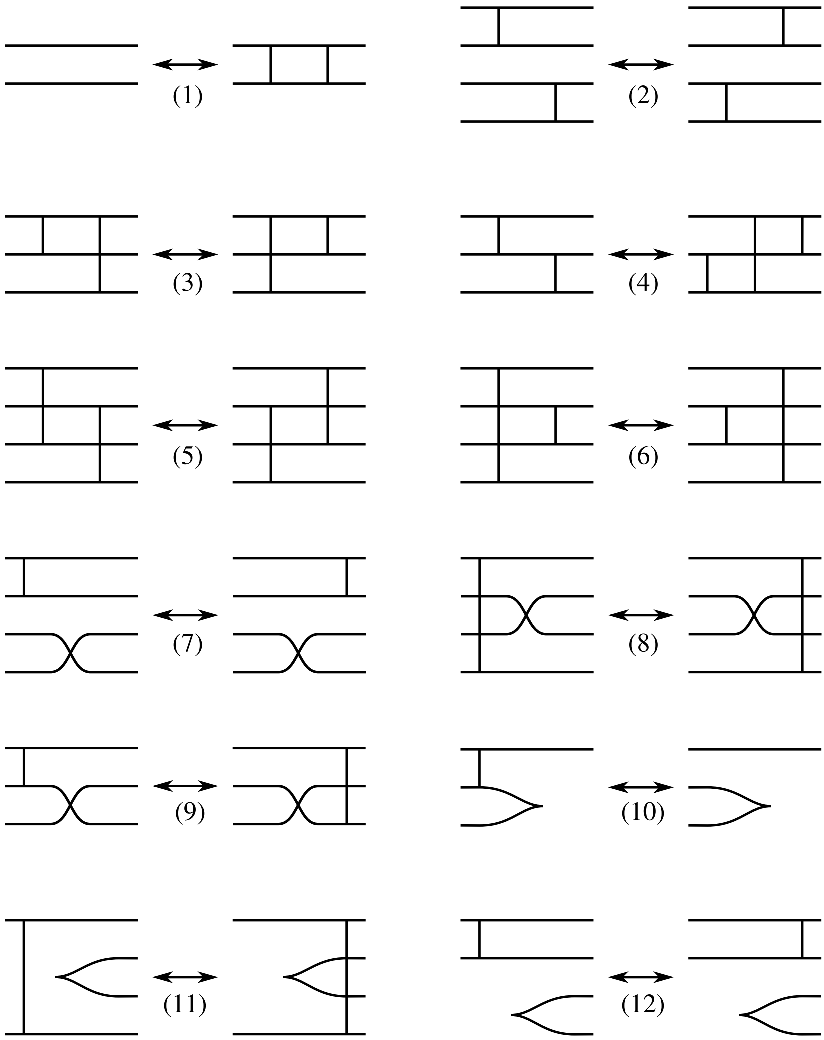


Figure 5: Handleslide modifications, called MCS moves, that result in an equivalent MCS

Definition 2.7 An MCS $\mathcal{C}$ in $\mathrm{MCS}_b(D)$ is in *A–form* if there exists a set R of degree 0 crossings so that just to the left of each q in R there is a handleslide with endpoints on the strands crossing at q and $\mathcal{C}$ has no other handleslides. A crossing q in R is said to be *marked*.

Figure 4 shows an MCS in A–form where R is the four left-most crossings. The subset $\mathrm{MCS}_A(D) \subset \mathrm{MCS}_b(D)$ consists of all A–form MCSs on D.

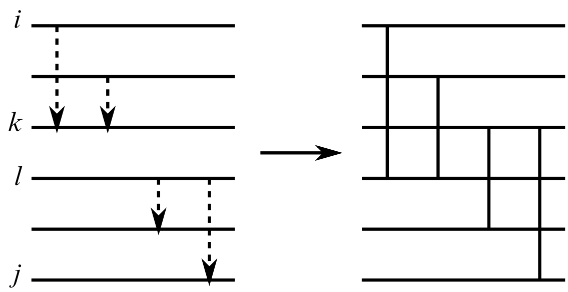


Figure 6: MCS move (13). On the left, a dotted arrow from strand α to strand β indicates that $\langle d_m e_\alpha, e_\beta\rangle$ is 1.

3 The main result

Suppose L is a Legendrian knot with σ–generic front diagram D, rotation number 0, and Maslov potential μ. The proof of Theorem 1.1 depends upon the following technical lemma, whose proof comprises most of this section.

Lemma 3.1 *Suppose D is a σ–generic front diagram and $\mathcal{C}$ and $\mathcal{C}'$ are in $\mathrm{MCS}_A(D)$ with corresponding augmentations $\epsilon_{\mathcal{C}}$ and $\epsilon_{\mathcal{C}'}$, respectively. If $\epsilon_{\mathcal{C}}$ and $\epsilon_{\mathcal{C}'}$ are homotopic, then $\mathcal{C}$ and $\mathcal{C}'$ are equivalent as MCSs.*

We now prove Theorem 1.1, assuming Lemma 3.1. Section 4 includes three corollaries of Theorem 1.1.

Proof of Theorem 1.1 By [10, Proposition 3.17], the natural inclusion of $\mathrm{MCS}_b(D)$ into $\mathrm{MCS}(D)$ induces a bijection from $\widehat{\mathrm{MCS}}_b(D)$ to $\widehat{\mathrm{MCS}}(D)$. Therefore, it suffices to construct a bijection from $\widehat{\mathrm{MCS}}_b(D)$ to $\mathrm{Aug}^{\mathrm{ch}}(D)$. In [10, Section 6], a surjective map $\widehat{\Psi}$ is constructed from $\widehat{\mathrm{MCS}}_b(D)$ to $\mathrm{Aug}^{\mathrm{ch}}(D)$. We will prove this map is injective. By [10, Theorem 1.6], every MCS is equivalent to an A–form MCS. Therefore, every MCS equivalence class contains an A–form representative. We give the definition of $\widehat{\Psi}$ in terms of A–form representatives and, in so doing, avoid most of the technical details of [10]. By [10, Corollary 6.21], given an MCS class $[\mathcal{C}]$ with A–form representative $\mathcal{C}$, $\widehat{\Psi}([\mathcal{C}])$ is the augmentation homotopy class $[\epsilon_{\mathcal{C}}]$, where a degree 0 crossing q is augmented by $\epsilon_{\mathcal{C}}$ if and only if q is marked by $\mathcal{C}$. Lemma 3.1 shows that if $\epsilon_{\mathcal{C}_1}$ is homotopic to $\epsilon_{\mathcal{C}_2}$, then $\mathcal{C}_1$ is equivalent to $\mathcal{C}_2$. It follows that $\widehat{\Psi}$ is injective. □

Before proving Lemma 3.1, we require two definitions and a lemma. Let D^2 be the disk of radius 1 centered at the origin in $\mathbb{R}^2$. Choose $m+2$ points on ∂D^2. Label the chosen points $\{b_0, \dots, b_{m+1}\}$ counter-clockwise. Let γ be the arc of ∂D^2 with endpoints b_{m+1} and b_0 and so that $b_1, \dots, b_m$ are not in γ. Given $x_0 \in \mathbb{R}$ that is not the x–coordinate of any crossing or cusp of D, we let $\{x_0\} \times [i, j]$ denote the vertical line segment with x–coordinate x_0 and endpoints on strands i and j of D, where the strands of D above $x = x_0$ are numbered $1, 2, \dots$ from top to bottom and $i < j$.

Definition 3.2 Let ϵ and ϵ' be homotopic augmentations in $\mathrm{Aug}(D)$ and let H be a chain homotopy from ϵ to ϵ'. An (ϵ, ϵ', H)*–half-disk* is a mapping of the two-disk D^2 into the xz–plane as in Definition 2.1, except for the following variations along the boundary:

(1) The arc γ maps to a vertical line $\{x_0\} \times [i, j]$; see Figure 2(f). We say the (ϵ, ϵ', H)–half-disk *originates at* $\{x_0\} \times [i, j]$.

(2) For exactly one $1 \leq j \leq m$, $f(b_j)$ is a degree -1 crossing q_{k_j}, $H(q_{k_j}) = 1$ holds, and f has a convex corner at $f(b_j)$; see Figure 2(d) or (e).

(3) If $1 \leq i < j$ (resp. $j < i \leq m$), $f(b_i)$ is a degree 0 crossing augmented by ϵ (resp. ϵ') and f has a convex corner at $f(b_i)$.

(4) The restriction of f to ∂D^2 is smooth except at $\{b_0, \dots, b_{m+1}\}$ as described in (1) and (2) and at points in $\partial D^2 \setminus (\{b_0, \dots, b_{m+1}\})$ where the image of f looks like Figure 2(b) or (c).

The set $\mathcal{H}(x_0, [i, j])$ consists of all (ϵ, ϵ', H)–half-disks originating at $\{x_0\} \times [i, j]$ up to reparametrization, and $\#\mathcal{H}(x_0, [i, j])$ is the mod 2 count of elements in $\mathcal{H}(x_0, [i, j])$.

Definition 3.3 Let ϵ be an augmentation in $\mathrm{Aug}(D)$. An ϵ*–half-disk* is a mapping f of the two-disk D^2 into the xz–plane as in Definition 3.2 except that conditions (2) and (3) are replaced with the requirement that all convex corners are at crossings that are augmented by ϵ.

The set $\mathcal{G}^\epsilon(x_0, [i, j])$ consists of all ϵ–half-disks originating at $\{x_0\} \times [i, j]$ up to reparametrization, and $\#\mathcal{G}^\epsilon(x_0, [i, j])$ is the mod 2 count of elements in $\mathcal{G}^\epsilon(x_0, [i, j])$.

As in Definition 2.1, the points in the vertical line $\{x_0\} \times [i, j]$ are the right-most points of either an (ϵ, ϵ', H)–half-disk or an ϵ–half-disk. It follows from the definitions that $\mu(i) = \mu(j)$ in the case of an (ϵ, ϵ', H)–half-disk and $\mu(i) = \mu(j) + 1$ in the case of an ϵ–half-disk

By [10, Corollary 6.21], the map $\Phi\colon \mathrm{MCS}_A(D) \to \mathrm{Aug}(D)$ defined as follows is a bijection. Given $\mathcal{C} \in \mathrm{MCS}_A(D)$ and a generator q of $\mathcal{A}(D)$, $\Phi(\mathcal{C})(q) = 1$ holds if and only if q is a marked crossing of $\mathcal{C}$. We let $\epsilon_{\mathcal{C}}$ be the augmentation $\Phi(\mathcal{C})$.

Lemma 3.4 below generalizes [11, Lemma 7.10] and [12, Lemma 5.4] by removing the assumption that the front diagram D is nearly plat. Note that “gradient paths” from [11, Lemma 7.10] correspond to $\epsilon_{\mathcal{C}}$–half-disks in our terminology, and that [12, Lemma 5.4] allows more general coefficients.

Lemma 3.4 *Suppose D is a σ–generic front diagram and $\mathcal{C} = (\{C_m, d_m\}, \{x_m\}, H)$ is in $\mathrm{MCS}_A(D)$. Suppose $\mathcal{C}$ has $M \in \mathbb{N}$ chain complexes, $p \in \{1, \dots, M\}$, and x_p is to the immediate right of a crossing or cusp. Then, for all $i < j$,*

$$\langle d_p e_i, e_j \rangle = \#\mathcal{G}^{\epsilon_{\mathcal{C}}}(x_p, [i, j]). \tag{2}$$

Proof We induct on p. The base case, $p = 1$, follows since there is a unique disk in $\mathcal{G}^{\epsilon_{\mathcal{C}}}(x_1, [1, 2])$, as in Figure 2(b), while $\langle d_1 e_1, e_2 \rangle = 1$ holds according to Definition 2.3(4) .

Assume now that x_p sits to the immediate right of a crossing or cusp and that the result is known for smaller values of p. We complete the inductive step by considering cases.

Left cusp Suppose x_p is to the right of a left cusp with the two strands that meet at the cusp labeled k and $k+1$ at x_p. Define $\tau\colon \{1, \dots, s_{p-1}\} \to \{1, \dots, s_p\}$ by

$$\tau(i) = \begin{cases} i & \text{if } i < k, \\ i+2 & \text{if } i \geq k. \end{cases}$$

(Note that $s_{p-1} = s_p - 2$.) For any $1 \leq i' < j' \leq s_{p-1}$ there is a bijection between $\mathcal{G}^{\epsilon_{\mathcal{C}}}(x_{p-1}, [i', j'])$ and $\mathcal{G}^{\epsilon_{\mathcal{C}}}(x_p, [\tau(i'), \tau(j')])$; see, for example, Figure 7. Moreover, Definition 2.3(5)(c) together with the requirement that $\mathcal{C}$ has simple left cusps give

$$\langle d_{p-1} e_{i'}, e_{j'} \rangle = \langle d_p e_{\tau(i')}, e_{\tau(j')} \rangle,$$

so (2) follows when $i = \tau(i')$ and $j = \tau(j')$.

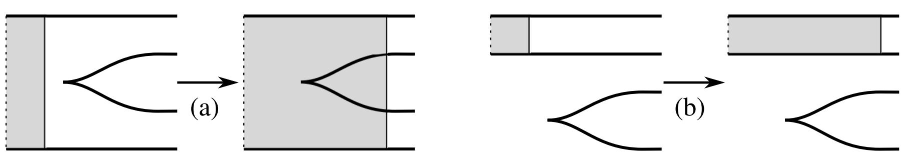


Figure 7: Possible extensions of an ϵ–half-disk or (ϵ, ϵ', H)–half-disk past a left cusp

It remains to consider those cases where $\{i, j\} \cap \{k, k+1\} \neq \varnothing$. Suppose that precisely one of i or j belongs to $\{k, k+1\}$. As $\mathcal{C}$ has simple left cusps, we have $\langle d_p e_i, e_j \rangle = 0$. In addition, the restriction on the behavior of an ϵ–half disk near a left cusp from Figure 2(b) gives that $\mathcal{G}^{\epsilon_{\mathcal{C}}}(x_p, [i, j]) = \varnothing$, so (2) holds. Finally, when $i = k$ and

$j = k + 1$, there is a unique ϵ–half disk in $\mathcal{G}^{\epsilon c}(x_p, [k, k+1])$. (This disk has no convex corners, so Definition 3.3(2) is vacuously satisfied.) Therefore, (2) follows in view of Definition 2.3(4).

Crossing When x_p sits immediately to the right of a crossing, the inductive step is achieved precisely as in [11, Lemma 7.10] or [12, Lemma 5.4] (the signs in the latter reference may be ignored). The arguments in these references apply regardless of whether or not the crossing is marked.

Right cusp Suppose a right cusp sits between x_p and x_{p-1} with the strands that meet at the cusp labeled k and $k+1$ at x_{p-1}. Let $a_{i,j}$ be $\langle d_{p-1}e_i, e_j\rangle$. In the quotient of (C_{p-1}, d_{p-1}) by the subcomplex spanned by e_k and $d_{p-1}e_k$, we have

$$0 = [d_{p-1}e_k] = [e_{k+1}] + \sum_{k+1<j} a_{k,j}[e_j],$$

so

$$d_{p-1}[e_i] = \sum_{i<j} a_{i,j}[e_j] = \sum_{i<j<k} a_{i,j}[e_j] + \sum_{k+1<j} (a_{i,j} + a_{i,k+1}\cdot a_{k,j})[e_j].$$

By Definition 2.3 (5) (b), this gives the computation of the differential in (C_p, d_p) as

$$\langle d_p e_i, e_j\rangle = \langle d_{p-1}e_{\pi(i)}, e_{\pi(j)}\rangle + \langle d_{p-1}e_{\pi(i)}, e_{k+1}\rangle \cdot \langle d_{p-1}e_k, e_{\pi(j)}\rangle, \tag{3}$$

where $\pi\colon \{1, \dots, s_p\} \to \{1, \dots, s_{p-1}\}$ is defined by

$$\pi(i) = \begin{cases} i & \text{if } i < k, \\ i+2 & \text{if } i \geq k. \end{cases}$$

We note that the second term on the right can be non-zero only if $i < k \leq j$; see Figure 8.

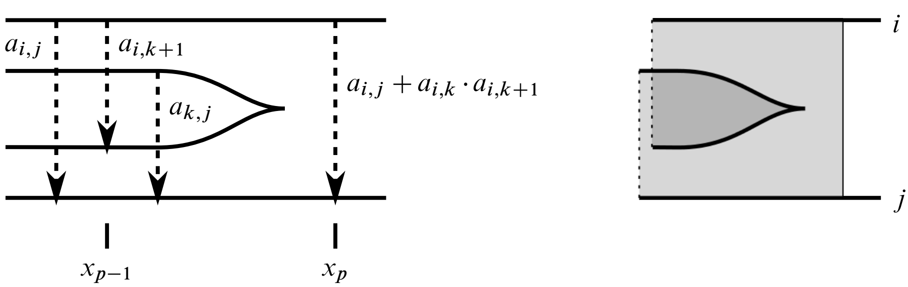


Figure 8: (Left) The relation between differentials at a right cusp. A dotted arrow at x_l pointing from strand i to strand j indicates the matrix coefficient $\langle d_l e_i, e_j\rangle$. (Right) The appearance of disks in $\mathcal{G}^{\epsilon c}(x_p, [i, j])$ with a boundary point at the right cusp between x_{p-1} and x_p.

To complete the proof, we combine Equation (3) with the observation that ϵ–half-disks satisfy a bijection

$$\mathcal{G}^{\epsilon c}(x_p,[i,j]) \cong \mathcal{G}^{\epsilon c}(x_{p-1},[\pi(i),\pi(j)]) \\ \cup\big(\mathcal{G}^{\epsilon c}(x_{p-1},[\pi(i),k+1]) \times \mathcal{G}^{\epsilon c}(x_{p-1},[k,\pi(j)])\big)$$

explained as follows. Those disks in $\mathcal{G}^{\epsilon c}(x_p,[i,j])$ whose boundaries do not intersect the cusp point are in bijection with $\mathcal{G}^{\epsilon c}(x_{p-1},[\pi(i),\pi(j)])$; see Figure 9. Disks in $\mathcal{G}^{\epsilon c}(x_p,[i,j])$ whose boundaries do intersect the cusp point appear between x_{p-1} and x_p as pictured in Figure 8. Removing the portion of the disk between x_{p-1} and x_p leaves a pair of initially overlapping disks from $\mathcal{G}^{\epsilon c}(x_{p-1},[\pi(i),k+1]) \times \mathcal{G}^{\epsilon c}(x_{p-1},[k,\pi(j)])$, and this correspondence is bijective. □

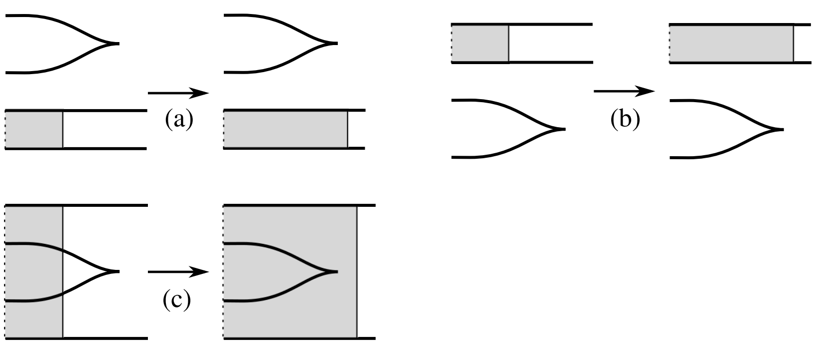


Figure 9: Possible extensions of an ϵ–half-disk or (ϵ,ϵ',H)–half-disk past a right cusp

We outline the central idea of Lemma 3.1 before proceeding to the proof. Recall that an augmentation ϵ has an associated A–form MCS $\mathcal{C}$ where a degree 0 crossing q is marked by $\mathcal{C}$ if and only if $\epsilon(q)$ is 1. The proof of Theorem 1.1 reduced to showing that if augmentations ϵ and ϵ' are homotopic, then their associated A–form MCSs $\mathcal{C}$ and $\mathcal{C}'$ are equivalent. This is accomplished in Lemma 3.1 where an algorithm is given to translate a chain homotopy H from ϵ to ϵ' into a sequence of MCS moves from $\mathcal{C}$ to $\mathcal{C}'$. In particular, for each degree -1 crossing p sent to 1 by H, we employ MCS move (13) just to the left of p to introduce new handleslides. We prove that these handleslides give the mod 2 count of certain (ϵ,ϵ',H)–half disks. Moving these handleslides to the right in the front diagram D, we find that a degree 0 crossing q is changed from marked to unmarked or from unmarked to marked if and only if there exists an odd number of (ϵ,ϵ',H)–half disks originating at q. Therefore, by Proposition 2.2 and the definition of $\mathcal{C}$ and $\mathcal{C}'$, q is changed from marked to unmarked or from unmarked to marked if and only if $\mathcal{C}$ and $\mathcal{C}'$ differ at q. We may therefore conclude that $\mathcal{C}$ and $\mathcal{C}'$ are equivalent.

Proof of Lemma 3.1 Suppose $\mathcal{C}$ and $\mathcal{C}'$ are A–form MCSs and $\epsilon_{\mathcal{C}}$ is homotopic to $\epsilon_{\mathcal{C}'}$. We simplify notation by letting ϵ be $\epsilon_{\mathcal{C}}$ and ϵ' be $\epsilon_{\mathcal{C}'}$. Since ϵ and ϵ' are homotopic, there exists a chain homotopy $H\colon \mathcal{A}(D) \to \mathbb{Z}/2\mathbb{Z}$. Label the degree -1 crossings sent to 1 by H, from left to right, $p_1, \dots, p_m$.

To prove the lemma, we will construct a sequence of MCSs $\mathcal{C}_0, \dots, \mathcal{C}_s$ so that $\mathcal{C}_0$ is $\mathcal{C}$ and $\mathcal{C}_s$ is $\mathcal{C}'$, and, for all $0 \le r < s$, $\mathcal{C}_r \simeq \mathcal{C}_{r+1}$ holds. The construction of the $\mathcal{C}_r$ is inductive, and each of the $\mathcal{C}_r$ will contain a (possibly empty) collection of handleslides V_r that are grouped together immediately to the right of a particular crossing or cusp.

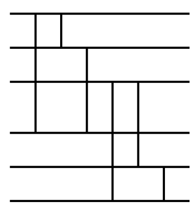


Figure 10: An ordered collection of handleslides

For our purposes it will be convenient to require that the handleslides in each of the V_r are ordered in the following sense. We say a collection of handleslides is *ordered* if, given two handleslides h and h' in the collection with endpoints on strands $i < j$ and $i' < j'$ respectively, h is right of h' if and only if $i > i'$ holds, or $i = i'$ and $j < j'$ hold; see Figure 10. We let $v_r^{i,j}$ be 1 if there exists a handleslide in V_r with endpoints on strands i and j, where $i < j$. Otherwise, $v_r^{i,j}$ is defined to be 0. In a slight abuse of notation, we also let $v_r^{i,j}$ refer to the handleslide in V_r with endpoints on i and j, if such a handleslide exists.

We will verify that Property 1 below holds for all $0 \le r \le s$ as we inductively construct MCSs $\mathcal{C}_r$ with ordered handleslide collections V_r.

Property 1 (a) The MCS $\mathcal{C}_r$ agrees with $\mathcal{C}'$ to the left of V_r and $\mathcal{C}$ to the right of V_r.

(b) For all $i < j$,
$$v_r^{i,j} = \#\mathcal{H}(x_r, [i,j]),$$
where x_r is the x–coordinate of the left-most handleslide in V_r.

Each time r increases, the collection of handleslides V_r is pushed to the right past one cusp or crossing. We continue this inductive process until we arrive at an MCS $\mathcal{C}_s$ with

V_s located just to the left of the right-most right cusp of D. Since the two strands of this cusp do not have the same Maslov potential, V_s must be empty. Then, Property 1 (a) shows that $\mathcal{C}_s$ is $\mathcal{C}'$. As $\mathcal{C} = \mathcal{C}_0$, and for $0 \le i < s$, $\mathcal{C}_i \simeq \mathcal{C}_{i+1}$ it will then follow that $\mathcal{C} \simeq \mathcal{C}'$ holds, as desired.

In the remainder of the proof we construct the sequence of MCSs $\mathcal{C}_0, \dots, \mathcal{C}_s$. Since a crossing q is augmented by ϵ (resp. ϵ') if and only if q is marked by $\mathcal{C}$ (resp. $\mathcal{C}'$), Proposition 2.2 implies $\mathcal{C}$ and $\mathcal{C}'$ differ at q if and only if there exists an odd number of (ϵ, ϵ', H)–admissible disks originating at q. Since q is the right-most point of an (ϵ, ϵ', H)–admissible disk originating at q, it follows that there are no admissible disks originating to the left of p_1 (which is the first crossing sent to 1 by H). Therefore, $\mathcal{C}$ and $\mathcal{C}'$ are identical to the left of p_1. We can then set $\mathcal{C}_0 = \mathcal{C}$ and define V_0 to be empty, but located just to the left of p_1. It follows that Property 1 holds for $\mathcal{C}_0$.

Given $\mathcal{C}_r$ and V_r, we will construct $\mathcal{C}_{r+1}$ and V_{r+1} by applying MCS moves to $\mathcal{C}_r$. We will prove that if Property 1 holds for $\mathcal{C}_r$, then it holds for $\mathcal{C}_{r+1}$ as well. We consider five cases depending on the type of crossing or cusp just to the right of V_r. Let q be the crossing or cusp point to the immediate right of V_r. Let x_r (resp. x_{r+1}) be an x–coordinate to the immediate left (resp. right) of q. In each of the five cases, we first analyze the (ϵ, ϵ', H)–half-disks in $\mathcal{H}(x_{r+1}, [i, j])$ before describing the sequence of MCS moves used to construct $\mathcal{C}_{r+1}$ from $\mathcal{C}_r$ and proving Property 1 holds for $\mathcal{C}_{r+1}$ and V_{r+1}.

In the first three cases considered, q is a crossing between strands k and $k+1$ where the strands of D have been numbered $1, \dots, s_r$, from top to bottom, just to the left of q. Let $\rho\colon \{1, \dots, s_r\} \to \{1, \dots, s_r\}$ be the permutation that transposes k and $k+1$.

Crossing q such that $|q| \neq 0$ and $H(q) \neq 1$ Since $|q|$ is non-zero, $\mathcal{H}(x_r, [k, k+1])$ and $\mathcal{H}(x_{r+1}, [k, k+1])$ are both empty. Given $1 \le i < j \le s_r$ such that $(i, j) \neq (k, k+1)$, one has that (ϵ, ϵ', H)–half-disks in $\mathcal{H}(x_{r+1}, [i, j])$ cannot have a convex corner at q, since $|q| \neq 0$ and $H(q) \neq 1$ hold. In fact, $\#\mathcal{H}(x_{r+1}, [i, j]) = \#\mathcal{H}(x_r, [\rho(i), \rho(j)])$ holds, since there is a natural bijection between $\mathcal{H}(x_{r+1}, [i, j])$ and $\mathcal{H}(x_r, [\rho(i), \rho(j)])$; see, for example, Figure 11.

We now define the sequence of MCS moves that create $\mathcal{C}_{r+1}$ from $\mathcal{C}_r$ and prove Property 1 holds for $\mathcal{C}_{r+1}$. Move all handleslides of V_r to the right of q using MCS moves (7)–(9). Since $|q|$ is non-zero, $v_r^{k,k+1}$ is 0, and therefore all handleslides of V_r can be moved past q and no new handleslides are created by doing so. The resulting collection can be ordered, using MCS moves, without creating new handleslides. The reordering requires rearranging handleslides with one endpoint on either strand k or $k+1$. Since $|q|$ is non-zero, there is no handleslide between k and $k+1$, and therefore the rearrangement can be done without using MCS move (4). The resulting

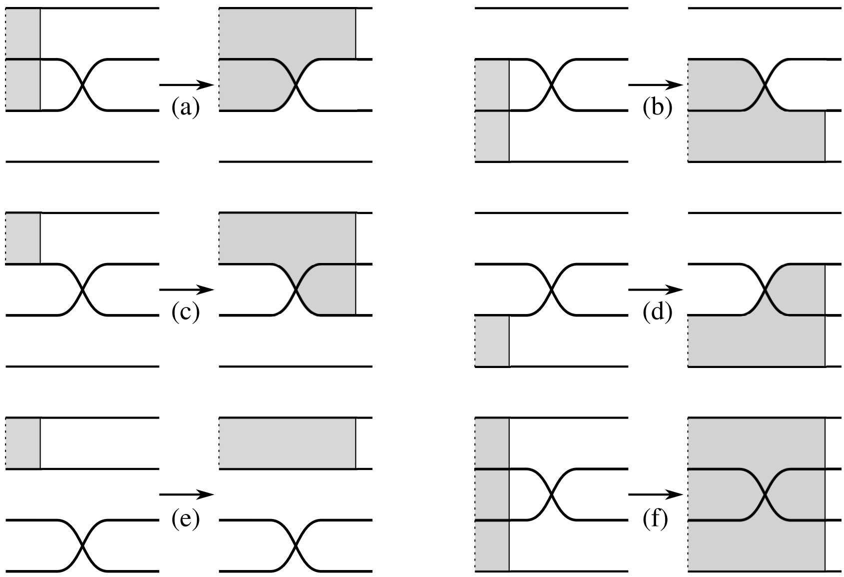


Figure 11: Possible extensions of an (ϵ, ϵ', H)–half-disk past a crossing

ordered collection is V_{r+1} and the MCS is $\mathcal{C}_{r+1}$, and $v_{r+1}^{i,j} = v_r^{\rho(i),\rho(j)}$ holds for all $1 \le i < j \le s_r$. By Property 1(b), $v_r^{\rho(i),\rho(j)} = \#\mathcal{H}(x_r, [\rho(i), \rho(j)])$ holds and as shown above, $\#\mathcal{H}(x_{r+1}, [i, j]) = \#\mathcal{H}(x_r, [\rho(i), \rho(j)])$ holds. Therefore, Property 1(b) holds for $\mathcal{C}_{r+1}$. Property 1(a) holds for $\mathcal{C}_r$ and since $|q|$ is non-zero, q is not marked by either $\mathcal{C}_{r+1}$ or $\mathcal{C}'$. Therefore, Property 1(a) holds for $\mathcal{C}_{r+1}$.

Crossing q such that $|q| = 0$ Let v_q be 1 if q is marked by $\mathcal{C}$ and 0 otherwise. Since Property 1(a) holds for $\mathcal{C}_r$, if q is marked by $\mathcal{C}$, then q is marked by $\mathcal{C}_r$ as well. We slightly abuse notation and also let v_q be the handleslide at q in $\mathcal{C}_r$ in the case such exists.

Suppose $i \neq k+1$ and $j \neq k$. Half-disks in $\mathcal{H}(x_{r+1}, [i, j])$ cannot have a convex corner at q, and therefore there is a bijection from $\mathcal{H}(x_r, [i, j])$ to $\mathcal{H}(x_{r+1}, [\rho(i), \rho(j)])$; see, for example, Figure 11(c)–(f). Since Property 1(b) holds for $\mathcal{C}_r$, $\#\mathcal{H}(x_{r+1}, [i, j]) = v_r^{\rho(i),\rho(j)}$ holds.

Note that $\mathcal{H}(x_{r+1}, [k, k+1])$ is empty. Suppose one of $i = k+1$ or $j = k$ holds. Half-disks in $\mathcal{H}(x_r, [\rho(i), \rho(j)])$ may be smoothly extended past q as in Figure 11(a) and (b). Therefore, there exists an injection from $\mathcal{H}(x_r, [\rho(i), \rho(j)])$ to $\mathcal{H}(x_{r+1}, [i, j])$. However, there may be half-disks in $\mathcal{H}(x_{r+1}, [i, j])$ that have a convex corner at q. If $j = k$ (resp. $i = k+1$) and q is marked by $\mathcal{C}'$ (resp. $\mathcal{C}$), then a half-disk in $\mathcal{H}(x_{r+1}, [i, j])$ can have a convex corner at q; see Figure 12(a) and (b) respectively.

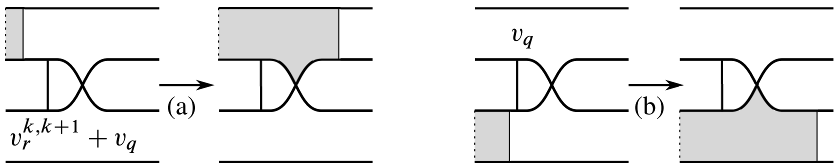


Figure 12: Extending an (ϵ,ϵ',H)–half-disk past a degree 0 crossing so as to have a convex corner at the crossing

Such half-disks are in bijection with half-disks in $\mathcal{H}(x_r,[i,j])$, as can be seen in Figure 12(a) and (b), and by Property 1(b), are counted by $v_r^{i,j}$. Since $v_r^{k,k+1}$ is 1 if and only if $\mathcal{C}'$ and $\mathcal{C}$ differ at q, q is marked by $\mathcal{C}'$ (resp. $\mathcal{C}$) if and only if $v_r^{k,k+1}+v_q$ is 1 (resp. v_q is 1). Therefore, if $j=k$ (resp. $i=k+1$), then the mod 2 count of half-disks in $\mathcal{H}(x_{r+1},[i,j])$ with a convex corner at q is $(v_r^{k,k+1}+v_q)\cdot v_r^{i,k}$ (resp. $v_q\cdot v_r^{k+1,j}$). In summary,

$$\#\mathcal{H}(x_{r+1},[i,j])=\begin{cases} v_r^{i,k+1}+(v_r^{k,k+1}+v_q)\cdot v_r^{i,k} & \text{if } j=k,\\ v_r^{k,j}+v_q\cdot v_r^{k+1,j} & \text{if } i=k+1,\\ 0 & \text{if } i=k \text{ and } j=k+1,\\ v_r^{\rho(i),\rho(j)} & \text{otherwise.}\end{cases} \tag{4}$$

We now define the sequence of MCS moves that create $\mathcal{C}_{r+1}$ from $\mathcal{C}_r$ and prove Property 1 holds for $\mathcal{C}_{r+1}$. We move each handleslide $v_r^{i,j}$ of V_r past q beginning with the right-most handleslide in V_r. If $i\ge k$ and $j\ne k+1$ hold, use MCS moves (2)–(9) to move $v_r^{i,j}$ past v_q, if v_q is 1, and then past the crossing q. If v_q is 1 and $i=k+1$, a new handleslide with endpoints on strands k and j is created when MCS move (4) is used to move $v_r^{i,j}$ past v_q. Move this handleslide to the right of q as well. It is not possible to move $v_r^{k,k+1}$ past q and so, for now, we simply leave $v_r^{k,k+1}$ to the left of q. If $i<k$ holds, use MCS moves (2)–(9) to move $v_r^{i,j}$ past $v_r^{k,k+1}$, if $v_r^{k,k+1}$ is 1, then past v_q, if v_q is 1, and then past the crossing q. If v_q is 1 or $v_r^{k,k+1}$ is 1, and $j=k$, a new handleslide with endpoints on strands i and $k+1$ is created when MCS move (4) is used to move $v_r^{i,j}$ past v_q or $v_r^{k,k+1}$. Move this handleslide to the right of q as well. Once all $v_r^{i,j}$, except $v_r^{k,k+1}$, have been moved past q, use MCS moves (5) and (1) to order the collection of handleslides just to the right of q and remove pairs of handleslides that have the same endpoints. The resulting collection is V_{r+1}. From our work above, we have:

$$v_{r+1}^{i,j}=\begin{cases} v_r^{i,k+1}+(v_r^{k,k+1}+v_q)\cdot v_r^{i,k} & \text{if } j=k,\\ v_r^{k,j}+v_q\cdot v_r^{k+1,j} & \text{if } i=k+1,\\ 0 & \text{if } i=k \text{ and } j=k+1,\\ v_r^{\rho(i),\rho(j)} & \text{otherwise.}\end{cases} \tag{5}$$

Use MCS move (1) to remove both $v_r^{k,k+1}$ and v_q, in the case that they both exist. The resulting MCS is $\mathcal{C}_{r+1}$. By Property 1, $v_r^{k,k+1}$ is 1 if and only if there is an odd number of (ϵ, ϵ', H)–half-disks originating at $\{x_r\} \times [k, k+1]$. There is a bijection between such disks and the (ϵ, ϵ', H)–admissible disks originating at q; see Figure 13. Therefore, by Proposition 2.2, $v_r^{k,k+1}$ is 1 if and only if $\mathcal{C}$ and $\mathcal{C}'$ differ at q. Therefore, Property 1(a) holds for $\mathcal{C}_{r+1}$. Finally, Equations (5) and (4) imply Property 1(b) holds for $\mathcal{C}_{r+1}$.

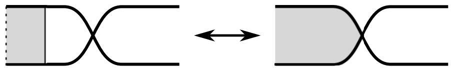


Figure 13: The bijection between (ϵ, ϵ', H)–half-disks originating at $\{x_r\} \times [k, k+1]$ and (ϵ, ϵ', H)–half-disks originating at a crossing q between strands k and $k+1$

Crossing p_i where $1 < i \leq m$ Suppose p_i is a degree -1 crossing between strands k and $k+1$ and $H(p_i) = 1$ holds. Suppose $i \neq k+1$ and $j \neq k$. Half-disks in $\mathcal{H}(x_{r+1}, [i, j])$ cannot have a convex corner at p_i and, therefore, there is a bijection from $\mathcal{H}(x_r, [i, j])$ to $\mathcal{H}(x_{r+1}, [\rho(i), \rho(j)])$; see, for example, Figure 11(c)–(f). Since Property 1(b) holds for $\mathcal{C}_r$, $\#\mathcal{H}(x_{r+1}, [i, j]) = v_r^{\rho(i), \rho(j)}$ holds.

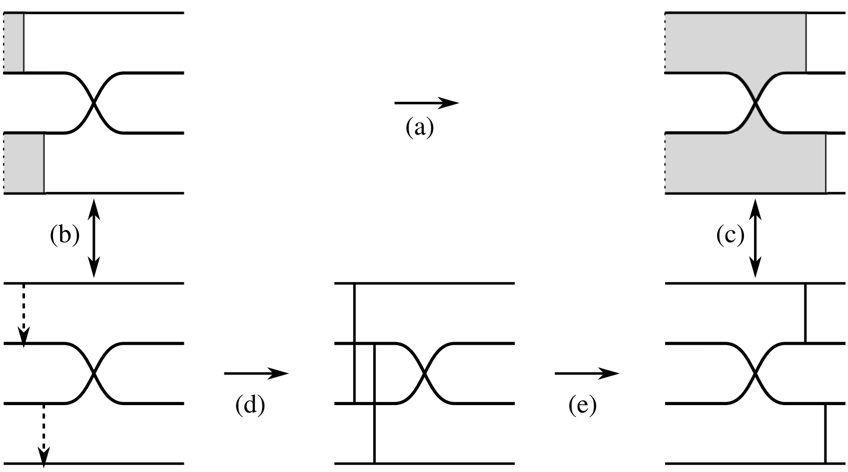


Figure 14: The correspondence between (ϵ, ϵ', H)–half-disks with a convex corner at the degree -1 crossing in the figure and handleslides introduced by MCS move (13) to the left of the crossing. In step (d), two handleslides are created by MCS move (13). These handleslides correspond to the two (ϵ, ϵ', H)–half-disks in the top right figure, each of which has a convex corner at the crossing.

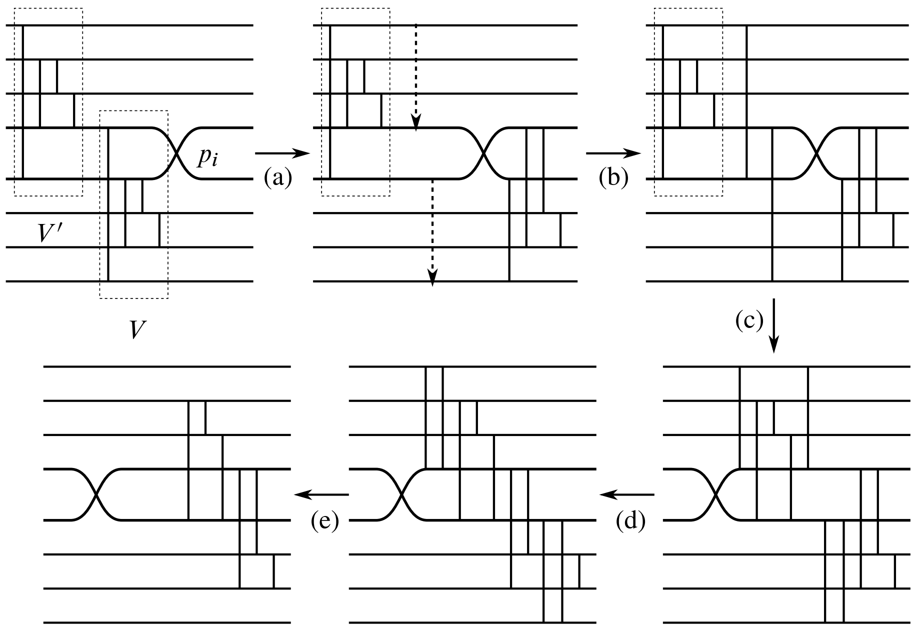


Figure 15: The sequence of MCS moves at a crossing p_i where $|p_i| = -1$ and $H(p_i) = 1$ both hold.

Note that $\mathcal{H}(x_{r+1}, [k, k+1])$ is empty. Suppose one of $i = k+1$ or $j = k$ holds. Half-disks in $\mathcal{H}(x_r, [\rho(i), \rho(j)])$ may be smoothly extended past p_i as in Figure 11(a) and (b). Therefore, there exists an injection from $\mathcal{H}(x_r, [\rho(i), \rho(j)])$ to $\mathcal{H}(x_{r+1}, [i, j])$. However, there may be half-disks in $\mathcal{H}(x_{r+1}, [i, j])$ that have a convex corner at q. In the case that $j = k$ (resp. $i = k+1$), such disks correspond to ϵ–half-disks (resp. ϵ'–half-disks) that have been extended past p_i so as to have a convex corner at p_i; see Figure 14(a)–(e). Let (C, d) (resp. (C', d')) be the chain complex of $\mathcal{C}$ (resp. $\mathcal{C}'$) just to the left of p_i. Property 1(a) implies that, in $\mathcal{C}_r$, (C, d) (resp. (C', d')) is the chain complex to the immediate right (resp. left) of V_r. By Lemma 3.4, the mod 2 count of such half-disks is $\langle de_i, e_j \rangle$ and $\langle d'e_i, e_j \rangle$ respectively. Since Property 1(b) holds for $\mathcal{C}_r$, we may summarize the work of the previous two paragraphs as follows:

$$\#\mathcal{H}(x_{r+1}, [i, j]) = \begin{cases} v_r^{\rho(i),\rho(j)} + \langle de_i, e_j \rangle & \text{if } j = k, \\ v_r^{\rho(i),\rho(j)} + \langle d'e_i, e_j \rangle & \text{if } i = k+1, \\ 0 & \text{if } i = k \text{ and } j = k+1, \\ v_r^{\rho(i),\rho(j)} & \text{otherwise.} \end{cases} \tag{6}$$

We now define the sequence of MCS moves that create $\mathcal{C}_{r+1}$ from $\mathcal{C}_r$ and prove Property 1 holds for $\mathcal{C}_{r+1}$. Let $V \subset V_r$ (resp. $V' \subset V_r$) be the handleslides $v_r^{i,j}$ in V_r

satisfying $i \geq k$ (resp. $i < k$). Since V_r is ordered, V is right of V'; see Figure 15. Let $(\overline{C}, \overline{d})$ be the chain complex of $\mathcal{C}_r$ between V' and V. Use MCS moves (7) and (9) to move the handleslides in V past p_i; see Figure 15(a). Since p_i has degree -1, $v_r^{k,k+1}$ is 0 and $\mu(k) = \mu(k+1) - 1$ holds. Therefore, strands k and $k+1$ satisfy the conditions of MCS move (13). Use MCS move (13) to introduce new handleslides between V' and p_i; see Figure 15(b). MCS move (13) introduces a handleslide with endpoints i and j if and only if either $j = k+1$ and $\langle \overline{d} e_i, e_k \rangle$ is 1, or $i = k$ and $\langle \overline{d} e_{k+1}, e_j \rangle$ is 1. Recall that (C, d) (resp. (C', d')) is the chain complex of $\mathcal{C}_r$ to the immediate right (resp. left) of V_r. Since V_r is ordered, the handleslides between (C, d) and $(\overline{C}, \overline{d})$ have upper endpoints on $k, \dots, s_r$ and the handleslides between (C', d') and $(\overline{C}, \overline{d})$ have upper endpoints on $1, \dots, k-1$. Because of the ordering of handleslides in V_r, the coefficient $\langle \overline{d} e_i, e_k \rangle$ (resp. $\langle \overline{d} e_{k+1}, e_j \rangle$) is unaffected by handleslides in V (resp. V'). As a consequence $\langle \overline{d} e_i, e_k \rangle = \langle d e_i, e_k \rangle$ holds for all $i < k$ and $\langle \overline{d} e_{k+1}, e_j \rangle = \langle d' e_{k+1}, e_j \rangle$ holds for all $k+1 < j$. Therefore, MCS move (13) introduces a handleslide with endpoints i and j if and only if either $j = k+1$ and $\langle d e_i, e_k \rangle$ is 1, or $i = k$ and $\langle d' e_{k+1}, e_j \rangle$ is 1. Move the handleslides created by MCS move (13) and the handleslides in V' past p_i using MCS moves (7)–(9); see Figure 15(c). Use MCS moves (1), (3), (5) and (6) to order the collection of handleslides to the right of p_i and remove pairs of handleslides with identical endpoints; see Figure 15 (d) and (e). In particular, this can be done without creating any new handleslides. The resulting ordered collection of handleslides is V_{r+1} and the MCS is $\mathcal{C}_{r+1}$. Since the only new handleslides created were those created by the single application of MCS move (13),

$$
v_{r+1}^{i,j} = \begin{cases} v_r^{\rho(i),\rho(j)} + \langle d e_i, e_j \rangle & \text{if } j = k, \\ v_r^{\rho(i),\rho(j)} + \langle d' e_i, e_j \rangle & \text{if } i = k+1, \\ 0 & \text{if } i = k \text{ and } j = k+1, \\ v_r^{\rho(i),\rho(j)} & \text{otherwise.} \end{cases} \tag{7}
$$

Equations (6) and (7) imply Property 1(b) holds for $\mathcal{C}_{r+1}$. Finally, Property 1(a) holds for $\mathcal{C}_r$ and $|q| \neq 0$ implies q is not marked by either $\mathcal{C}_{r+1}$ or $\mathcal{C}'$. Therefore, Property 1(a) holds for $\mathcal{C}_{r+1}$.

Left cusp Suppose q is a left cusp. Number the strands of D, from top to bottom, $1, \dots, s_r$ (resp. $1, \dots, s_{r+1}$) just to the left (resp. right) of q. Define $\tau\colon \{1, \dots, s_r\} \to \{1, \dots, s_{r+1}\}$ by

$$
\tau(i) = \begin{cases} i & \text{if } i < k, \\ i+2 & \text{if } i \geq k. \end{cases}
$$

(Note that $s_r = s_{r+1} - 2$.) For any $1 \leq i' < j' \leq s_r$, there is a bijection between $\mathcal{H}(x_{r+1}, [\tau(i'), \tau(j')])$ and $\mathcal{H}(x_r, [i', j'])$; see, for example, Figure 7(a) and (b). If

$\{i,j\} \cap \{k,k+1\}$ is non-empty, then $\mathcal{H}(x_{r+1},[i,j])$ is empty. Therefore, since Property 1(b) holds for $\mathcal{C}_r$,

$$\#\mathcal{H}(x_{r+1},[i,j]) = \begin{cases} v_r^{\tau^{-1}(i),\tau^{-1}(j)} & \text{if } \{i,j\}\cap\{k,k+1\}=\varnothing, \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

Use MCS moves (11) and (12) to move each handleslide in V_r past q. The resulting collection V_{r+1} is ordered and the resulting MCS is $\mathcal{C}_{r+1}$. The endpoints of a handleslide remain on the same strands of D as it is moved past q. Therefore, we have

$$v_{r+1}^{i,j} = \begin{cases} v_r^{\tau^{-1}(i),\tau^{-1}(j)} & \text{if } \{i,j\}\cap\{k,k+1\}=\varnothing, \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

Equations (8) and (9) imply Property 1(b) holds for $\mathcal{C}_{r+1}$. Since q is not a crossing and Property 1(a) holds for $\mathcal{C}_r$, it must hold for $\mathcal{C}_{r+1}$ as well.

Right cusp Suppose q is a right cusp between strands k and $k+1$. Let (C,d) (resp. (C',d')) be the chain complex of $\mathcal{C}$ (resp. $\mathcal{C}'$) just to the left of q. Property 1(a) implies that, in $\mathcal{C}_r$, (C,d) (resp. (C',d')) is the chain complex to the immediate right (resp. left) of V_r. Number the strands of D, from top to bottom, by $1,\dots,s_{r+1}$ (resp. by $1,\dots,s_r$) just to the right (resp. left) of q. Define $\pi\colon \{1,\dots,s_{r+1}\} \to \{1,\dots,s_r\}$ by

$$\pi(i) = \begin{cases} i & \text{if } i<k, \\ i+2 & \text{if } i \geq k. \end{cases}$$

(Note that $s_r = s_{r+1}+2$.)

If $j<k$ or $i \geq k$, then

$$\#\mathcal{H}(x_{r+1},[i,j]) = v_r^{\pi(i),\pi(j)} \tag{10}$$

holds, since Property 1(b) holds for $\mathcal{C}_r$ and there is a bijection from $\mathcal{H}(x_r,[\pi(i),\pi(j)])$ to $\mathcal{H}(x_{r+1},[i,j])$; see Figure 9(a) and (b).

When $j \geq k$ and $i<k$, we claim that there is a bijection

$$\begin{aligned} \mathcal{H}(x_{r+1},[i,j]) \cong \mathcal{H}(x_r,[\pi(i),\pi(j)]) & \\ \cup\, (\mathcal{G}^{\epsilon}(x_r,[\pi(i),k+1]) \times \mathcal{H}(x_r,[k,\pi(j)])) & \\ \cup\, (\mathcal{H}(x_r,[\pi(i),k+1]) \times \mathcal{G}^{\epsilon'}(x_r,[k,\pi(j)])). & \end{aligned} \tag{11}$$

Suppose $j \geq k$ and $i<k$. Half-disks in $\mathcal{H}(x_r,[\pi(i),\pi(j)])$ may be smoothly extended past q as in Figure 9(c). Therefore, there exists an injection from $\mathcal{H}(x_r,[\pi(i),\pi(j)])$ to $\mathcal{H}(x_{r+1},[i,j])$. However, there may be half-disks in $\mathcal{H}(x_{r+1},[i,j])$ whose boundary intersects the cusp point; see Figure 16(a) and Figure 17(a).

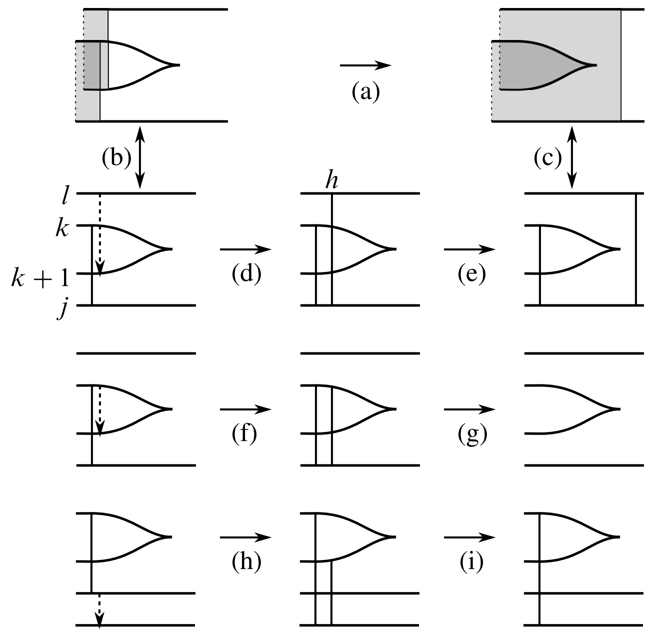


Figure 16: (a)–(e): The correspondence between Type 1 (ϵ, ϵ', H)–half-disks whose boundary intersects the right cusp in the figure and handleslides introduced by MCS move (13) to the left of the crossing. (f), (g): A handleslide introduced by MCS move (13) that is removed, along with $v_r^{k,j}$, by an MCS (1) move. (h), (i): A handleslide introduced by MCS move (13) that is removed by an MCS (10) move.

We divide half-disks whose boundary intersects q into two types as follows. Any such half-disk has one convex corner at a degree -1 crossing, which we denote p. Trace the boundary of such a half-disk counter-clockwise beginning at the vertical line $\{x_{r+1}\} \times [i, j]$. In a Type 1 (resp. Type 2) half-disk, p appears after (resp. before) q. A Type 1 (resp. Type 2) half-disk can be uniquely decomposed into an (ϵ, ϵ', H)–half-disk and an ϵ–half-disk (resp. ϵ'–half-disk) as in Figure 16(a) and (b) (resp. Figure 17(a) and (b)). Therefore, the set in the second (resp. third) line of Equation (11) is in bijection with Type 1 (resp. Type 2) half-disks. Since Property 1(b) holds for $\mathcal{C}_r$ and Lemma 3.4 holds for both (C, d) and (C', d'), the mod 2 count of Type 1 half-disks is $\langle de_{\pi(i)}, e_{k+1}\rangle \cdot v_r^{k,\pi(j)}$ and the mod 2 count of Type 2 half-disks is $v_r^{\pi(i),k+1} \cdot \langle d'e_k, e_{\pi(j)}\rangle$. Therefore, for $j \geq k$ and $i < k$, we have the formula

$$\#\mathcal{H}(x_{r+1}, [i, j]) = v_r^{\pi(i),\pi(j)} + v_r^{\pi(i),k+1} \cdot \langle d'e_k, e_{\pi(j)}\rangle + \langle de_{\pi(i)}, e_{k+1}\rangle \cdot v_r^{k,\pi(j)}. \tag{12}$$

We now define the sequence of MCS moves that create $\mathcal{C}_{r+1}$ from $\mathcal{C}_r$ and prove Property 1 holds for $\mathcal{C}_{r+1}$. We move the handleslides of V_r past q iteratively beginning with the right-most handleslide. Suppose $v_r^{i,j}$ is the right-most handleslide of V_r that has yet to be moved past q. If $i > k+1$ or $j < k$, use MCS move (12) to move $v_r^{i,j}$ past q. If $i < k$ and $j > k+1$, use MCS move (11) to move $v_r^{i,j}$ past q. If $i = k+1$ or $j = k$, use MCS move (10) to remove $v_r^{i,j}$. Since $\mu(k) = \mu(k+1)+1$ and a handleslide has endpoints on strands with the same Maslov potential, $v_r^{k,k+1}$ must be 0. It remains to consider the two cases $i = k, j > k+1$ and $i < k, j = k+1$.

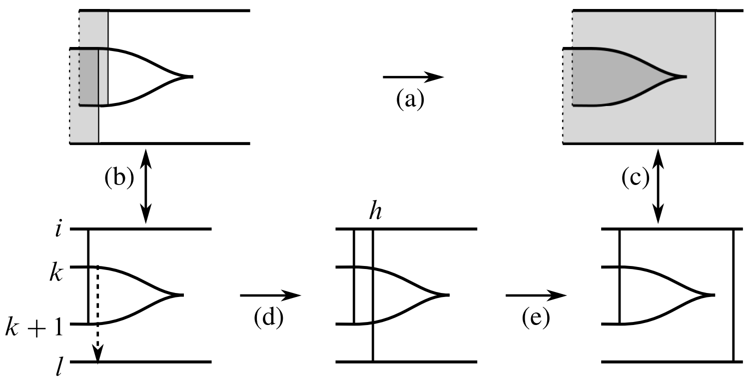


Figure 17: (a)–(e): The correspondence between Type 2 (ϵ, ϵ', H)–half-disks whose boundary intersects the right cusp in the figure and handleslides introduced by MCS move (13) to the left of the crossing.

Suppose $v_r^{i,j}$ is $v_r^{k,j}$ where $j > k+1$. Since $\mu(k) = \mu(j)$ and $\mu(k) = \mu(k+1)+1$ both hold, $\mu(k+1) = \mu(j) - 1$ holds and, thus, strands $k+1$ and j satisfy the conditions of MCS move (13). Use MCS move (13) to create new handleslide marks; see the arrow directed to the right in Figure 6. Let $(\overline{C}, \overline{d})$ be the chain complex of $\mathcal{C}_r$ just to the right of $v_r^{k,j}$. The handleslides created are of three types. By Definition 2.3(4), $\langle \overline{d} e_k, e_{k+1} \rangle$ is 1. Therefore, MCS move (13) introduces a handleslide with endpoints k and j; see Figure 16(f). Use MCS move (1) to remove this handleslide and $v_r^{k,j}$; see Figure 16(g). For each l such that $\langle \overline{d} e_j, e_l \rangle$ is 1, MCS move (13) introduces a handleslide with endpoints $k+1$ and l; see Figure 16(h). Use MCS move (10) to remove this handleslide; see Figure 16(i). Suppose $l < k$ and $\langle \overline{d} e_l, e_{k+1} \rangle$ is 1. The third type of handleslide introduced by MCS move (13) has endpoints l and j; see Figure 16(d). Let h be this handleslide. Use MCS move (11) to move h past q; see Figure 16(e). Recall that (C, d) is the chain complex of $\mathcal{C}_r$ to the immediate right of V_r. Since V_r is ordered, the handleslides between $(\overline{C}, \overline{d})$ and (C, d) have endpoints on strands $k+1, \dots, s_p$. The coefficient $\langle \overline{d} e_l, e_{k+1} \rangle$ is unaffected by such handleslides and, thus, $\langle \overline{d} e_l, e_{k+1} \rangle = \langle d e_l, e_{k+1} \rangle$ holds. Therefore, h exists if and only if $\langle de_l, e_{k+1}\rangle \cdot v_r^{k,j}$ is 1. As we noted earlier, $\langle de_l, e_{k+1}\rangle \cdot v_r^{k,j}$ is 1 if and only if the mod 2 count of Type 1 half-disks in $\mathcal{H}(x_{r+1}, [l, j+2])$ is 1. Therefore, h exists if and only if the mod 2 count of Type 1 half-disks in $\mathcal{H}(x_{r+1}, [l, j+2])$ is 1.

Suppose $v_r^{i,j}$ is $v_r^{i,k+1}$ where $i < k$. Note that $\mu(i) = \mu(k) - 1$ holds and, thus, strands i and k satisfy the conditions of MCS move (13). Use MCS move (13) to create new handleslides to the immediate right of $v_r^{i,k+1}$. Suppose $l > k+1$ and $\langle \bar{d} e_k, e_l\rangle$ is 1. MCS move (13) introduces a handleslide with endpoints i and l; see Figure 17(d). Let h be this handleslide. Use MCS move (11) to move h past q; see Figure 17(e). Following an analogous argument as was used in the case of a Type 1 half-disk, h exists if and only if the mod 2 count of Type 2 half-disks in $\mathcal{H}(x_{r+1}, [i, l+2])$ is 1. MCS move (13) also introduces handleslides analogous to those in Figure 16(f) and (h), which are removed in same manner as was done in Figure 16(g) and (i).

Once we have applied the above algorithm to each handleslide in V_r, we are left with a collection of handleslides V to the right of q. The ordering of V_r ensures the only new handleslides were those introduced by applications of MCS move (13). Therefore, given $1 \le i < j \le s_{r+1}$, there may be up to 3 handleslides in V with endpoints on i and j; one counts (ϵ, ϵ', H)–half-disks extended past q as in Figure 9, one counts Type 1 half-disks as in Figure 16(a)–(e), and the third counts Type 2 half-disks as in Figure 17(a)–(e). Use MCS moves (1), (3), (5), and (6) to remove pairs of handleslides with identical endpoints and order V. In particular, V can be ordered without creating new handleslides. The resulting ordered collection of handleslides is V_r and the MCS is $\mathcal{C}_{r+1}$. If $j < k$ or $i \ge k$, then

$$v_{r+1}^{i,j} = v_r^{\pi(i),\pi(j)}$$

holds and, if $j \ge k$ and $i < k$, then

$$v_{r+1}^{i,j} = v_r^{\pi(i),\pi(j)} + v_r^{\pi(i),k+1} \cdot \langle d' e_k, e_{\pi(j)}\rangle + \langle de_{\pi(i)}, e_{k+1}\rangle \cdot v_r^{k,\pi(j)}$$

holds. These equations, along with Equations (10) and (12), imply Property 1(b) holds for $\mathcal{C}_{r+1}$. Finally, since q is not a crossing and Property 1(a) holds for $\mathcal{C}_r$, it must hold for $\mathcal{C}_{r+1}$ as well.

This completes the construction of the MCSs $\mathcal{C}_0, \dots, \mathcal{C}_s$. □

4 Corollaries to Theorem 1.1

In the following corollaries to Theorem 1.1, D is the σ–generic front diagram of a Legendrian knot with rotation number 0. Recall that an augmentation ϵ in $\mathrm{Aug}(D)$ has a corresponding A–form MCS $\mathcal{C}$ where, for a degree 0 crossing q, $\epsilon(q) = 1$ holds if and only if q is marked by $\mathcal{C}$.

Corollary 4.1 *The count of MCS classes of a Legendrian knot is a Legendrian isotopy invariant.*

Corollary 4.1 follows from the fact that the count of homotopy classes of augmentations is a Legendrian isotopy invariant and every Legendrian knot is Legendrian isotopic to a Legendrian knot with σ–generic front diagram by an arbitrarily small Legendrian isotopy. Corollary 4.1 is stated and a proof is briefly sketched by Petya Pushkar in a letter to Dmitry Fuchs from 2000. The proposed proof investigates the effect of Legendrian Reidemeister moves on the number of MCS classes and is different from the approach in this article.

Given the Chekanov–Eliashberg algebra $(\mathcal{A}(D), \partial)$, the differential $\partial^\epsilon\colon \mathcal{A}(D) \to \mathcal{A}(D)$ is $\phi^\epsilon \circ \partial \circ (\phi^\epsilon)^{-1}$, where $\phi^\epsilon\colon \mathcal{A}(D) \to \mathcal{A}(D)$ is the algebra map defined on generators by $\phi^\epsilon(q) = q + \epsilon(q)$. The group $\mathrm{LCH}(\epsilon)$, called the *linearized contact homology* of ϵ, is the homology of the chain complex $(A(D), \partial_1^\epsilon)$, where $\partial_1^\epsilon(q)$ is the length 1 monomials of $\partial^\epsilon(q)$. By [2], the set $\{\mathrm{LCH}(\epsilon)\}_{\epsilon \in \mathrm{Aug}(D)}$ is a Legendrian isotopy invariant, which we will call the *LCH invariant*.

Corollary 4.2 *If ϵ and ϵ' are homotopic as augmentations, then LCH(ϵ) and LCH(ϵ') are isomorphic as homology groups. Therefore, augmentation homotopy classes have well-defined linearized contact homology groups.*

Proof We will apply two theorems from [10]. In order to do so, the front diagram must be "nearly plat". A front diagram is *plat* if all left cusps have the same x–coordinate, all right cusps have the same x–coordinate, and no two crossings have the same x–coordinate. A front diagram is *nearly plat* if it is the result of perturbing a plat front diagram slightly so that no two cusps have the same x–coordinate.

We now deduce the corollary in the case that D is nearly plat. Suppose ϵ and ϵ' are homotopic. By Lemma 3.1, the A–form MCSs $\mathcal{C}$ and $\mathcal{C}'$ corresponding to ϵ and ϵ' are equivalent as MCSs. In [11], differential graded algebras $(\mathcal{A}_{\mathcal{C}}, d)$ and $(\mathcal{A}_{\mathcal{C}'}, d')$ are assigned to $\mathcal{C}$ and $\mathcal{C}'$, respectively. The linear level of each algebra is a chain complex $(A_{\mathcal{C}}, d_1)$ and $(A_{\mathcal{C}'}, d_1')$, respectively. By [11, Theorem 5.5], $(A_{\mathcal{C}}, d_1)$ and $(A_{\mathcal{C}'}, d_1')$ are isomorphic. By [11, Theorem 7.3], $(A(D), \partial_1^\epsilon)$ is isomorphic to $(A_{\mathcal{C}}, d_1)$ and $(A(D), \partial_1^{\epsilon'})$ is isomorphic to $(A_{\mathcal{C}'}, d_1')$. Therefore, $\mathrm{LCH}(\epsilon)$ and $\mathrm{LCH}(\epsilon')$ are isomorphic as homology groups.

For the general case of a Chekanov–Eliashberg algebra $(\mathcal{A}, \partial)$ assigned to a front (or Lagrangian) diagram that is not nearly plat, we argue as follows. By [1], the Chekanov–Eliashberg algebras assigned to Legendrian isotopic Legendrian knots are stable tame isomorphic. Any Legendrian knot is Legendrian isotopic to a knot with nearly plat front diagram, therefore $(\mathcal{A}, \partial)$ is stable tame isomorphic to a DGA that satisfies the property stated in Corollary 4.2. We then verify that $(\mathcal{A}, \partial)$ also satisfies the corollary in two steps.

Step 1 The corollary holds for a stabilization $(S(\mathcal{A}), \partial')$ of a DGA $(\mathcal{A}, \partial)$ if and only if it holds for $(\mathcal{A}, \partial)$.

Here, $S(\mathcal{A})$ is obtained from $\mathcal{A}$ by adding two generators x and y in successive degrees, and the differential satisfies $\partial'|_{\mathcal{A}} = \partial$ and $\partial' x = y$. Restricting augmentations of $S(\mathcal{A})$ to $\mathcal{A}$ provides a surjection from the set of augmentations of $S(\mathcal{A})$ to the set of augmentations of $\mathcal{A}$, and this gives a well-defined bijection between homotopy classes of augmentations of $S(\mathcal{A})$ and $\mathcal{A}$. Moreover, for any augmentation $\epsilon\colon S(\mathcal{A}) \to \mathbb{Z}/2$, the linearized homology groups associated to ϵ and $\epsilon|_{\mathcal{A}}$ are isomorphic, so Step 1 follows.

Step 2 If $\varphi\colon (\mathcal{A}_1, \partial_1) \to (\mathcal{A}_2, \partial_2)$ is an isomorphism of DGAs, then the corollary holds for $(\mathcal{A}_1, \partial_1)$ if and only if it holds for $(\mathcal{A}_2, \partial_2)$.

To see this, observe that $\epsilon_2 \mapsto \epsilon_1 \circ \varphi$ gives a bijection from augmentations of $(\mathcal{A}_2, \partial_2)$ to augmentations of $(\mathcal{A}_1, \partial_1)$ that preserves homotopy classes and linearized homology groups. $\square$

Corollary 4.2 provides a means for strengthening the LCH invariant. The set

$$\{\mathrm{LCH}(\epsilon)\}_{\epsilon \in \mathrm{Aug}(D)},$$

along with a count of the number of augmentation homotopy classes associated with each group, is a Legendrian isotopy invariant. The authors are currently unaware of an example where this refinement is able to distinguish knots that are not already distinguished by the LCH invariant taken without regard to multiplicity.

Corollary 4.3 *If ϵ and ϵ' are homotopic, then ϵ and ϵ' are mapped to the same graded normal ruling by the many-to-one map from augmentations to graded normal rulings defined in [16].*

Proof Suppose ϵ and ϵ' are homotopic. By Lemma 3.1, the A–form MCSs $\mathcal{C}$ and $\mathcal{C}'$ corresponding to ϵ and ϵ' are equivalent. By [10, Lemma 3.14], every MCS determines a graded normal ruling. By [10, Proposition 3.15], equivalent MCSs determine the same graded normal ruling. Therefore, $\mathcal{C}$ and $\mathcal{C}'$ determine the same graded normal ruling. In [16], there is an algorithmically defined many-to-one map Ω from $\mathrm{Aug}(D)$ to the set of graded normal rulings of D. In the case of an augmentation ϵ and its corresponding A–form MCS $\mathcal{C}$, $\Omega(\epsilon)$ is the same as the graded normal ruling determined by $\mathcal{C}$ in [10, Lemma 3.14]. Therefore, Ω maps ϵ and ϵ' to the same graded normal ruling. $\square$

References

[1] **Y Chekanov**, *Differential algebra of Legendrian links*, Invent. Math. 150 (2002) 441–483 MR1946550

[2] **Y V Chekanov**, *Invariants of Legendrian knots*, from: "Proceedings of the International Congress of Mathematicians, II", (T Li, editor), Higher Ed. Press, Beijing (2002) 385–394 MR1957049

[3] **Y V Chekanov**, **P E Pushkar'**, *Combinatorics of fronts of Legendrian links, and Arnol'd's 4–conjectures*, Uspekhi Mat. Nauk 60 (2005) 99–154 MR2145660 In Russian; translated in Russian Math. Surveys 60 (2005) 95–149

[4] **T Ekholm**, **K Honda**, **T Kálmán**, *Legendrian knots and exact Lagrangian cobordisms*, preprint `arXiv:1212.1519`

[5] **Y Eliashberg**, **A Givental**, **H Hofer**, *Introduction to symplectic field theory*, Geom. Funct. Anal. (2000) 560–673 MR1826267

[6] **Y Félix**, **S Halperin**, **J-C Thomas**, *Differential graded algebras in topology*, from: "Handbook of algebraic topology", (I M James, editor), North-Holland, Amsterdam (1995) 829–865 MR1361901

[7] **D Fuchs**, *Chekanov–Eliashberg invariant of Legendrian knots: Existence of augmentations*, J. Geom. Phys. 47 (2003) 43–65 MR1985483

[8] **D Fuchs**, **T Ishkhanov**, *Invariants of Legendrian knots and decompositions of front diagrams*, Mosc. Math. J. 4 (2004) 707–717, 783 MR2119145

[9] **D Fuchs**, **D Rutherford**, *Generating families and Legendrian contact homology in the standard contact space*, J. Topol. 4 (2011) 190–226 MR2783382

[10] **M B Henry**, *Connections between Floer-type invariants and Morse-type invariants of Legendrian knots*, Pacific J. Math. 249 (2011) 77–133 MR2764942

[11] **M B Henry**, **D Rutherford**, *A combinatorial DGA for Legendrian knots from generating families*, Commun. Contemp. Math. 15 (2013) MR3037194

[12] **M B Henry**, **D Rutherford**, *Ruling polynomials and augmentations over finite fields*, J. Topol. 8 (2015) 1–37 MR3335247

[13] **J Jordan**, **L Traynor**, *Generating family invariants for Legendrian links of unknots*, Algebr. Geom. Topol. 6 (2006) 895–933 MR2240920

[14] **T Kálmán**, *Contact homology and one parameter families of Legendrian knots*, Geom. Topol. 9 (2005) 2013–2078 MR2209366

[15] **L L Ng**, *Computable Legendrian invariants*, Topology 42 (2003) 55–82 MR1928645

[16] **L L Ng**, **J M Sabloff**, *The correspondence between augmentations and rulings for Legendrian knots*, Pacific J. Math. 224 (2006) 141–150 MR2231655

[17] **J M Sabloff**, *Augmentations and rulings of Legendrian knots*, Int. Math. Res. Not. 2005 (2005) 1157–1180 MR2147057

[18] **J M Sabloff**, **L Traynor**, *Obstructions to Lagrangian cobordisms between Legendrians via generating families*, Algebr. Geom. Topol. 13 (2013) 2733–2797 MR3116302

[19] **L Traynor**, *Legendrian circular helix links*, Math. Proc. Cambridge Philos. Soc. 122 (1997) 301–314 MR1458235

[20] **L Traynor**, *Generating function polynomials for Legendrian links*, Geom. Topol. 5 (2001) 719–760 MR1871403

Department of Mathematics, Siena College
515 Loudon Road, Loudonville, NY 12211, USA

Department of Mathematics, Ball State University
2000 W University Ave, Muncie, IN 47306, USA

mbhenry@siena.edu, rutherford@bsu.edu

Received: 25 July 2014 Revised: 10 April 2015

Algebraic & Geometric Topology 15 (2015) 3355–3369

On finite derived quotients of 3–manifold groups

WILL CAVENDISH

This paper studies the set of finite groups appearing as $\pi_1(M)/\pi_1(M)^{(n)}$, where M is a closed, orientable 3–manifold and $\pi_1(M)^{(n)}$ denotes the n^{th} term of the derived series of $\pi_1(M)$. Our main result is that if M is a closed, orientable 3–manifold, $n \geq 2$, and $G \cong \pi_1(M)/\pi_1(M)^{(n)}$ is finite, then the cup-product pairing $H^2(G) \otimes H^2(G) \to H^4(G)$ has cyclic image C, and the pairing $H^2(G) \otimes H^2(G) \xrightarrow{\smile} C$ is isomorphic to the linking pairing $H_1(M)_{\text{Tors}} \otimes H_1(M)_{\text{Tors}} \to \mathbb{Q}/\mathbb{Z}$.

57M10; 57M60

1 Introduction

One of the most elementary invariants of a connected topological space M is its first homology group $H_1(M)$, which, via Hurewitz's theorem, may be expressed as $\pi_1(M)/\pi_1(M)^{(1)}$, the quotient of the fundamental group of M by its derived subgroup. Somewhat more mysterious are the topological invariants given by the higher derived quotients of the fundamental group, $\pi_1(M)/\pi_1(M)^{(n)}$, where $\pi_1(M)^{(n)}$ denotes the n^{th} term of the derived series of $\pi_1(M)$. This paper studies the groups that appear as $\pi_1(M)/\pi_1(M)^{(n)}$ when M is a closed, orientable 3–manifold in the special case that $\pi_1(M)/\pi_1(M)^{(n)}$ is finite. Our interest in these groups is motivated by the following well-known question, which was conjectured by Roushon in [15] to have a positive answer in the case that $\pi_1(M)$ is torsion free:

Question 1 *Let M be a closed, orientable 3–manifold. If $[\pi_1(M) : \pi_1(M)^{(n)}]$ is finite for all n, does the derived series of $\pi_1(M)$ stabilize, ie is $\pi_1(M)^{(i)}$ a perfect group for some i?*

A positive answer to this conjecture would supply an alternative proof of the virtual positive Betti number conjecture for hyperbolic 3–manifolds, which was resolved by Ian Agol in [1], building on work of Kahn and Markovic in [10] and Wise in [16]. Indeed, when the derived series of $\pi_1(M)$ stabilizes, it is well known that $\pi_1(M)/\pi_1(M)^{(i)}$ is a solvable group with 4–periodic cohomology for sufficiently large i. A group of this form is extremely rare, and it is isomorphic to the product of a trivial, dihedral, or

Published: 12 January 2016 DOI: 10.2140/agt.2015.15.3355

generalized quaternion group with a cyclic group of relatively prime order; see [13]. All of these groups have abelianization with p–rank at most three for every prime p. It is a well-known result of Lubotzky (see [12]) that every hyperbolic 3–manifold has a finite-sheeted covering space with arbitrarily large p–rank for all but finitely many p; so if the above conjecture is true, we can pass to a finite-sheeted covering space N of any hyperbolic 3–manifold M such that $\pi_1(N)/\pi_1(N)^{(n)}$ is infinite for some n. It follows by a simple group-theoretic argument that M has a finite-sheeted covering space with positive first Betti number.

The main thrust of this paper is that the finite groups that appear as $\pi_1(M)/\pi_1(M)^{(n)}$ satisfy restrictive group-theoretic constraints. These constraints do not appear when $n=1$ since, by taking connected sums of lens spaces, one can easily show that any finite abelian group appears as $\pi_1(M)/\pi_1(M)^{(1)} \cong H_1(M)$ for a closed orientable 3–manifold M. The question of which finite groups appear as $\pi_1(M)/\pi_1(M)^{(2)}$, however, is already more interesting. Note that if $\pi_1(M)/\pi_1(M)^{(2)}$ is finite, then $\pi_1(M)^{(1)}/\pi_1(M)^{(2)}$ is a finite group, and therefore the maximal abelian cover of M has trivial first Betti number. Reznikov showed in [14] that the fundamental groups of 3–manifolds with this property satisfy a number of nontrivial constraints, and consequently not every metabelian group can appear as $\pi_1(M)/\pi_1(M)^{(2)}$. The restrictions Reznikov discovered are especially interesting in light of the main result of Cooper and Long in [5], which shows that any finite group appears as the group of deck transformations of a regular covering $\rho\colon M' \to M$ of closed 3–manifolds where $b_1(M')=0$. This shows that Reznikov's constraints do not arise from obstructions to fixed-point free group actions on rational homology 3–spheres.

In this paper we build on the themes explored by Reznikov in [14] by showing that the cohomology ring of a finite group of the form $\pi_1(M)/\pi_1(M)^{(n)}$ for $n \geq 2$ directly reflects information about the *linking pairing* on $H_1(M)$, which is a nondegenerate bilinear form $H_1(M)_{\mathrm{Tors}} \otimes H_1(M)_{\mathrm{Tors}} \to \mathbb{Q}/\mathbb{Z}$ whose definition we now recall. Given an element $[a] \in H_1(M)_{\mathrm{Tors}}$ and a loop γ_a representing $[a]$, there exists an integer $n \in \mathbb{N}$ such that $n \cdot [a] = 0$. Since the 1–cycle $n \cdot \gamma_a$ is homologically trivial, there exists an immersed oriented surface Σ_a in M such that the oriented boundary of Σ_a is equal to $n \cdot \gamma_a$. Given another class $[b] \in H_1(M)_{\mathrm{Tors}}$, there exists a loop γ_b representing b such that γ_b is transverse to Σ_a. The value of the linking pairing $\langle [a],[b] \rangle$ is defined by $(\gamma_b \pitchfork \Sigma_a)/n \in \mathbb{Q}/\mathbb{Z}$, the algebraic intersection number of γ_b and Σ_a divided by n.

The following theorem shows that the linking pairing on $H_1(M)_{\mathrm{Tors}}$ is isomorphic to the 2–dimensional cup-product pairing in $H^*(G)$ for any finite quotient of $q\colon \pi_1(M) \to G$ such that $\ker(q) \subseteq \pi_1(M)^{(2)}$.

Theorem 1.1 *Let M be a closed, orientable 3–manifold, let $\Gamma \cong \pi_1(M)$, and let $q\colon \Gamma \to G$ be a surjective homomorphism such that* $\ker(q) \subseteq \Gamma^{(2)}$*. If G is finite, then the cup-product pairing $H^2(G) \otimes H^2(G) \to H^4(G)$ is nondegenerate and has cyclic image $C < H^4(G)$. Furthermore, there exists an embedding $i\colon C \to \mathbb{Q}/\mathbb{Z}$ such that for any $\omega_1, \omega_2 \in H^2(G)$,*

$$i(\omega_1 \smile \omega_2) = \langle [M] \frown \widetilde{q}^*(\omega_1), [M] \frown \widetilde{q}^*(\omega_2) \rangle,$$

where $[M] \in H_3(M)$ denotes the fundamental class of1 M, $\langle -, - \rangle$ denotes the linking pairing on $H_1(M)_{\mathrm{Tors}}$, and $\widetilde{q}\colon M \to BG$ is a continuous map from M to the classifying space of G such that $\widetilde{q}_*\colon \pi_1(M) \to \pi_1(BG) \cong G$ is equal to q.

We remark that Theorem 1.1 does not hold when the quotient group $\Gamma/\Gamma^{(2)}$ is infinite, even when $H_1(M) = \Gamma/\Gamma^{(1)}$ is finite. To see this, let M be homeomorphic to $\mathbb{RP}^3 \# \mathbb{RP}^3$, the connected sum of two copies of real projective 3–space, and let $G = \pi_1(M)/\pi_1(M)^{(2)}$. The fundamental group $\pi_1(M)$ is isomorphic to the infinite dihedral group $D_\infty \cong 11\mathbb{Z}/2 * \mathbb{Z}/2$, and hence $H_1(M) \cong \mathbb{Z}/2 \oplus \mathbb{Z}/2$. The commutator subgroup of $\mathbb{Z}/2 * \mathbb{Z}/2$ is isomorphic to $\mathbb{Z}$, so $\pi_1(M)^{(2)}$ is trivial and therefore $G = \pi_1(M) \cong \pi_1(M)/\pi_1(M)^{(2)} \cong \mathbb{Z}/2 * \mathbb{Z}/2$. Recall that the infinite dimensional real projective space $\mathbb{RP}^\infty$ is a classifying space for $\mathbb{Z}/2$, and that the cohomology ring $H^*(\mathbb{RP}^\infty)$ is generated by a single degree 2 element of order 2. It follows that the wedge sum $\mathbb{RP}^\infty \vee \mathbb{RP}^\infty$ is a classifying space for $\mathbb{Z}/2 * \mathbb{Z}/2$. Since the cohomology ring of a wedge sum of connected spaces is isomorphic to the direct sum of the cohomology rings of the summands modulo the identification of the zeroth cohomology groups, $H^*(G) \cong H^*(\mathbb{RP}^\infty \vee \mathbb{RP}^\infty) \cong \mathbb{Z}[x, y]/(xy, 2x, 2y)$, where $\deg(x) = \deg(y) = 2$. The elements x^2 and y^2 are linearly independent in $H^4(G)$, so the image of the cup-product pairing $H^2(G) \otimes H^2(G) \to H^4(G)$ is not cyclic. This example can be modified to produce aspherical (indeed hyperbolic) examples of such 3–manifolds using the techniques of Baker, Boileau and Wang in [3].

As a sample application of Theorem 1.1, we demonstrate how it can be used to derive the following result of Reznikov [14, Theorem 12.5] from well-known results about 2–groups and their cohomology rings.

Theorem 1.2 (Reznikov) *Let Γ be the fundamental group of a closed, orientable 3–manifold M such that $H_1(M) \cong \mathbb{Z}/2 \oplus \mathbb{Z}/2$. If $\Gamma/\Gamma^{(2)}$ is finite and $\langle c, c \rangle = 0$ for all nontrivial $c \in H_1(M)$, then the Sylow-2 subgroup of $\Gamma/\Gamma^{(2)}$ is equal to the quaternion group of order 8. If $\Gamma/\Gamma^{(2)}$ is finite and $\langle c, c \rangle$ is nontrivial for some nontrivial $c \in H_1(M)$, then the Sylow-2 subgroup of $\Gamma/\Gamma^{(2)}$ is isomorphic to a generalized quaternion group Q_{2^k} for $k > 3$.*

Proof Let G denote $\Gamma/\Gamma^{(2)}$, let S denote the Sylow-2 subgroup of G, and let K denote $G^{(1)}$. Note that since G is metabelian, it follows that K is abelian. Let $K_{(2)}$ denote the 2–part of K and let K' be the complementary subgroup of K so that $K \cong K' \oplus K_{(2)}$. Note that K' is a characteristic subgroup of G and is therefore normal, and that the order of G/K' is equal to the order of S. Since S is a Sylow subgroup, $G/K' \cong S$, and the inclusion $S \hookrightarrow G$ therefore has a right inverse $r\colon G \to S$. This shows that $r^*\colon H^*(S) \to H^*(G)$ is injective. It is a simple consequence of the universal coefficients theorem (see Lemma 2.2 in Section 2 below) that $r^*\colon H^2(S) \to H^2(G)$ is an isomorphism, and it follows from naturality of the cup product that the cup-product pairing on $H^2(S)$ is isomorphic to the cup-product pairing on $H^2(G)$. Applying Theorem 1.1, the cup-product pairing on $H^2(S)$ is therefore isomorphic to the linking form on $H_1(M)_{\mathrm{Tors}}$.

We now examine the possibilities for the group S. If S is abelian, then $S \cong \mathbb{Z}/2 \oplus \mathbb{Z}/2$ and the cup-product pairing $H^2(S) \otimes H^2(S) \to H^4(S)$ has 3–dimensional image by the Künneth theorem. We may therefore assume that S is nonabelian. Since $H_1(S) \cong \mathbb{Z}/2 \oplus \mathbb{Z}/2$, S is a 2–group of maximal class (see [7, Section 5.4]), and is therefore isomorphic to either a dihedral group, a semidihedral group, or a quaternion group. The cup-product pairing on the second cohomology of a dihedral group of order $4k$ has image with rank larger than one (see [8]), and is degenerate on any semidihedral group; see [6]. It follows that S is isomorphic to Q_{2^k}, a generalized quaternion group of order 2^k.

The cohomology ring $H^*(Q_8)$ of the quaternion group of order 8 has the feature that any $\alpha \in H^2(Q_8)$ satisfies $\alpha^2 = 0$ (see [2]), whereas the cohomology ring of $H^*(Q_{2^k})$ for $k > 3$ has 2–dimensional elements with nontrivial squares; see [9]. Since the cup-product pairing on $H^2(S)$ is isomorphic to the linking pairing on $H_1(M)_{\mathrm{Tors}}$, it follows that if $\langle c, c\rangle = 0$ for all $c \in H_1(M)_{\mathrm{Tors}}$, then $S \cong Q_8$; otherwise $S \cong Q_{2^k}$ for some $k > 3$. $\square$

Given a manifold M, let $\widetilde{M}_{ab}$ denote the maximal abelian cover of M. It is interesting to note that there are only two isomorphism types of nondegenerate pairings $(\mathbb{Z}/2)^2 \otimes (\mathbb{Z}/2)^2 \to \mathbb{Q}/\mathbb{Z}$, and that both types appear as pairings on $H_1(M)_{\mathrm{Tors}}$ for a closed orientable 3–manifold M such that $b_1(\widetilde{M}_{ab}) = 0$. Examples of such manifolds are given by the spaces S^3/Q_8 and S^3/Q_{16}, where S^3 is viewed as the group of unit quaternions and Q_{2n} is realized as the subgroup of S^3 generated by $e^{i\pi/n}$ and j. It has been shown by Kawauchi and Kojima in [11] that every nondegenerate pairing $A \otimes A \to \mathbb{Q}/\mathbb{Z}$ on a finite abelian group A appears as the linking form of a 3–manifold with $b_1(M) = 0$. Given this result, it is interesting to ask the following:

Question 2 *Given a finite abelian group A, does every nondegenerate bilinear pairing $A \otimes A \to \mathbb{Q}/\mathbb{Z}$ appear as the linking pairing of a 3–manifold M such that $H_1(M) = A$ and $b_1(\widetilde{M}_{ab}) = 0$?*

Note that by Theorem 1.1, any pairing that appears in this way also appears as the cup-product pairing on $H^2(G)$ for the finite metabelian group $G \cong \pi_1(M)/\pi_1(M)^{(2)}$.

The proof of Theorem 1.1 breaks into two parts. The first part, carried out in Section 2, consists of showing that the linking pairing on $H_1(M)$ can be factored through the cup-product pairing $H^2(G) \otimes H^2(G) \to H^4(G)$. One consequence of this factorization is the following theorem, which applies to quotients of $\pi_1(M)$ whose abelianization has maximal order.

Theorem 1.3 *Let Γ be the fundamental group of a closed, orientable 3–manifold M, let $[M] \in H_3(M)$ denote the fundamental class of M, and let $q\colon \Gamma \to G$ be a surjective homomorphism onto a finite group G such that $\ker(q) \subseteq \Gamma^{(1)}$. Then:*

(i) $H^2(G) \otimes H^2(G) \to H^4(G)$ *is nondegenerate.*

(ii) $\mathrm{ord}(\widetilde{q}_*([M])) \cdot H_1(M) = 0$, *where $\widetilde{q}\colon M \to BG$ is a continuous map such that $\widetilde{q}_*\colon \pi_1(M) \to \pi_1(BG) \cong G$ is equal to q.*

Note that this theorem provides information about how the fundamental class of the manifold M behaves under finite quotient maps. Indeed, in the special case that $G \cong H_1(M)$, this theorem shows that the image of the homomorphism $\widetilde{q}_*\colon H_3(M) \to H_3(G)$ has maximal order, since the annihilator of $H_3(G)$ is equal to the annihilator of $H_1(G)$ when G is an abelian group.

The second part of the proof of Theorem 1.1, carried out in Section 3, establishes the following result using an argument known as spectral sequence comparison:

Lemma 1.4 *Let Γ be the fundamental group of a closed, orientable 3–manifold M, and let $\rho\colon \pi_1(M) \to G$ be a surjective homomorphism. If $\ker(\rho) \subseteq \Gamma^{(2)}$, then the image of the cup-product pairing $H^2(G) \otimes H^2(G) \to H^4(G)$ is cyclic.*

As we will show in section 4, Theorem 1.1 follows easily from Lemma 1.4 together with the relationship between the cup-product pairing and the linking pairing established in Section 2.

We remark that the methods of this paper are purely algebraic, and also apply to Poincaré duality groups of dimension 3.

2 The linking pairing on $H_1(M)_{\mathrm{Tors}}$ and cup products in quotients of $\pi_1(M)$ with maximal abelianization

Throughout this paper, we will let M be a closed orientable 3–manifold, Γ denote $\pi_1(M)$, $\langle -,-\rangle$ denote the linking pairing on $H_1(M)_{\mathrm{Tors}}$, and $[M]\in H_3(M)$ denote the fundamental class of M. For functoriality reasons, we will regard the image of the linking pairing $\langle -,-\rangle$ as an element of $H_0(M,\mathbb{Q}/\mathbb{Z})$, rather than as an element of the abstract group $\mathbb{Q}/\mathbb{Z}$. It will also be convenient to work the dual pairing $\lambda\colon H^2(M)_{\mathrm{Tors}}\otimes H^2(M)_{\mathrm{Tors}}\to H_0(M,\mathbb{Q}/\mathbb{Z})$, defined by $\lambda(\omega_1,\omega_2)=\langle [M]\frown\omega_1,[M]\frown\omega_2\rangle$. This pairing satisfies the well-known identity

$$\lambda(\omega_1,\omega_2)=[M]\frown(\omega_1\smile\beta^{-1}(\omega_2)), \tag{2-1}$$

where $\beta\colon H^1(M,\mathbb{Q}/\mathbb{Z})\to H^2(M,\mathbb{Z})$ denotes the Bockstein homomorphism arising from the short exact sequence $0\to\mathbb{Z}\to\mathbb{Q}\to\mathbb{Q}/\mathbb{Z}\to 0$.

The following lemma shows how the linking pairing on $H_1(M)_{\mathrm{Tors}}$ relates to the cup-product pairing $H^2(G)\otimes H^2(G)\to H^4(G)$.

Lemma 2.1 *Let $q\colon\Gamma\to G$ be a surjective homomorphism onto a finite group G, and let $\widetilde{q}\colon M\to BG$ be a continuous map from M to the classifying space of G such that $\widetilde{q}_*\colon\pi_1(M)\to\pi_1(BG)\cong G$ equals q. Given $\omega_1,\omega_2\in H^2(G)$,*

$$\widetilde{q}_*\big(\lambda(\widetilde{q}^*(\omega_1),\widetilde{q}^*(\omega_2))\big)=\widetilde{q}_*([M])\frown\beta^{-1}(\omega_1\smile\omega_2).$$

Proof We begin by noting that the expressions on the left-hand side of the above identity are well defined, since $H^2(G)$ is a finite group, and therefore $\widetilde{q}^*(\omega)\in H^2(M)_{\mathrm{Tors}}$. Note also that, since G is finite, $H^i(G,\mathbb{Q})=0$ for all $i>0$. It follows that $\beta\colon H^1(G,\mathbb{Q}/\mathbb{Z})\to H^2(G,\mathbb{Z})$ is an isomorphism, and therefore the map β^{-1} appearing on the right-hand side of the above equation is well defined as well.

Recall that the cup-product pairing $H^i(G,A)\otimes H^j(G,\mathbb{Z})\to H^{i+j}(G,A\otimes\mathbb{Z})\cong H^{i+j}(G,A)$ equips $H^*(G,A)$ with the structure of a right $H^*(G)$–module. Given a short exact sequence $0\to A\to B\to C\to 0$ of G–modules, the connecting homomorphisms in the long exact sequence

$$\cdots\to H^i(G,A)\to H^i(G,B)\to H^i(G,C)\xrightarrow{\delta}H^{i+1}(G,A)\to\cdots$$

fit together to give an $H^*(G)$–module homomorphism $\delta\colon H^*(G,C)\to H^*(G,A)$; ie given $\alpha\in H^*(G,C)$ and $\omega\in H^*(G)$, $\delta(\alpha)\smile\omega=\delta(\alpha\smile\omega)$; see [4, Chapter V.3]. Since the Bockstein homomorphism $\beta\colon H^*(G,\mathbb{Q}/\mathbb{Z})\to H^*(G,\mathbb{Z})$ is given by the connecting homomorphism in the long exact sequence in cohomology arising from the short exact sequence $0 \to \mathbb{Z} \to \mathbb{Q} \to \mathbb{Q}/\mathbb{Z} \to 0$, given $\omega \in H^1(G, \mathbb{Q}/\mathbb{Z})$ and $\eta \in H^1(G, \mathbb{Z})$, $\beta(\omega) \smile \eta = \beta(\omega \smile \eta)$. It follows that given $\eta_1, \eta_2 \in H^2(G, \mathbb{Z})$, $\eta_1 \smile \eta_2 = \beta(\beta^{-1}(\eta_1) \smile \eta_2)$, and therefore

$$\beta^{-1}(\eta_1 \smile \eta_2) = \beta^{-1}(\eta_1) \smile \eta_2. \tag{2-2}$$

Recall that for a continuous map $f\colon X \to Y$ between topological spaces, the cap product satisfies the following naturality property for $c \in H_i(X), \eta \in H^j(Y)$:

$$f_*(c \frown f^*(\eta)) = f_*(c) \frown \eta. \tag{2-3}$$

Applying these identities together with the identity (2-1) for the pairing λ and naturality of the cup product and the Bockstein homomorphism, we obtain

$$\begin{aligned} \tilde{q}_*\big(\lambda\big(\tilde{q}^*(\omega_1), \tilde{q}^*(\omega_2)\big)\big) &\overset{\text{(2-1)}}{=} \tilde{q}_*\big([M] \frown \big(\tilde{q}^*(\omega_1) \smile \beta^{-1}(\tilde{q}^*(\omega_2))\big)\big) \\ &= \tilde{q}_*\big([M] \frown \big(\tilde{q}^*(\omega_1) \smile \tilde{q}^*(\beta^{-1}(\omega_2))\big)\big) \\ &= \tilde{q}_*\big([M] \frown \tilde{q}^*\big(\omega_1 \smile \beta^{-1}(\omega_2)\big)\big) \\ &\overset{\text{(2-3)}}{=} \tilde{q}_*([M]) \frown (\omega_1 \smile \beta^{-1}(\omega_2)) \\ &\overset{\text{(2-2)}}{=} \tilde{q}_*([M]) \frown \beta^{-1}(\omega_1 \smile \omega_2). \end{aligned}$$
□

We now turn to the proof of Theorem 1.3, which will require several preliminary lemmas.

Lemma 2.2 *Let $f\colon X \to Y$ be a continuous map between topological spaces such that $H_2(X)$ and $H_2(Y)$ are finite. If $f_*\colon H_1(X) \to H_1(Y)$ is an isomorphism, then $f^*\colon H^2(Y) \to H^2(X)$ is an isomorphism.*

Proof By naturality of the universal coefficients exact sequence, we have the following commutative diagram of exact sequences:

$$\begin{tikzcd} 0 \arrow[r] & \mathrm{Ext}(H_1(Y), \mathbb{Z}) \arrow[r] \arrow[d, "f^*"] & H^2(Y) \arrow[r] \arrow[d, "f^*"] & \mathrm{Hom}(H_2(Y), \mathbb{Z}) \arrow[r] \arrow[d, "f^*"] & 0 \\ 0 \arrow[r] & \mathrm{Ext}(H_1(X), \mathbb{Z}) \arrow[r] & H^2(X) \arrow[r] & \mathrm{Hom}(H_2(X), \mathbb{Z}) \arrow[r] & 0 \end{tikzcd}$$

Since $H_2(Y)$ and $H_2(X)$ are finite, the Hom terms in the above diagram are trivial and therefore $f^*\colon H^2(Y) \to H^2(X)$ is completely determined by the morphism $f^*\colon \mathrm{Ext}(H_1(Y), \mathbb{Z}) \to \mathrm{Ext}(H_1(X), \mathbb{Z})$. If $f_*\colon H_1(X) \to H_1(Y)$ is an isomorphism, then $f^*\colon \mathrm{Ext}(H_1(Y), \mathbb{Z}) \to \mathrm{Ext}(H_1(X), \mathbb{Z})$ is an isomorphism by functoriality, so the lemma follows. □

Note that since finite groups have finite second homology groups, Lemma 2.2 can be applied to any homomorphism between finite groups that induces an isomorphism on the level of abelianizations. We will use the following simple consequence of this lemma several times in what follows.

Lemma 2.3 *Let M be a rational homology 3–sphere, let $q\colon \pi_1(M) \to G$ be a surjective homomorphism onto a finite group, and let $\widetilde{q}\colon M \to BG$ be a continuous homomorphism from M to the classifying space of G such that $\widetilde{q}_*\colon \pi_1(M) \to \pi_1(BG) \cong G$ is equal to q. If $\ker(q) \subseteq \pi_1(M)^{(1)}$, then $\widetilde{q}^*\colon H^2(G) \to H^2(M)$ is an isomorphism.*

Proof We claim that the hypotheses of Lemma 2.2 hold in this setting. To see that $q_*\colon H_1(M) \to H_1(G)$ is an isomorphism, note that since $\ker(q) \subseteq \Gamma^{(1)}$, the abelianization map $\Gamma \to \Gamma/\Gamma^{(1)} \cong H_1(M)$ factors through q. It follows that the homomorphism $\widetilde{q}_*\colon H_1(M) \to H_1(G)$ is injective. Since the map q is surjective, the induced map $\widetilde{q}_*\colon H_1(M) \to H_1(G)$ is surjective as well.

Since $H_2(G)$ is finite for any finite group G, it remains to check that $H_2(M)$ is finite. This is a simple consequence of Poincaré duality and the universal coefficients theorem, since $H_2(M) \cong H^1(M) \cong \mathrm{Hom}(H_1(M), \mathbb{Z})$, and $\mathrm{Hom}(H_1(M), \mathbb{Z}) = 0$ since $H_1(M)$ is a torsion group. □

The next lemma we will need is the following well-known result on the values taken by the linking form.

Lemma 2.4 *Let M be a 3–manifold. Given $a \in H^2(M)_{\mathrm{Tors}}$, there exists $b \in H^2(M)_{\mathrm{Tors}}$ such that $\mathrm{ord}(\lambda(a,b)) = \mathrm{ord}(a)$.*

Proof Let A denote $H^2(M)_{\mathrm{Tors}}$. It is a well-known consequence of Poincaré duality that $\lambda\colon A \times A \to \mathbb{Q}/\mathbb{Z}$ is nondegenerate, so the homomorphism $A \to \mathrm{Hom}(A, \mathbb{Q}/\mathbb{Z})$ given by $a \mapsto \lambda(a, -)$ is injective. Given $b \in A$, let $\mu_b = \mathrm{ord}(a)/\mathrm{ord}(\lambda(a,b))$. Let d be the greatest common divisor of $\{\mu_b \mid b \in A\}$. Then

$$\frac{\mathrm{ord}(a)}{d} \cdot \lambda(a,b) = \frac{\mu_b \cdot \mathrm{ord}(\lambda(a,b))}{d} \cdot \lambda(a,b) = \frac{\mu_b}{d} \cdot \mathrm{ord}(\lambda(a,b)) \cdot \lambda(a,b) = 0$$

for all b, so $(\mathrm{ord}(a)/d) \cdot \lambda(a,-) = 0$, and therefore $\mathrm{ord}(\lambda(a,-))$ divides $\mathrm{ord}(a)/d$. Since $a \mapsto \lambda(a,-)$ is an isomorphism, $\mathrm{ord}(\lambda(a,-)) = \mathrm{ord}(a)$, so $d = \pm 1$. It follows that for each prime p dividing $\mathrm{ord}(a)$, there exists an element b_p such that μ_{b_p} is coprime to p, and hence the p–part of $\mathrm{ord}(a)$ divides $\mathrm{ord}(\lambda(a,b))$. By taking a multiple of b_p if necessary, we can assume that $\mathrm{ord}(\lambda(a,b_p))$ is exactly equal to the

p–part of $\mathrm{ord}(a)$. Since the sum of a set of elements with pairwise coprime orders $n_1, n_2, \dots, n_\ell$, in an abelian group has order given by $n_1 \cdot n_2 \cdot \cdots \cdot n_\ell$,

$$\begin{aligned}\mathrm{ord}(\lambda(a,b)) = \mathrm{ord}\Big(\lambda(a, \sum_p b_p)\Big) &= \mathrm{ord}\Big(\sum_p \lambda(a, b_p)\Big) \\ &= \prod_p \mathrm{ord}\ \lambda(a, b_p) = \mathrm{ord}(a).\end{aligned}$$

□

The proof of Theorem 1.3 follows easily from the above lemmas.

Proof of Theorem 1.3 Let $q\colon \Gamma \to G$ be a surjective homomorphism onto a finite group such that $\ker(q) \subseteq \Gamma^{(1)}$. Note that By Lemma 2.3, $\widetilde{q}^*\colon H^2(G) \to H^2(M)$ is an isomorphism.

We first show that the cup-product pairing $H^2(G) \otimes H^2(G) \to H^4(G)$ is nondegenerate. Given a nontrivial element $\omega_1 \in H^2(G)$, $\widetilde{q}^*(\omega_1)$ gives a nontrivial element of $H^2(M)_{\mathrm{Tors}}$ since $\widetilde{q}^*$ is injective. By Lemma 2.4, there exists an element $\eta \in H^2(M)$ such that the order of $\lambda(\widetilde{q}^*(\omega_1), \eta)$ is equal to the order of ω_1, and since $\widetilde{q}^*\colon H^2(G) \to H^2(M)$ is surjective, there exists an element $\omega_2 \in H^2(G)$ such that $\widetilde{q}^*(\omega_2) = \eta$. Applying Lemma 2.1 together with the fact that $\widetilde{q}_*\colon H_0(M, \mathbb{Q}/\mathbb{Z}) \to H_0(G, \mathbb{Q}/\mathbb{Z})$ is an isomorphism, we have

$$0 \neq \widetilde{q}_*(\lambda(\widetilde{q}^*(\omega_1), \eta)) = \widetilde{q}_*(\lambda(\widetilde{q}^*(\omega_1), \widetilde{q}^*(\omega_1))) = \widetilde{q}_*([M]) \frown \beta^{-1}(\omega_1 \smile \omega_2).$$

This shows that $\omega_1 \smile \omega_2 \neq 0$, and since ω_1 was an arbitrary nontrivial element of $H^2(G)$, the cup product pairing $H^2(G) \otimes H^2(G) \to H^4(G)$ is nondegenerate.

Note that the order of $\widetilde{q}_*([M]) \frown \beta^{-1}(\omega_1 \smile \omega_2)$ divides the order of $\widetilde{q}_*([M])$, so since

$$\mathrm{ord}(\omega_1) = \mathrm{ord}(\lambda(\widetilde{q}^*(\omega_1), \eta)) = \mathrm{ord}\big(\widetilde{q}_*([M]) \frown \beta^{-1}(\omega_1 \smile \omega_2)\big),$$

the order of ω_1 divides $\mathrm{ord}(\widetilde{q}_*([M]))$ for all $\omega_1 \in H^2(G)$. This shows that $\mathrm{ord}(\widetilde{q}_*([M])) \cdot H^2(G) = 0$. Since $H^2(G) \cong H^2(M) \cong H_1(M)$, $\mathrm{ord}(\widetilde{q}_*([M]))$ annihilates $H_1(M)$ as well. □

3 Cup products in $H^*(G)$ for finite quotients $G \cong \pi_1(M)/N$ with $N < \pi_1(M)^{(2)}$

In this section we prove Lemma 1.4 from the introduction. Throughout this section, we will let G be a finite group, and we will let $\rho\colon \pi_1(M) \cong \Gamma \to G$ be a surjective homomorphism such that $\ker(\rho) \subseteq \Gamma^{(2)}$. We will also let $Q = H_1(M)$, $N = [\Gamma, \Gamma]$, $K = [G, G]$, and we will refer to the maps labeled in the following commutative diagram of exact sequences:

$$\begin{tikzcd}
1 \arrow[r] & N \arrow[r] \arrow[d, "r"] & \Gamma \arrow[r, "q\circ\rho"] \arrow[d, "\rho"] & Q \arrow[r] \arrow[d, "\mathrm{id}"] & 1 \\
1 \arrow[r] & K \arrow[r] & G \arrow[r, "q"] & Q \arrow[r] & 1
\end{tikzcd}$$

The above commutative diagram corresponds to a commutative diagram of continuous maps of the form

$$\begin{tikzcd}
\widetilde{M} \arrow[r] \arrow[d, "\widetilde{r}"] & M \arrow[r, "\widetilde{q}\circ\widetilde{\rho}"] \arrow[d, "\widetilde{\rho}"] & BQ \arrow[d, "\mathrm{id}"] \\
BK \arrow[r] & BG \arrow[r, "\widetilde{q}"] & BQ,
\end{tikzcd}$$

where $\widetilde{M}$ is the regular covering space of M corresponding to $N < \pi_1(M)$. The map $\widetilde{\rho}^*$ induces a morphism between the Lyndon–Hochschild–Serre spectral sequence for the extension $1 \to K \to G \to Q \to 1$ and the Cartan–Serre spectral sequence for the regular cover $\widetilde{M} \to M$. The proof of Lemma 1.4 proceeds by analyzing this morphism. We will require several preliminary lemmas.

Lemma 3.1 *The cover $\widetilde{M}$ is a rational homology 3–sphere and $\widetilde{r}^*\colon H^2(K) \to H^2(\widetilde{M})$ is an isomorphism.*

Proof Since Γ surjects onto G, $N \cong \pi_1(\widetilde{M})$ surjects onto K and $H_1(\widetilde{M})$ surjects onto $H_1(K)$. By assumption $\ker(\rho) \subseteq \Gamma^{(2)} \cong [N, N] = N^{(1)}$ and $r = \rho|_N$, so $\ker(r) \subseteq N^{(1)}$. The abelianization map $N \to N/N^{(1)} \cong H_1(\widetilde{M})$ therefore factors through r, and so $\widetilde{r}_*\colon H_1(\widetilde{M}) \to H_1(K)$ is an isomorphism. Since K is a finite group, its abelianization $H_1(K)$ is also finite, so $H_1(M, \mathbb{Q}) \cong H_1(K) \otimes \mathbb{Q}$ is trivial, and therefore $\widetilde{M}$ is a rational homology 3–sphere. The result then follows by Lemma 2.3 in Section 2. □

To set some notation for the next lemmas, let $(E_k)^\alpha$, $(d_k)^\alpha$ and $(E_k)^\beta$, $(d_k)^\beta$ denote the pages and differentials in the cohomological spectral sequences $H^r(Q, H^s(K)) = \Rightarrow H^{r+s}(G)$ and $H^r(Q, H^s(\widetilde{M})) \Longrightarrow H^{r+s}(M)$ respectively.

Lemma 3.2 *There exists a commutative diagram with exact rows of the form*

$$
\begin{tikzcd}
0 \arrow[r] & (E_4^{1,2})^\alpha \arrow[r] \arrow[d, "\rho_4^*"] & (E_3^{1,2})^\alpha \arrow[r, "(d_3^{1,2})^\alpha"] \arrow[d, "\rho_3^*"] & (E_3^{4,0})^\alpha \arrow[r] \arrow[d, "\rho_3^*"] & (E_4^{4,0})^\alpha \arrow[r] \arrow[d, "\rho_4^*"] & 0 \\
0 \arrow[r] & (E_4^{1,2})^\beta \arrow[r] & (E_3^{1,2})^\beta \arrow[r, "(d_3^{1,2})^\beta"] & (E_3^{4,0})^\beta \arrow[r] & (E_4^{4,0})^\beta \arrow[r] & 0
\end{tikzcd}
\tag{3-1}
$$

where the middle two homomorphisms labeled ρ_3^* are isomorphisms.

Proof Note that since the first cohomology group with integral coefficients is trivial for any finite group, $(E_2)^\alpha$ has the following form:

$$
\begin{array}{ccccc}
H^0(Q,H^3(K)) & H^1(Q,H^3(K)) & H^2(Q,H^3(K)) & H^3(Q,H^3(K)) & H^4(Q,H^3(K)) \\
H^0(Q,H^2(K)) & H^1(Q,H^2(K)) & H^2(Q,H^2(K)) & H^3(Q,H^2(K)) & H^4(Q,H^2(K)) \\
0 & 0 & 0 & 0 & 0 \\
\mathbb{Z} & 0 & H^2(Q) & H^3(Q) & H^4(Q)
\end{array}
$$

By assumption the manifold M is orientable, so Q acts on $\widetilde{M}$ by orientation preserving homeomorphisms. It follows that Q acts on $H^3(\widetilde{M},\mathbb{Z}) \cong \mathbb{Z}$ trivially, and hence $H^i(Q, H^3(\widetilde{M},\mathbb{Z})) \cong H^i(Q)$ for all i. Furthermore, since $\widetilde{M}$ is a rational homology 3–sphere and $H^1(\widetilde{M}) \cong \operatorname{Hom}(H_1(\widetilde{M}),\mathbb{Z}) = 0$, $(E_2)^\beta$ has the following form:

$$
\begin{array}{ccccc}
\mathbb{Z} & 0 & H^2(Q) & H^3(Q) & H^4(Q) \\
H^0(Q,H^2(\widetilde{M})) & H^1(Q,H^2(\widetilde{M})) & H^2(Q,H^2(\widetilde{M})) & H^3(Q,H^2(\widetilde{M})) & H^4(Q,H^2(\widetilde{M})) \\
0 & 0 & 0 & 0 & 0 \\
\mathbb{Z} & 0 & H^2(Q) & H^3(Q) & H^4(Q)
\end{array}
$$

The map $\widetilde{\rho}$ induces a morphism between these two spectral sequences, ie a sequence of homomorphisms $\rho_k^*\colon (E_k)^\alpha \to (E_k)^\beta$ such that $\rho_k^* \circ (d_k)^\alpha = (d_k)^\beta \circ \rho_k^*$, ρ_{k+1}^* is the map induced on homology by ρ_k^*, and the map $\rho_2^*\colon H^r(Q, H^s(K)) \to H^r(Q, H^s(\widetilde{M}))$ is induced by the Q-module homomorphism $q^*\colon H^s(K) \to H^s(\widetilde{M})$. Since $q^*\colon H^2(K) \to H^2(\widetilde{M})$ is an isomorphism by Lemma 3.1 and $q^*\colon H^s(K) \to H^s(\widetilde{M})$ is trivially an isomorphism for $s \in \{0,1\}$, the maps $\rho_2^*\colon (E_2^{r,s})^\alpha \to (E_2^{r,s})^\beta$ are isomorphisms for all pairs (r,s) such that $s \le 2$.

Since the first row of each spectral sequence vanishes, the differentials $(d_2^{i,2})^\alpha$, $(d_2^{i,2})^\beta$ and $(d_2^{i,1})^\alpha$, $(d_2^{i,1})^\beta$, whose domain or range lie in the first row, are trivial for all i. It follows that 0^{th} row of the E_3–page of each spectral sequence is identical to the 0^{th} row of the E_2–page, ie $(E_3^{i,0})^\alpha \cong (E_2^{i,0})^\alpha$ and $(E_3^{i,0})^\beta \cong (E_2^{i,0})^\beta$ for all i. Since none of the d_2–differentials have image lying in the 0^{th} or first columns, it also follows that for $j \in \{0,1\}$, $(E_3^{j,2})^\alpha \cong (E_2^{j,2})^\alpha$ and $(E_3^{j,2})^\beta \cong (E_2^{j,2})^\beta$.

Note that the $E_3^{1,2}$ terms of both spectral sequences are outside the range of the d_3 differential, and that the d_3 differential vanishes on the $E_3^{4,0}$ terms. This implies that the rows in the following commutative diagram are exact:

$$\begin{tikzcd}
0 \arrow[r] & (E_4^{1,2})^\alpha \arrow[r] \arrow[d, "\rho_4^*"] & (E_3^{1,2})^\alpha \arrow[r, "(d_3^{1,2})^\alpha"] \arrow[d, "\rho_3^*"] & (E_3^{4,0})^\alpha \arrow[r] \arrow[d, "\rho_3^*"] & (E_4^{4,0})^\alpha \arrow[r] \arrow[d, "\rho_4^*"] & 0 \\
0 \arrow[r] & (E_4^{1,2})^\beta \arrow[r] & (E_3^{1,2})^\beta \arrow[r, "(d_3^{1,2})^\beta"] & (E_3^{4,0})^\beta \arrow[r] & (E_4^{4,0})^\beta \arrow[r] & 0
\end{tikzcd}$$

The fact that ρ_2^* induces isomorphisms on $E_2^{r,s}$ for $s \leq 2$ implies that ρ_3^* induces isomorphisms $(E_3^{1,2})^\alpha \to (E_3^{1,2})^\beta$ and $(E_3^{4,0})^\alpha \to (E_3^{4,0})^\beta$, since ρ_3^* is induced by ρ_2^* and each of these groups are isomorphic to the corresponding entries on the E_2–page of the spectral sequence. This shows that the middle two homomorphisms in the above commutative diagram are isomorphisms. □

Lemma 3.3 *The term $(E_4^{1,2})^\beta$ is trivial.*

Proof The term $(E_4^{1,2})^\beta$ is isomorphic to $(E_\infty^{1,2})^\beta$, so it suffices to show that $(E_\infty^{1,2})^\beta$ is trivial. The term $(E_\infty^{1,2})^\beta$ lies on the third diagonal of the E_∞–page for the spectral sequence $H^r(Q, H^s(\widetilde{M})) \Longrightarrow H^{r+s}(M)$. Since $H^r(Q, H^s(\widetilde{M}))$ is annihilated by $|Q|$ for any $r \geq 1$, the group $E_k^{r,s}$ is torsion for all $r \geq 1$. The groups $(E_\infty^{i,3-i})^\beta$ give successive quotients in the filtration

$$(E_\infty^{3,0})^\beta \cong F_3^3 \subseteq F_2^3 \subseteq F_1^3 \subseteq F_0^3 = H^3(M) \cong \mathbb{Z}.$$

Since $\mathbb{Z}$ is torsion free and $(E_\infty^{3,0})^\beta$ is torsion, $(E_\infty^{3,0})^\beta \cong 0$. This implies that

$$(E_\infty^{2,1})^\beta \cong F_2^3/F_3^3 \cong F_2^3/(E_\infty^{3,0}) \cong F_2^3,$$

so $F_2^3 \subseteq \mathbb{Z}$ is torsion and hence trivial as well. Applying this argument once more, we find that $(E_\infty^{1,2})^\beta \cong F_1^3/F_2^3$ must vanish as well. □

Lemma 3.4 *The term $(E_4^{4,0})^\alpha$ is isomorphic to $(E_4^{4,0})^\beta$.*

Proof By Lemma 3.3, the term in bottom left of the commutative diagram (3-1) from Lemma 3.2 is trivial. The commutativity of the diagram together with the fact that the second vertical map is an isomorphism shows that $(E_4^{1,2})^\alpha$ is also trivial. Since the first 3 maps in this commutative diagram of exact sequences are isomorphisms, a straightforward diagram chasing argument shows that the last map $\rho_4^*\colon (E_4^{4,0})^\alpha \to (E_4^{4,0})^\beta$ is an isomorphism as well. □

Lemma 3.5 *The term $(E_4^{4,0})^\beta$ is cyclic.*

Proof Note that there is an exact sequence

$$(E_4^{0,3})^\beta \xrightarrow{(d_4^{4,0})^\beta} (E_4^{4,0})^\beta \longrightarrow (E_5^{4,0})^\beta \longrightarrow 0.$$

Since $E_5^{4,0} \cong E_\infty^{4,0}$ and $H^4(\Gamma) = 0$, it follows that $(d_4^{4,0})^\beta$ is surjective. The group $(E_4^{0,3})^\beta$ is isomorphic to a subgroup of

$$(E_2^{0,3})^\beta \cong H^0(Q, H^3(\widetilde{M})) \cong \mathbb{Z},$$

so since $(E_4^{4,0})^\beta$ is isomorphic to a quotient of $(E_4^{0,3})^\beta$, $(E_4^{4,0})^\beta$ is cyclic. □

Lemma 3.6 *The map $q^*\colon H^4(Q) \to H^4(G)$ has cyclic image.*

Proof Recall that the image of $q^*\colon H^4(Q) \to H^4(G)$ is isomorphic to $(E_\infty^{4,0})^\alpha$, and that $(E_\infty^{4,0})^\alpha$ is isomorphic to a quotient of $(E_4^{4,0})^\alpha$. By Lemma 3.4,

$$(E_4^{4,0})^\alpha \cong (E_4^{4,0})^\beta,$$

and by Lemma 3.5, $(E_4^{4,0})^\beta$ is cyclic. Since quotients of cyclic groups are cyclic, the result follows. □

We are now ready to prove Lemma 1.4, which is an immediate consequence of Lemma 2.2 from the previous section and Lemma 3.6.

Proof of Lemma 1.4 Let ω_1 and ω_2 be elements of $H^2(G)$. By Lemma 2.2, $q^*\colon H^2(Q) \to H^2(G)$ is surjective, so there exist α_1, $\alpha_2 \in H^2(Q)$ such that $q^*(\alpha_1) = \omega_1$ and $q^*(\alpha_2) = \omega_2$. It follows that

$$\omega_1 \smile \omega_2 = q^*(\alpha_1) \smile q^*(\alpha_2) = q^*(\alpha_1 \smile \alpha_2),$$

so any cup product of elements in $H^2(G)$ lies in $q^*(H^4(Q))$. By Lemma 3.6, the image of $q^*\colon H^4(Q) \to H^4(G)$ is cyclic. □

4 The proof of Theorem 1.1

We now turn to the proof of the main theorem.

Proof of Theorem 1.1 Since $\ker(q) \subseteq \Gamma^{(2)}$ and $\Gamma^{(2)} \subseteq \Gamma^{(1)}$, the nondegeneracy of the cup-product pairing follows from Theorem 1.3. The cyclicity of the image C of the cup-product pairing $H^2(G) \otimes H^2(G) \to H^4(G)$ follows from Lemma 3.6, so it remains to demonstrate the existence of the desired embedding $i\colon C \to \mathbb{Q}/\mathbb{Z}$.

Let $\psi\colon H^4(G) \to H_0(G, \mathbb{Q}/\mathbb{Z})$ denote the map $\alpha \mapsto \widetilde{q}_*([M]) \frown \beta^{-1}(\alpha)$, and let i denote the restriction of ψ to C. By Lemma 2, given ω_1 and ω_2 in $H^2(G)$,

$$i(\omega_1 \smile \omega_2) = \widetilde{q}_*([M]) \frown \beta^{-1}(\omega_1 \smile \omega_2) = \widetilde{q}_*(\lambda(\widetilde{q}^*(\omega_1), \widetilde{q}^*(\omega_2))). \tag{4-1}$$

After identifying $H_0(G, \mathbb{Q}/\mathbb{Z})$ and $H_0(M, \mathbb{Q}/\mathbb{Z})$ with $\mathbb{Q}/\mathbb{Z}$ in the natural way, the last term of this equation is equal to $\langle [M] \frown \widetilde{q}^*(\omega_1), [M] \frown \widetilde{q}^*(\omega_2)\rangle$.

It remains to check that $i\colon C \to \mathbb{Q}/\mathbb{Z}$ is injective. Given an abelian group A, let $\exp(A)$ denote the maximal order of an element of A. Note that for a finite abelian group A, $\exp(A) = \exp(A \otimes A)$. Since C is cyclic and is isomorphic to a quotient of $H^2(G) \otimes H^2(G)$, it follows that the order of C divides $\exp(H^2(G) \otimes H^2(G)) = \exp(H^2(G))$. It therefore suffices to show that $i(C)$ contains an element of order $\exp(H^2(G))$.

By Lemma 2.4 there exist elements $\eta_1, \eta_2 \in H^2(M)$ such that the order of $\lambda(\eta_1, \eta_2)$ is equal to $\exp(H^2(M))$. Since $\ker(q) \subseteq \Gamma^{(1)}$, $q^*\colon H^2(G) \to H^2(M)$ is an isomorphism by Lemma 2.3, so $\exp(H^2(G)) = \exp(H^2(M))$, and there exist elements ω_1 and ω_2 such that $q^*(\omega_i) = \eta_i$. Equation (4-1) above therefore shows that

$$i(\omega_1 \smile \omega_2) = \widetilde{q}_*(\lambda(\eta_1, \eta_2)),$$

and since $\widetilde{q}_*\colon H_0(M, \mathbb{Q}/\mathbb{Z}) \to H_0(G, \mathbb{Q}/\mathbb{Z})$ is an isomorphism,

$$\operatorname{ord}(i(\omega_1 \smile \omega_2)) = \operatorname{ord}(\widetilde{q}_*(\lambda(\eta_1, \eta_2))) = \operatorname{ord}(\lambda(\eta_1, \eta_2)) = \exp(H^2(G)). \qquad \square$$

References

[1] **I Agol**, *The virtual Haken conjecture*, Doc. Math. 18 (2013) 1045–1087 MR3104553

[2] **M F Atiyah**, *Characters and cohomology of finite groups*, Inst. Hautes Études Sci. Publ. Math. 9 (1961) 23–64 MR0148722

[3] **M Baker**, **M Boileau**, **S Wang**, *Towers of covers of hyperbolic 3–manifolds*, Rend. Istit. Mat. Univ. Trieste 32 (2001) 35–43 MR1893391

[4] **K S Brown**, *Cohomology of groups*, Graduate Texts in Mathematics 87, Springer, New York (1994) MR1324339

[5] **D Cooper**, **D D Long**, *Free actions of finite groups on rational homology 3–spheres*, Topology Appl. 101 (2000) 143–148 MR1732066

[6] **L Evens**, **S Priddy**, *The cohomology of the semidihedral group*, from: "Conference on algebraic topology in honor of Peter Hilton", (R Piccinini, D Sjerve, editors), Contemp. Math. 37, Amer. Math. Soc. (1985) 61–72 MR789794

[7] **D Gorenstein**, *Finite groups*, 2nd edition, Chelsea, New York (1980) MR569209

[8] **D Handel**, *On products in the cohomology of the dihedral groups*, Tohoku Math. J. 45 (1993) 13–42 MR1200878

[9] **T Hayami**, **K Sanada**, *Cohomology ring of the generalized quaternion group with coefficients in an order*, Comm. Algebra 30 (2002) 3611–3628 MR1922301

[10] **J Kahn**, **V Markovic**, *Immersing almost geodesic surfaces in a closed hyperbolic three manifold*, Ann. of Math. 175 (2012) 1127–1190 MR2912704

[11] **A Kawauchi**, **S Kojima**, *Algebraic classification of linking pairings on 3–manifolds*, Math. Ann. 253 (1980) 29–42 MR594531

[12] **A Lubotzky**, **D Segal**, *Subgroup growth*, Progress in Mathematics 212, Birkhäuser, Basel (2003) MR1978431

[13] **J Milnor**, *Groups which act on S^n without fixed points*, Amer. J. Math. 79 (1957) 623–630 MR0090056

[14] **A Reznikov**, *Three-manifolds class field theory (homology of coverings for a nonvirtually b_1–positive manifold)*, Selecta Math. 3 (1997) 361–399 MR1481134

[15] **S K Roushon**, *Topology of 3–manifolds and a class of groups, II*, Bol. Soc. Mat. Mexicana 10 (2004) 467–485 MR2200344

[16] **D T Wise**, *Research announcement: The structure of groups with a quasiconvex hierarchy*, Electron. Res. Announc. Math. Sci. 16 (2009) 44–55 MR2558631

McKinsey and Company, 110 Charles Street
Toronto ON, M5S 1K9, Canada

wpcavendish@gmail.com

Received: 30 July 2014 Revised: 13 April 2015

Algebraic & Geometric Topology 15 (2015) 3371–3408

A colored operad for string link infection

JOHN BURKE
ROBIN KOYTCHEFF

Budney constructed an operad that encodes splicing of knots and further showed that the space of (long) knots is generated over this splicing operad by the space of torus knots and hyperbolic knots. This generalized the satellite decomposition of knots from isotopy classes to the level of the space of knots. Infection by string links is a generalization of splicing from knots to links. We construct a colored operad that encodes string link infection. We prove that a certain subspace of the space of 2–component string links is generated over a suboperad of our operad by its subspace of prime links. This generalizes a result from joint work with Blair from isotopy classes of string links to the space of string links. Furthermore, all the relations in the monoid of 2–string links (as determined in our joint work with Blair) are captured by our infection operad.

57M25, 18D50, 55P48, 57R40, 57R52

1 Introduction

This paper concerns operations on knots and links, particularly infection by string links. Classically, knots and links are considered as isotopy classes of embeddings of a 1–manifold into a 3–manifold, such as $\mathbb{R}^3$, D^3, or S^3. Instead of considering just isotopy classes, we consider the whole *space* of links, that is the space of embeddings of a certain 1–manifold into a certain 3–manifold. We also consider spaces parametrizing the operations and organize all of these spaces via the concept of an operad (or colored operad). The operad framework is in turn convenient for studying spaces of links and generalizing statements about isotopy classes to the space level. Finding such statements to generalize was the motivation for recent work of the authors and R Blair on isotopy classes of string links [1].

Our work closely follows the work of Budney. Budney first showed that the little 2–cubes operad $\mathcal{C}_2$ acts on the space $\mathcal{K}$ of (long) knots, which implies the well-known commutativity of connect-sum of knots on isotopy classes. He showed that $\mathcal{K}$ is freely generated over $\mathcal{C}_2$ by the space $\mathcal{P}$ of prime knots, generalizing the prime decomposition of knots of Schubert from isotopy classes to the level of the space of knots [2]. Later,

Published: 12 January 2016 DOI: 10.2140/agt.2015.15.3371

he constructed a splicing operad $\mathcal{SP}$ which encodes splicing of knots. He showed that $\mathcal{K}$ is freely generated over a certain suboperad of $\mathcal{SP}$ by the subspace of torus and hyperbolic knots, thus generalizing the satellite decomposition of knots from isotopy classes to the space level [4].

Infection by string links is a generalization of splicing from knots to links. This operation is most commonly used in studying knot concordance. One instance where string link infection arises is in the clasper surgery of Habiro [15], which is related to finite-type invariants of knots and links. In another vein, Cochran, Harvey, and Leidy observed that iterating the infection operation gives rise to a fractal-like structure [9]. This motivated our work, and we provide another perspective on the structure arising from string link infection. We do this by constructing a colored operad which encodes this infection operation. We then prove a statement that decomposes part of the space of 2–component string links via our colored operad.

Splicing and infection are both generalizations of the connect-sum operation. The latter is always a well defined operation on isotopy classes of knots, but if one considers long knots, it is even well defined on the knots themselves. This connect-sum operation (ie "stacking") is also well defined for long (aka string) links with any number of components. Thus we restrict our attention to string links.

1.1 Basic definitions and remarks

Let $I = [-1, 1]$ and let $D^2 \subset \mathbb{R}^2 \cong \mathbb{C}$ be the unit disk with boundary.

Definition 1.1 A *c–component string link* (or *c–string link*) is a proper embedding of c disjoint intervals

$$\coprod_c I \hookrightarrow I \times D^2$$

whose values and derivatives of all orders at the boundary points agree with those of a fixed embedding i_c. For concreteness, we take i_c to be the map which on the i^{th} copy of I is given by $t \mapsto (t, x_i)$, where $x_i = ((i-1)/c, 0)$. We will call i_c the trivial string link. Another example of a string link is shown in Figure 1.

In our work [1], our definition of string links allowed more general embeddings, and the ones defined above were called "pure string links". We choose the definition above in this paper because infection by string links behaves more nicely with this more restrictive notion of string link. (Specifically, it preserves the number of components in the infected link.)

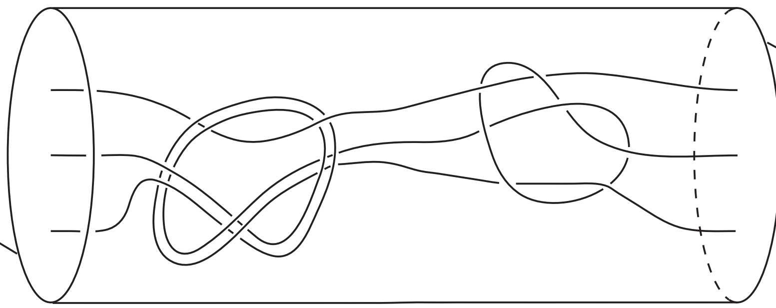


Figure 1: A string link

The condition on derivatives is not always required in the literature.[1] We impose it because this allows us to identify a c–string link with an embedding $\coprod_c \mathbb{R} \hookrightarrow \mathbb{R} \times D^2$ which agrees with a fixed embedding outside of $I \times D^2$. Let $\mathcal{L}_c = \mathrm{Emb}(\coprod_c \mathbb{R}, \mathbb{R} \times D^2)$ denote the space of c–string links, equipped with the C^∞ Whitney topology. An isotopy of string links is a path in this space, so the path components of $\mathcal{L}_c$ are precisely the isotopy classes of c–string links. Often we will write $\mathcal{K}$ for the space $\mathcal{L}_1$ of long knots.

The braids which qualify as string links under Definition 1.1 are precisely the pure braids. There is a map from $\mathcal{L}_c$ to the space $\mathrm{Emb}(\coprod_c S^1, \mathbb{R}^3)$ of closed links in $\mathbb{R}^3$ by taking the closure of a string link. When $c = 1$, this map is an isomorphism on π_0. In other words, isotopy classes of long knots correspond to isotopy classes of closed knots. In general, this map is easily seen to be surjective on π_0, but it is not injective on π_0. For example, any string link and its conjugation by a pure braid yield isotopic closed links, and for $c \geq 3$, there are conjugations of string links by braids which are not isotopic to the original string link. We will sometimes write just "link" rather than "string link" or "closed link" when the type of link is either clear from the context or unimportant.

1.2 Main results

Our first main result is the construction of a colored operad encoding string link infection. An operad $\mathcal{O}$ consists of spaces $\mathcal{O}(n)$ of n–ary operations for all $n \in \mathbb{N}$. Roughly, an operad acts on a space X if each $\mathcal{O}(n)$ can parametrize ways of multiplying n elements in X. (We provide thorough definitions in Section 3.) A colored operad arises when different types of inputs must be treated differently. In our case, we have to treat string links with different numbers of components differently, so the colors in

[1] The homotopy type of the space of such embeddings would be unchanged by omitting the condition on derivatives, since the space of possible tangent vectors and higher-order derivatives at the boundary is contractible.

our colored operad are the natural numbers. This theorem is proven as Theorem 5.6 and Proposition 6.3.

Theorem 1 *There is a colored operad $\mathcal{I}$ which encodes the infection operation and acts on spaces of string links $\mathcal{L}_c$ for $c = 1, 2, 3, \ldots$.*

- *When restricting to the color 1, the (ordinary) operad $\mathcal{I}_{\{1\}}$ which we recover is Budney's splicing operad, and the action of $\mathcal{I}_{\{1\}}$ on $\mathcal{K}$ is the same as Budney's splicing operad action.*
- *For any c, the operad $\mathcal{I}_{\{c\}}$ obtained by restricting to c is an operad which admits a map from the little intervals operad $\mathcal{C}_1$. The resulting $\mathcal{C}_1$–action on $\mathcal{L}_c$ encodes the operation of stacking string links.*
- *On the level of π_0, our infection operad encodes all the relations in the whole 2–string link monoid.*

We then use our colored operad to decompose part of the space of string links. We rely on an analogue of prime decomposition for 2–string links proven in our joint work with R Blair [1], so we must restrict to $c = 2$. We consider a "stacking operad" $\mathcal{I}_{\#}$, which is a suboperad of $\mathcal{I}_{\{2\}}$ and which is homeomorphic to the little intervals operad. This operad simply encodes the operation of stacking 2–string links in $I \times D^2$, with the little intervals acting in the I factor. The theorem below is proven as Theorem 6.8.

Theorem 2 *Let $\pi_0\mathcal{S}_2$ denote the submonoid of $\pi_0\mathcal{L}_2$ generated by those prime 2–string links which are not central. (By [1], this monoid is free.) Let $\mathcal{S}_2$ be the subspace of $\mathcal{L}_2$ consisting of the path components of $\mathcal{L}_2$ that are in $\pi_0\mathcal{S}_2$. Then $\pi_0\mathcal{S}_2$ is freely generated as a monoid over the stacking suboperad $\mathcal{I}_{\#}$. The generating space is the subspace consisting of those components in $\mathcal{S}_2$ which correspond to prime string links.*

1.3 Organization of the paper

In Section 2, we review the definition of string link infection.

In Section 3, we review the definitions of an operad and the particular example of the little cubes operad. We then give the more general definition of a colored operad.

In Section 4, we review Budney's operad actions on the space of knots. This includes his action of the little 2–cubes operad, as well as the action of his splicing operad.

In Section 5, we define our colored operad for infection and prove Theorem 1. We make some remarks about our operad related to pure braids and rational tangles, and we briefly discuss a generalization to embedding spaces of more general manifolds.

In Section 6, we focus on the space of 2–string links. We prove Theorem 2, which decomposes part of the space of 2–string links in terms of a suboperad of our infection colored operad. We conclude with several other statements about the homotopy type of certain components of the space of 2–string links.

Notation

- $\coprod_c X$ means $\overbrace{X \sqcup \cdots \sqcup X}^{c \text{ times}}$.
- $f|A$ means the restriction of f to A.
- $\overline{X}$ denotes the closure of X; $\mathring{X}$ denotes the interior of X.
- $[a]$ denotes the equivalence class represented by an element a; $[a_1, \dots, a_n]$ denotes the equivalence class of a tuple $(a_1, \dots, a_n)$.

Acknowledgments The authors thank Tom Goodwillie for useful comments and conversations. They thank Ryan Budney for useful explanations and especially for his work which inspired this project. They thank Ryan Blair for useful conversations and for invaluable contributions in their joint work with him, on which Theorem 2 depends. They thank a referee for a careful reading of the paper and useful comments. They thank Connie Leidy for suggesting the rough idea of this project. They thank Slava Krushkal for suggesting terminology and for pointing out the work of Habiro. Finally, they thank David White for introducing the authors to each other. The second author was supported partly by NSF grant DMS-1004610 and partly by a PIMS Postdoctoral Fellowship.

2 Infection

Infection is an operation which takes a link with additional decoration together with a string link and produces a link. This operation is a generalization of splicing which in turn is a generalization of the connect-sum operation. Infection has been called multi-infection by Cochran, Friedl, and Teichner [8], infection by a string link by Cochran [7] and tangle sum by Cochran and Orr [10]. Special cases of this construction have been used extensively since the late 1970s, for example in the work of Gilmer [14], Livingston [22], Cochran, Orr, and Teichner [11; 12], Harvey [16], and Cimasoni [6]. The operad we define in this paper will encode a slightly more general operation than the infection operation that has been defined in previous literature. This section is meant to inform the reader of the definition in previous literature and provide motivation for the infection operad.

2.1 Splicing

Consider a link $R \in S^3$ and a closed curve $\eta \in S^3 \setminus R$ such that η bounds an embedded disk in S^3 (η is unknotted in S^3) which intersects the link components transversely. Given a knot K, one can create a new link $R_\eta(K)$, with the same number of components as R, called the result of splicing R by K at η. Informally, the splicing process is defined by taking the disk in S^3 bounded by η, cutting R along the disk, grabbing the cut strands, tying them into the knot K (with no twisting among the strands) and regluing. The result of splicing given a particular R, η and K is show in Figure 2. Note that if η simply linked one strand of R then the result of the splicing would be isotopic to the connect-sum of R and K.

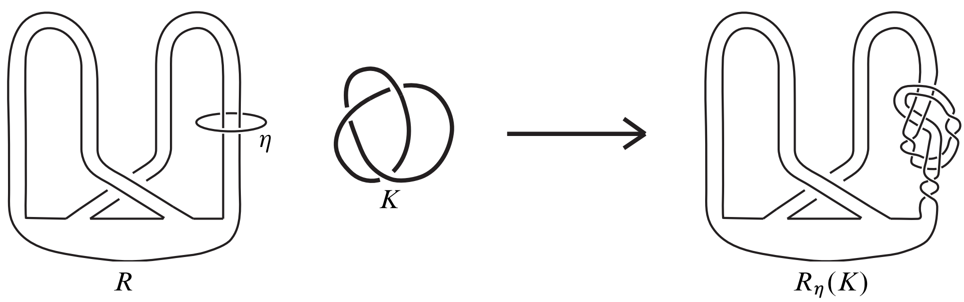


Figure 2: An example of the splicing operation

Formally, $R_\eta(K)$ is arrived at by first removing a tubular neighborhood, $N(\eta)$, of η from S^3. Note $S^3 \setminus N(\eta) \subset S^3$ is a solid torus with R embedded in its interior. Let C_K denote the complement in S^3 of a tubular neighborhood of K. Since the boundary of C_K is also a torus, one can identify these two manifolds along their boundary. In order to specify the identification, we use the terminology of meridians and longitudes. Recall that the meridian of a knot is the simple closed curve, up to ambient isotopy, on the boundary of the complement of the knot which bounds a disk in the closure of the tubular neighborhood of the knot and has $+1$ linking number with the knot. Also recall that the longitude of a knot is the simple closed curve, up to ambient isotopy, on the boundary of the complement of the knot which has $+1$ intersection number with the meridian of the knot and has zero linking number with the knot.

The gluing of $S^3 \setminus N(\eta)$ to C_K is done so that the meridian of the boundary of $S^3 \setminus N(\eta)$ is identified with the meridian of K in the boundary of C_K. Note that this process describes a Dehn surgery with surgery coefficient ∞ along $K \subset S^3$ where the solid torus glued in is $S^3 \setminus N(\eta)$. Thus, the resulting manifold will be a 3–sphere with a subset of disjoint embedded circles whose union is $R_\eta(K)$ (the image of R under this identification). Although the embedding of $R_\eta(K)$ in S^3 depends on the identification of the surgered 3–manifold with S^3, its isotopy class is independent of this choice of identification.

2.2 String link infection

Although there is a well-studied generalization of the connect-sum operation from closed knots to closed links, there is no generalization of splicing by a closed link. There is, however, a generalization of splicing called infection by a string link, which we will now define. See the work of Cochran, Friedl, and Teichner [8, Section 2.2] for a thorough reference.

By an r*–multi-disk* $\mathbb{D}$ we mean the oriented disk D^2 together with r ordered embedded open disks $D_1, \dots D_r$ (see Figure 3). Given a link $L \subset S^3$ we say that an embedding $\varphi\colon \mathbb{D} \to S^3$ of an r–multi-disk into S^3 is *proper* if the image of the multi-disk, denoted by $\mathcal{D}$, intersects the link components transversely and only in the images of the disks $D_1, \dots D_r$ as in Figure 3. We will refer to the image of the boundary curves of $\varphi(D_1), \dots, \varphi(D_r)$ by $\eta_1, \dots, \eta_r$.

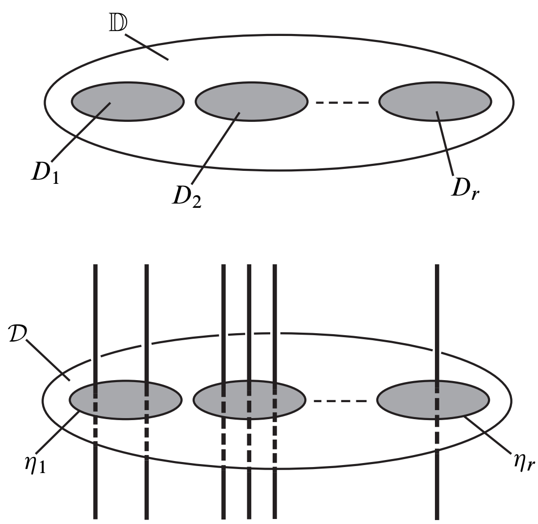


Figure 3: An r–multi-disk and a properly embedded multi-disk

Suppose $R \subset S^3$ is link, $\mathcal{D} \subset S^3$ is the image of a properly embedded r–multi-disk, and L is an r–component string link. Then informally, the infection of R by L at $\mathcal{D}$, denoted by $R_{\mathcal{D}}(L)$, is the link obtained by tying the r collections of strands of R that intersect the disks $\varphi(D_1), \dots, \varphi(D_r)$ into the pattern of the string link L, where the strands linked by η_i are identified with the i^{th} component of L, such that the i^{th}

collection of strands are parallel copies of the i^{th} component of L. Figure 5 shows an example of this operation.

We now define this operation formally. Given a string link $L\colon \coprod_r I \hookrightarrow I \times D^2$, let C_L denote the complement of a tubular neighborhood of (the image of) L in $I \times D^2$. In Figure 4, an example of a string link is shown with its complement to the right. The meridian of a component of a string link is the simple closed curve, up to ambient isotopy, on the $I \times \partial D^2$ boundary of the closure of the tubular neighborhood of the component which bounds a disk and has $+1$ linking number with the component. We call the set of such meridians the meridians of the string link. The longitude of a component of a string link is a properly embedded line segment $f\colon I \to I \times \partial D^2$, up to ambient isotopy, on the $I \times \partial D^2$ boundary of the closure of the tubular neighborhood of the component; it is required to have $+1$ intersection number with the meridian of that component, to have zero linking number with that component, and to satisfy $f(0) = (1,0) \in \{0\} \times \partial D^2$ and $f(1) = (1,1) \in \{1\} \times \partial D^2$. We call the set of such longitudes the longitudes of the string link. In Figure 4 the meridians μ_i and longitudes ℓ_i are shown on the boundary of the complement. Note that the boundary of the complement of any r–component string link is homeomorphic to a genus-r orientable surface.

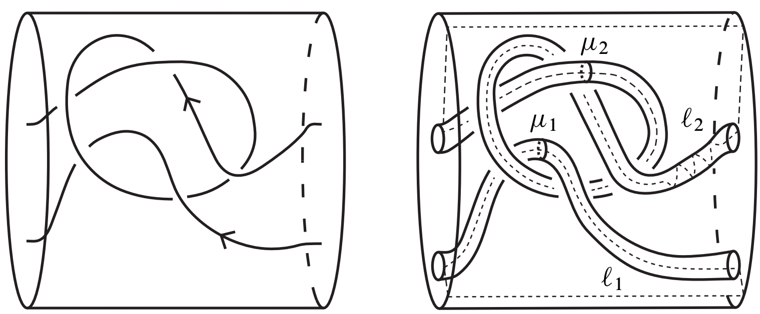


Figure 4: A string link and its complement

Let $R \subset S^3$ be a link, and let $L\colon \coprod_r I \hookrightarrow I \times D^2 \subset S^3$ be a string link. Fix a proper embedding of a thickened r–multidisk $I \times \mathcal{D}$ in $S^3 \setminus R$. Formally the infection of R by L at $\mathcal{D}$ is obtained by removing $I \times (\mathcal{D} \setminus \bigsqcup_i \varphi(D_i))$ from S^3 and gluing in the complement of L. Note that $I \times (\mathcal{D} \setminus \bigsqcup_i \varphi(D_i))$ is the complement of a r–component trivial string link T (see Figure 5), and thus the boundary of $S^3 \setminus (I \times (\mathcal{D} \setminus \bigsqcup_i \varphi(D_i)))$ is a genus-r orientable surface. One identifies the boundary of $S^3 \setminus (I \times (\mathcal{D} \setminus \bigsqcup_i \varphi(D_i)))$ and the boundary of the complement of L, ∂C_L, by

- identifying $I \times \partial\mathcal{D} \subset S^3 \setminus (I \times (\mathcal{D} \setminus \bigsqcup_i \varphi(D_i)))$ with $I \times \partial D^2 \subset \partial C_L$,
- identifying $\{0,1\} \times (\mathcal{D} \setminus \bigsqcup_i \varphi(D_i)) \subset S^3 \setminus (I \times (\mathcal{D} \setminus \bigsqcup_i \varphi(D_i)))$ with $(\{0,1\} \times D^2) \setminus N(L) \subset \partial C_L$, and

- identifying $I \times \partial\varphi(D_i) \subset S^3 \setminus (I \times (\mathcal{D} \setminus \bigsqcup_i \varphi(D_i)))$ with the boundaries of the tubular neighborhoods of the components of L in ∂C_L in such a way that the meridians μ_i and longitudes ℓ_i of L are identified with $\{\frac{1}{2}\} \times \partial\varphi(D_i)$ and $I \times \{0\}$ respectively.

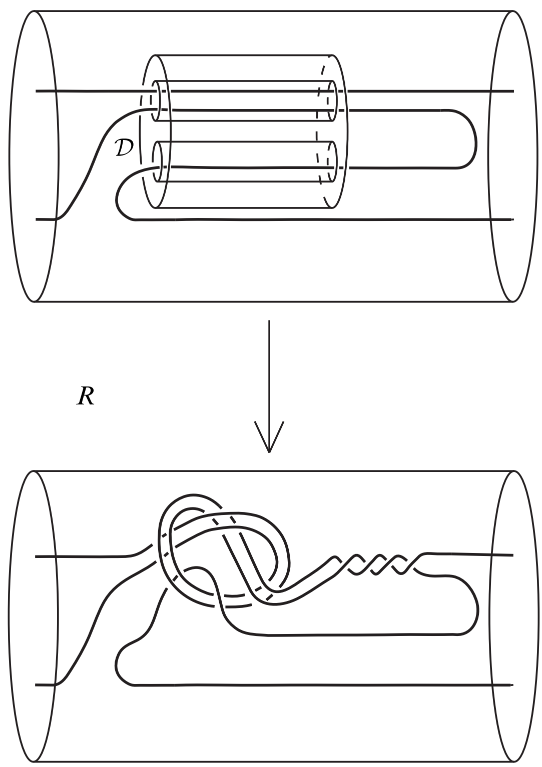


Figure 5: Infection of the string link R along $\mathcal{D}$ by the string link L from Figure 4

We claim that the resulting manifold is S^3 containing a link $R_{\mathcal{D}}(L)$ (which is the image of R under this identification). The resulting manifold is homeomorphic to S^3 because

$$S^3 \setminus \mathrm{Int}\left(I \times \left(\mathcal{D} \setminus \bigsqcup_i \varphi(D_i)\right)\right) \cup (I \times D^2) \setminus N(L)$$
$$= \big(S^3 \setminus (I \times \mathcal{D})\big) \cup \left(I \times \bigsqcup_i \varphi(D_i) \cup (I \times D^2) \setminus N(L)\right) \cong S^3,$$

where the last homeomorphism follows form the observation that the previous space is the union of two 3–balls. Again, the specific embedding of $R_{\mathcal{D}}(L)$ will depend on the choice of homeomorphism, but all choices will yield isotopic embeddings.

3 Operads

We start by reviewing the definitions of an operad $\mathcal{O} = \{\mathcal{O}(n)\}_{n\in\mathbb{N}}$, and an action of $\mathcal{O}$ on X (aka an algebra X over $\mathcal{O}$). We then proceed to colored operads. Technically, the definition of a colored operad subsumes the definition of an ordinary operad, but for ease of readability, we first present ordinary operads. Readers familiar with these concepts may safely skip this section.

3.1 Operads

Operads can be defined in any symmetric monoidal category, but we will only consider the category of topological spaces. In this case, the rough idea is as follows. An algebra X over an operad $\mathcal{O}$ is a space with a multiplication $X \times X \to X$, and the space $\mathcal{O}(n)$ parametrizes ways of multiplying n elements of X, ie maps $X^n \to X$. In other words, $\mathcal{O}(n)$ captures homotopies between different ways of multiplying the elements, as well as homotopies between these homotopies, etc. Thus an element of $\mathcal{O}(n)$ is an operation with n inputs and one output. This can be visualized as a tree with n leaves and a root, and in fact, free operads are certain spaces of decorated trees. For a more detailed introduction, the reader may wish to consult the book of Markl, Shnider, and Stasheff [23], May's book [24], or the expository paper of McClure and Smith [25].

Definition 3.1 An operad $\mathcal{O}$ (in the category of spaces) consists of

- a space $\mathcal{O}(n)$ for each $n = 1, 2, \dots$ with an action of the symmetric group Σ_n,
- structure maps

(1) $$\mathcal{O}(n) \times \mathcal{O}(k_1) \times \cdots \times \mathcal{O}(k_n) \to \mathcal{O}(k_1 + \cdots + k_n),$$

such that the following three conditions are satisfied:

Associativity The following diagram commutes:

$$\begin{tikzcd}
\mathcal{O}(n) \times \prod_{i=1}^{n} \mathcal{O}(k_i) \times \prod_{i=1}^{n} \prod_{j=1}^{k_i} \mathcal{O}(\ell_{i,j}) \arrow[r] \arrow[d] & \mathcal{O}(n) \times \prod_{i=1}^{n} \mathcal{O}\Big(\sum_{j=1}^{k_i} \ell_{i,j}\Big) \arrow[d] \\
\mathcal{O}(k_1 + \cdots + k_n) \times \prod_{i=1}^{n} \prod_{j=1}^{k_i} \mathcal{O}(\ell_{i,j}) \arrow[r] & \mathcal{O}\Big(\sum_{i=1}^{n} \sum_{j=1}^{k_i} \ell_{i,j}\Big)
\end{tikzcd}$$

Symmetry Let $\sigma \times \sigma$ denote the diagonal action on the product

$$\mathcal{O}(n) \times (\mathcal{O}(k_1) \times \cdots \times \mathcal{O}(k_n))$$

coming from the actions of Σ_n on $\mathcal{O}(n)$ and on $\mathcal{O}(k_1) \times \cdots \times \mathcal{O}(k_n)$ by permuting the factors. For a partition $\vec{k} = (k_1, \ldots, k_n)$ of a natural number $k_1 + \cdots + k_n$, let $\sigma_{\vec{k}} \in \Sigma_{k_1+\cdots+k_n}$ denote the "block permutation" induced by σ and $\vec{k}$.

We require that the following composition agrees with the map of Equation (1):

$$\mathcal{O}(n) \times \prod_{i=1}^{n} \mathcal{O}(k_i) \xrightarrow{\sigma\times\sigma} \mathcal{O}(n) \times \prod_{i=1}^{n} \mathcal{O}(k_{\sigma(i)}) \longrightarrow \mathcal{O}\Big(\sum_{i=1}^{n} k_i\Big) \xrightarrow{\sigma_{\vec{k}}^{-1}} \mathcal{O}\Big(\sum_{i=1}^{n} k_i\Big).$$

We also require that for $\tau_i \in \Sigma_{k_i}$ for $i = 1, \ldots, n$, the following diagram commutes:

$$\begin{tikzcd}
\mathcal{O}(n) \times \prod_{i=1}^{n} \mathcal{O}(k_i) \arrow[r] \arrow[d, "\mathrm{id}\times\tau_1\times\cdots\times\tau_n"'] & \mathcal{O}\Big(\sum_{i=1}^{n} k_i\Big) \arrow[d, "\tau_1\times\cdots\times\tau_n"] \\
\mathcal{O}(n) \times \prod_{i=1}^{n} \mathcal{O}(k_i) \arrow[r] & \mathcal{O}\Big(\sum_{i=1}^{n} k_i\Big)
\end{tikzcd}$$

Identity There exists an element $1 \in \mathcal{O}(1)$ (ie a map $* \to \mathcal{O}(1)$) which induces the identity on $\mathcal{O}(k)$ via

$$\begin{aligned} \mathcal{O}(1) \times \mathcal{O}(k) &\to \mathcal{O}(k), \\ (1, L) &\mapsto L, \end{aligned}$$

and which induces the identity on $\mathcal{O}(n)$ via

$$\begin{aligned} \mathcal{O}(n) \times \mathcal{O}(1) \times \mathcal{O}(1) \times \cdots \times \mathcal{O}(1) &\to \mathcal{O}(n), \\ (L, 1, 1, \ldots, 1) &\mapsto L. \end{aligned}$$

Some authors define the structure maps via $\circ_i$ operations, ie plugging in just one operation into the i^{th} input, as opposed to n operations into all n inputs. These $\circ_i$ maps can be recovered from the above definition by setting $k_j = 1$ for all $j \neq i$ and using the identity element in $\mathcal{O}(1)$.

Definition 3.2 Given an operad $\mathcal{O}$, an *action of $\mathcal{O}$ on X* (also called an *algebra X over $\mathcal{O}$*) is a space X together with maps

$$\mathcal{O}(n) \times X^n \to X$$

such that the following conditions are satisfied:

Associativity The following diagram commutes:

$$\begin{tikzcd}
\mathcal{O}(n)\times\mathcal{O}(k_1)\times\ldots\times\mathcal{O}(k_n)\times X^{k_1+\cdots+k_n} \arrow[r] \arrow[d] & \mathcal{O}(n)\times X^n \arrow[d] \\
\mathcal{O}(k_1+\cdots+k_n)\times X^{k_1+\cdots+k_n} \arrow[r] & X
\end{tikzcd}$$

Symmetry For each n, the action map is Σ_n–invariant, where Σ_n acts on $\mathcal{O}(n)$ by definition, on X^n by permuting the factors, and on the product diagonally. In other words, the action map descends to a map

$$\mathcal{O}(n)\times_{\Sigma_n} X^n \to X.$$

Identity The identity element $1\in\mathcal{O}(1)$ together with the map

$$\mathcal{O}(1)\times X\to X$$

induce the identity map on X, ie the map takes $(1,x)\mapsto x$.

3.2 The little cubes operad

Our infection colored operad extends Budney's splicing operad, which in turn was an extension of Budney's action of the little 2–cubes operad on the space of long knots. Thus the little 2–cubes operad is of interest here.

Definition 3.3 The *little j–cubes operad $\mathcal{C}_j$* is the operad with $\mathcal{C}_j(n)$ the space of maps

$$(L_1,\ldots,L_n)\colon \coprod_n I^j \hookrightarrow I^j$$

which are embeddings when restricted to the interiors of the I^j and which are increasing affine-linear maps in each coordinate. The structure maps are given by composition:

$$\begin{aligned}
&\mathcal{C}_j(n)\times\mathcal{C}_j(k_1)\times\ldots\times\mathcal{C}_j(k_n)\longrightarrow\mathcal{C}_j(k_1+\cdots+k_n),\\
&(L_1,\ldots,L_n),(L^1_1,\ldots,L^1_{k_1}),\ldots,(L^n_1,\ldots,L^n_{k_n})\\
&\qquad\qquad\mapsto (L_1\circ(L^1_1,\ldots,L^1_{k_1}),\ldots,L_n\circ(L^n_1,\ldots,L^n_{k_n})).
\end{aligned}$$

Note that for all $j\ge 2$, the multiplication induced by choosing (any) element in $\mathcal{C}_j(2)$ is commutative *up to homotopy*, which can be seen via the same picture that shows that $\pi_j X$ is abelian for $j\ge 2$.

3.3 Colored operads

Now we present the precise definitions of a colored operad and an action of a colored operad on a space. This generalization of an operad is necessary to generalize Budney's operad from splicing of knots to infection by links.

Definition 3.4 A *colored operad* $\mathcal{O} = (\mathcal{O}, C)$ (in the category of spaces) consists of:

- A set of colors C.
- A space $\mathcal{O}(c_1, \dots, c_n; c)$ for each $(n+1)$–tuple $(c_1, \dots, c_n, c) \in C$ together with compatible maps $\mathcal{O}(c_1, \dots, c_n; c) \to \mathcal{O}(c_{\sigma(1)}, \dots, c_{\sigma(n)}; c)$ for each $\sigma \in \Sigma_n$.
- (Continuous) maps

$$\mathcal{O}(c_1, \dots, c_n; c) \times \mathcal{O}(d_{1,1}, \dots, d_{1,k_1}; c_1) \times \dots \times \mathcal{O}(d_{n,1}, \dots, d_{n,k_n}; c_n) \longrightarrow \mathcal{O}(d_{1,1}, \dots, d_{n,k_n}; c).$$

Here the maps satisfy the following three conditions:

Associativity The map below is the same regardless of whether one first applies the structure maps to the first two factors or the last two factors on the left-hand side:

$$\mathcal{O}(c_1, \dots, c_n; c) \times \prod_{i=1}^{n} \mathcal{O}(d_{i,1}, \dots, d_{i,k_i}; c_i) \times \prod_{i=1}^{n} \prod_{j=1}^{k_i} \mathcal{O}(e_{i,j,1}, \dots, e_{i,j,\ell_{i,j}}; d_{i,j}) \longrightarrow \mathcal{O}(e_{1,1,1}, \dots, e_{n,k_n,\ell_{1,k_n}}).$$

Symmetry The following diagram below commutes. The vertical map is induced by σ in both the first factor and the last n factors, and $\sigma_{\vec{k}} \in \Sigma_{k_1+\dots+k_n}$ is the block permutation induced by σ and the partition $(k_1, \dots, k_n)$:

$$\begin{tikzcd}
\mathcal{O}(c_1, \dots, c_n; c) \times \prod_{i=1}^{n} \mathcal{O}(d_{i,1}, \dots, d_{i,k_i}; c_i) \arrow[r] \arrow[d, "\sigma\times\sigma"'] & \mathcal{O}(d_{1,1}, \dots, d_{n,k_n}; c) \arrow[d, "\sigma_{\vec{k}}"] \\
\mathcal{O}(c_{\sigma(1)}, \dots, c_{\sigma(n)}; c) \times \prod_{i=1}^{n} \mathcal{O}(d_{\sigma(i),1}, \dots, d_{\sigma(i),k_{\sigma(i)}}; c_{\sigma(i)}) \arrow[r] & \mathcal{O}(d_{1,1}, \dots, d_{n,k_n}; c)
\end{tikzcd}$$

We also require that, for $\tau_i \in \Sigma_{k_i}$, $i = 1, \dots, n$, the following diagram commutes:

$$\begin{tikzcd}
\mathcal{O}(c_1,\dots,c_n;c) \times \prod_{i=1}^{n} \mathcal{O}(d_{i,1},\dots,d_{i,k_i};c_i) \arrow[r] \arrow[d, "\mathrm{id}\times\tau_1\times\cdots\times\tau_n"'] & \mathcal{O}(d_{1,1},\dots,d_{n,k_n};c) \arrow[d, "\tau_1\times\cdots\times\tau_n"] \\
\mathcal{O}(c_1,\dots,c_n;c) \times \prod_{i=1}^{n} \mathcal{O}(d_{i,1},\dots,d_{i,k_i};c_i) \arrow[r] & \mathcal{O}(d_{1,1},\dots,d_{n,k_n};c)
\end{tikzcd}$$

Identity For every $c \in C$, there is an element $1_c \in \mathcal{O}(c;c)$ which together with

$$\mathcal{O}(c;c) \times \mathcal{O}(c_1,\dots,c_n;c) \to \mathcal{O}(c_1,\dots,c_n;c)$$

induces the identity map on $\mathcal{O}(c_1,\dots,c_n;c)$. We also require that the elements $1_{c_1},\dots,1_{c_n}$ together with

$$\mathcal{O}(c_1,\dots,c_n;c) \times \mathcal{O}(c_1;c_1) \times \cdots \times \mathcal{O}(c_n;c_n) \to \mathcal{O}(c_1,\dots,c_n;c)$$

induce the identity map on $\mathcal{O}(c_1,\dots,c_n;c)$.

The colors $c_1,\dots,c_n$ can be thought of as the colors of the inputs, while c is the color of the output. A colored operad with $C = \{c\}$ is just an operad, where

$$\mathcal{O}(\underbrace{c,\dots,c}_{n \text{ times}};c)$$

is $\mathcal{O}(n)$. Sometimes, for brevity, we write "operad" to mean "colored operad".

Note that if we have a colored operad $\mathcal{O}$ with colors C and a subset $C' \subset C$, we can restrict to another colored operad $\mathcal{O}_{C'}$ consisting of just the spaces $\mathcal{O}(c_1,\dots,c_n;c)$ with $c_i, c \in C'$ (and the same structure maps as $\mathcal{O}$).

Definition 3.5 Given a colored operad $\mathcal{O} = (\mathcal{O}, C)$, an *action* of $\mathcal{O}$ on A (also called an *$\mathcal{O}$–algebra* A) consists of a collection of spaces $\{A_c\}_{c\in C}$ together with maps

$$\mathcal{O}(c_1,\dots,c_n;c) \times A_{c_1} \times \cdots \times A_{c_n} \to A_c \tag{2}$$

satisfying the following conditions:

Associativity The following diagram commutes:

$$\begin{array}{ccc}
\mathcal{O}(c_1,\dots,c_n;c) \times \prod_{i=1}^{n} \mathcal{O}(d_{i,1},\dots,d_{i,k_i};c_i) \times \prod_{j=1}^{n} A_{d_{j,k_j}} & \longrightarrow & \mathcal{O}(c_1,\dots,c_n;c) \times \prod_{i=1}^{n} A_{c_i} \\
\downarrow & & \downarrow \\
\mathcal{O}(d_{1,1},\dots,d_{n,k_n};c) \times \prod_{j=1}^{n} A_{d_{j,k_j}} & \longrightarrow & A_c
\end{array}$$

Symmetry For each $\sigma \in \Sigma_n$, the following diagram commutes, where the vertical map is induced by the Σ_n–action and permuting the factors of A:

$$\begin{tikzcd}
\mathcal{O}(c_1,\dots,c_n;c) \times A_{c_1} \times \cdots \times A_{c_n} \arrow[r] \arrow[d] & A_c \\
\mathcal{O}(c_{\sigma(1)},\dots,c_{\sigma(n)};c) \times A_{c_{\sigma(1)}} \times \cdots \times A_{c_{\sigma(n)}} \arrow[ur] &
\end{tikzcd}$$

Identity The map induced by $1_c \in \mathcal{O}(c,c)$ together with $\mathcal{O}(c;c) \times A_c \to A_c$ is the identity on A_c.

If we have a subset $C' \subset C$, the restriction colored operad $\mathcal{O}_{C'}$ acts on the collection of spaces $\{A_c\}_{c \in C'}$.

Example 3.6 A *planar algebra* as in the work of Jones [20] is an algebra over a certain colored operad. In fact, planar diagrams form a colored operad called the *planar operad* $\mathcal{P}$. The colors C are the even natural numbers, and $\mathcal{P}(c_1,\dots,c_n;c)$ is the space of diagrams with n holes, c_i strands incident to the i^{th} boundary circle, and c strands incident to the outer boundary circle. If A_c denotes the space of tangle diagrams in D^2 with c endpoints on ∂D^2, then the collection $\{A_c\}_{c \in C}$ is an example of an algebra over $\mathcal{P}$ (aka a planar algebra).

4 A review of Budney's operad actions

4.1 Budney's 2–cubes action

The operation of connect-sum of knots is always well defined on isotopy classes of knots. If one considers *long* knots, one can further define connect-sum (or stacking) of knots themselves, rather than just the isotopy classes. That is, there is a well-defined map

$$\#\colon \mathcal{K}\times\mathcal{K}\to\mathcal{K},$$

where $\mathcal{K}=\mathrm{Emb}(\mathbb{R},\mathbb{R}\times D^2)$ is the space of long knots. If one descends to isotopy classes, this operation is commutative, ie # is *homotopy-commutative*. See Budney's paper [2, page 4, Figure 2] for a beautiful picture of the homotopies involved. This picture suggests that one can parametrize the operation # by $S^1\simeq\mathcal{C}_2(2)$. Thus it suggests that the little 2–cubes operad $\mathcal{C}_2$ acts on $\mathcal{K}$.

Budney succeeded in constructing such a 2–cubes action, but to do so, he had to consider a space of *fat* long knots

$$\mathrm{EC}(1,D^2):=\{f\colon \mathbb{R}^1\times D^2\hookrightarrow\mathbb{R}^1\times D^2 \mid \mathrm{supp}(f)\subset I\times D^2\},$$

where $\mathrm{supp}(f)$ is defined as the closure of $\{x\in\mathbb{R}^1\times D^2 \mid f(x)\neq x\}$. The notation $\mathrm{EC}(1,D^2)$ stands for (self-)*embeddings* of $\mathbb{R}^1\times D^2$ with *cubical* support. This space is equivalent to the space of *framed* long knots, but one can restrict to the subspace where the linking number of the curves $f|_{\mathbb{R}\times(0,0)}$ and $f|_{\mathbb{R}\times(0,1)}$ is zero; this subspace is then equivalent to the space of long knots.

The advantage of $\mathrm{EC}(1,D^2)$ is that one can compose elements. In the 2–cubes action on this space, the first coordinate of a cube acts on the $\mathbb{R}$ factor in $\mathbb{R}\times D^2$, while the second factor dictates the order of composition of embeddings. Precisely, the action is defined as follows. For one little cube L, let L^y be the embedding $I\hookrightarrow I$ given by projecting to the last factor. Let L^x be the embedding $I\hookrightarrow I$ given by projecting to the first factor(s). Let $\sigma\in\Sigma_n$ be a permutation (thought of as an ordering of $\{1,\dots,n\}$) such that $L^y_{\sigma(1)}(0)\leq\cdots\leq L^y_{\sigma(n)}(0)$. The action

$$\mathcal{C}_2(n)\times\mathrm{EC}(1,D^2)^n\to\mathrm{EC}(1,D^2)$$

is given by

$$(L_1,\dots,L_n)\cdot(f_1,\dots,f_n)\mapsto L^x_{\sigma(n)}\circ f_{\sigma(n)}\circ(L^x_{\sigma(n)})^{-1}\circ\cdots\circ L^x_{\sigma(1)}\circ f_{\sigma(1)}\circ(L^x_{\sigma(1)})^{-1}.$$

4.2 The splicing operad

In the above 2–cubes action, the second coordinate is only used to order the embeddings. Thus instead of the 2–cubes operad, one could consider an operad of "overlapping intervals" $\mathcal{C}_1'$. An element in $\mathcal{C}_1'(n)$ is n intervals in the unit interval, not necessarily disjoint, but with an order dictating which interval is above the other when two intervals do overlap. Precisely, an element of $\mathcal{C}_1'(n)$ is an equivalence class $(L_1,\dots,L_n,\sigma)$ where each L_i is an embedding $I\hookrightarrow I$ and where $\sigma\in\Sigma_n$. Elements $(L_1,\dots,L_n,\sigma)$

and $(L'_1, \dots, L'_n, \sigma')$ are equivalent if $L_i = L'_i$ for all i and if whenever L_i and L_j intersect, $\sigma^{-1}(i) \leq \sigma^{-1}(j) \iff (\sigma')^{-1}(i) \leq (\sigma')^{-1}(j)$. It is not hard to see what the structure maps for the operad are (and they are given in Budney's paper [4]). Budney then easily recasts his 2–cubes action as an action of the overlapping intervals operad $\mathcal{C}'_1$.

The splicing operad $\mathcal{SC}_1^{D^2}$ (which we abbreviate for now as $\mathcal{SC}$) is formally similar to the overlapping intervals operad, in that $\mathcal{SC}(n)$ consists of equivalence classes of elements $(L_0, L_1, \dots, L_n, \sigma)$ with the same equivalence relation as for $\mathcal{C}'_1$. In the splicing operad, however, L_0 is in $\mathrm{EC}(1, D^2)$, $L_1, \dots, L_n$ are embeddings

$$L_i \colon I \times D^2 \hookrightarrow I \times D^2,$$

and all the L_i are required to satisfy a "continuity constraint", as follows. One considers $\sigma \in \Sigma_n$ as an element of $\Sigma_{n+1} = \mathrm{Aut}\{0, \dots, n\}$ which fixes 0. If $\sigma^{-1}(i) < \sigma^{-1}(k)$ one can think of L_i as inner (or first in order of composition) with respect to L_k. One wants the "round boundary" of L_k not to touch L_i, but for the operad to have an identity element, one needs to allow for L_k to be flush around L_i. The precise requirement needed is that for $0 \leq \sigma^{-1}(i) < \sigma^{-1}(k)$,

$$\overline{\operatorname{im} L_i \setminus \operatorname{im} L_k} \cap L_k(\mathring{I} \times \partial D^2) = \varnothing.$$

Note that $\mathcal{SC}$ is a much "bigger" operad than $\mathcal{C}'_1$. One can think of L_0 as the "starting (thickened long) knot" for the splicing operation and of the other L_i as n "hockey pucks" with which one grabs L_0 and ties up into n knots. This gives a map

$$\mathcal{SC}(n) \times \mathcal{K}^n \to \mathcal{K}$$

which will define the action of the splicing operad on $\mathcal{K}$. To fully construct $\mathcal{SC}$ as an operad, one needs the operad structure maps, which also come from the map above. Roughly speaking, the structure maps are as follows. Given one splicing diagram with n pucks and n other splicing diagrams each with k_i pucks ($i = 1, \dots, n$), put the i^{th} splicing diagram into the i^{th} puck by composing the "starting knots" L_0 and "taking the pucks along for the ride". For a precise definition and pictures, the reader may either consult [4] or read the next section, which closely follows Budney's construction and subsumes the splicing operad.

5 The infection colored operad

Definition 5.1 Fix for each $c = 1, 2, 3, \dots$ a trivial c–component fat string link

$$i_c \colon \coprod_c I \times D^2 \hookrightarrow I \times D^2,$$

with image denoted $S_c := \mathrm{im}(i_c) \subset I \times D^2$.

We will be more concerned with this image of the fixed trivial fat string link rather than the embedding itself.

A convenient way of choosing an i_c is to fix an embedding $\coprod_c D^2 \hookrightarrow D^2$ and then take the product with the identity map on I. For $c \geq 2$, we choose an embedding which takes the centers of the c copies of D^2 to the points $x_1, \dots, x_c$ from our Definition 1.1 of string links. Beyond that, we remain agnostic about this fixed embedding. For $c = 1$, we choose i_1 to be the identity map. This will recover Budney's splicing operad from our colored operad when all the colors are 1.

Now we define the space of c–component fat string links to be

$$\mathrm{FSL}_c := \left\{ f \colon \coprod_c I \times D^2 \hookrightarrow I \times D^2 \;\middle|\; f \text{ agrees with } i_c \text{ on } \partial I \times D^2 \right\}.$$

These are the spaces on which the infection colored operad will act. An element of FSL_3 is displayed in Figure 6. By our condition on the fixed trivial fat string link, we can restrict f to the cores of the solid cylinders to obtain an ordinary string link $f|_{I \times \{x_1, \dots, x_c\}}$ as in Definition 1.1.

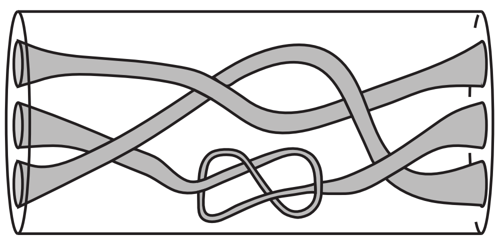


Figure 6: A fat string link, or more precisely, an element of FSL_3

5.1 The definition of the infection colored operad

We now define our colored operad $\mathcal{I} = (\mathcal{I}, C)$. We put $C = \mathbb{N}^+$, so each color c is a positive natural number.

Definition 5.2 (The spaces in the colored operad $\mathcal{I}$) An *infection diagram* is a tuple $(L_0, L_1, \dots, L_n, \sigma)$ with $L_0 \in \mathrm{FSL}_c$, $\sigma \in \Sigma_n$, and L_i an embedding $L_i \colon I \times D^2 \hookrightarrow I \times D^2$ (for $i = 1, \dots, n$) satisfying a certain continuity constraint. The constraint is that if $0 \leq \sigma^{-1}(i) < \sigma^{-1}(k)$, then

$$(\dagger) \qquad \overline{L_i(I \times D^2) \setminus L_k(S_{c_k})} \cap L_k(\mathring{I} \times (D^2 \setminus \mathring{S}_{c_k})) = \varnothing,$$

where S_{c_k} is the image of a fixed trivial string link, as in Definition 5.1. As in the splicing operad, we think of $\sigma \in \Sigma_n$ as a permutation in $\Sigma_{n+1} = \mathrm{Aut}\{0, 1, \ldots, n\}$ which fixes 0.

The space $\mathcal{I}(c_1, \ldots, c_n; c)$ is the space of equivalence classes $[L_0, \ldots, L_n, \sigma]$ of infection diagrams, where $(L_0, \ldots, L_n, \sigma)$ and $(L'_0, \ldots, L'_n, \sigma')$ are equivalent if $L_i = L'_i$ for all i, and if whenever the images of L_i and L_k intersect, $\sigma^{-1}(i) \leq \sigma^{-1}(k)$ if and only if $(\sigma')^{-1}(i) \leq (\sigma')^{-1}(k)$. $\square$

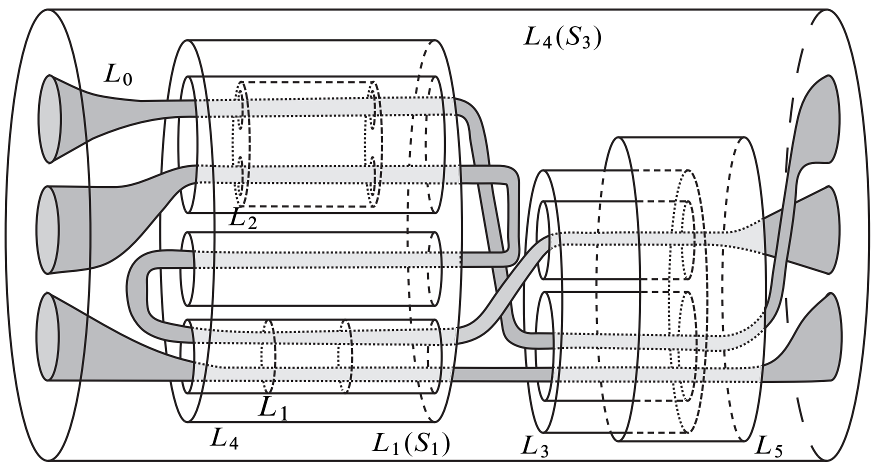


Figure 7: An infection diagram, or more precisely, an element of $\mathcal{I}(1, 2, 2, 3, 1; 3)$

Informally, the L_i are like the hockey pucks in Budney's splicing operad, and the permutation σ is a map that sends the order of composition to the index i of L_i. The difference is that instead of re-embedding a hockey puck into itself, we will re-embed the image of S_{c_i}, a subspace of thinner inner cylinders, into the puck. Thus we keep track of the image of S_{c_i}, and our pucks can be thought of as having cylindrical holes drilled in them, the holes with which we will grab the string link (or long knot) L_0. Following a suggestion of V Krushkal, we call the restrictions of the

$$L_i \quad \text{to} \quad (I \times D^2) \setminus \mathring{S}_{c_i}$$

"mufflers" (motivated by the picture for $c_i = 2$).

The generalization of Budney's continuity constraint to the constraint (†) is the key technical ingredient in defining our colored operad. The need for this constraint is explained precisely in Remark 5.4 below. The rough meaning of this condition is that a muffler which acts earlier should be inside a hole of a muffler that acts later; in other words, the "solid part" of a higher L_k (which remains after drilling out the trivial string link) should not intersect any part of a lower L_i, where "higher" and "lower" are in the semi-linear ordering determined by σ. However, we must allow for the possibility of the boundaries of the mufflers intersecting in certain ways. Figure 8 displays the cross-section of a set of mufflers which satisfy constraint (†).

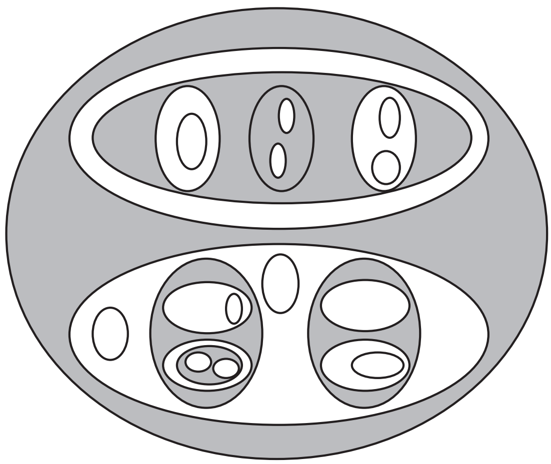


Figure 8: The cross-section of a set of thirteen mufflers, including seven one-holed mufflers (or hockey pucks), satisfying the constraint (†). Each grey area is the "forbidden region" $L_k(\mathring{I} \times (D^2 \setminus \mathring{S}_{c_k}))$ of the k^{th} muffler, ie the region where no other muffler may lie.

So far we haven't finished defining the operad, since we haven't defined the structure maps. We start by defining the action on the space of fat string links. Only after that will we define the structure maps and check that they form a colored operad and that the definition below is a colored operad action.

Definition 5.3 (The action of $\mathcal{I}$ on fat string links) Consider $[L_0, L_1 \dots, L_n, \sigma] \in \mathcal{I}(c_1, \dots, c_n; c)$ and fat string links $f_1, \dots, f_n$ where $f_k \in \mathrm{FSL}_{c_k}$. Let L_k^{in} be the map obtained from L_k by restricting the domain to S_{c_k} and restricting the codomain to its image. We use the shorthand notation $L_k \cdot f_k$ to denote the map

$$L_k \circ f_k \circ (L_k^{in})^{-1} \colon L_k(S_{c_k}) \to I \times D^2.$$

Then we define

$$\begin{aligned} &\mathcal{I}(c_1, \dots, c_n; c) \times \mathrm{FSL}_{c_1} \times \cdots \times \mathrm{FSL}_{c_n} \longrightarrow \mathrm{FSL}_c, \\ &\qquad ([L_0, L_1, \dots, L_n, \sigma], f_1, \dots, f_n) \mapsto (L_{\sigma(n)} \cdot f_{\sigma(n)}) \circ \cdots \circ (L_{\sigma(1)} \cdot f_{\sigma(1)}) \circ L_0. \end{aligned}$$

Remark 5.4 Strictly speaking, each map $L_{\sigma(k)} \cdot f_{\sigma(k)}$ is only defined on

$$L_{\sigma(k)}(S_{c_{\sigma(k)}}) = \operatorname{im} L^{in}_{\sigma(k)},$$

so one might worry whether the above composition is well defined. We claim that the conditions on the support of the $f_{\sigma(k)}$ and the continuity constraint (†) guarantee that we can continuously extend each $L_{\sigma(k)} \cdot f_{\sigma(k)}$ by the identity on $\operatorname{im} L_0 \setminus \operatorname{im} L^{in}_{\sigma(k)}$.

In fact, first write

$$\partial\big(\operatorname{im} L^{in}_{\sigma(k)}\big) = \Big(\partial I \times \coprod_{c_k} D^2\Big) \cup \Big(I \times \partial \coprod_{c_k} D^2\Big).$$

Since each $f_{\sigma(k)}$ is the identity on the $\partial I \times \coprod_{c_k} D^2$ part of its domain (the "flat boundary"), the map $L_{\sigma(k)} \cdot f_{\sigma(k)}$ is the identity on the $\partial I \times \coprod_{c_k} D^2$ part of $\operatorname{im} L^{in}_{\sigma(k)}$.

The constraint (†) says that

$$\overline{\operatorname{im} L_0 \setminus \operatorname{im} L^{in}_{\sigma(k)}} \cap L_{\sigma(k)}\Big(\mathring{I} \times \partial \coprod_{c_k} D^2\Big) = \varnothing,$$

hence

$$\overline{\operatorname{im} L_0 \setminus \operatorname{im} L^{in}_{\sigma(k)}} \cap \operatorname{im} L_{\sigma(k)} \subseteq \partial I \times \coprod_{c_k} D^2.$$

So the continuity constraint guarantees that we don't need to worry about extending past the $I \times \partial \coprod D^2$ part of the boundary (the "round boundary").

Hence this defines the composition on the whole image of L_0.

Definition 5.5 (The structure maps in $\mathcal{I}$) The structure maps

$$\begin{aligned} (3) \quad & \mathcal{I}(c_1, \dots, c_n; c) \times \mathcal{I}(d_{1,1}, \dots, d_{1,k_1}; c_1) \times \cdots \times \mathcal{I}(d_{n,1}, \dots, d_{n,k_n}; c_n) \\ & \qquad\qquad \longrightarrow \mathcal{I}(d_{1,1}, \dots, d^{n,k_n}; c), \\ & (J_0, \dots, J_n, \rho) \times (L_{1,0}, \dots, L_{1,k_1}, \sigma_1) \times \cdots \times (L_{n,0}, \dots, L_{n,k_n}, \sigma_n) \\ & \qquad\qquad \mapsto ((J \cdot \vec{L})_0, (J \cdot \vec{L})_{1,1}, \dots, (J \cdot \vec{L})_{n,k_n}, \tau) \end{aligned}$$

are defined as follows. (Here $\vec{L} = (L_{1,*}, \dots, L_{n,*})$, which can be thought of as n infection diagrams, and $J \cdot \vec{L}$ is just shorthand for the result on the right-hand side.) The "starting" fat string link is

$$\begin{aligned} (J \cdot \vec{L})_0 &:= \Big(\bigcirc_{i=1}^{n} J_{\rho(i)} \cdot L_{\rho(i),0}\Big) \circ J_0 \\ &:= (J_{\rho(n)} \circ L_{\rho(n),0} \circ (J^{in}_{\rho(n)})^{-1}) \circ \cdots \circ (J_{\rho(1)} \circ L_{\rho(1),0} \circ (J^{in}_{\rho(1)})^{-1}) \circ J_0. \end{aligned}$$

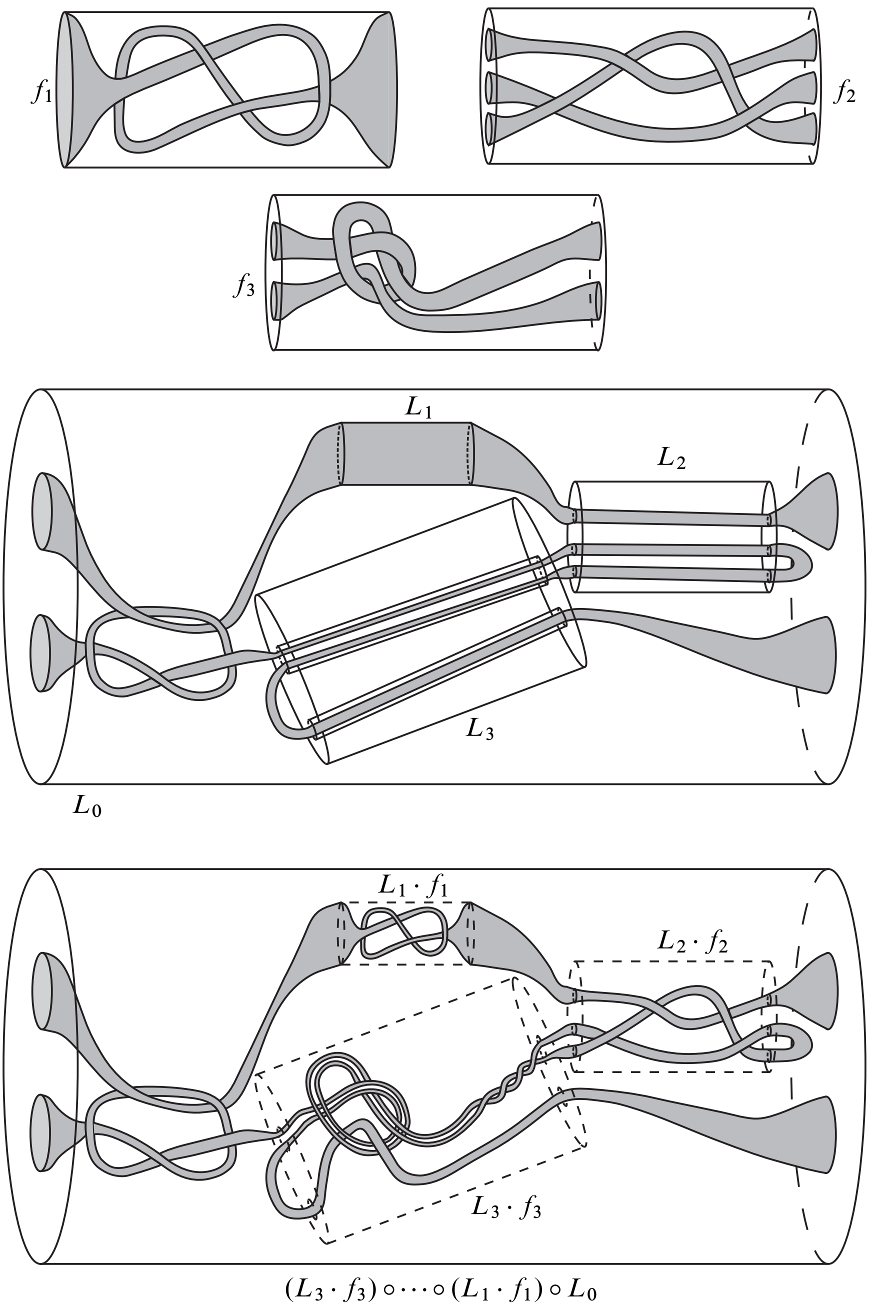


Figure 9: The action of an infection diagram on three fat string links via the map $\mathcal{I}(1,3,2;2) \times \mathrm{FSL}_1 \times \mathrm{FSL}_2 \times \mathrm{FSL}_3 \to \mathrm{FSL}_2$. The 2–component fat string link at the bottom is the result of this action.

Given $a \in \{1, \dots, n\}$ and $b \in \{1, \dots, k_a\}$, the $(a,b)^{\text{th}}$ puck is

$$(J \cdot \vec{L})_{a,b} := \left(\bigcirc_{i=\rho^{-1}(a)+1}^{n} J_{\rho(i)} \cdot L_{\rho(i),0} \right) \circ (J_a \circ L_{a,b}).$$

Finally, the permutation τ associated to $J \cdot \vec{L}$ is given by

$$\tau^{-1}(a,b) := \tau^{-1}\left(b + \sum_{i=1}^{a-1} k_i \right) := \sigma_a^{-1}(b) + \sum_{i=1}^{\rho(a)-1} k_{\rho(i)}.$$

In other words

$$\begin{aligned}(4) \quad \tau^{-1} \colon & (1,1), (1,2), \dots, (n,k_n) \\ & \longmapsto (\rho^{-1}(1), \sigma_1^{-1}(1)), (\rho^{-1}(1), \sigma_1^{-1}(2)), \dots, (\rho^{-1}(n), \sigma_n^{-1}(k_n)),\end{aligned}$$

where the set acted on can be thought of as a set of ordered pairs (though not a cartesian product) with a lexicographical ordering as on the left.

Notice that the action maps are just special cases of the structure maps. In fact, FSL_c is precisely $\mathcal{I}(\varnothing; c)$ where $\varnothing$ is 0–tuple of positive integers (or the sequence of zero elements). Thus each action map can be written as

$$\mathcal{I}(c_1, \dots, c_n; c) \times \mathcal{I}(\varnothing; c_1) \times \dots \times \mathcal{I}(\varnothing; c_n) \to \mathcal{I}(\varnothing; c).$$

Thus we can make just a slight modification to Figure 9 to produce a picture of a structure map (that is not an action map), as in Figure 10.

Theorem 5.6 (A) *The spaces and maps in Definitions 5.2 and 5.5 make $\mathcal{I}$ a colored operad with an action on the space of fat string links given by Definition 5.3.*

(B) *When restricting to the single color $c = 1$, one recovers Budney's splicing operad $\mathcal{SC}_1^{D^2}$. Thus $\mathcal{C}_2$ maps to this part of the colored operad.*

(C) *There is a map of the little intervals operad $\mathcal{C}_1$ to the restriction $\mathcal{I}_{\{c\}}$ of $\mathcal{I}$ to* any *single color c.*

Proof For (A), we can first see that a composed operation (ie an infection diagram on the right-hand side $\mathcal{I}(d_{1,1}, \dots, d_{n,k_n}; c)$ of Equation (3)) satisfies the constraint (†), as follows. Any two non-disjoint mufflers in the composed diagram are the images of mufflers in some $\mathcal{I}(d_{i,1}, \dots, d_{i,k_i}; c_i)$ under a composition of embeddings. But if the constraint (†) holds for L_i, L_k, then it holds for the compositions of L_i, L_k with these embeddings, since "image under an embedding" commutes with complement, closure, and intersection.

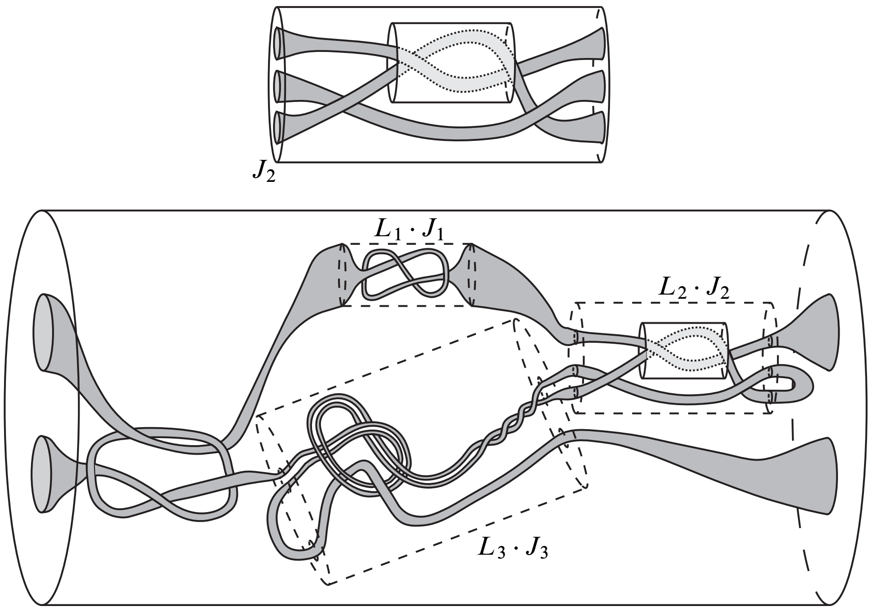


Figure 10: A slight variation of Figure 9, using the same (L_0, L_1, L_2, L_3) but replacing the fat string link f_2 in Figure 9 by the infection diagram J_2 shown above, gives an example of the operad structure maps. The infection diagrams J_1 and J_3 have zero mufflers, and their 0^{th} components are respectively f_1 and f_3. Thus the picture above is the image of $((L_0, \dots, L_3), J_1, J_2, J_3)$ under the structure map $\mathcal{I}(1,3,2;2) \times \mathcal{I}(\varnothing;1) \times \mathcal{I}(1;3) \times \mathcal{I}(\varnothing;2) \to \mathcal{I}(1;2)$.

Now we need to check the conditions of (a) associativity, (b) symmetry, and (c) identity for the structure maps. The corresponding conditions for the action maps will then follow because the action maps are special cases of the structure maps.

(a) Suppose we have

$$J = (J_0, \dots, J_j, \sigma), \quad \vec{L} = ((L_{1,0}, \dots, L_{1,\ell_1}, \sigma_1), \dots, (L_{j,0}, \dots, L_{j,\ell_j}, \sigma_j)),$$
$$\vec{M} = ((M_{1,1,0}, \dots, M_{1,1,m_{1,1}}, \tau_{1,1}), \dots, (M_{j,\ell_j,0}, \dots, M_{j,\ell_j,m_{j,\ell_j}}, \tau_{j,\ell_j})).$$

Then $(J \cdot \vec{L}) \cdot \vec{M}$ has 0^{th} component

$$\left(\bigcirc_{(h,k)=\nu^{-1}(1,1)}^{\nu^{-1}(j,\ell_j)} \left(\left(\bigcirc_{i=\rho^{-1}(h)+1}^{j} J_{\rho(i)} \cdot L_{\rho(i),0} \right) \circ J_h \circ L_{h,k} \right) \cdot M_{h,k,0} \right)$$
$$\circ \left(\bigcirc_{i=1}^{j} J_{\rho(i)} \cdot L_{\rho(i),0} \right) \circ J_0,$$

where ν is the permutation for JL, and where the order of the terms in the leftmost composition is given by the indices $\nu^{-1}(1,1), \nu^{-1}(1,2), \ldots, \nu^{-1}(j,\ell_j)$. On the other hand, the 0^{th} component of $J \cdot (\vec{L} \cdot \vec{M})$ is

$$\bigcirc_{i=1}^{j} J_{\rho(i)} \cdot \left(\left(\bigcirc_{k=1}^{\ell_{\rho(i)}} L_{\rho(i),\sigma(k)} \cdot M_{\rho(i),\sigma(k),0} \right) \circ L_{\rho(i),0} \right) \circ J_0.$$

These two expressions agree by canceling adjacent terms

$$J_{\rho(i)},\ (J^{in}_{\rho(i)})^{-1} \quad \text{and} \quad L_{\rho(i),0},\ (L_{\rho(i),0})^{-1}$$

in the expression for $((J \cdot \vec{L}) \cdot \vec{M})_0$. For example, if

$$J = (J_0, J_1, J_2, \iota), \quad \vec{L} = ((L_{1,0}, L_{1,1}), (L_{2,0}, L_{2,1}), \iota),$$
$$\vec{M} = ((M_{1,1,0}, M_{1,1,1}), (M_{2,1,0}, M_{2,1,1}), \iota),$$

with ι denoting the identity permutation, then

$$\begin{aligned}
&((J \cdot \vec{L}) \cdot \vec{M})_0 \\
&= [(J \cdot \vec{L})_{2,1} \cdot M_{2,1,0}] \circ [(J \cdot \vec{L})_{1,1} \cdot M_{1,1,0}] \circ [J_2 \cdot L_{2,0}] \circ [J_1 \cdot L_{1,0}] \circ J_0 \\
&= [J_2 \circ L_{2,1} \circ M_{2,1,0} \circ (L^{in}_{2,1})^{-1} \circ \cancel{J_2^{-1}}] \\
&\quad \circ [\cancel{J_2} \circ L_{2,0} \circ (J_2^{in})^{-1} \circ J_1 \circ L_1 \circ M_{1,1,0} \circ (L^{in}_{1,1})^{-1} \circ \cancel{J_1^{-1}} \circ \cancel{J_2} \circ \cancel{(L_{2,0})^{-1}} \circ \cancel{J_2^{-1}}] \\
&\qquad\qquad \circ [\cancel{J_2} \circ \cancel{L_{2,0}} \circ \cancel{(J_2^{in})^{-1}}] \circ [\cancel{J_1} \circ L_{1,0} \circ (J_1^{in})^{-1}] \circ J_0 \\
&= [J_2 \cdot (\vec{L} \cdot \vec{M})_{2,1,0}] \circ [J_1 \cdot (\vec{L} \cdot \vec{M})_{1,1,0}] \circ J_0 \\
&= (J \cdot (\vec{L} \cdot \vec{M}))_0.
\end{aligned}$$

Checking that the $(a,b,c)^{\text{th}}$ mufflers of these two infection diagrams agree similarly involves canceling adjacent terms in the expression for $((J \cdot \vec{L}) \cdot \vec{M})_{a,b,c}$. (Also cf [4].)

Finally, to check that the permutations for these two infection diagrams agree, note that the inverse of either one is given (with notation as in Equation (4)) by

$$(i,k,h) \mapsto (\rho^{-1}(i), \sigma_i^{-1}(k), \tau_{i,k}^{-1}(h)).$$

(b) We need to check that both diagrams in the symmetry condition in Definition 3.4 commute for $\mathcal{O} = \mathcal{I}$. The maps involved consist of permutations of labels on mufflers and labels on infection diagrams. The commutativity of these diagrams is easily verified.

(c) The identity $1_c \in \mathcal{I}(c;c)$ is an element $[L_0, L_1, e]$ with L_0 the fixed trivial c–component fat string link, L_1 the identity map on $I \times D^2$, and e the element in Σ_1.

Part (B) of the theorem follows quickly from our definitions. One can check that by choosing the identity map for the trivial fat 1–string link, our constraint (†) reduces to Budney's continuity constraint. The rest of our definitions are then exactly as in Budney's splicing operad.

For part (C), the map $\mathcal{C}_1 \to \mathcal{I}_{\{c\}}$ is easy to construct. An element of $\mathcal{C}_1(n)$ is $(a_1, \dots, a_n)$ where each $a_i \colon I \hookrightarrow I$ is the restriction of an affine-linear, increasing map. The map $\mathcal{C}_1 \to \mathcal{I}_{\{c\}}$ is given by $(a_1, \dots, a_n) \mapsto (i_c, a_1 \times \mathrm{id}_{D^2}, \dots, a_n \times \mathrm{id}_{D^2}, \iota)$ where i_c was the trivial fat c–string link, and where ι is the identity permutation. (Actually, we could choose any permutation since the mufflers are disjoint.) $\square$

Remark 5.7 For $c \neq 1$, it is clear that $\mathcal{C}_2$ cannot map to the operad $\mathcal{I}_c$, for then connect-sum of string links would be (homotopy-)commutative. But this is not the case. For $c \geq 3$, the pure braid group is not abelian, and for $c = 2$, the monoid of string links up to isotopy is nonabelian. The latter result can be deduced either from our recent results on the structure of this monoid [1] or from work of Le Dimet in the late 1980s [21] on the group of string links up to cobordism.

Just as Budney's fat long knots are equivalent to framed long knots, our fat string links are equivalent to framed string links. In more detail, given a fat string link $L \in \mathrm{FSL}_c$, we can restrict to the "cores of the tubes" to get an ordinary string link $L|(I \times \{x_1, \dots, x_c\})$. Thus we have a map $\mathrm{FSL}_c \to \mathcal{L}_c$, which is a fibration, since in general restriction maps are fibrations. The fiber Fib_L over L is the space of tubular neighborhoods of $\operatorname{im} L$ which are fixed at the boundaries. We express such a neighborhood as a map $\eta \colon \coprod_c I \times D^2 \to I \times D^2$ and associate to η a collection of c loops in $\mathrm{SO}(2)$; these are obtained by taking the derivative at $(0,0)$ of the map $\{t\} \times D^2 \to I \times D^2$, for each $t \in \coprod_c I$. Thus we can map the fiber Fib_L to $(\Omega\, \mathrm{SO}(2))^c$. This "derivative map" is a homotopy equivalence (by shrinking η to a small neighborhood of $\coprod_c I \times \{0\}$). Since $\Omega\, \mathrm{SO}(2) \cong \mathbb{Z}$, we can write the fibration as

$$\mathbb{Z}^c \longrightarrow \mathrm{FSL}_c \longrightarrow \mathcal{L}_c.$$

For $L \in \mathrm{FSL}_c$, there are c framing numbers $\omega_1, \dots, \omega_c$, one for each component. The j^{th} framing number is given by the linking number of $I_j \times (0,0)$ with $I_j \times (1,0)$, where I_j is the j^{th} copy of I in $\coprod_c I$. The map $\omega_1 \times \cdots \times \omega_c \colon \mathrm{FSL}_c \to \mathbb{Z}^c$ gives a splitting of the above fibration. Then we consider the product fibration $\mathbb{Z}^c \to \mathcal{L}_c \times \mathbb{Z}^c \to \mathcal{L}_c$ and the map from the above fibration to this one induced by the splitting. The long exact sequence of homotopy groups for a fibration together with the five lemma imply that the map from FSL_c to $\mathcal{L}_c \times \mathbb{Z}^c$ is a weak equivalence, hence a homotopy equivalence. Thus $\widehat{\mathcal{L}}_c := (\omega_1 \times \cdots \times \omega_c)^{-1}\{(0,0,\dots,0)\}$ is equivalent to $\mathcal{L}_c$.

Corollary 5.8 *By restricting to the subspaces $\widehat{\mathcal{L}}_c \subset \mathrm{FSL}_c$ of fat string links with zero framing number in every component, we obtain an action of $\mathcal{I}$ on spaces homotopy equivalent to the spaces of c–component string links.*

5.2 Mufflers, rational tangles, and pure braids

We now briefly discuss how general an infection our operad $\mathcal{I}$ encodes. Informally, one might wonder how twisted the inner cylinders (ie the holes) L^{in} of a muffler could be. Clearly, a fat string link can appear as L^{in} if and only if the pair $(I \times D^2, L^{in})$ is homeomorphic to the pair $(I \times D^2, i_c)$ where i_c is the trivial fat c–string link. The purpose of the following well-known proposition is just to show an alternative and perhaps more intuitive way of thinking about such string links. Recall from Definition 5.1 that S_c is the image of i_c.

Proposition 5.9 *The following are equivalent:*

(i) *There is a diffeomorphism of pairs* $(I \times D^2, S_c) \xrightarrow{\cong} (I \times D^2, \mathrm{im}(L))$.

(ii) *There is an isotopy from* L *to the trivial link which takes* $\partial\big(\coprod_c I\big)$ *into* $\partial(I \times D^2)$. *Note that the isotopy need not* fix *the endpoints of* $\coprod_c I$.

Proof (i) $\Longrightarrow$ (ii) Suppose we have a diffeomorphism of pairs h as in (i). It suffices to show that the identity can be connected to this diffeomorphism by a path of diffeomorphisms of $I \times D^2$, for then we can restrict to S_c to obtain the desired isotopy.

By Cerf's theorem [5], the space of diffeomorphisms of S^3 is connected. As a corollary, so is the space of diffeomorphisms of D^3 whose values and derivatives agree with the identity on the boundary. In fact, this follows by considering the fibration

$$\mathrm{Diff}(D^3, \partial D^3) \to \mathrm{Diff}(S^3) \to \mathrm{Emb}(D^3, S^3)$$

given by restricting to a hemisphere of S^3. The base space is homotopy-equivalent to $\mathrm{SO}(3)$, which is connected, while the fiber is the space of diffeomorphisms of D^3 fixed on the boundary.

Now a diffeomorphism $\varphi\colon (I \times D^2, S_c) \xrightarrow{\cong} (I \times D^2, \mathrm{im}(L))$ is clearly isotopic to one that is the identity outside of a ball D^3 contained in $I \times D^2$. Combining this with Cerf's Theorem, we get a path from φ to the identity, as desired.

(ii) $\Longrightarrow$ (i) By the isotopy extension theorem (see for example Hirsch's text [19]), an isotopy as in (ii) can be extended to a diffeotopy of the whole space $I \times D^2$. The diffeotopy at time 1 then gives the desired diffeomorphism. □

The 2–string links which satisfy the above condition(s) are by definition precisely the 2–string links which are also *rational 2–tangles*. (Here we consider only string links, not arbitrary tangles; the reader may consult the work of Conway [13] for more details about rational tangles in general.) Note that pure braids are examples of rational 2–tangles, since it is easy to see that a pure braid satisfies (ii) above. We immediately have the following result, which informally says that "a muffler can grab the string link in the shape of any rational 2–tangle".

Proposition 5.10 *A fat 2–string link L^{in},*

$$S_c \cong \coprod_c I \times D^2 \xhookrightarrow{L^{in}} I \times D^2,$$

extends to a diffeomorphism L of $I \times D^2$ if and only if the core $L^{in}|(I \times \{x_1, x_2\})$ of L^{in} is a rational tangle. $\square$

5.3 Generalizations to other embedding spaces

For $j \in \mathbb{N}^+$ and M a compact manifold with boundary, let $\mathrm{EC}(j, M)$ be the space of "cubical embeddings" $\mathbb{R}^j \times M \hookrightarrow \mathbb{R}^j \times M$, that is, all such embeddings which are the identity outside $I^j \times M$. Budney constructs the actions of the little 2–cubes operad $\mathcal{C}_j$ and the splicing operad $\mathcal{SC}_1^{D^2}$ on the space of long knots as special cases of actions of the operads $\mathcal{C}_j$ and $\mathcal{SC}_j^M$ on $\mathrm{EC}(j, M)$. Our extension of the splicing operad to string links also gives an extension of the more general splicing operad $\mathcal{SC}_j^M$ to a colored operad acting on spaces of embeddings $I^j \times \coprod_c M \hookrightarrow I^j \times M$.

For each $c \in \mathbb{N}^+$ fix an embedding

$$i_c\colon \coprod_c I^j \times M \hookrightarrow \coprod_c I^j \times M$$

by fixing an embedding $\coprod_c M \hookrightarrow M$. Let S_c be the image of i_c.

Let

$$\mathrm{EC}^{\coprod_c}(j, M) := \left\{ f\colon \coprod_c I^j \times M \hookrightarrow I^j \times M \;\middle|\; f \text{ agrees with } i_c \text{ on } \partial I \times M \right\}.$$

Definition 5.11 (The spaces in the colored operad $\mathcal{I}_j^M$) An element in

$$\mathcal{I}_j^M(c_1, \dots, c_n; c)$$

is an equivalence class of tuples $(L_0, L_1, \dots, L_n, \sigma)$, where $L_0 \in \mathrm{EC}^{\coprod_c}(j, M)$, $\sigma \in \Sigma_n$, and for $i = 1, \dots, n$, L_i is an embedding $L_i\colon I^j \times M \hookrightarrow I^j \times M$ subject to the constraint that for $0 \le \sigma^{-1}(i) < \sigma^{-1}(k)$,

$$\overline{\mathrm{im}\, L_i \setminus L_k(S_c)} \cap L_k(\mathring{I}^j \times (M \setminus \mathring{S}_c)) = \varnothing.$$

Here we think of $\sigma \in \Sigma_n$ as a permutation in $\Sigma_{n+1} = \mathrm{Aut}\{0, 1, \ldots, n\}$ which fixes 0.

Tuples $(L_0, \ldots, L_n, \sigma)$ and $(L'_0, \ldots, L'_n, \sigma')$ are equivalent if $L_i = L'_i$ for all i and if whenever the images of L_i and L_k intersect, $\sigma^{-1}(i) \leq \sigma^{-1}(k)$ if and only if $(\sigma')^{-1}(i) \leq (\sigma')^{-1}(k)$.

The structure maps of $\mathcal{I}_j^M$, and an action of $\mathcal{I}_j^M$ on the spaces $\{\mathrm{EC}^{\coprod_c}(j, M)\}_{c \in \mathbb{N}+}$, can be defined exactly as in the special case where $j = 1$ and $M = D^2$.

6 Decomposing the space of 2–string links using the infection operad

6.1 The monoid of 2–string links

Note that given any monoid $\mathcal{M}$ and subset $\mathcal{C}$ of central elements, the quotient monoid $\mathcal{M}/\mathcal{C}$ is well defined. We are interested in the monoid $\mathcal{M} = \pi_0\mathcal{L}_c$ of isotopy classes of c–string links, especially for $c = 2$. The units in $\pi_0\mathcal{L}_c$ are precisely the pure braids [1, Proposition 2.7]. We say that a non-unit c–string link L is *prime* if $L = L_1 \# L_2$ implies that either L_1 or L_2 is a unit (pure braid).

Definition 6.1 (i) A string link L is *split* if there exists a properly embedded 2–disk $(D, \partial D) \hookrightarrow (I \times D^2, \partial(I \times D^2))$ whose image is disjoint from L and such that the two 3–balls into which D separates $I \times D^2$ each contain component(s) of L. Such a 2–disk is called a *splitting disk*. See Figure 11.

(ii) A 1–*strand cable* is a string link L which has a neighborhood $N \cong I \times D^2$ such that L considered as a link in N is a (pure) braid B. In other words, "all the strands are tied into a knot". We call $\partial N \setminus \partial(I \times D^2)$ a *cabling annulus* for L. See Figure 12.

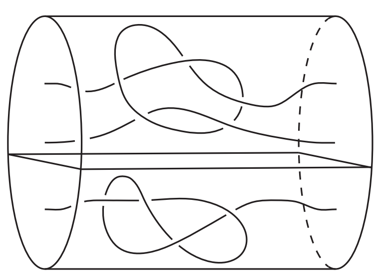


Figure 11: An example of a split link

Since we are now focusing on 2–string links, we need not consider (or even define) k–strand cables for $k > 1$. Hence we will often refer to 1–strand cables as just *cables*.

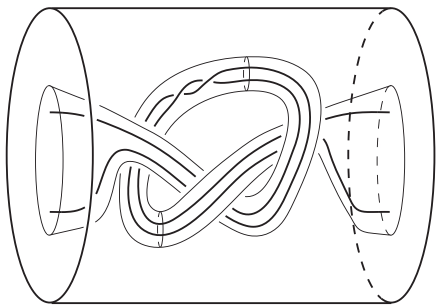


Figure 12: An example of a 1–strand cable, shown together with a cabling annulus

Theorem 6.2 (Proven in [1]) *The monoid $\pi_0\mathcal{L}_2$ has center $\mathcal{C}$ generated by the pure braids, split links, and 1–strand cables. The quotient $\pi_0\mathcal{L}_2/\mathcal{C}$ is free. Furthermore, every 2–component string link can be written as a product of prime factors*

$$L_1 \# \cdots \# L_m \# K_1 \# \cdots \# K_{n-m},$$

where the K_i are precisely the factors which are in the center. Such an expression is unique up to reordering the K_i and multiplying any of the factors by units (pure braids).

6.2 Removing twists

Next note that the linking number gives a map $\ell\colon \mathcal{L}_2 \to \mathbb{Z}$ which descends to a monoid homomorphism $\pi_0\ell\colon \pi_0\mathcal{L}_2 \to \mathbb{Z}$. For $n \in \mathbb{Z}$, let $\mathcal{L}_2^n = \ell^{-1}\{n\}$. We might like to think of this as $0 \to \mathcal{L}_2^0 \to \mathcal{L}_2 \to \mathbb{Z} \to 0$, though if we wanted this to be a short exact sequence of monoids, we should instead write

$$0 \to \pi_0\mathcal{L}_2^0 \to \pi_0\mathcal{L}_2 \to \mathbb{Z} \to 0,$$

since $\mathcal{L}_2$ is only a monoid up to homotopy. There is an action of $\mathbb{Z}$ on $\mathcal{L}_2$ where the generator $1 \in \mathbb{Z}$ acts by following the embedding in $\mathcal{L}_2$ by the map $D^2 \times I \to D^2 \times I$ given by $(z,t) \mapsto (e^{2\pi i t}z, t)$. The action of any $m \in \mathbb{Z}$ thus gives a continuous map[2] $\mathcal{L}_2^n \to \mathcal{L}_2^{n+m}$ with continuous inverse given by the action of $-m$. Thus $\mathcal{L}_2 \cong \mathcal{L}_2^0 \times \mathbb{Z}$, and it suffices to study $\mathcal{L}_2^0$ to understand $\mathcal{L}_2$. Note that Theorem 6.2 above implies that an element of $\pi_0\mathcal{L}_2^0$ can be written as a product of primes $L_1\#\cdots\#L_m\#K_1\#\cdots\#K_{n-m}$ which is unique up to *only reordering the* K_i. We similarly define a subspace $\widehat{\mathcal{L}}_2^0 \subset \widehat{\mathcal{L}}_2$ in the space of fat string links with zero framing number; note that $\widehat{\mathcal{L}}_2^0 \simeq \mathcal{L}_2^0$.

[2]We can see that the map does indeed have this codomain because the resulting twisted link can be taken by an isotopy to a link where the twists are on one end, in which case the linking number is clearly increased by m.

Proposition 6.3 *The isotopies that yield the commutativity relations in $\pi_0\mathcal{L}_2^0$ (which by Theorem 6.2 are* all *the relations in $\pi_0\mathcal{L}_2^0$) can be realized as paths in the spaces $\mathcal{I}(c_1,\dots,c_n;2)$, where $c_i\in\{1,2\}$.*

Proof Note that by Theorem 6.2 any 2–string link can be obtained from infections of the trivial 2–string link by prime knots and non-central prime 2–string links; these infections can be chosen to commute with each other (so that they can be carried out "all at once"). In terms of fat string links in $\widehat{\mathcal{L}}_2^0$ ($\simeq\mathcal{L}_2^0$), we can express these operations using a relatively small class of 2–holed mufflers and hockey pucks, as follows.

Recall that an element $a\in\mathcal{C}_1(1)$ is just an affine-linear map $I\hookrightarrow I$. Let $e_1,e_2\colon D^2\hookrightarrow D^2$ denote the restrictions of the trivial fat 2–string link $i_2\colon\coprod_2 I\times D^2\hookrightarrow I\times D^2$ to the two components in the 0–time slice:

$$e_1\sqcup e_2\colon(\{0\}\times D^2)\sqcup(\{0\}\times D^2)\hookrightarrow\{0\}\times D^2.$$

(Equivalently, e_1,e_2 are the restrictions of i_2 to the two components of a time-slice at any time $t\in I$). Consider infection diagrams $(L_0,M_1,\dots,M_n,\sigma)$ representing classes in $\mathcal{I}(c_1,\dots,c_n;2)$ which satisfy the following three conditions (see Figure 13):

- L_0 is the trivial 2–string link.
- If $c_i=1$, then either
  - (A1) $L_i=a_i\times e_1$ for some $a_i\in\mathcal{C}_1(1)$, or
  - (A2) $L_i=a_i\times e_2$ for some $a_i\in\mathcal{C}_1(1)$, or
  - (B) $L_i=a_i\times\mathrm{id}\colon I\times D^2\hookrightarrow I\times D^2$ for some $a_i\in\mathcal{C}_1(1)$.
- If $c_i=2$, then $L_i=a_i\times\mathrm{id}\colon I\times D^2\hookrightarrow I\times D^2$ for some $a_i\in\mathcal{C}_1(1)$.

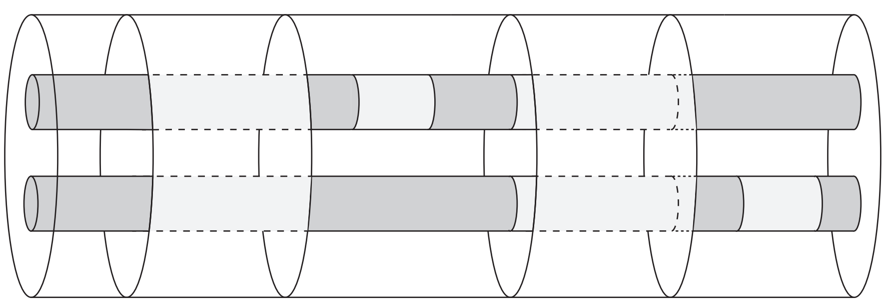


Figure 13: An element of $\mathcal{I}(1,1,2,1;2)$, where the L_i are ordered from left to right. In this element, we have a hockey puck of type (B), then a hockey puck of type (A1), then a two-holed muffler, then a hockey puck of type (A2).

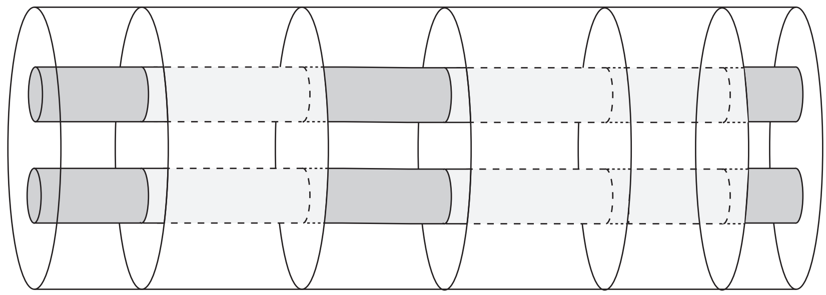


Figure 14: An element of $\mathcal{I}_\#(3)$ ($\cong \mathcal{C}_1(3)$)

Notice that plugging knots into pucks of types (A1) and (A2) produces a split link, while plugging a knot into a puck of type (B) produces a cable. Hockey pucks of types (A1) and (A2) can move through the inside of the two-holed mufflers, while the pucks of type (B) can move through the two-holed mufflers on the outside. These two motions correspond to the centrality of split links and cables, which by Theorem 6.2 are all the commutativity relations in $\pi_0 \mathcal{L}_2^0$. This proves the proposition. (Since $\mathcal{L}_2 \cong \mathcal{L}_2^0 \times \mathbb{Z}$, this is fairly close to a statement about all of $\mathcal{L}_2$.) □

6.3 A suboperad of the 2–colored restriction

Let $\mathcal{I}_{\{2\}}$ denote the suboperad of $\mathcal{I}$ corresponding to the color $\{2\} \subset \mathbb{N}^+$. Note that $\mathcal{I}_{\{2\}}$ is an ordinary operad.

Definition 6.4 We define the *stacking suboperad* $\mathcal{I}_\# \subset \mathcal{I}_{\{2\}}$ as the suboperad where each space $\mathcal{I}_\#(n)$ consists of elements of $\mathcal{I}(2, \dots, 2; 2)$ represented by infection diagrams $(L_0, M_1, \dots, M_n, \sigma)$ satisfying the following conditions:

- L_0 is the trivial 2–string link.
- $M_i = a_i \times \mathrm{id}\colon I \times D^2 \hookrightarrow I \times D^2$ for some $a_i \in \mathcal{C}_1(1)$.

See Figure 14 for a picture of an element of this suboperad.

The following is obvious:

Proposition 6.5 *For each n, the space $\mathcal{I}_\#(n)$ is homeomorphic to $\mathcal{C}_1(n)$. Thus $\mathcal{I}_\#(n)$ has contractible components and is equivalent to Σ_n.* □

6.4 A decomposition theorem

Recall that $\widehat{\mathcal{L}}_c$ is the space of fat c–string links with zero framing number in each component; $\widehat{\mathcal{L}}_c^0 \subset \widehat{\mathcal{L}}_c$ is the subspace where the linking number is 0 (defined in Section 6.2); and we have homotopy equivalences $\widehat{\mathcal{L}}_c \simeq \mathcal{L}_c$ and $\widehat{\mathcal{L}}_c^0 \simeq \mathcal{L}_c^0$. Let $\mathcal{P}_c \subset \widehat{\mathcal{L}}_c$ be the subspace of prime c–component fat string links. We will decompose a certain subspace of $\widehat{\mathcal{L}}_2$ in terms of our infection operad and the prime links in this subspace.

Definition 6.6 Define $\mathcal{S}_2$ to be the subspace of $\widehat{\mathcal{L}}_2$ consisting of certain components of $\widehat{\mathcal{L}}_2$: the component of $\widehat{\mathcal{L}}_2$ corresponding to a string link L is in $\mathcal{S}_2$ if and only if L is a product of prime string links, each of which is *not* in the center of $\pi_0\mathcal{L}_2$. (In other words, each prime factor of L is neither a split link nor a cable.) Let $\mathcal{PS}_2 := \mathcal{P}_2 \cap \mathcal{S}_2$, let $\mathcal{S}_2^0 := \mathcal{S}_2 \cap \widehat{\mathcal{L}}_2^0$, and let $\mathcal{PS}_2^0 := \mathcal{PS}_2 \cap \widehat{\mathcal{L}}_2^0 = \mathcal{P}_2 \cap \mathcal{S}_2 \cap \widehat{\mathcal{L}}_2^0$.

Before stating our decomposition theorem, we review a useful lemma, well known to embedding theorists. Before proving the lemma, we need to set some more definitions.

- Let $\widehat{\mathcal{L}} = \coprod_{c=1}^{\infty} \widehat{\mathcal{L}}_c$.
- For $L \in \widehat{\mathcal{L}}$, let $\widehat{\mathcal{L}}(L)$ denote the component of L in $\widehat{\mathcal{L}}$.
- Recall that if L is an embedding of a 3–manifold with boundary into $I \times D^2$,
$$C_L := D^3 \setminus (\mathrm{i}\overset{\circ}{\mathrm{m}}\, L),$$
where we identify $I \times D^2$ with D^3.
- For a manifold with boundary M, let $\mathrm{Diff}(M;\partial)$ denote the space of diffeomorphisms of M which are the identity on the boundary.
- For a group G, let BG denote the classifying space of G.

Lemma 6.7 *For any* $L \in \widehat{\mathcal{L}}$, $\widehat{\mathcal{L}}(L) \simeq B\mathrm{Diff}(C_L;\partial)$.

Proof Given a diffeomorphism in $\mathrm{Diff}(D^3,\partial)$, we can restrict to the image of L to get a fibration
$$\mathrm{Diff}(C_L;\partial) \longrightarrow \mathrm{Diff}(D^3;\partial) \longrightarrow \widehat{\mathcal{L}}(L).$$
Hatcher showed that $\mathrm{Diff}(D^3;\partial)$ is contractible (the Smale conjecture [18]), which implies the result. □

Theorem 6.8 *The subspace* $\mathcal{S}_2^0$ *is freely generated over the stacking suboperad* $\mathcal{I}_\#$ *by its subspace* $\mathcal{PS}_2^0$ *of non-split, non-cable prime string links. More precisely,*
$$\mathcal{S}_2^0 \simeq \mathcal{I}_\#(\mathcal{PS}_2^0 \sqcup \{*\}) := \coprod_{n=0}^{\infty} \mathcal{I}_\#(n) \times_{\Sigma_n} (\mathcal{PS}_2^0 \sqcup \{*\})^n \ \left(\simeq \coprod_{n=0}^{\infty} (\mathcal{PS}_2^0 \sqcup \{*\})^n\right),$$

where $\{*\}$ corresponds to the component of the trivial 2–string link. Furthermore $\mathcal{S}_2 \cong \mathcal{S}_2^0 \times \mathbb{Z}$.

Proof First note that by Theorem 6.2 we have a bijection on π_0. In fact, a prime decomposition $L = L_1 \# \cdots \# L_n$ corresponds to an isotopy class of an equivalence class of infection diagram in $\mathcal{I}_\#$ with n mufflers exactly as in Definition 6.4.

Now we will check that we have an equivalence on each component of $\mathcal{S}_2^0$. So fix $L \in \mathcal{S}_2^0$. Let C_L denote the complement of L in D^3, as above.

Definition 6.9 For a c–component string link L, a *decomposing disk* $D \subset C_L$ is a 2–disk with c open 2–disks removed which is properly embedded in C_L in such a way that c of its boundary components are (isotopic to) the c meridians of L.

Note that a decomposing disk D is incompressible in C_L [1, Lemma 2.9].

A prime decomposition $L = L_1 \# \cdots \# L_n$ corresponds to a maximal collection of decomposing disks $D_1, \dots, D_{n-1}$ such that no two D_i are isotopic. Thus the decomposing disks $D_1, \dots, D_{n-1}$ cut C_L into n pieces that are precisely $C_{L_1}, \dots, C_{L_n}$. Recall the uniqueness of prime decompositions for $L \in \mathcal{S}_2^0$ given by Theorem 6.2. The proof of this theorem implies that (the image of) such a maximal collection of decomposing disks is unique up to isotopy. Note that the prime factors of $L \in \mathcal{S}_2^0$ cannot even be reordered.

Now consider the fibration

$$\mathrm{Diff}\Big(\coprod_{i=1}^{n} C_{L_i}; \partial\Big) \longrightarrow \mathrm{Diff}(C_L; \partial) \longrightarrow \mathrm{Emb}\Big(\coprod_{i=1}^{n} D_i,\ C_L\Big). \tag{5}$$

Hatcher proved [17] that for a 3–manifold M and a properly embedded incompressible surface $S \subset M$, the space $\mathrm{Emb}(S, M)$ has contractible components unless S is a torus. (Strictly speaking, Hatcher proves this for connected S, but for $S = \coprod_{i=1}^n S_i$ with each S_i a connected surface, one can use the fibration

$$\mathrm{Emb}\Big(S_n,\ M \setminus \Big(\coprod_{i=1}^{n-1} S_i\Big)\Big) \longrightarrow \mathrm{Emb}\Big(\coprod_{i=1}^{n} S_i,\ M\Big) \longrightarrow \mathrm{Emb}\Big(\coprod_{i=1}^{n-1} S_i,\ M\Big)$$

and induction on n to get the result, noting that Hatcher's theorem applies when the 3–manifold is a component of $M \setminus (\coprod_{i=1}^{n-1} S_i)$.)

Thus the components of the base space in Equation (5) are contractible. Since the images of the D_i are determined up to isotopy, we may replace $\mathrm{Emb}(\coprod_{i=1}^n D_i, C_L)$

by $\mathrm{Diff}(\coprod_{i=1}^n D_i)$ (since the latter space also has contractible components). Note that the fiber in Equation (5) is $\prod_{i=1}^n \mathrm{Diff}(C_{L_i})$. So we have

$$\mathrm{Diff}\Big(\coprod_{i=1}^{n} D_i\Big) \longrightarrow \mathrm{Diff}(C_L;\partial) \longrightarrow \mathrm{Diff}\Big(\coprod_{i=1}^{n} D_i\Big).$$

Now apply the classifying space functor $B(-)$ to the above fibration. By Lemma 6.7, we get

$$\prod_{i=1}^{n} \widehat{\mathcal{L}}(L_i) \longrightarrow \widehat{\mathcal{L}}(L) \longrightarrow \prod_{i=1}^{n} \mathrm{Conf}_2(D^2),$$

where $\mathrm{Conf}_2(D^2)$ is the space of ordered distinct pairs in D^2 (or the classifying space of the braid group on two strands). The base space is a $\mathrm{K}(\pi,1)$, ie it has trivial π_i for $i>1$. We claim that on π_1, the fibration is the zero map: in fact, if $\alpha \in \pi_1(\widehat{\mathcal{L}}(L))$ produced a nontrivial braid (say, in the i^{th} factor), then in $\alpha(1)$, at least one of the two summands determined by D_i would have nonzero ℓ (number of twists), contradicting the fact that α is a loop (in $\mathcal{S}_2^0$).

So by the long exact sequence in homotopy groups for a fibration, the map from fiber to total space is an isomorphism on π_i for all $i \geq 0$. Then by the Whitehead theorem,

$$\widehat{\mathcal{L}}(L) \simeq \prod_{i=1}^{n} \widehat{\mathcal{L}}(L_i).$$

The right-hand space can be rewritten as $\Sigma_n \times_{\Sigma_n} \prod_{i=1}^n \widehat{\mathcal{L}}(L_i)$, which by Proposition 6.5 is equivalent to $\mathcal{I}_\#(n) \times_{\Sigma_n} \prod_{i=1}^n \widehat{\mathcal{L}}(L_i)$. This proves the main assertion of the theorem.

The remaining assertion, that $\mathcal{S}_2 \cong \mathcal{S}_2^0 \times \mathbb{Z}$, follows immediately from Section 6.2. □

6.5 Final remarks and future directions

We have described the components of links in $\mathcal{S}_2$ in terms of the components of the prime links in $\mathcal{S}_2$. In general, we do not have descriptions of the components of the prime links in $\mathcal{S}_2$ themselves. However, we can describe some components of $\mathcal{L}_2$. We believe that at least some of these descriptions have been known to experts.

Proposition 6.10 *The component of a 2–string link $R \in \widehat{\mathcal{L}}_2$ which is a rational tangle is contractible.*

Proof We have a fibration

$$\mathrm{Diff}(C_R;\partial) \longrightarrow \mathrm{Diff}(D^3;\partial) \longrightarrow \widehat{\mathcal{L}}(R) \tag{6}$$

given by restricting to the image of R. The total space is contractible by the Smale conjecture. So it suffices to show that the fiber $\mathrm{Diff}(C_R;\partial)$ is contractible. Note that C_R is a genus-2 handlebody.

We claim that for any 3–dimensional handlebody H, $\mathrm{Diff}(H;\partial)$ is contractible. This can be proven by induction on the genus. The basis case of genus 0 is the Smale conjecture. For the induction step, let S be a meridional disk in H. Consider the fibration

$$F \longrightarrow \mathrm{Diff}(H;\partial) \longrightarrow \mathrm{Emb}(S,H),$$

where the base is the space of proper embeddings of S with fixed behavior on ∂S. The fiber F is the space of diffeomorphisms of a handlebody whose genus is 1 less than that of H, and it is contractible by the induction hypothesis. Hatcher's result on incompressible surfaces says that $\mathrm{Emb}(S,H)$ has contractible components. Furthermore, we claim that any two such embeddings of S are isotopic; this can be proven using the fact that handlebodies are irreducible (ie every 2–sphere in H bounds a 3–ball) and standard "innermost disk" arguments from 3–manifold theory. Hence $\mathrm{Emb}(S,H)$ is connected, hence contractible. Thus $\mathrm{Diff}(H;\partial)$ is contractible. Thus the base space in the fibration of Equation (6) is also contractible. □

Recall the definitions of split links and splitting disks from Definition 6.1.

Proposition 6.11 *If L is a split string link which splits as links L_1, L_2, then*

$$\widehat{\mathcal{L}}(L) \simeq \widehat{\mathcal{L}}(L_1) \times \widehat{\mathcal{L}}(L_2).$$

Proof Let D be a splitting disk for L. Consider the fibration

$$\mathrm{Diff}(C_{L_1};\partial) \times \mathrm{Diff}(C_{L_2};\partial) \longrightarrow \mathrm{Diff}(C_L;\partial) \longrightarrow \mathrm{Emb}(D, C_L),$$

where $\mathrm{Emb}(D, C_L)$ is the space of embeddings of D which agree on ∂D with the given embedding of D. By Hatcher's theorem on incompressible surfaces, this space has contractible components. Irreducibility of C_L implies further that any two such embeddings of D in C_L are isotopic, showing that the base space is connected, hence contractible. This gives us the desired equivalence. □

If we restrict our attention to 2–string links, the split links are just those links which are obtained by tying a knot in one or both strands. So Budney's work [3] together with Proposition 6.11 gives a description of the homotopy type of each such component of $\mathcal{L}_2$.

We conclude by mentioning two open problems that immediately stand out as follow-ups to Theorem 6.8:

Problem 1 To determine the homotopy types of components of prime non-central 2–string links.

Problem 2 To understand how different types of 2–string links interact, ie find a generalization of Theorem 6.8 from the subspace $\mathcal{S}_2$ to the space of all 2–string links.

References

[1] **R Blair**, **J Burke**, **R Koytcheff**, *A prime decomposition theorem for the* 2*–string link monoid*, J. Knot Theory Ramifications 24 (2015) MR3334658

[2] **R Budney**, *Little cubes and long knots*, Topology 46 (2007) 1–27 MR2288724

[3] **R Budney**, *Topology of knot spaces in dimension* 3, Proc. Lond. Math. Soc. 101 (2010) 477–496 MR2679699

[4] **R Budney**, *An operad for splicing*, J. Topol. 5 (2012) 945–976 MR3001316

[5] **J Cerf**, *Sur les difféomorphismes de la sphère de dimension trois* $(\Gamma_4 = 0)$, Lecture Notes in Mathematics 53, Springer, Berlin (1968) MR0229250

[6] **D Cimasoni**, *Slicing Bing doubles*, Algebr. Geom. Topol. 6 (2006) 2395–2415 MR2286030

[7] **T D Cochran**, *Noncommutative knot theory*, Algebr. Geom. Topol. 4 (2004) 347–398 MR2077670

[8] **T Cochran**, **S Friedl**, **P Teichner**, *New constructions of slice links*, Comment. Math. Helv. 84 (2009) 617–638 MR2507256

[9] **T D Cochran**, **S Harvey**, **C Leidy**, *Primary decomposition and the fractal nature of knot concordance*, Math. Ann. 351 (2011) 443–508 MR2836668

[10] **T D Cochran**, **K E Orr**, *Homology boundary links and Blanchfield forms: Concordance classification and new tangle-theoretic constructions*, Topology 33 (1994) 397–427 MR1286923

[11] **T D Cochran**, **K E Orr**, **P Teichner**, *Knot concordance, Whitney towers and* L^2*–signatures*, Ann. of Math. 157 (2003) 433–519 MR1973052

[12] **T D Cochran**, **K E Orr**, **P Teichner**, *Structure in the classical knot concordance group*, Comment. Math. Helv. 79 (2004) 105–123 MR2031301

[13] **J H Conway**, *An enumeration of knots and links, and some of their algebraic properties*, from: "Computational Problems in Abstract Algebra", (J Leech, editor), Pergamon, Oxford (1970) 329–358 MR0258014

[14] **P M Gilmer**, *Slice knots in* S^3, Quart. J. Math. Oxford Ser. 34 (1983) 305–322 MR711523

[15] **K Habiro**, *Claspers and finite type invariants of links*, Geom. Topol. 4 (2000) 1–83 MR1735632

[16] **S L Harvey**, *Homology cobordism invariants and the Cochran–Orr–Teichner filtration of the link concordance group*, Geom. Topol. 12 (2008) 387–430 MR2390349

[17] **A Hatcher**, *Homeomorphisms of sufficiently large P^2–irreducible 3–manifolds*, Topology 15 (1976) 343–347 MR0420620

[18] **A E Hatcher**, *A proof of the Smale conjecture,* $\mathrm{Diff}(S^3) \simeq \mathrm{O}(4)$, Ann. of Math. 117 (1983) 553–607 MR701256

[19] **M W Hirsch**, *Differential topology*, Graduate Texts in Mathematics 33, Springer, New York (1976) MR0448362

[20] **V F R Jones**, *Planar algebras, I*, preprint (1999) arXiv:math/9909027

[21] **J-Y Le Dimet**, *Cobordisme d'enlacements de disques*, Mém. Soc. Math. France 32, Soc. Math. France, Paris (1988) MR971415

[22] **C Livingston**, *A survey of classical knot concordance*, from: "Handbook of knot theory", (W Menasco, M Thistlethwaite, editors), Elsevier, Amsterdam (2005) 319–347 MR2179265

[23] **M Markl**, **S Shnider**, **J Stasheff**, *Operads in algebra, topology and physics*, Mathematical Surveys and Monographs 96, Amer. Math. Soc. (2002) MR1898414

[24] **J P May**, *The geometry of iterated loop spaces*, Lecture Notes in Mathematics 271, Springer, Berlin (1972) MR0420610

[25] **J E McClure**, **J H Smith**, *Operads and cosimplicial objects: An introduction*, from: "Axiomatic, enriched and motivic homotopy theory", (J P C Greenlees, editor), NATO Sci. Ser. II Math. Phys. Chem. 131, Kluwer, Dordrecht, The Netherlands (2004) 133–171 MR2061854

Department of Mathematics, Rhode Island College
Providence, RI 02908-1991, USA

Department of Mathematics and Statistics
University of Massachusetts, Amherst, MA 01003, USA

jburke@ric.edu, koytcheff@math.umass.edu

http://people.math.umass.edu/~koytcheff

Received: 11 August 2014 Revised: 25 February 2015

msp

Algebraic & Geometric Topology 15 (2015) 3409–3433

Systoles and kissing numbers of finite area hyperbolic surfaces

FEDERICA FANONI
HUGO PARLIER

We study the number and the length of systoles on complete finite area orientable hyperbolic surfaces. In particular, we prove upper bounds on the number of systoles that a surface can have (the so-called kissing number for hyperbolic surfaces). Our main result is a bound which only depends on the topology of the surface and which grows subquadratically in the genus.

30F10; 32G15, 53C22

1 Introduction

In analogy with classical sphere packing problems in $\mathbb{R}^n$, Schmutz Schaller named and studied *kissing numbers* for hyperbolic surfaces. This is a particular instance of a more general analogy between the study of n–dimensional lattices (and their parameter spaces) and the study of hyperbolic surfaces (and their parameter spaces). Both are natural generalizations of the study of 2–dimensional flat tori. The natural parameter spaces of these tori are $\mathbb{H}$ and $\mathbb{H}/\operatorname{PSL}_2(\mathbb{Z})$; their higher-dimensional analogues include on the one hand the spaces of lattices and on the other Teichmüller and moduli spaces.

The classical kissing number problem is to bound the number of disjoint open unit balls that can be tangent to a fixed unit ball; the *lattice kissing number* is the same problem but where one asks that the centers of the spheres lie on some lattice. This is in fact an equivalent problem to counting the number of systoles (up to isotopy) of the underlying lattice. Another classical topic for flat tori is the study of Hermite constants. This involves finding sharp upper bounds on the length of shortest non-trivial lattice vectors or, in other words, bounds on the systole length of the quotient tori. Both of these problems make perfect sense for finite-area hyperbolic surfaces and have been studied by a variety of authors including Bavard [3] and Schmutz Schaller [17; 19].

Schmutz Schaller provided a variety of results on the length and the number of systoles for complete hyperbolic surfaces in both the closed and finite area cases. Lower bounds for either of these quantities can be found using arithmetic methods. Buser

Published: 12 January 2016 DOI: 10.2140/agt.2015.15.3409

and Sarnak [8] were the first to show that there exist families S_k of closed surfaces of genus g_k with $g_k \to \infty$ as $k \to \infty$ whose systole length grows like

$$\mathrm{sys}(S_k) \geq \tfrac{4}{3} \log g_k.$$

Katz, Schaps and Vishne [11] generalized this construction to principal congruence subgroups of arbitrary arithmetic surfaces. Makisumi [12] showed that, in some sense, this is the best one can hope for via arithmetic constructions. Schmutz Schaller [19] found analogous results for kissing numbers: for any $\varepsilon > 0$, there is a family of closed surfaces T_k of genus g_k with $g_k \to \infty$ as $k \to \infty$ whose number of systoles grows like

$$\mathrm{Kiss}(T_k) \geq g_k^{\frac{4}{3}-\varepsilon}.$$

For surfaces with cusps, families reaching these lower bounds (for both quantities) are directly obtainable by considering principal congruence subgroups of $\mathrm{PSL}_2(\mathbb{Z})$ (see Schmutz Schaller [18], Brooks [5] and Balacheff, Makover and Parlier [2]). The number of cusps in these examples grows roughly like $g^{2/3}$.

Upper bounds for these quantities have also been studied, in particular for closed surfaces. Via an easy area argument, one can obtain an upper bound on the systole length of closed surfaces of genus g that grows like $2 \log g$. This complements Buser and Sarnak's lower bound to show that the rough growth is logarithmic, but the discrepancy between the $\frac{4}{3}$ and the 2 remains mysterious. Schmutz Schaller, using a disk packing argument of Fejes Tóth, proved a very nice upper bound on systole length which is actually sharp for the congruence subgroups of $\mathrm{PSL}_2(\mathbb{Z})$ (see also Adams [1] and Bavard [4]). We use this result in an essential way and give the exact formulation in the sequel (Theorem 2.4).

For kissing numbers, the best known upper bounds are results of the second author [13]. In particular, there is a bound which depends only on the genus g and which grows at most subquadratically in function of g. Again, there is a discrepancy between the $g^{4/3}$ lower bound and the g^2 upper bound (although the latter cannot be sharp). Upper bounds for kissing numbers of non-closed finite-area complete surfaces (ie surfaces with cusps) have yet to be approached. Filling this gap is the main goal of our article.

One of the main consequences of what we obtain is the following:

Theorem 4.11 *There exists a universal constant $C > 0$ such that, for any $S \in \mathcal{M}_{g,n}$, $g \geq 1$, its kissing number satisfies*

$$\mathrm{Kiss}(S) \leq C(g+n)\frac{g}{\log(g+1)}.$$

We obtain this result as a consequence of a number of results concerning the length and the topological configurations of systoles.

In particular, concerning the length of systoles, we show the following:

Theorem 2.3 *There exists a universal constant $K<8$ such that every $S\in\mathcal{M}_{g,n}$ $(g\neq 0)$ satisfies*
$$\mathrm{sys}(S)\leq 2\log g+K.$$

The result is not surprising in view of the results for closed surfaces and Schmutz Schaller's bound, but it is interesting to note that it is asymptotically a stronger bound when the growth of the number of cusps is bounded above by $g^{1/2}$.

Our results on topological configurations of systoles can be summarized as follows:

Propositions 3.2 and 3.3 and Lemma 3.5 *If α and β are systoles of a surface $S\in\mathcal{M}_{g,n}$, then*
$$i(\alpha,\beta)\leq 2$$
and, if $i(\alpha,\beta)=2$, then either α or β surrounds two cusps. Furthermore, for every genus $g\geq 0$, there exists $n(g)\in\mathbb{N}$ and a surface S_g of genus g with $n(g)$ cusps which has systoles that intersect twice.

The above result is in contrast with closed surfaces, where systoles can intersect at most once.

Finally, we obtain the following bound, which relates systolic lengths and kissing numbers.

Theorem 4.10 *If $S\in\mathcal{M}_{g,n}$ has systole of length $\mathrm{sys}(S)=\ell$, then*
$$\mathrm{Kiss}(S)\leq 20n\cosh\left(\tfrac{1}{4}\ell\right)+200\frac{e^{\ell/2}}{\ell}(2g-2+n).$$

The article is organized as follows. In Section 2 we prove our upper bounds on systole length. Section 3 is dedicated to the study of the topological configurations of systoles. In Section 4 we prove Theorems 4.10 and 4.11.

Acknowledgements The authors would like to thank Bram Petri for useful discussions and the reviewer for her or his pertinent remarks and suggestions. Both authors were fully supported by Swiss National Science Foundation grant number PP00P2_128557. In addition, the authors acknowledge support from US National Science Foundation grants DMS 1107452, 1107263, 1107367 "RNMS: GEometric structures And Representation varieties" (the GEAR Network). The first author was also supported by Swiss National Science Foundation grant number P2FRP2_161723.

2 Bounds on lengths of systoles

We denote by $\mathcal{M}_{g,n}$ the *moduli space* of surfaces of signature (g, n), by which we mean the space of all complete finite area hyperbolic surfaces of genus g with n cusps up to isometry. We shall always assume that g and n satisfy $3g-3+n>0$. A *systole* of a surface $S \in \mathcal{M}_{g,n}$ is a shortest closed geodesic. We think of systoles — and closed geodesics in general — as being non-oriented. Given a surface S, we denote its systole length (the length of one of its systoles) by $\operatorname{sys}(S)$. The main objective of this section is to show that every surface of genus $g \geq 1$, with or without cusps, has systole length bounded above by a function which only depends on the genus.

For any cusp c, let H_c be the associated open horoball region of area 2. By the collar lemma (see for instance Chapter 4 of Buser [7]), two such regions are disjoint.

For any cusp c and any non-negative r, define the set $D_r(c)$ to be

$$D_r(c) := \{p \in S \mid d(p, H_c) < r\} \cup H_c.$$

If $D_r(c)$ is homeomorphic to a once-punctured disk, we can compute its area, which is

$$\operatorname{area}(D_r(c)) = 2e^r.$$

Lemma 2.1 (a) *If there are two cusps c and c' such that $D_r(c)$ and $D_r(c')$ are tangent, then the simple closed geodesic forming a pair of pants with them has length $4\operatorname{arccosh} e^r$, so*

$$\operatorname{sys}(S) \leq 4\operatorname{arccosh}(e^r).$$

(b) *If $D_r(c)$ is tangent to itself for some $r \geq \log 2$, then*

$$\operatorname{sys}(S) \leq 2\operatorname{arccosh}(e^r - 1).$$

Proof (a) Consider the pair of pants determined by the two cusps and the simple closed geodesic γ surrounding them. Cut it along the orthogonal from γ to itself, the shortest geodesic between the cusps and the perpendiculars from the cusps to γ. Consider one of the four obtained quadrilaterals; we denote its vertices by q, s, t and c and the intersection point of ∂H_c with a side by p, as in Figure 1.

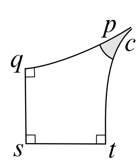


Figure 1: One of the quadrilaterals

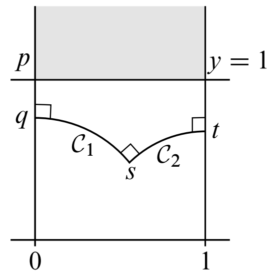


Figure 2: In the upper half-plane

Draw the quadrilateral in the upper half-plane, choosing infinity as the ideal point; see Figure 2. We fix the two geodesics containing qc and tc to be $x=0$ and $x=1$. The area of H_c intersected with the quadrilateral is 1, so ∂H_c is given by $y=1$ and $p=i$. Moreover, $d(p,q)=\frac{1}{2}d(H_c,H_{c'})=r$, so $q=ie^{-r}$. Consider $\mathcal{C}_1$ and $\mathcal{C}_2$, the Euclidean circles representing the geodesics through q and s and through s and t.

Since $\mathcal{C}_1\perp\{x=0\}$, $\mathcal{C}_2\perp\{x=1\}$ and $\mathcal{C}_1\perp\mathcal{C}_2$, they have equations

$$\begin{aligned}\mathcal{C}_1:&\qquad x^2+y^2=R^2,\\ \mathcal{C}_2:&\quad (x-1)^2+y^2=1-R^2\end{aligned}$$

for some R. As $q\in\mathcal{C}_1$, we have $R=e^r$. By imposing $d(t,s)=\frac{1}{4}\ell$, we obtain $\ell=4\operatorname{arccosh}(e^r)$.

(b) The cusp c with the curve of length $2r$ from H_c and back determines a pair of pants with at least one simple closed geodesic as boundary.

If the pair of pants has two cusps and a boundary curve α, we can cut it along the geodesic between the two cusps, the shortest geodesics between the cusps and α and the geodesic containing curve of length $2r$. We get two right-angled triangles with two ideal vertices and $\frac{1}{2}\pi$ and two quadrilaterals with three right angles and an ideal vertex, as in Figure 3.

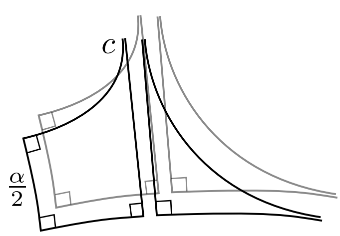


Figure 3: The cut pair of pants with two cusps

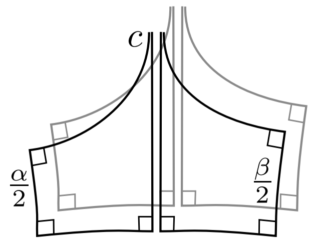


Figure 4: The cut pair of pants with one cusp

By direct computation similar to before, we obtain

$$\ell(\alpha) = 2\operatorname{arccosh}(e^r - 1).$$

If the pair of pants has two boundary curves, we denote them by α and β and we suppose that $\ell(\alpha) \leq \ell(\beta)$. We cut along the orthogonal from α to β, the shortest geodesics from α and β to the horoball and the geodesic containing the curve of length $2r$. We obtain four quadrilaterals, with three right angles and an ideal vertex, pairwise isometric; see Figure 4.

Again by direct computation we have

$$\ell(\alpha) = 2\operatorname{arccosh}(ae^r),$$
$$\ell(\beta) = 2\operatorname{arccosh}((1-a)e^r),$$

where a is the area of H_c intersected with one of the two quadrilaterals containing a part of α. Since $\ell(\alpha) \leq \ell(\beta)$, we have $a \leq \frac{1}{2}$. Moreover, α is longest when a is maximum, that is, when $a = \frac{1}{2}$. In this case

$$\ell(\alpha) = \ell(\beta) = 2\operatorname{arccosh}\left(\tfrac{1}{2}e^r\right).$$

Since by assumption $r \geq \log 2$, we get that in both cases the curve α satisfies

$$\ell(\alpha) \leq 2\operatorname{arccosh}(e^r - 1). \qquad \square$$

Remark 2.2 From the proof of the lemma we also have that, if $D_r(c)$ is tangent to itself for some $r \leq \log 2$, then $\operatorname{sys}(S) \leq 2\operatorname{arcsinh} 1$.

We can now prove our bound on systole length for surfaces of genus $g \geq 1$.

Theorem 2.3 *There exists a universal constant $K < 8$ such that every $S \in \mathcal{M}_{g,n}$ satisfies*

$$\operatorname{sys}(S) \leq 2\log g + K.$$

Proof Set $\ell = \mathrm{sys}(S)$. We begin by recalling the well-known situation where $n = 0$ (and thus $g \geq 2$). As the surface is closed, any open disk $D_{\ell/2}(p)$ of radius $\frac{1}{2}\ell$ is embedded in the surface and thus

$$\mathrm{area}(D_{\ell/2}(p)) = 2\pi\big(\cosh\big(\tfrac{1}{2}\ell\big) - 1\big) \leq \mathrm{area}(S) = 2\pi(2g-2),$$

which in turn implies

$$\ell \leq 2\log g + 2\log 4.$$

Suppose now that $n \geq 1$. We split the proof into three non-mutually exclusive cases. The first situation we consider is when there are "many" cusps (how many will be made explicit); in this case, two of the $D_c(r)$ have to meet for a "small" r and will determine a short curve. In the second case, we assume that there are two cusps which are close to each other and the systole length will be bounded by the length of the curve surrounding them. In the final situation, there are "few" cusps and we further assume any two are far away; in this case, we show that there is a cusp with a short loop from its horoball to itself, which in turn determines a short curve.

Case 1 $$n \geq \sqrt{2\pi g}.$$

If the sets $D_r(c)$ are pairwise disjoint for different cusps c and each homeomorphic to a once-punctured disk, then

$$\mathrm{area}\Big(\bigcup_{c \text{ cusp}} D_r(c)\Big) = 2ne^r \leq \mathrm{area}(S) = 2\pi(2g+n-2);$$

thus,

$$e^r \leq \frac{\pi(2g-2+n)}{n}.$$

Since $n \geq \sqrt{2\pi g}$, this implies

$$e^r \leq \frac{\sqrt{2\pi}(g-1)}{\sqrt{g}} + \pi.$$

So, for some $r \leq \log(\sqrt{2\pi}(g-1)/\sqrt{g} + \pi)$, either two $D_r(c)$ are tangent to each other or one is tangent to itself. Lemma 2.1 now implies

$$\ell \leq 4\,\mathrm{arccosh}\Big(\frac{\sqrt{2\pi}(g-1)}{\sqrt{g}} + \pi\Big).$$

Case 2 There are distinct cusps c_1 and c_2 with $d(H_{c_1}, H_{c_2}) \leq \log(2\pi(g-1+\sqrt{2\pi g}))$.

By Lemma 2.1,

$$\ell \leq 4\,\mathrm{arccosh}\big(\sqrt{2\pi(g-1+\sqrt{2\pi g})}\big),$$

and we are done.

Case 3 $0 < n < \sqrt{2\pi g}$ and any two cusps c_1 and c_2 satisfy

$$d(H_{c_1}, H_{c_2}) > \log(2\pi(g - 1 + \sqrt{2\pi g})).$$

We fix a cusp c. Since any two cusps are far away, for $r \leq \log(2\pi(g - 1 + \sqrt{2\pi g}))$ the set $D_r(c)$ is disjoint from any other $H_{c'}$. If it is also an embedded, once-punctured disk, then

$$\operatorname{area}(D_r(c)) = 2e^r \leq \operatorname{area}(S) < 4\pi(g - 1 + \sqrt{2\pi g}),$$

so

$$r \leq \log(2\pi(g - 1 + \sqrt{2\pi g})).$$

We deduce that, for some $r \leq \log(2\pi(g - 1 + \sqrt{2\pi g}))$, $D_r(c)$ is tangent to itself. By Remark 2.2, if $r \leq \log 2$ then $\ell \leq 2 \operatorname{arcsinh} 1$. Otherwise, by Lemma 2.1, we obtain

$$\ell \leq 2 \operatorname{arccosh}(2\pi(g - 1 + \sqrt{2\pi g}) - 1).$$

Now any surface with $n > 0$ will be in one of the three cases detailed above and, as such, we can deduce

$$\begin{aligned} \ell \leq \max\Big\{ & 4 \operatorname{arccosh}(\sqrt{2\pi}(g - 1)/\sqrt{g} + \pi), 4 \operatorname{arccosh}(\sqrt{2\pi(g - 1 + \sqrt{2\pi g})}), \\ & 2 \operatorname{arccosh}(2\pi(g - 1 + \sqrt{2\pi g}) - 1)\Big\} \\ < {} & 2 \log g + 8. \end{aligned}$$

□

Applying the techniques of the above theorem to punctured spheres, one can show that the systole length of a punctured sphere is bounded by a uniform constant (which doesn't depend on the number of cusps). This is also a consequence of a theorem of Schmutz Schaller, who provided a different bound for the systole length of punctured surfaces.

Theorem 2.4 [17] *For $S \in \mathcal{M}_{g,n}$ with $n \geq 2$ we have*

$$\operatorname{sys}(S) \leq 4 \operatorname{arccosh} \frac{6g - 6 + 3n}{n}.$$

For $n \sim g^\alpha$, Schmutz Schaller's bound grows roughly like $4(1 - \alpha) \log g$. So our bound is stronger for $\alpha < \frac{1}{2}$, while Schmutz Schaller's is better for $\alpha \geq \frac{1}{2}$.

3 Intersection properties of systoles

It is well known, via a simple cutting and pasting argument, that systoles on closed surfaces pairwise intersect at most once. On surfaces with cusps, this is not necessarily the case. For instance, on punctured spheres it is not difficult to see that systoles can intersect twice (the simplest case is a four-times punctured sphere with at least two systoles — they necessarily intersect and the minimal intersection number between two distinct curves is 2). This phenomenon also occurs for surfaces with positive genus. An example of this can be derived from Buser's hairy torus (see [7, Chapter 5]) with cusps instead of boundary curves and explicit examples in all genera are given in the sequel. On the other hand, since systole length is bounded within each moduli space, it follows from the collar lemma that the intersection number between any two systoles is also bounded. This can be considerably sharpened: the first main result of this section will be that two systoles on punctured surfaces can intersect at most twice.

We begin with some notation and well-known preliminary results. A curve is *non-trivial* if it represents a non-trivial element of the fundamental group. A non-trivial curve is *essential* if it does not bound a cusp. In particular, systoles are the shortest essential curves of a surface. Given two closed curves α and β, we denote by $i(\alpha,\beta)$ their *geometric intersection number* (the minimum number of transversal intersection points among representatives in the isotopy classes $[\alpha]$ and $[\beta]$). Two curves are said to intersect *minimally* if they intersect minimally among all representatives of their respective isotopy classes. The unique geodesics in the isotopy classes of simple closed curves are also simple and intersect minimally.

Let α and β be simple closed geodesics on a surface S with $i(\alpha,\beta) \geq 2$ and fix orientations on them. The curve α divides β into arcs between consecutive intersection points. We say such an arc is of *type I* if the orientations at the two intersection points are different and of *type II* if the orientations are the same; see Figure 5.

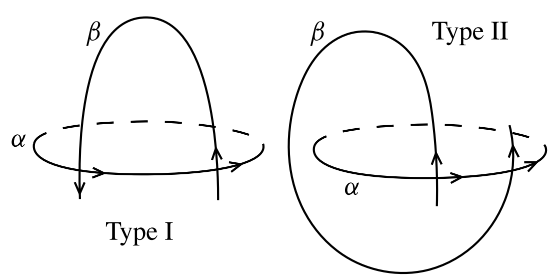


Figure 5: The two kinds of arcs

Note that the orientation at each intersection point depends on the choice of orientations of α and β, but being of type I or II is independent of the choice of orientations.

Lemma 3.1 *If α and β are systoles of a surface $S \in \mathcal{M}_{g,n}$ with $i(\alpha,\beta) \geq 2$, all arcs between consecutive intersection points are of type I.*

Proof By contradiction, suppose that β contains arcs of type II. If there are at least two of them, there exists one, say β_1, of length at most $\frac{1}{2}\,\mathrm{sys}(S)$. Since β_1 divides α into two arcs, one of the two is of length at most $\frac{1}{2}\,\mathrm{sys}(S)$. Call this arc α_1 and consider the curve $\alpha_1 \cup \beta_1$.

If $\alpha_1 \cup \beta_1$ were essential, its geodesic representative would be shorter than $\mathrm{sys}(S)$, which is impossible. Thus $\alpha_1 \cup \beta_1$ must be non-essential. However, one can construct a curve γ homotopic to $\alpha_1 \cup \beta_1$ such that $|\gamma \cap \alpha| = 1$, so via the bigon criterion (see for instance Farb and Margalit [9]) γ and α intersect minimally. Thus

$$i(\gamma,\alpha) = 1$$

and as such γ is non-trivial in homology and is therefore essential, a contradiction.

If there is exactly one arc β_1 of type II, there should be at least two (consecutive) arcs β_2 and β_3 of type I. Then, if $\ell(\beta_1) \leq \frac{1}{2}\,\mathrm{sys}(S)$, we can argue as before to obtain a contradiction. If not, then $\ell(\beta_2 \cup \beta_3) \leq \frac{1}{2}\,\mathrm{sys}(S)$. The arcs β_2, β_3 and α determine an embedded four-holed sphere with a non-trivial curve of length at most $\mathrm{sys}(S)$. By construction, the geodesic in the isotopy class of this curve is strictly shorter than the systole, a contradiction. □

Proposition 3.2 *If α and β are systoles of $S \in \mathcal{M}_{g,n}$, then $i(\alpha,\beta) \leq 2$.*

Proof Suppose by contradiction that $i(\alpha,\beta) > 2$. By Lemma 3.1, all arcs between consecutive intersection points are of type I, so $i(\alpha,\beta)$ is even. Thus there are at least four intersection points and at least four arcs of β between consecutive intersection points. This implies that there is an intersection point and two arcs β_1 and β_2 departing from it with $\ell(\beta_1 \cup \beta_2) \leq \frac{1}{2}\,\mathrm{sys}(S)$. We argue as in the proof of Lemma 3.1: β_1, β_2 and α determine an embedded four-holed sphere with a non-trivial curve of length at most $\frac{1}{2}\,\mathrm{sys}(S)$. By construction, the geodesic in the isotopy class of this curve is strictly shorter than the systole, a contradiction. □

The next proposition shows that if two systoles intersect twice, there is a constraint on the topological configuration of the two curves.

Proposition 3.3 *If two systoles α and β intersect twice, one of them bounds two cusps.*

Proof The two curves cut each other into arcs α_1, α_2 and β_1, β_2. Without loss of generality, we can assume $\ell(\alpha_1) \leq \ell(\beta_1) \leq \frac{1}{2}\operatorname{sys}(S)$. Consider $\gamma_1 = \alpha_1 \cup \beta_1$ and $\gamma_2 = \alpha_1 \cup \beta_2$. As γ_1 and γ_2 do not surround bigons, they cannot be trivial and, as they can be represented by curves of length strictly less than $\operatorname{sys}(S)$, they must both bound a cusp. Hence β bounds two cusps. □

An obvious consequence of Proposition 3.3 is that systoles on surfaces with at most one cusp intersect at most once. In the case of tori this can be improved to show that a surface with twice-intersecting systoles has at least three cusps.

Lemma 3.4 *If $S \in \mathcal{M}_{1,2}$, and α and β are systoles of S, then $i(\alpha, \beta) \leq 1$.*

Proof Suppose two systoles α and β intersect twice. Then $\operatorname{sys}(S) \geq 4 \operatorname{arcsinh} 1$ (see Gauglhofer and Semmler [10]) and, by Proposition 3.3, one of the two curves bounds two cusps. Cut the surface along α and consider the one-holed torus component. The length of the shortest closed geodesic γ in the one-holed torus which doesn't intersect α satisfies (see Parlier [14])

$$\cosh\left(\tfrac{1}{2}\ell(\gamma)\right) \leq \cosh\left(\tfrac{1}{6}\ell(\alpha)\right) + \tfrac{1}{2}$$

and $\ell(\gamma) \geq \operatorname{sys}(S) = \ell(\alpha)$, so

$$\cosh\left(\tfrac{1}{2}\ell(\alpha)\right) \leq \cosh\left(\tfrac{1}{6}\ell(\alpha)\right) + \tfrac{1}{2},$$

which contradicts $\ell(\alpha) \geq 4 \operatorname{arcsinh} 1$. □

On the other hand, we can prove that for every genus there is a punctured surface with systoles intersecting twice. The constructions will involve gluing ideal hyperbolic triangles. Any such triangle has a unique maximal embedded disk tangent to all three sides. We say that two such triangles are glued *without shear* if their embedded disks are tangent.

Lemma 3.5 *For every $g \geq 0$, there exists $n(g) \in \mathbb{N}$ and a surface $S \in \mathcal{M}_{g,n(g)}$ with two systoles intersecting twice.*

Proof For $g = 0$, we can set $n(0) = 4$, as mentioned at the beginning of Section 3: any four-times punctured sphere with at least two systoles will satisfy the requirement. To show the existence of such a surface, pick any $S \in \mathcal{M}_{0,4}$. If it has only one systole γ,

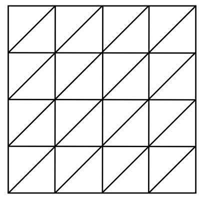


Figure 6: The triangulation of the square

increase the length of γ so that it is still a systole and there is another simple closed geodesic on the surface of the same length.

For $g \geq 1$, we use a building block constructed as follows. Consider a square and a triangulation of it with 32 triangles, given by first subdividing the square into a grid of 16 squares and then adding one diagonal for all squares, as in Figure 6.

Each of the triangles in the square will be replaced by an ideal hyperbolic triangle and all gluings will be without shear.

For $g = 1$, glue opposite sides of the square (again triangles are glued without shear) to obtain a torus with $n(1) = 16$ cusps.

For $g \geq 2$, consider a polygon obtained by gluing a $1 \times (g-1)$ rectangle and a $1 \times 2(g-1)$ rectangle along the long sides, as in Figure 7.

Think of this polygon as a $4g$–gon (with sides corresponding to sides of the squares). Fix an orientation and choose a starting side, to identify the $4g$ sides following the standard pattern $a_1 b_1 a_1^{-1} b_1^{-1} \cdots a_g b_g a_g^{-1} b_g^{-1}$ to obtain a genus-g surface. If we now replace each 1×1 square by the building block (always gluing adjacent triangles without shear), we get a surface of genus g with a decomposition into $32 \cdot 3(g-1)$ ideal triangles. Since it is a triangulation, the number of edges is $\frac{3}{2} \cdot 32 \cdot 3(g-1)$. By an Euler characteristic argument, this implies that the surface has $n(g) = 46g - 46$ cusps.

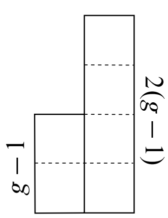


Figure 7: The polygon for $g = 3$

For any $g \geq 1$, consider the set $\mathcal{C}_g$ of curves surrounding pairs of cusps which are connected by an edge between vertices of degree 6 in the triangulation of the surface. By construction, each of these intersects another such curve twice and we defer the proof that these curves are systoles to Lemma 3.6. □

We now prove our claim that the curves in $\mathcal{C}_g$ are indeed systoles.

Lemma 3.6 *For all $g \geq 1$, the curves in $\mathcal{C}_g$ are systoles.*

Proof Consider the triangulation of the surface. For $g = 1$, all vertices are of degree 6. When $g \geq 2$, the pasting scheme associates all exterior vertices of the $4g$–gon and the point in the quotient has degree $12g - 6$ (to see this, simply apply the hand-shaking lemma to the graph given by the triangulation). The remaining vertices are all of degree 6. We denote by Γ the graph dual to the triangulation. From what we have just said, for $g = 1$, cutting the surface along Γ decomposes the surface into hexagons. When $g \geq 2$, cutting along Γ decomposes the surface into hexagons and a single $(12g{-}6)$–gon.

Any simple closed oriented geodesic γ on the surface can be homotoped to a curve on Γ. At every vertex crossed by the curve, the orientations on the surface and on the curve give us a notion of "going left" or "going right". We can associate to γ a word w in the matrices $L = \left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right)$ and $R = \left(\begin{smallmatrix} 1 & 0 \\ 1 & 1 \end{smallmatrix}\right)$, where each L corresponds to a left turn and each R to a right turn. This way of understanding curves on "zero shear surfaces" is fully explained in [6]. In particular, Brooks and Makover show how to compute the length of these curves in terms of the associated word:

$$\ell(\gamma) = 2\operatorname{arccosh}\bigl(\tfrac{1}{2}\operatorname{Tr}(w)\bigr).$$

Each curve in $\mathcal{C}_g$ corresponds to the word $w_0 = RL^4RL^4$ (or LR^4LR^4 or any cyclic permutation of these, depending on the choice of an orientation and a starting point on the curve), which, via a simple computation, has trace 34. To show that the curves in $\mathcal{C}_g$ are systoles, it is enough to show that all other words corresponding to simple closed geodesics have trace at least 34.

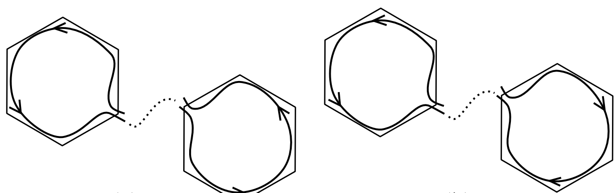
We use the following remark (see for instance [15]):

Remark 3.7 If a word can be written as $w = \cdots w_1 \cdots w_2 \cdots w_k \cdots$, then

$$\operatorname{Tr}(w) \geq \operatorname{Tr}(w_{\sigma(1)} \cdots w_{\sigma(k)})$$

for any cyclic permutation σ of $1, \dots, k$.

Let γ be a simple closed geodesic which is not in $\mathcal{C}_g$. First we observe that we only need to consider curves represented by circuits in Γ. Indeed, if γ corresponds to a closed path which contains an essential (ie not corresponding to a curve going around a cusp) circuit γ', a word representing γ will contain a word representing γ'. By Remark 3.7, γ' is at most as long as γ and we can consider γ' instead. Otherwise, if γ is formed from non-essential circuits, it should contain at least two of them. Note that, since non-essential circuits surround a cusp, they trace a hexagon or a $(12g-6)$–gon. If both these circuits surround hexagons, we are in one of the following situations:



(a) (b)

In case (a) a word associated to the curve contains $RL^5 \cdots RL^5$ and in case (b) it contains $LR^5 \cdots RL^5$. In both cases, by Remark 3.7 and a simple computation, their traces are bigger than 34. Now, if one of the two circuits surrounds the vertex of the triangulation of degree $12g-6$, the curve is even longer.

Suppose then that γ is represented by an essential circuit. If it passes through five consecutive edges of a hexagon (said differently, a corresponding word contains L^4) and is not in $\mathcal{C}_g$, the following modification of the curve (see Figure 8) provides an essential circuit. A word of the curve on the left contains LR^4L, while the one of the curve on the right contains R^2, so the trace decreases (again by Remark 3.7) and we obtain a shorter curve.

We now assume a word w representing γ does not contain L^4 or R^4 and as such it is made of blocks of type L^iR^j for $1 \le i, j \le 3$. If w is made of four or more such blocks, then

$$\operatorname{Tr} w \geq \operatorname{Tr}((LR)^4) > 34.$$

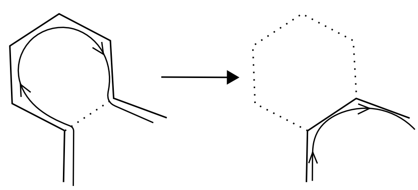


Figure 8: Shortening a curve

Moreover, the length of w is at least 7, as the shortest circuits in Γ are of length 6 and correspond to curves surrounding cusps. With this in hand, one needs to check the finite set of words w made of blocks as above, of length at least 7, and of trace at most 33. To do this one can proceed as follows. The conditions on w give two systems of equations for the exponents of L and R (a system for the words made of two blocks as above and one for words made of three blocks). These systems can be solved to get the set of words we are interested in. It is then straightforward to check that the curves corresponding to these words do not correspond to simple closed geodesics on the surface. $\square$

4 Kissing number bound

In this section we will prove an upper bound for the kissing number depending on the systole length. We then deduce a universal upper bound depending only on the signature of the surface. To do so, we separate the systoles into three sets and we give separate bounds for each of their cardinalities.

For a surface S, let $\mathfrak{S}(S)$ be the set of its systoles and $\mathrm{Kiss}(S) := |\mathfrak{S}(S)|$ be the *kissing number* of S. We say that α and β bound a cusp if they form a pair of pants with a cusp. We define:

$A(S) := \{\alpha \in \mathfrak{S}(S) \mid \alpha \text{ bounds two cusps}\}.$

$B(S) := \{\alpha \in \mathfrak{S}(S) \setminus A(S) \mid \alpha \text{ and } \beta \text{ bound a cusp for some } \beta \in \mathfrak{S}(S) \setminus A(S)\}.$

$C(S) := \mathfrak{S}(S) \setminus (A(S) \cup B(S)).$

Note that by Proposition 3.3 two systoles in $\mathfrak{S}(S) \setminus A(S)$ intersect at most once.

4.1 Bounds on $|A(S)|$

As seen in Lemma 2.1, a curve of length ℓ bounds two cusps c and c' if and only if the distance between H_c and $H_{c'}$ is

$$d(\ell) = 2 \log \cosh\left(\tfrac{1}{4}\ell\right),$$

To bound $|A(S)|$ we will bound the number of pairs of cusps at distance $d(\mathrm{sys}(S))$.

Lemma 4.1 *Let S be a surface with $\mathrm{sys}(S) = \ell$ and c a cusp of S. There are at most $\lfloor 2 \cosh(\frac{1}{4}\ell) \rfloor$ cusps c' which satisfy $d(H_c, H_{c'}) = d(\ell)$.*

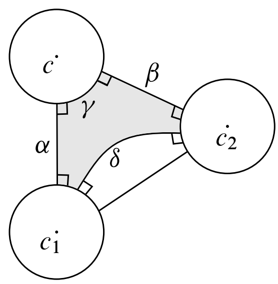
Proof Suppose c_1 and c_2 are two cusps such that

$$d(H_c, H_{c_1}) = d(H_c, H_{c_2}) = d(\ell).$$

Since $\operatorname{sys}(S) = \ell$, the distance between H_{c_1} and H_{c_2} is at least $d(\ell)$. Consider

- the segment α realizing the distance between H_c and H_{c_1},
- the segment β realizing the distance between H_c and H_{c_2},
- the shortest arc γ of ∂H_c bounded by the endpoints of α and β.

Let δ be the unique geodesic segment freely homotopic with endpoints on ∂H_{c_1} and ∂H_{c_2} to the curve $\alpha \cup \beta \cup \gamma$. Then its length is at least $d(\ell)$:



By a direct computation on the (non-geodesic) hexagon determined by α, β, δ and the three horocycles, one can show that

$$\ell(\gamma) \geq \frac{1}{\cosh\left(\frac{1}{4}\ell\right)}.$$

Since ∂H_c has length 2, the number of cusps around c at distance $d(\ell)$ is bounded above by

$$\frac{2}{1/\cosh\left(\frac{1}{4}\ell\right)},$$

which proves the claim as we are bounding an integer. □

As a consequence, we get the following:

Proposition 4.2 *For $S \in \mathcal{M}_{g,n}$ with $\operatorname{sys}(S) = \ell$,*

$$|A(S)| \leq \tfrac{1}{2} n \left\lfloor 2\cosh\left(\tfrac{1}{4}\ell\right)\right\rfloor.$$

Proof There are n cusps, each of which can be surrounded by at most $\lfloor 2\cosh(\frac{1}{4}\ell)\rfloor$ cusps at distance $d(\ell)$. The result follows as each curve surrounds two cusps. □

Remark 4.3 We can get another upper bound for $A(S)$ using the Euler characteristic as follows.

Consider the set of punctures; if there is a systole bounding two of them, we join them with a simple geodesic lying in the pair of pants determined by the systole. We complete this set of geodesics into an ideal triangulation (decomposition into ideal triangles) of the surface. The number of vertices of the triangulation is the number of punctures n. If e is the number of edges, the number of triangles is $\frac{2}{3}e$. The Euler characteristic of the compactified surface is $2g-2$, so

$$n - e + \tfrac{2}{3}e = 2 - 2g.$$

From how we constructed the triangulation, it is clear that $|A(S)| \leq e$, so we get

$$|A(S)| \leq 3(n + 2g - 2).$$

Interestingly, this bound can also be seen as a corollary of the above proposition. If we use Schmutz Schaller's upper bound on systole length (Theorem 2.4) in Proposition 4.2 above, this is exactly the resulting bound.

For surfaces of genus at least one, we will use the bound from the remark above, but for punctured spheres we will use Proposition 4.2 directly.

4.2 Bound on $|B(S)|$

Consider a cusp c; we define two associated sets:

$$B(c) := \{\alpha \in B(S) \mid \alpha \text{ and } \beta \text{ bound } c \text{ for some } \beta \in B(S)\},$$
$$B(c)^{(2)} := \{(\alpha, \beta) \in B(S) \times B(S) \mid \ \alpha \text{ and } \beta \text{ bound } c\}.$$

Suppose (α, β), $(\gamma, \delta) \in B(c)^{(2)}$. Then γ has to pass through the pair of pants given by α, β and c, so γ must intersect α or β. Since curves in $\mathfrak{S}(S) \setminus A(S)$ pairwise intersect at most once, $i(\alpha, \gamma) = i(\beta, \gamma) = 1$ (and the same for δ).

Any curve $\alpha \in B(c)$ is at a fixed distance $D(\ell)$ from H_c. By a direct computation in the pair of pants bounded by α and β, one obtains

$$D(\ell) = \log\left(2\frac{\cosh\left(\frac{1}{2}\ell\right)}{\sinh\left(\frac{1}{2}\ell\right)}\right).$$

When curves in $B(S)$ intersect they do so exactly once and we can obtain a lower bound on their angle of intersection of curves in $B(S)$. (Note that the lemma holds for any pair of systoles that intersect once.)

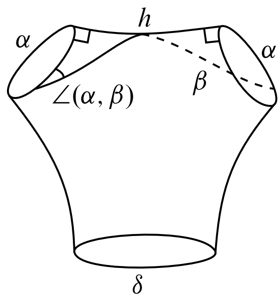


Figure 9: The result of cutting the torus T along α

Lemma 4.4 *Let S be a surface of signature $(g,n) \neq (1,1)$. If α and β are systoles of length ℓ intersecting once, their angle of intersection satisfies*

$$\sin\angle(\alpha,\beta) \geq \sin\theta_\ell := \begin{cases} 2/\sqrt{5}, & \ell < 2\operatorname{arccosh}\frac{3}{2}, \\ \sqrt{2\cosh\left(\frac{1}{2}\ell\right)+1}/(\cosh\left(\frac{1}{2}\ell\right)+1), & \ell \geq 2\operatorname{arccosh}\frac{3}{2}. \end{cases}$$

In particular, the angle of intersection is bounded below by a function θ_ℓ that behaves like $e^{-\ell/4}$ as ℓ goes to infinity.

Note that [13, Lemma 2.4] also gives a lower bound on the angle of intersection, with the same order of growth.

Proof Consider the two systoles and the one-holed torus T they determine. Since $(g,n) \neq (1,1)$, the boundary component δ of T is a simple closed geodesic.

As α and β are systoles of S, they are also systoles of T. As such they satisfy the systole bound for T that depends on the length of δ, namely

$$\cosh\left(\tfrac{1}{6}\ell(\delta)\right) \geq \cosh(\ell) - \tfrac{1}{2}.$$

We first consider the case when $\ell \geq 2\operatorname{arccosh}\frac{3}{2}$. We have $\cosh(\frac{1}{2}\ell) - \frac{1}{2} \geq 1$ and the condition stated above is non-empty. Cut T along α (see Figure 9) and consider the shortest curve h connecting the two copies of α. By hyperbolic trigonometry, using $\cosh(\frac{1}{6}\ell(\delta)) \geq \cosh(\frac{1}{2}\ell) - \frac{1}{2}$, a direct computation provides

$$\cosh h \geq \frac{4\cosh\left(\frac{1}{2}\ell\right)^2 - \cosh\left(\frac{1}{2}\ell\right) - 1}{\cosh\left(\frac{1}{2}\ell\right)+1}.$$

Now consider one of the two right-angled triangles determined by arcs of α, β and h. We have

$$\frac{\sinh\left(\frac{1}{2}h\right)}{\sin\angle(\alpha,\beta)} = \sinh\left(\tfrac{1}{2}\ell\right),$$

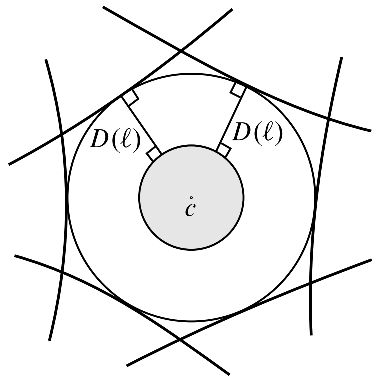


Figure 10: Geodesics around a horoball

which, together with the estimate on h, yields

$$\sin\angle(\alpha,\beta) \geq \frac{\sqrt{2\cosh\left(\frac{1}{2}\ell\right)+1}}{\cosh\left(\frac{1}{2}\ell\right)+1}.$$

If $\ell < 2\operatorname{arccosh}\frac{3}{2}$, we deduce the inequality $\sin\angle(\alpha,\beta) \geq 2/\sqrt{5}$ by arguing as above, but replacing the estimate $\cosh(\frac{1}{6}\ell(\delta)) \geq \cosh(\frac{1}{2}\ell) - \frac{1}{2}$ by $\ell(\delta) \geq \ell$. $\square$

Fix $(\alpha,\beta) \in B(c)^{(2)}$ and denote by $\mathcal{P}$ the pair of pants they determine with c. As they form a pair of pants with two boundary curves of the same length, there is an isometric involution φ of $\mathcal{P}$ that sends α to β (a rotation of angle π around the cusp). Note that, for any $(\gamma,\delta) \in B(c)^{(2)}$, the involution sends $\gamma \cap \mathcal{P}$ to $\delta \cap \mathcal{P}$ because of the symmetry of the pair of pants determined by γ, δ and p. If we quotient $\mathcal{P}$ by φ and we consider the image of $B(c)$, we get a set of geodesics at distance $D(\ell)$ from a horoball of area 1, all pairwise intersecting with angle at least θ_ℓ. This observation is crucial to show the following result.

Lemma 4.5 *If $(g,n) \neq (1,1)$, the number of elements in $B(c)$ is bounded above by*

$$m(\ell) := \frac{\cosh\left(\frac{1}{2}\ell\right)}{\sinh\left(\frac{1}{2}\ell\right)} \frac{2}{\sin\left(\frac{1}{2}\theta_\ell\right)}.$$

Proof The situation is as in Figure 10, which locally represents the elements of $B(c)$ under the quotient by φ. Note that every element in the quotient by φ represents two elements from $B(c)$.

The inner circle (which we'll refer to as the inner horocycle) represents the quotient horoball of area 1 and the external one is the horocycle at distance $D(\ell)$ from the horoball of area 1. By looking at the unique orthogonal geodesics between elements of $B(c)/\varphi$ and the inner horocycle, we can determine a cyclic ordering on the elements of $B(c)/\varphi$. Two neighboring geodesics with respect to this ordering determine a *subarc* on the inner horocycle as follows. We consider the orthogonal geodesic between them and the inner horocycle and take the subarc of the horocycle which forms a pentagon with the two geodesics and the orthogonal (see Figure 10). By a direct computation, using the lower bound on the angle of intersection, this subarc on the inner horocycle is of length at least

$$\frac{\sinh\left(\frac{1}{2}\ell\right)}{\cosh\left(\frac{1}{2}\ell\right)}\sin\left(\tfrac{1}{2}\theta_\ell\right).$$

These subarcs are all disjoint and are of the same number as the elements of $B(c)/\varphi$ (keep in mind that *any* two elements of $B(c)/\varphi$ intersect).

From this we deduce an upper bound on $|B(c)/\varphi|$:

$$\frac{1}{\left(\sinh\left(\frac{1}{2}\ell\right)/\cosh\left(\frac{1}{2}\ell\right)\right)\sin\left(\frac{1}{2}\theta_\ell\right)}.$$

Now $2|B(c)/\varphi| = |B(c)|$, which completes the proof. □

As a consequence, we obtain an upper bound on $|B(S)|$.

Proposition 4.6 *If $S \in \mathcal{M}_{g,n}$, $(g,n) \neq (1,1)$, has systole of length $\operatorname{sys}(S) = \ell$, then*

$$|B(S)| \leq nm(\ell).$$

Proof We have

$$B(S) = \bigcup_{c \text{ cusp}} B(c)$$

and, for every cusp c,

$$|B(c)| \leq m(\ell).$$

□

4.3 Bound on $|C(S)|$

By definition, elements of $C(S)$ are systoles that satisfy

- two curves in $C(S)$ intersect at most once, and
- two disjoint curves in $C(S)$ do not bound a cusp.

We follow a similar argument to one found in [13] to obtain an upper bound on $|C(S)|$. In particular, we will need a collar lemma for systoles.

Lemma 4.7 *Let* $\operatorname{sys}(S) = \ell$ *and consider* $\alpha, \beta \in C(S)$. *If* α *and* β *do not intersect, then they are at distance at least* $2r(\ell)$, *where*

$$r(\ell) = \operatorname{arcsinh} \frac{1}{2\sinh\left(\frac{1}{4}\ell\right)}.$$

Proof Fix a pair of pants with α and β as boundary and consider the third boundary component γ. Since α and β are in $C(S)$, they do not bound a cusp, so γ is a simple closed geodesic of length at least ℓ. The result follows by a standard trigonometric computation. □

As a consequence, if α and β in $C(S)$ pass through the same disk of radius $r(\ell)$ then they intersect.

Moreover, we have seen in Lemma 4.4 that there is a lower bound on the angle of intersection of systoles intersecting once. With this in hand we prove the following:

Lemma 4.8 *If* $(g,n) \neq (1,1)$, $\operatorname{sys}(S) = \ell$, *and* α *and* β *in* $C(S)$ *pass through a disk of center* p *and radius* $r(\ell)$, *then the distance between* p *and the point* q *of intersection between* α *and* β *satisfies*

$$d(p,q) \leq R(\ell),$$

where

$$\sinh(R(\ell)) = \begin{cases} 5/\left(8\sinh\left(\frac{1}{4}\ell\right)\right), & \ell < 2\operatorname{arccosh}\frac{3}{2}, \\ \left(\cosh\left(\frac{1}{2}\ell\right)+1\right)/\left(2\sinh\left(\frac{1}{4}\ell\right)\sqrt{2\cosh\left(\frac{1}{2}\ell\right)+1}\right), & \ell \geq 2\operatorname{arccosh}\frac{3}{2}. \end{cases}$$

Note that $R(\ell)$ is bounded for $\ell \geq 2\operatorname{arcsinh} 1$.

Proof The proof is analogous to the proof of [13, Lemma 2.6]. Fix $p_\alpha \in \alpha$ and $p_\beta \in \beta$ lying in $D_{r(\ell)}(p)$. We have two triangles of vertices p, p_α, q and p, p_β, q, and the sum of the two angles θ_α and θ_β at q is the angle of intersection $\angle(\alpha,\beta)$. Suppose $\theta_\alpha \geq \frac{1}{2}\angle(\alpha,\beta)$ and consider the angle η of the triangle p, p_α, q at p_α; see Figure 11. Then

$$\frac{\sin(\eta)}{\sinh(d(p,q))} = \frac{\sin(\theta_\alpha)}{\sinh(d(p,p_\alpha))}.$$

Using $\theta_\alpha \geq \frac{1}{2}\angle(\alpha,\beta)$, $d(p,p_\alpha) < r(\ell)$ and Lemma 4.4, we obtain the result. □

We are now in a position to obtain a bound on $|C(S)|$.

Proposition 4.9 *If* $S \in \mathcal{M}_{g,n}$, $g \neq 0$ *and* $(g,n) \neq (1,1)$, *has systole of length* $\operatorname{sys}(S) = \ell$, *then*

$$|C(S)| \leq 200\frac{e^{\ell/2}}{\ell}(2g-2+n).$$

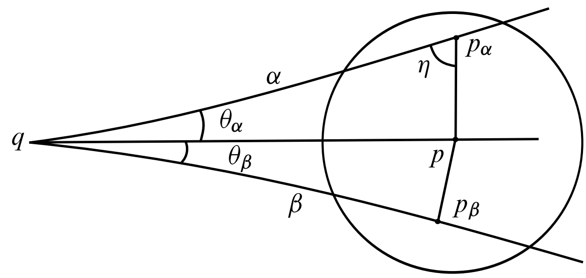


Figure 11: α and β passing though a disk of radius $r(\ell)$

Proof If $\ell \le 2 \operatorname{arcsinh} 1$, then all systoles are pairwise disjoint, so

$$|C(S)| \le \operatorname{Kiss}(S) \le 3g-3+n.$$

We now suppose that $\ell > 2 \operatorname{arcsinh} 1$. Consider $\widetilde{S} = S \setminus \bigcup_{c \text{ cusp}} D_{w(\ell)}(c)$, where

$$w(\ell) = \operatorname{arcsinh} \frac{1}{\sinh\left(\frac{1}{2}\ell\right)}$$

is the width of a collar around a systole. By the collar lemma, each curve of $C(S)$ is contained in $\widetilde{S}$. We cover $\widetilde{S}$ with disks of radius $r(\ell)$. Then the cardinality of $C(S)$ is bounded above by

$$\frac{F(S)G(S)}{H(S)},$$

where

$$F(S) = \#\{\text{balls of radius } r(\ell) \text{ needed to cover } \widetilde{S}\},$$
$$G(S) = \#\{\text{curves in } C(S) \text{ crossing a ball of radius } r(\ell)\},$$
$$H(S) = \#\{\text{number of balls of radius } r(\ell) \text{ a curve in } C(S) \text{ must cross}\}.$$

To bound $|C(S)|$, we need to give upper bounds for $F(S)$ and $G(S)$ and a lower bound for $H(S)$.

Upper bound for $F(S)$ We have

$$\begin{aligned} F(S) &\le \max \#\left\{\text{embedded balls of radius } \tfrac{1}{2}r(\ell) \text{ which are pairwise disjoint}\right\} \\ &\le \frac{\operatorname{area}(\widetilde{S})}{\operatorname{area}\left(\text{ball of radius } \frac{1}{2}r(\ell)\right)} \le \frac{\operatorname{area}(S)}{2\pi\left(\cosh\left(\frac{1}{2}r(\ell)\right)-1\right)} \le 8(2g-2+n)e^{\ell/2}. \end{aligned}$$

Upper bound for $G(S)$ We proceed as in the proof of [13, Theorem 2.9], by reasoning in the universal cover and estimating how many geodesics, pairwise intersecting at an angle of at least θ_ℓ, can intersect a disk of radius $r(\ell)$. We obtain

$$G(S) \le \frac{\pi}{2}\frac{\sinh(R(\ell) + \operatorname{arcsinh} 1)}{\operatorname{arcsinh}\sin(\theta_\ell)} \le \frac{5\pi}{2\operatorname{arcsinh}\sin(\theta_\ell)}.$$

Lower bound for $H(S)$ To cover a curve of length ℓ with disks of radius $r(\ell)$ we need at least $\ell/(2r(\ell))$. So

$$H(S) \ge \frac{\ell}{2\operatorname{arcsinh}\big(1/\big(2\sinh\big(\frac{1}{4}\ell\big)\big)\big)} \ge \ell\sinh\big(\tfrac{1}{4}\ell\big).$$

By putting the three bounds together and considering that $\sinh(\frac{1}{4}\ell)\operatorname{arcsinh}\sin(\theta_\ell)$ is bounded below by $\frac{1}{3}$ for $\ell > 2\operatorname{arcsinh} 1$ we obtain the claimed result. □

4.4 Proof of main results

Using Propositions 4.2, 4.6 and 4.9, we get an upper bound for the kissing number of a surface in terms of its signature and its systole length.

Theorem 4.10 *If $S \in \mathcal{M}_{g,n}$ with $g \ge 1$, $(g,n) \ne (1,1)$, has systole of length $\operatorname{sys}(S) = \ell$, then*

$$\operatorname{Kiss}(S) \le 20n\cosh\big(\tfrac{1}{4}\ell\big) + 200\frac{e^{\ell/2}}{\ell}(2g - 2 + n).$$

As a consequence, we can get a bound on the kissing number which is independent of the systole length.

Theorem 4.11 *There exists a universal constant C (which we can take to be 2×10^4) such that, for any $S \in \mathcal{M}_{g,n}$, $g \ge 1$, its kissing number satisfies*

$$\operatorname{Kiss}(S) \le C(g+n)\frac{g}{\log(g+1)}.$$

Proof This follows from the bounds in Theorem 4.10 and bounds on systole lengths. Precisely, we insert the Schmutz Schaller bound (Theorem 2.4) into the term $\cosh(\frac{1}{4}\ell)$ and we use Theorem 2.3 for the $e^{\ell/2}/\ell$ term. For $(g,n) = (1,1)$, we recall the well-known fact that $\operatorname{Kiss}(S) \le 3$ (there can be at most 3 distinct curves that pairwise intersect at most once on a one-holed torus). □

Remark 4.12 Przytycki [16] obtained an upper bound for the number of simple closed curves pairwise intersecting at most once. Using this and our bound on $|A(S)|$ (Proposition 4.2), one can obtain an upper bound for the kissing number which is cubic in the Euler characteristic. Our upper bound, on the other hand, is subquadratic in $|\chi(S)|$, like the one for closed surfaces in [13].

The upper bound of Theorem 4.11 is linear in the number of cusps if we fix the genus. For punctured spheres we can obtain a more meaningful bound.

Theorem 4.13 *For every $S \in \mathcal{M}_{0,n}$, the number of systoles satisfies*

$$\mathrm{Kiss}(S) \le \tfrac{7}{2}n - 5.$$

Proof By Proposition 4.2 and Schmutz Schaller's upper bound for the systole, we have

$$|A(S)| \le \frac{n}{2}\left\lfloor \frac{2(3n-6)}{n} \right\rfloor = \frac{n}{2}\left\lfloor 6 - \frac{12}{n} \right\rfloor \le \frac{5}{2}n.$$

Moreover, systoles are separating, so can only intersect an even number of times. This implies that systoles in $\mathfrak{S}(S) \setminus A(S)$ are pairwise disjoint and hence part of a pants decomposition. Note that any pants decomposition of a sphere contains at least two curves bounding two cusps; indeed, the dual graph to the pants decomposition is a tree, so it has at least two leaves, which correspond to curves bounding two cusps. This implies that

$$|\mathfrak{S}(S) \setminus A(S)| \le \#\,\text{curves in a pants decomposition} - 2 = n - 5. \qquad \square$$

By using short pants decompositions where every curve is of equal length, it is easy to obtain a family of punctured spheres with a number of systoles that grows linearly in the number of cusps. Matching the $\frac{7}{2}n$ upper bound from this theorem seems much more challenging.

References

[1] **C Adams**, *Maximal cusps, collars, and systoles in hyperbolic surfaces*, Indiana Univ. Math. J. 47 (1998) 419–437 MR1647904

[2] **F Balacheff**, **E Makover**, **H Parlier**, *Systole growth for finite area hyperbolic surfaces*, Ann. Fac. Sci. Toulouse Math. 23 (2014) 175–180 MR3204736

[3] **C Bavard**, *Systole et invariant d'Hermite*, J. Reine Angew. Math. 482 (1997) 93–120 MR1427658

[4] **C Bavard**, *Anneaux extrémaux dans les surfaces de Riemann*, Manuscripta Math. 117 (2005) 265–271 MR2154250

[5] **R Brooks**, *Platonic surfaces*, Comment. Math. Helv. 74 (1999) 156–170 MR1677565

[6] **R Brooks**, **E Makover**, *Random construction of Riemann surfaces*, J. Differential Geom. 68 (2004) 121–157 MR2152911

[7] **P Buser**, *Geometry and spectra of compact Riemann surfaces*, Progress in Mathematics 106, Birkhäuser, Boston (1992) MR1183224

[8] **P Buser**, **P Sarnak**, *On the period matrix of a Riemann surface of large genus*, Invent. Math. 117 (1994) 27–56 MR1269424

[9] **B Farb**, **D Margalit**, *A primer on mapping class groups*, Princeton Mathematical Series 49, Princeton Univ. Press (2012) MR2850125

[10] **T Gauglhofer**, **K-D Semmler**, *Trace coordinates of Teichmüller space of Riemann surfaces of signature* $(0,4)$, Conform. Geom. Dyn. 9 (2005) 46–75 MR2133805

[11] **M G Katz**, **M Schaps**, **U Vishne**, *Logarithmic growth of systole of arithmetic Riemann surfaces along congruence subgroups*, J. Differential Geom. 76 (2007) 399–422 MR2331526

[12] **S Makisumi**, *A note on Riemann surfaces of large systole*, J. Ramanujan Math. Soc. 28 (2013) 359–377 MR3113388

[13] **H Parlier**, *Kissing numbers for surfaces*, J. Topol. 6 (2013) 777–791 MR3100890

[14] **H Parlier**, *Simple closed geodesics and the study of Teichmüller spaces*, from: "Handbook of Teichmüller theory, IV", (A Papadopoulos, editor), IRMA Lect. Math. Theor. Phys. 19, Eur. Math. Soc., Zürich (2014) 113–134 MR3289700

[15] **B Petri**, *Random regular graphs and the systole of a random surface*, preprint (2013) `arXiv:1311.5140v1`

[16] **P Przytycki**, *Arcs intersecting at most once*, Geom. Funct. Anal. 25 (2015) 658–670 MR3334237

[17] **P Schmutz**, *Congruence subgroups and maximal Riemann surfaces*, J. Geom. Anal. 4 (1994) 207–218 MR1277506

[18] **P Schmutz**, *Arithmetic groups and the length spectrum of Riemann surfaces*, Duke Math. J. 84 (1996) 199–215 MR1394753

[19] **P Schmutz Schaller**, *Extremal Riemann surfaces with a large number of systoles*, from: "Extremal Riemann surfaces", (J R Quine, P Sarnak, editors), Contemp. Math. 201, Amer. Math. Soc. (1997) 9–19 MR1429190

Mathematics Institute, Zeeman Building
University of Warwick, Coventry CV4 7AL, UK

Department of Mathematics, University of Fribourg
Ch. du Musée 23, CH-1700 Fribourg, Switzerland

`federica.fanoni@gmail.com`, `hugo.parlier@unifr.ch`

Received: 1 September 2014 Revised: 3 March 2015

msp

Algebraic & Geometric Topology 15 (2015) 3435–3465

Combinatorial cohomology of the space of long knots

ARNAUD MORTIER

The motivation of this work is to define cohomology classes in the space of knots that are both easy to find and to evaluate, by reducing the problem to simple linear algebra. We achieve this goal by defining a combinatorial graded cochain complex such that the elements of an explicit submodule in the cohomology define algebraic intersections with some "geometrically simple" strata in the space of knots. Such strata are endowed with explicit co-orientations that are canonical in some sense. The combinatorial tools involved are natural generalisations (degeneracies) of usual methods using arrow diagrams.

57M25; 55N33, 57N80

The paper is organised as follows.

In Section 1 we build a prototypical cochain complex which contains all the essential combinatorics while using the most simple input, namely a finite collection of finite subsets of $\mathbb{R}$ (a *coloured leaf diagram*). The point of this preliminary is not only theoretical, it serves to point out clearly that this part of our construction does not depend on the material introduced later.

In Section 2 we show that the incidence signs of the previous cochain complex are of a topological nature, as they are an essential ingredient in the computation of the boundary of the meridian discs of some "geometrically simple" strata in the space of knots, provided that these discs are correctly oriented. This property canonically defines a co-orientation of simple strata.

Simple strata are represented by means of degenerated Gauss diagrams, ie whose arrows are allowed to meet on the base circle. In Section 3, similarly to Polyak and Viro's formulas for finite-type invariants, we define cochains by counting subconfigurations in those diagrams, with weights given by products of writhes. A little twist appears here: we do not count the signs of arrows that participate in singularities; these signs contribute implicitly, via the definition of the canonical co-orientation.

At the end of Section 3 we define the maps of the main cochain complex, as a slightly twisted version of those of Section 1, and construct a Stokes formula relating it with the boundary maps from Section 2, which model the meridians of simple strata. The announced result follows.

Published: 12 January 2016 DOI: 10.2140/agt.2015.15.3435

Lastly, Section 4 is a review of examples, including new formulas for the low-degree Vassiliev invariants obtained by integrating 1– and 2–cocycles over some canonical 1– and 2–chains. In particular we give a method for integrating our 1–cocycle formulas into knot invariants without any computations, over the two main canonical cycles in the space of knots; namely the Gramain loop, and the Fox–Hatcher loop.

Acknowledgements

I sincerely thank the referee for a very careful reading and many helpful suggestions, substantially improving clarity. I also express my full gratitude to Seiichi Kamada for his hospitality and the opportunity he gave me to pursue research for one year in excellent conditions in Osaka City University Advanced Mathematical Institute. I am grateful to Allen Hatcher, Thomas Fiedler, Victoria Lebed and Ryan Budney for helpful discussions and comments, and to Akio Kawauchi for inspiring conversations in OCAMI and in Korea.

1 Cohomology of coloured leaf diagrams in $\mathbb{R}$

1.1 Polygons

A *polygon* is a finite subset of the oriented based circle $\mathbb{S}^1 = \mathbb{R} \cup \{\infty\}$. We make no distinction between a polygon and the corresponding singular 0–chain in $C_0(\mathbb{S}^1, \mathbb{Z}_2)$. It is said to be *even* or *odd* according to the parity of its cardinality; in other words, odd polygons are those representing the non-trivial homology class in $H_0(\mathbb{S}^1, \mathbb{Z}_2)$.

Let P and P' be two disjoint even polygons. Then they have a well-defined (*mod* 2) *linking number*, denoted by $\mathrm{lk}(P, P') \in \mathbb{Z}_2 = \{\pm 1\}$, which is the algebraic intersection between P' and any 1–chain in $C_1(\mathbb{S}^1, \mathbb{Z}_2)$ whose boundary is P. The map lk is symmetric, and bilinear in the sense that if P and P' are disjoint, as well as P and P'', then:

$$\mathrm{lk}(P, P' + P'') = \mathrm{lk}(P, P') \cdot \mathrm{lk}(P, P'').$$

If P' is odd and P has two elements, again with $P \cap P' = \varnothing$, we extend the notation by setting

$$\mathrm{lk}(P, P') = \mathrm{lk}(P', P) \overset{\text{def}}{=} (-1)^{\sharp[\min(P),\max(P)]\cap P'},$$

where we agree that the point ∞ is *greater* than any real number. Note that the same formula holds when P' is even. The map lk can then be extended by symmetry and bilinearity to any pair of disjoint polygons at least one of which is even.

We define a partial order on the set of polygons by setting

$$P < P' \iff \inf(P) < \inf(P').$$

1.2 Coloured leaf diagrams

A *(coloured) leaf diagram* in $\mathbb{R} = \mathbb{S}^1 \setminus \{\infty\}$ is a finite collection of pairwise disjoint polygons, none of which contains ∞. The elements of the polygons are called *leaves* of the diagram and two leaves from the same polygon are said to have the same *colour*. The terminology is inspired by the fact that such diagrams are meant to later be completed into *tree diagrams* by connecting all leaves of a same colour by an abstract tree. We define two $\mathbb{Z}$–valued complexities associated with a leaf diagram L:

- The *Gauss degree* $\deg(L)$, which is the total number of leaves minus the number of colours (polygons) in L.
- The *codimension* $\iota(L)$, or *cohomological degree*, which is the total number of leaves minus *twice* the number of colours in L.

The term "Gauss degree" comes from the theory of chord diagrams, where it denotes the number of chords. For instance, a leaf diagram with d polygons, all of which have cardinality 2, has Gauss degree d and codimension 0.

Leaf diagrams are regarded up to orientation-preserving homeomorphisms of the real line $\mathbb{S}^1 \setminus \{\infty\}$. The $\mathbb{Z}$–module freely generated by equivalence classes of leaf diagrams of degree d and codimension i is denoted by $\mathcal{L}^i_d$. Note that $\mathcal{L}^i_d$ is always finitely generated and is trivial whenever i is greater than $d-1$.

Remark 1.1 Special attention should be paid to polygons with only one leaf. Such a polygon contributes -1 to the codimension and has no effect on the Gauss degree. They are actually the only reason why the cohomological degree is not bounded and $\mathbb{N}$–valued. In our main application for this theory, they are naturally excluded, and the spaces of diagrams with fixed Gauss degree are finitely generated. However, it is harmless to allow them in the prototypical cochain complex and there may be a theoretical interest in studying their meaning and the relations between the main and "reduced" cohomology theories.

1.3 The ε signs and prototypical complex

Let L be a leaf diagram. An *edge* of L is a closed connected part of the circle that lies between two neighbouring leaves of L; in particular, an edge cannot contain a leaf in its interior and it cannot contain ∞. An edge is called *admissible* if its two boundary points have different colours. From such an edge, we construct a new leaf diagram L_e in the following way. The polygons of L_e are the polygons of L, except for the two that have a leaf at the boundary of e: those two are merged into a single polygon in L_e and one of the two boundary points of e is removed from it (which one exactly has no effect on the resulting diagram up to positive homeomorphism of $\mathbb{R}$).

One easily checks the relations

$$\deg(L_e) = \deg(L) \quad \text{and} \quad \iota(L_e) = \iota(L) + 1.$$

Consider the linear maps $\mathcal{L}_d^{i-1} \to \mathcal{L}_d^i$ defined on each generator by the formula:

$$L \mapsto \sum_{e \text{ admissible}} L_e.$$

It is easy to see that using $\mathbb{Z}_2$ coefficients, these maps turn the collection of spaces $\mathcal{L}_d^i$ into a graded cochain complex. Our goal is to define signs to lift this complex over $\mathbb{Z}$.

The global sign Let P be an odd polygon in a leaf diagram L. We define the *odd index* of P as the parity of the number of odd polygons in L that are greater than P. Using the convention that a boolean expression has value -1 when it is true and 1 otherwise, this can be written as

$$\operatorname{Odd}(P, L) \stackrel{\text{def}}{=} \prod_{P' \text{ odd}} (P < P').$$

We extend this definition to all polygons by setting $\operatorname{Odd}(P) = 1$ whenever P is even.

Let e be an admissible edge in L, bounded by the leaves v and w lying in the polygons P_v and P_w, respectively. Also, denote by P_{vw} the polygon of L_e that results from the merging of P_v and P_w.

We define the *global sign* associated with the edge e in L by

$$\sigma_{\text{glo}}(e, L) \stackrel{\text{def}}{=} \operatorname{Odd}(P_v, L) \cdot \operatorname{Odd}(P_w, L) \cdot \operatorname{Odd}(P_{vw}, L_e).$$

This will be the only contribution to the signs in the coboundary maps that depends on polygons located far from e.

Remark 1.2 When P_v and P_w are both odd, both booleans $(P_v < P_w)$ and $(P_w < P_v)$ appear in σ_{glo}, which results in a minus sign.

The local sign From now on, for the sake of lightness, we will not mention that every sign depends on L, since other diagrams like L_e will not contribute any more.

If x is a leaf in L, we denote by P_x the polygon that contains it. Define the *evenisation* of P_x with respect to x as

$$P_x^{(x)} = \begin{cases} P_x + x & \text{if } P_x \text{ is odd,} \\ P_x & \text{if } P_x \text{ is even.} \end{cases}$$

As a set, $P_x + x$ corresponds to $P_x \setminus \{x\}$, so that the polygon $P_x^{(x)}$ is always even.

As previously, let e be an admissible edge in L, bounded by the leaves v and w. Recall the convention that a boolean expression takes value -1 when it is true and 1 otherwise. We define:

Linking number of e $\quad \mathrm{lk}(e) = \mathrm{lk}(P_v^{(v)}, P_w^{(w)})$.

Even index of P_v with respect to e

$$\mathrm{E}(P_v, e) = \begin{cases} \mathrm{lk}(v + \infty, P_w) & \text{if } P_v \text{ is even,} \\ 1 & \text{otherwise.} \end{cases}$$

Odd consistency of e

$$\psi(e) = \begin{cases} (v < w)(P_v < P_w) & \text{if both } P_v \text{ and } P_w \text{ are odd,} \\ 1 & \text{otherwise.} \end{cases}$$

The *local sign* associated with the edge e is

$$\sigma_{\mathrm{loc}}(e, L) = \mathrm{lk}(e)\psi(e)\,\mathrm{E}(P_v, e)\,\mathrm{E}(P_w, e).$$

Finally, we set

$$\varepsilon_L(e) = \sigma_{\mathrm{loc}}(e, L)\sigma_{\mathrm{glo}}(e, L),$$
$$\delta_d^i(L) = \sum_{e \text{ admissible}} \varepsilon_L(e) \cdot L_e,$$

and extend this formula into a linear map $\delta_d^i \colon \mathcal{L}_d^{i-1} \to \mathcal{L}_d^i$.

Theorem 1.3 *For each $d \geq 1$, the collection of spaces $\mathcal{L}_d^*$ and maps δ_d^* forms a cochain complex of $\mathbb{Z}$–modules.*

Proof Let e and e' be two edges in a leaf diagram L such that e is admissible. Then e' is admissible in L_e if and only if e' is admissible in L and e is admissible in $L_{e'}$. We call such a couple *bi-admissible*. To prove the theorem, it is enough to show that for any bi-admissible couple, the contribution of e and e' in the computation of $\delta^2 L$ is 0 or, in other words, that the product $\varepsilon_L(e)\varepsilon_L(e')\varepsilon_{L_{e'}}(e)\varepsilon_{L_e}(e')$ is always equal to -1.

Parities of the polygons				Total contribution of σ_{glo} to $\varepsilon_L(e)\varepsilon_{L_{e'}}(e)$
P_v	P_w	$P_{v'}$	$P_{w'}$	
0	0	0	0	$(P_{vw} < P_{v'w'})$
0	0	0	1	$(P_{vw} < P_{w'})$
0	0	1	1	$(P_{vw} < P_{v'})(P_{vw} < P_{w'})(P_{vw} < P_{v'w'})$
0	1	0	1	$(P_w < P_{w'})$
0	1	1	1	$(P_w < P_{v'})(P_w < P_{w'})(P_w < P_{v'w'})$
1	1	1	1	$(P_v < P_{v'})(P_w < P_{v'})(P_{vw} < P_{v'})(P_v < P_{w'})(P_w < P_{w'})$ $\times (P_{vw} < P_{w'})(P_v < P_{v'w'})(P_w < P_{v'w'})(P_{vw} < P_{v'w'})$

Table 1: Computation of δ^2 when $\sharp\{P_v, P_w, P_{v'}, P_{w'}\} = 4$. By symmetry, there are only six cases to consider. Note that the minus sign due to P_v and P_w being odd in the last line (Remark 1.2) appears twice and cancels out.

If e and e' are bounded respectively by v, w and v', w', then (e, e') is bi-admissible if and only if e and e' are admissible and the leaves v, w, v' and w' represent at least 3 different colours. We split the proof into two parts, accordingly.

First, assume that all leaves have pairwise different colours. In this case, every contribution from σ_{loc} appears twice and cancels out. So do the contributions of σ_{glo} involving other polygons than those neighbouring e and e'. The remaining contributions of σ_{glo} are summarised in Table 1; 0 stands for "even", 1 for "odd". We show only the contribution to $\varepsilon_L(e)\varepsilon_{L_{e'}}(e)$: the contribution to $\varepsilon_L(e')\varepsilon_{L_e}(e')$ contains exactly the opposite boolean expressions. So the point is that in each row, there is an odd number of booleans.

We now assume that v, w, v' and w' represent 3 colours, and without loss of generality that w and w' share the same one. We need not discuss the special case when w is actually equal to w', since the following computations apply equally well in that case. Table 2 details the contribution of each factor to the product $\varepsilon_L(e)\varepsilon_L(e')\varepsilon_{L_{e'}}(e)\varepsilon_{L_e}(e')$. The proof that the product of all contributions is always -1 is straightforward, using the bilinearity of lk and the formula

$$\mathrm{lk}(a + \infty, b) = (a < b) \quad \text{for all } a, b \in \mathbb{R}. \qquad \square$$

2 Simple singularities in the space of knot diagrams

2.1 Germs and the associated strata

By the *space of long knots* $\mathcal{K}$ we mean the (arbitrarily high, but finite)-dimensional affine approximation of the space of all smooth maps $\mathbb{R} \to \mathbb{R}^3$ with prescribed asymptotical

Parities			Contributions	
P_v	P_w	$P_{v'}$	lk	E
0	0	0	$\mathrm{lk}(P_v+P_{v'}, w+w')$	$\mathrm{lk}(P_v+P_{v'}, w+w')\,\mathrm{lk}(v+\infty, w')\,\mathrm{lk}(v'+\infty, w)$
0	0	1	$\mathrm{lk}(P_{v'}+v', w+w')$	$\mathrm{lk}(P_{v'}, w+w')\,\mathrm{lk}(w+\infty, w')$
1	0	1	1	1
0	1	0	$\mathrm{lk}(P_v+P_{v'}, w+w')$	$\mathrm{lk}(P_v+P_{v'}, w+w')\,\mathrm{lk}(v+\infty, w')\,\mathrm{lk}(v'+\infty, w)$
0	1	1	$\mathrm{lk}(P_{v'}+v', w+w')$	$\mathrm{lk}(P_{v'}, w+w')\,\mathrm{lk}(v+\infty, v')$
1	1	1	1	1

Parities			Contributions	
P_v	P_w	$P_{v'}$	ψ	σ_{glo}
0	0	0	1	1
0	0	1	$(v' < w')(P_{v'} < P_{vw})$	$(P_{v'} < P_{vw})$
1	0	1	1	-1
0	1	0	1	1
0	1	1	$(v' < w')(P_{v'} < P_w)$	$(P_{v'} < P_w)$
1	1	1	$\pm(P_v < P_w)(P_{v'} < P_w)(P_v < P_{v'w})(P_{v'} < P_{vw})$	

Table 2: Contribution of each factor after obvious simplifications, in the case $P_w = P_{w'}$. By symmetry, there are only six cases to consider. In the last line, ψ and σ_{glo} have the same contribution up to sign, which is $+$ for ψ and $-$ for σ_{glo}.

behaviour, as defined by Vassiliev [23]. The *discriminant* Σ is the subset of all maps in $\mathcal{K}$ that are not embeddings. A projection $p\colon \mathbb{R}^3 \to \mathbb{R}^2$ endows $\mathcal{K} \setminus \Sigma$ with a stratification, whose strata are defined by certain semi-algebraic varieties in multijet spaces (see Example 2.4, and see David [5], Wall [25], Fiedler and Kurlin [8] and references therein for an introduction to stratified spaces and the simplest examples used in knot theory). Those strata can be represented by Gauss diagrams with additional information of geometrical nature, ie involving derivatives (see Vassiliev [24]). We will call such a stratum *simple* if the only geometric data are the writhes of the crossings, and *geometric* otherwise.

Definition 2.1 An *abstract germ* is the datum of a finite number of complete oriented graphs, together with an embedding of the disjoint union of their vertices into $\mathbb{R} = \mathbb{S}^1 \setminus \{\infty\}$, such that

(1) each graph has at least two vertices,

(2) no graph has oriented cycles,

(3) each edge of each graph is decorated with a sign $+$ or $-$.

An abstract germ γ has an underlying leaf diagram $L(\gamma)$, from which it inherits the terminology of polygons, leaves, colours, edges, as well as the Gauss and cohomological degrees deg and ι. The edges of the graphs in γ are called (*signed*) *arrows*, while the word "edge" keeps the same meaning as in Section 1.3.

Condition (2) above implies that a germ induces a total order on each of its polygons, and a partial order on the set of all of its leaves, denoted by $<_\gamma$. A knot is said to *respect* γ, and is called a γ*–knot*, if it maps any two leaves with the same colour to a classical crossing with over/under datum given by the order $<_\gamma$, and writhe given by the sign of the arrow between those leaves. These conditions may be inconsistent, so that no knot can respect γ; otherwise, γ is called a *topological germ*, or more simply a *germ*. In that case the diagram of a generic γ–knot is uniquely determined near each imposed crossing up to local diagram isotopy. Out of the $2^{\binom{n}{2}}$ ways to put signs on a complete graph (consistently oriented) with n leaves, exactly $2^{n-1}(n-1)!$ are topological.

The $\mathbb{Z}$–module freely generated by (topological) germs with cohomological degree i is denoted by $\mathcal{G}_i$, because we will essentially think of meridians (that is, i–discs transverse to the stratification) for which i is the dimension.

Remark 2.2 A γ–knot may very well have crossings besides those required by γ. However, in a *generic* γ–knot these additional crossings cannot be multiple.

If the leaves of γ are fixed, the set of all γ–knots in $\mathcal{K}$ is an affine subspace of codimension $2\deg(\gamma)$, because there are $(\sharp P - 1)$ affine equations for each polygon P (which are independent if $\dim \mathcal{K}$ is large enough), and because the writhe conditions are open, hence 0–codimensional. If the leaves are set free, ie the germ is regarded up to positive homeomorphism of the real line, then the codimension drops to $2\deg(\gamma) - (\text{number of leaves})$, which is equal to $\iota(\gamma)$.

Definition 2.3 The $\iota(\gamma)$–codimensional subspace of all knots in $\mathcal{K}$ that respect γ up to positive homeomorphism of the real line is denoted by $\mathcal{K}_\gamma$ and called the *simple stratum* associated with γ.

In Section 2.3, we will show that a germ γ defines canonically a co-orientation of $\mathcal{K}_\gamma$ (that is, an orientation of its meridian disc). That is the reason for calling it a *germ*.

Example 2.4 The strata of codimension 1 are described by the classical Reidemeister moves. R-I and R-II strata are geometric, and R-III is simple. Indeed, choose a basis

(e_1, e_2, e_3) of $\mathbb{R}^3$ such that e_3 is the axis of the projection p. This splits a knot parametrisation $f\colon \mathbb{R} \to \mathbb{R}^3$ into three coordinate functions $f_{1,2,3}$. Reidemeister strata are then defined by writhe data together with the conditions (for example):

R-I $$\exists x \quad f_1'(x) = f_2'(x) = 0.$$

R-II $$\exists x < y \quad \begin{array}{l} f_1(x) = f_1(y) \\ f_2(x) = f_2(y) \\ f_3(x) < f_3(y) \end{array} \quad \text{and} \quad \det\begin{pmatrix} f_1'(x) & f_1'(y) \\ f_2'(x) & f_2'(y) \end{pmatrix} = 0.$$

R-III $$\exists y < x < z \quad \begin{array}{l} f_1(x) = f_1(y) = f_1(z), \\ f_2(x) = f_2(y) = f_2(z), \\ f_3(x) < f_3(y) < f_3(z). \end{array}$$

Note that the conditions do not depend on the choice of a basis for the projection plane, (e_1, e_2); this is a general observation, the stratification depends only on p. Also, this stratification should not be confused with that of Σ used by Vassiliev [23] to define finite-type cohomology classes. That one will not be used in the present paper.

Remark 2.5 When a germ is regarded up to homeomorphism, it may happen that a knot respects it in several different ways. Note however that a generic γ–knot cannot have more singularities than imposed by γ, so that the only source of multiplicity lies in two-leaved polygons, which give 0–codimensional constraints. Rather than the strata $\mathcal{K}_\gamma$, one may consider *simplicial chains*, whose local weight near a given γ–knot is equal to the number of ways that knot respects γ; this is the implicit choice in Vassiliev's calculus [24]. Here, algebraic intersection with such chains will be modelled by means of the map $\mathrm{T} \circ \mathrm{I}$ which is defined in Section 3.2.

2.2 Boundary of simple strata

The boundary of a stratum $\mathcal{K}_\gamma$ is defined by the generic ways for its constraints to degenerate. There are essentially six basic ways, from which all others can be built. They can be interpreted by thinking of a generic γ–knot as a knot diagram some of whose crossings, *including all multiple crossings*, are coloured in red (as in Figure 1).

Type Σ Two leaves of γ that are consecutive with respect to the order $<_\gamma$ tend to be mapped to the same point in $\mathbb{R}^3$. The corresponding piece of boundary lies in Σ, so it is harmless for our purposes (namely understanding the cohomology of $\mathcal{K} \setminus \Sigma$, which is the relative homology of $(\mathcal{K}, \Sigma)$).

Type 1 One edge of γ whose boundary points have the same colour collapses into a point x, accompanied by the condition $f_1'(x) = f_2'(x) = 0$.

Type 2-1 Two branches of a red crossing tend to have the same direction in the knot diagram; from the point of view of γ, it results in a writhe not being well-defined any more, and replaced with either a condition of positive or negative collinearity of derivatives.

Type 2-2 Two edges of γ that bound a bigon in the knot diagram collapse simultaneously. This produces the same geometric condition as in Type 2-1.

Type 3-2 One edge whose boundary points have distinct colours collapses.

Type 3-3 Three edges that bound a triangle in the knot diagram collapse simultaneously.

Types 1 to 3-3 correspond to generalised Reidemeister moves, in that the crossings are allowed to be multiple. They are sorted according to how many red crossings they involve.

Besides these basic types, it can happen that types 2-2, 3-2 and 3-3 are accompanied by the simultaneous collapsing of an arbitrarily large number of triangles of type 3-3. Indeed, in all of these cases, one can see on the knot diagram that a number of crossings are locally present although they may not be imposed by the germ (red). Now these crossings may also actually be present in the germ, in which case they can either be regarded as far (which yields a basic type as above) or close, in which case they participate in the collapsing. Then, these extra crossings may themselves be multiple crossings from the beginning, and this phenomenon may repeat itself.

We are now ready to define precisely which kind of degeneracies will be of interest in this paper.

Definition 2.6 We call a degeneracy $\gamma \rightsquigarrow \gamma'$ *type* 0 if it is of basic type 3-2 together with finitely many *non-multiple* extra crossings as above, ie at most two polygons with more than two leaves can be involved in the collapsing. *It may involve type* 3-3 *degeneracies (it does as soon as there are extra crossings) but it is not regarded as such.* If the two polygons of the underlying type 3-2 degeneracy have m and n leaves, respectively, then there are at most $(m-1)(n-1)$ extra arrows. Degeneracies of basic type 3-2 with extra *multiple* crossings are naturally considered to fall down into type 3-3. See examples in Figures 1 and 2.

Reidemeister farness We now define a class of germs that will allow us to avoid bad geometric strata as well as the type 3-3 frenzy described earlier.

Definition 2.7 Let γ be a germ. We say that two leaves in γ are *neighbours* if they are the two boundary points of an edge. Then γ is called:

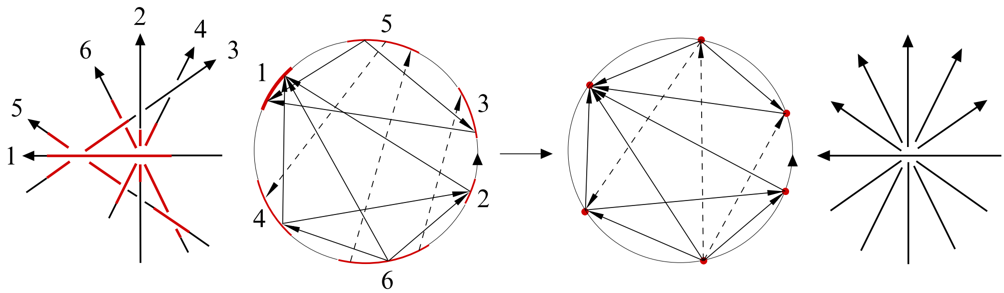


Figure 1: A type 0 degeneracy with three out of possibly six extra arrows involved, represented here with dashed lines for the sake of clarity. The result, a germ with missing arrows, will be called a subgerm in Section 3. It is not a new kind of stratum, rather it should be thought of as the formal sum of all ways to complete it into an actual germ. The branches are numbered from the highest to the lowest. In the underlying type 3-2 degeneracy, only the edge labelled "1" collapses.

(1) *R–I–close* if it contains an arrow (v, w) such that v and w are neighbours.

(2) *R–II–close* if it contains four distinct leaves v, w, x and y such that
- v and w are neighbours, and so are x and y;
- v and w have distinct colours;
- $v <_\gamma x$ and $w <_\gamma y$.

(3) *R–III–close* if it contains six distinct leaves v, w, x, y, z and t such that
- $\{v, w\}$, $\{x, y\}$ and $\{z, t\}$ are pairs of neighbours;
- v, w and y have pairwise distinct colours;
- $v <_\gamma x$, $y <_\gamma z$ and $w <_\gamma t$.

We define *R–farness* of germs, and therefore of simple strata, as the negation of all of these properties. In other words, γ is R–far if no generic γ–knot can be subject to a generalised Reidemeister move involving only red crossings, that is, basic types 1, 2-2 and 3-3 cannot occur. For instance, the germ in Figure 1 is three times R-III–close.

2.3 Meridian systems and the ∂_i map

Roughly speaking, our goal is to define cohomology classes in the space of knots as intersection forms with R–far simple strata. This requires us to understand in which meridian spheres these strata occur. By the previous discussion we mainly need to consider the meridians of simple strata. The geometric strata resulting from 2-1 degeneracies will later prove to be completely harmless (see Lemma 3.10).

Let f be a knot respecting an i–germ γ (that is, a germ with codimension $\iota(\gamma) = i$), and $\mathbb{D}^i_f$ a piecewise linear (PL) i–disc about f in $\mathcal{K}$, transverse to the stratification.

Then the boundary of $\mathbb{D}^i_f$ intersects finitely many $(i-1)$–strata, at points $f_1, \dots, f_p$, and can be covered with PL discs $\mathbb{D}^{i-1}_{f_k}$ with pairwise disjoint interiors. Since γ is simple, every meridian stratum γ_k is necessarily simple, and the degeneracy $\gamma_k \rightsquigarrow \gamma$ is either of type 0 (see Definition 2.6) or of type 3-3.

Definition 2.8 (Reduced meridian system) The cellular boundary map (over $\mathbb{Z}$) associated with the above decomposition of $\partial\mathbb{D}^i_f$ depends only on γ. It is called the *meridian system* of γ. The *reduced meridian system* of γ is the induced map with target restricted to those C_{γ_k} such that $\gamma_k \rightsquigarrow \gamma$ is of type 0. We denote it by

$$\partial_\gamma \colon C_\gamma \to \bigoplus_0 C_{\gamma_k}.$$

When $i = 0$, $\mathbb{D}^0_f$ consists of a single point and has a canonical orientation, ie there is a canonical generator of $C_\gamma \cong \mathbb{Z}$, which we denote by 1_γ.

We now show that the signs ε used to construct the cochain complex from Section 1 provide a combinatorial realisation of this boundary map, and a preferred generator for each module C_γ.

Definition 2.9 (k–splittings) Let γ be a germ and Γ one of the graphs of γ with n leaves, $n \geq 3$. A *splitting of γ along* Γ is a germ γ_s together with the datum of a type 0 degeneracy $\gamma_s \rightsquigarrow \gamma$ resulting in the creation of the graph Γ. It has to involve two graphs Γ_1 and Γ_2 with k and $n+1-k$ leaves, respectively (we assume $k \leq n+1-k$), together with $(k-1)(n-k)$ two-leaved graphs. If $k \geq 3$, γ_s has a favourite edge $e(s)$ which is the only edge bounded by Γ_1 and Γ_2 that gets shrunk in the degeneracy. In Figure 1, with $k = 3$ and $n = 6$, $e(s)$ would be the edge labelled "1", and all visible crossings should be red, so that the six-vertex graph on the right is complete.

When $k = 2$, the choice of Γ_1 (and also Γ_2 if $n = 3$) and therefore $e(s)$ is not unique; see an example in Figure 2. However, k is uniquely defined and we have a notion of *k–splitting*.

Definition 2.10 Let Γ be a graph with two leaves in a germ γ. We define the *consistency* $\chi(\Gamma)$ to be $+1$ if the order defined by $\mathbb{R}$ and that defined by $<_\gamma$ agree on Γ, and -1 otherwise. We let $\chi w(\Gamma)$ denote the product of $\chi(\Gamma)$ with the sign of the unique arrow in Γ. The maps χ and w are set to $+1$ for graphs with more than two leaves.

Lemma 2.11 *Let γ_s be a (2–)splitting of γ, and let $L(s)$ be the underlying leaf diagram of γ_s. Then the sign*

$$\chi w(\Gamma_1)\chi w(\Gamma_2)\varepsilon_{L(s)}(e(s))$$

does not depend on the choice of Γ_1 *and* Γ_2.

This lemma is the key ingredient to show that our signs ε are of a topological nature. It will be proved at the end of this section.

We set

$$\partial_i(\gamma) = \sum_{\text{all splittings}} \chi w(\Gamma_1)\chi w(\Gamma_2)\varepsilon_{L(s)}(e(s)) \cdot \gamma_s$$

and extend this into a linear map $\partial_i \colon \mathcal{G}_i \to \mathcal{G}_{i-1}$. The reason for this map to not be graded with respect to the Gauss degree lies essentially in the bunch of two-leaved polygons that appear as a result of splitting a germ.

Theorem 2.12 (1) *The maps* ∂_i *and spaces* $\mathcal{G}_i$ *together form a chain complex.*

(2) *There is a unique collection of maps*

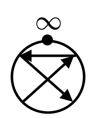
$$\phi_\gamma \colon C_\gamma \hookrightarrow \mathcal{G}_{\iota(\gamma)}$$

such that $\phi_\gamma(1_\gamma) = \gamma$ *if* $\iota(\gamma) = 0$, *and such that all the following diagrams commute:*

$$\begin{tikzcd} \mathcal{G}_{\iota(\gamma)} \arrow[r, "\partial_{\iota(\gamma)}"] & \mathcal{G}_{\iota(\gamma)-1} \\ C_\gamma \arrow[u, hook, "\phi_\gamma"] \arrow[r, "\partial_\gamma"'] & \bigoplus_0 C_{\gamma_k} \arrow[u, hook, "\bigoplus_0 \phi_{\gamma_k}"'] \end{tikzcd}$$

(3) *The map* ϕ_γ *maps* C_γ *isomorphically onto the submodule* $\mathbb{Z}\gamma \subset \mathcal{G}_{\iota(\gamma)}$. *Hence the preimage* $\phi_\gamma^{-1}(\gamma)$ *defines a canonical co-orientation of the simple stratum* $\mathcal{K}_\gamma$.

Figure 9 in Section 4.2 shows the co-orientation of the 2–stratum



depending on the signs of the arrows.

Proof The proof of the first point is identical to that of Theorem 1.3. One only has to notice that the signs χw always appear twice and cancel themselves in the computation of $\partial \circ \partial$, and that the collection of two-leaved polygons that result from a splitting does not affect the computations, because they are even polygons.

We prove points (2) and (3) simultaneously, by induction on i.

When $i=0$, there is nothing to prove. The case $i=1$ also needs to be treated separately. Here it suffices to notice that on the two sides of a Reidemeister III move, the sign $\chi w(\Gamma_1)\chi w(\Gamma_2)\varepsilon_{L(s)}(e(s))$ takes opposite values; indeed, such a move switches the signs $\mathrm{lk}(P_1,P_2)$, $\mathrm{E}(P_1,e)$ and $\mathrm{E}(P_2,e)$, and leaves the remaining signs unchanged. So ϕ_γ is defined uniquely by mapping to γ the generator of C_γ that is oriented from the negative side to the positive side. Point (3) is then satisfied, and it implies that the direct sum of any collection of maps ϕ_γ is injective.

Now let $i\geq 2$ and assume that (2) and (3) hold up to $i-1$. The crucial point is the following:

Lemma 2.13 *If $i\geq 2$, then in the cell decomposition of a meridian sphere $\mathbb{S}_\gamma^{i-1}$ made of meridian discs, the union of all $(i-1)$–discs corresponding to type 0 degeneracies is connected.*

Assuming this lemma, consider a germ $\gamma\in\mathcal{G}_i$. By definition, $\partial_i\gamma$ lies in $\bigoplus_0 C_{\gamma_k}$, so by induction (3) it has a unique preimage x by $\bigoplus_0\phi_{\gamma_k}$. By part (1) of the theorem and by induction (2), x lies in the kernel of $\bigoplus_0\partial_{\gamma_k}$. In other words, it is a relative cycle in $(\mathbb{S}_\gamma,\mathbb{S}_\gamma\setminus\bigcup_0\mathbb{D}_{\gamma_k})$. Also, it has local weight ± 1, so it is a generator of $H_{i-1}(\mathbb{S}_\gamma,\mathbb{S}_\gamma\setminus\bigcup_0\mathbb{D}_{\gamma_k})$, which by Lemma 2.13 is canonically isomorphic to $H_{i-1}(\mathbb{S}_\gamma)\cong H_i(\mathbb{D}_\gamma,\mathbb{S}_\gamma)\cong C_\gamma$. By pushing x through these isomorphisms, we obtain a generator of C_γ, and ϕ_γ is uniquely defined by the fact that it must map this generator to γ. This finishes the proof of Theorem 2.12, up to Lemma 2.13.

Proof of Lemma 2.13 If γ has at least two graphs Γ and Γ' with more than two leaves, then any two splittings along Γ and Γ', respectively, have a piece of boundary in common. If γ has only one graph with $n>2$ leaves, then n must be at least 4 so that $i=n-2\geq 2$. Here, any two 2–splittings sliding different branches of the knot away have a common piece of boundary, and any k–splitting ($k\geq 3$) has a common boundary piece with $n-1$ distinct 2–splittings. □

Proof of Lemma 2.11 We first prove the result in one particular case, then proceed by induction, using a number of "moves" that allow one to join any splitting of any germ.

First note that by symmetry of the formula in $\{\Gamma_1,\Gamma_2\}$ we need not check separately the case $n=3$, even though Γ_2 is not uniquely determined. Figure 2 shows a splitting $\gamma^+(2,n-1)$ of a germ with only one graph, n leaves and only $+$ signs. The orientations of the arrows in the $(n-1)$–gon depend on the way to connect virtually the branches of the $(n-1)$–crossing; they are not shown because the sign ε only depends on the underlying polygon. One easily sees that $\chi w(\Gamma_1)$ is -1 in $\gamma^+(2,n-1)$ for any choice of Γ_1, and only the maps lk and E can contribute non-trivially in ε. Let P_1 and P_2 denote the underlying polygons to Γ_1 and Γ_2 (so that $\sharp P_1=2$).

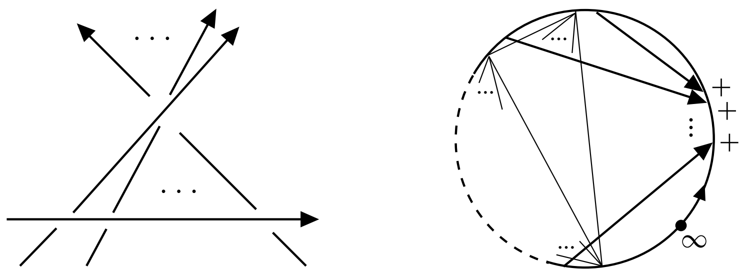


Figure 2: The germ $\gamma^+(2, n-1)$, a 2–splitting of the positive n–branch crossing

Now:

- $\mathrm{lk}(P_1, P_2)$ is $+1$ if Γ_1 is the topmost arrow (Γ_1–candidate) in the diagram in the right of Figure 2, and alternates up to $(-1)^n$ for the bottom arrow.
- $\mathrm{E}(P_1, e(s))$ has the same alternating property and is -1 for the bottom arrow
- If $n-1$ is even, $\mathrm{E}(P_2, e(s))$ has the same value $+1$ for any choice of Γ_1 (and this also holds obviously if $n-1$ is odd).

This proves that the sign $\chi w(\Gamma_1)\chi w(\Gamma_2)\varepsilon_{L(s)}(e(s))$ is $(-1)^n$ for any choice of Γ_1 in $\gamma^+(2, n-1)$.

We now prove the invariance of the result under the following moves:

(1) Adding a bystander graph.
(2) Making one crossing change in the $(n-1)$–crossing.
(3) Making one crossing change at one of the $n-1$ Γ_1–candidates.
(4) Reversing the orientation of a branch of the $(n-1)$–crossing.
(5) Sliding the branch that was split away from the n–crossing to the other side of the $(n-1)$–crossing.
(6) Changing the order in which the n local branches are virtually connected.
(7) Moving the point ∞ to an adjacent region.

Note that reversing the orientation of the branch of the knot that was split away can be formally realised by move (5) followed by $n-1$ moves of type (4).

We always neglect the orientation and sign changes on Γ_2, which are harmless. Move (1) may only modify the contribution of σ_{glo}, but it does so in the same way for all choices of Γ_1, essentially because Γ_1 is always even. Move (2) has no effect at all. Move (3) only changes $\chi(\Gamma_1)$ and $w(\Gamma_1)$ into their opposite, so that $\chi w(\Gamma_1)$ remains the same.

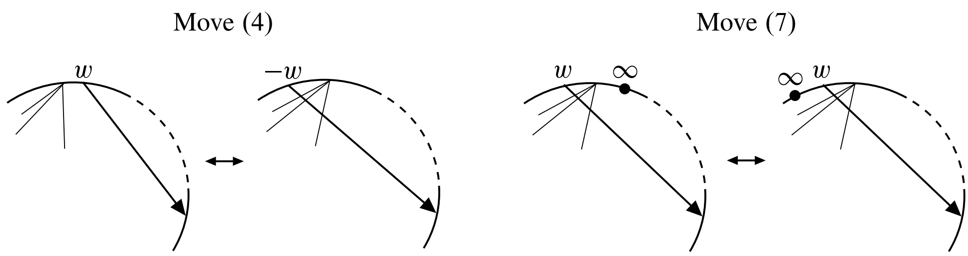


Figure 3: The effect of moves (4) and (7) on germs

Move (6) commutes with the other moves, so it suffices to see that it does not affect the result for $\gamma^+(2, n-1)$.

The effect of move (5) on the germ is identical to changing the sign of all Γ_1–candidates, and then formally applying the effect of $n-1$ moves of type (4). So we are left with the two moves (4) and (7), shown in Figure 3. Move (4) does not affect any signs for choices of Γ_1 other than the one in the picture; for this one, χw is changed into its opposite, and so is $\mathrm{E}(P_1, e(s))$. If $n-1$ is odd, nothing else changes; otherwise, both the linking number of $e(s)$ and the even index of P_2 are also reversed.

For all choices of Γ_1 except the one visible in Figure 2, the only effect of move (7) is to change $E(P_1, e(s))$. For the choice of Γ_1 in Figure 2, it changes χ, and nothing else if $n-1$ is odd; otherwise it also changes both even indices of P_1 and P_2. $\square$

3 Main result

We now introduce a degenerate version of arrow diagrams, designed to count *subgerms* in the spirit of Polyak and Viro [20]. Subgerms are the algebraic artefact that allows one to see whether a knot respects a germ, and in how many ways. They also appear naturally as the result of type 0 degeneracies.

3.1 Tree diagrams

Let P be a polygon in $\mathbb{S}^1 \setminus \{\infty\}$ of cardinality greater than 1. A *spanning tree* for P is a maximal collection of ordered pairs (v, w) with v, $w \in P$, still called *arrows*, such that the corresponding abstract oriented graph is a tree. The number of arrows in a spanning tree is always equal to the cardinality of the underlying polygon minus 1.

A *tree diagram* is a finite collection of pairwise disjoint polygons in $\mathbb{S}^1 \setminus \{\infty\}$ endowed with spanning trees. We keep denoting such diagrams by the letter "A" to respect the tradition of arrow diagrams, and save "T" for single spanning trees. Tree diagrams naturally inherit the Gauss and cohomological degrees defined for leaf diagrams, namely:

- The Gauss degree $\deg(A)$ of a tree diagram is equal to its total number of arrows.
- The cohomological degree $\iota(A)$ is the Gauss degree minus the number of colours (trees).

Again, tree diagrams are regarded up to positive homeomorphisms of the real line $\mathbb{S}^1 \setminus \{\infty\}$. The $\mathbb{Z}$–module freely generated by equivalence classes of tree diagrams of degree d and codimension i is denoted by $\mathcal{A}_d^i$. Note that $\mathcal{A}_d^i$ is trivial whenever i is greater than $d-1$, and whenever i or d is negative (see Remark 1.1).

The triangle relation Observe that a spanning tree T defines a partial order on the underlying polygon: say that $v <_T w$ if T contains the arrow (v, w), and extend this definition by transitivity, which is possible because T is a tree. We say that T is *monotonic* if the relation $<_T$ is total. Accordingly, a tree diagram is called *monotonic* if all of its trees are so. Monotonic spanning trees for a given polygon P are in one-to-one correspondence with total orders on P. Denote by $\nabla(T)$ the set of all monotonic spanning trees that correspond to total orders compatible with $<_T$.

Definition 3.1 The *triangle relation* is the equivalence relation on $\mathcal{A}_d^i$ generated by the equalities

$$A = \sum_{T' \in \nabla(T)} A_{T'}, \tag{3-1}$$

where A is a tree diagram that contains T as a spanning tree and $A_{T'}$ is the diagram obtained from A by replacing T with T'. We denote the quotient $\mathbb{Z}$–module by $\widetilde{\mathcal{A}}_d^i$. It is naturally isomorphic to the subspace of $\mathcal{A}_d^i$ spanned by monotonic tree diagrams.

Remark 3.2 This relation originated in the work of Polyak [18; 19] and also Polyak and Viro [21] on arrow diagrams. It owes its name to the fact that it is locally generated by the relation schematically depicted in Figure 4.

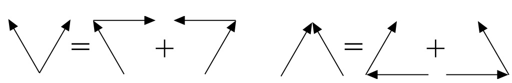


Figure 4: Local triangle relations. Only a part of a spanning tree is shown; the remaining invisible parts should be identical for all three diagrams in a given equality.

Definition 3.3 The *Reidemeister farness* of monotonic diagrams is defined similarly to that of germs (see Definition 2.7).The submodule of $\widetilde{\mathcal{A}}_d^i$ generated by R–far monotonic diagrams is denoted by $\widetilde{\mathcal{A}}_{d,\mathrm{far}}^i$. This definition makes sense since any $\alpha \in \widetilde{\mathcal{A}}_d^i$ has a unique representative involving only monotonic diagrams.

3.2 The pairing of tree diagrams with germs

Definition 3.4 (Partial germs and signed tree diagrams) A *partial germ* is a leaf diagram in which every polygon is enhanced into a *connected* abstract graph with oriented and signed arrows. The difference with germs is that here the graphs need not be complete. A partial germ in which every graph is a tree is called a *signed tree diagram.*

Partial germs inherit the degrees deg and ι from their underlying leaf diagrams. The corresponding $\mathbb{Z}$–modules of signed tree diagrams are denoted by $\mathcal{T}_{i,d}$ and $\mathcal{T}_i$.

Definition 3.5 A *subgerm* of a germ γ is the result of forgetting an arbitrary number of its arrows in such a way that every graph corresponding to a polygon with more than two leaves remains connected, although two-leaved polygons may completely disappear. This condition means that subgerms must remember the codimension of γ, *but the Gauss degree may drop.*

We set $\mathrm{I}(\gamma)$ to be the formal sum of all subgerms of γ that are signed tree diagrams. It is understood that subgerms are counted with multiplicity if the removal of distinct sets of arrows yields homeomorphic results. This defines a linear map

$$\mathrm{I}\colon \mathcal{G}_i \to \mathcal{T}_i.$$

If τ is a signed tree diagram, we define $\mathrm{T}(\tau)$ as the underlying tree diagram, multiplied by the product of the signs of all arrows of two-leaved polygons. Again this extends into a linear map

$$\mathrm{T}\colon \mathcal{T}_{i,d} \to \mathcal{A}_d^i.$$

Remark 3.6 The fact that the map T disregards the signs of arrows associated with polygons that have more than two leaves should be interpreted this way: for these polygons, the signs of the arrows have already contributed by entering the co-orientation defined by germs on their associated strata. In other words, when a simple crossing merges with others into a multiple crossing, we stop regarding its writhe as making sense individually. See Lemma 2.11 and Theorem 2.12.

Definition 3.7 For $\alpha \in \widetilde{\mathcal{A}}_d^i$ and $\gamma \in \mathcal{G}_i$, we set

$$\langle\!\langle \alpha, \gamma \rangle\!\rangle = \langle \alpha, \mathrm{T} \circ \mathrm{I}(\gamma) \rangle,$$

where $\langle \cdot, \cdot \rangle$ is the Kronecker delta on tree diagrams, extended by bilinearity.

We have to prove that this is a good definition, that is:

Lemma 3.8 *Let $\nabla \in \mathcal{A}_d^i$ be a triangle relator, ie the difference between the two sides of* (3-1). *Then*

$$\langle\!\langle \nabla, \gamma \rangle\!\rangle = 0 \quad \textit{for all } \gamma \in \mathcal{G}_i.$$

Proof The result follows immediately from the fact that the graphs of γ are complete and consistently oriented, considering the generating relations from Figure 4. □

This elementary proof should be compared with that of Lemma 1.9 of Mortier [16]. There, the result was deeply related to the fact that the germ was topological. Here, all the topology is confined to the implicit co-orientation associated with germs, and this lemma also holds for abstract germs.

Definition 3.9 Let c be a PL i–chain in $\mathcal{K} \setminus \Sigma$ that is transverse to the stratification. Then c intersects finitely many simple i–strata γ_p, with intersection numbers η_p defined by the co-orientation from Theorem 2.12. For $\alpha \in \widetilde{\mathcal{A}}_d^i$, we set

$$\{\alpha, c\} = \left\langle\!\!\left\langle \alpha, \sum_p \eta_p \gamma_p \right\rangle\!\!\right\rangle.$$

We are now in a position to see why degeneracies of type 2-1 do not deserve particular attention.

Lemma 3.10 *Let ζ be an* almost simple stratum*, that is, a boundary component of a simple i–stratum corresponding to a type 2-1 degeneracy. Let $\mathbb{S}_\zeta$ be the meridian sphere of ζ. Then*

$$\{\alpha, \mathbb{S}_\zeta\} = 0 \quad \textit{for all } d \geq 0 \textit{ and } \alpha \in \widetilde{\mathcal{A}}_{d,\mathrm{far}}^i.$$

This means that the cocyclicity condition for R–far cochains is empty around such strata.

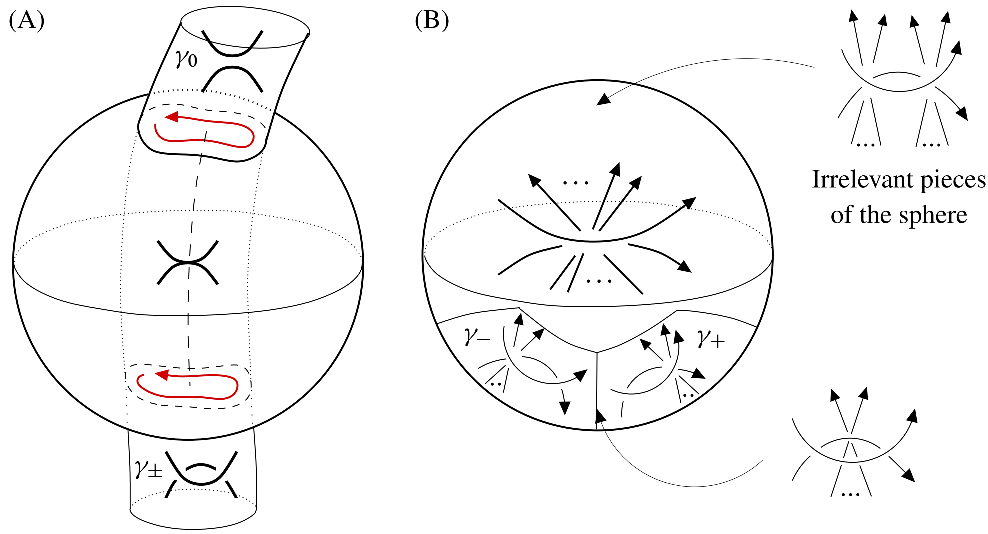


Figure 5: Meridian of an almost simple stratum for a simple (A) and multiple (B) crossing

Proof Denote by a the arrow in the i–germ γ that is subject to a 2-1 degeneracy. The situation is quite different according to whether or not a is part of a multiple crossing.

First assume that a is isolated. Then $\mathbb{S}_\zeta$ intersects exactly two simple strata γ_0 and $\gamma_\pm$, corresponding respectively to γ with the arrow a forgotten, and γ with the arrow a duplicated into two arrows with opposite writhe, which intersect or not depending on the geometric condition of ζ. We have precisely the two sides of a usual Reidemeister II move. Moreover, since the co-orientation of a germ depends only on the configuration of its graphs with more than two leaves, γ_0 and $\gamma_\pm$ induce opposite orientations on $\mathbb{S}_\zeta$ (see Figure 5(A)), so that, up to sign,

$$\{\alpha, \mathbb{S}_\zeta\} = \langle\langle \alpha, \gamma_0 - \gamma_\pm \rangle\rangle.$$

The result follows by classical arguments.

Now assume that a is part of a multiple crossing, with $k \geq 3$ branches (two of which have tangent projections). This time $\mathbb{S}_\zeta$ intersects 2^{k-2} simple i–strata, obtained from ζ by duplicating a into two arrows with opposite sign, and then forming two new multiple crossings by sharing the remaining $k-2$ branches among those two. However, one of the two arrows a_+ and a_- must remain isolated so that subgerms stand a chance to be R-II–far. Hence, only two diagrams may contribute, γ_+ and γ_-, as indicated by Figure 5(B). One can see in the picture that they have a piece of boundary in common (in fact, two); that is the key allowing us to compare their orientations.

Indeed, direct computation shows that γ_+ and γ_- induce the same orientation on their common boundary, hence they induce opposite orientations on $\mathbb{S}_\zeta$, and again, up to sign,

$$\{\alpha, \mathbb{S}_\zeta\} = \langle\langle \alpha, \gamma_+ - \gamma_- \rangle\rangle.$$

Now since α is R-II–far, the isolated duplicate of a must be deleted for a subgerm to contribute, so that the relevant subgerms in γ_+ are also subgerms in γ_-, with the only difference given by the sign of a. But this sign is disregarded by $\langle\langle \cdot, \cdot \rangle\rangle$, because a is a part of a multiple crossing. $\square$

3.3 Cohomology of tree diagrams and of the space of knots

Given a tree diagram A, an edge is called admissible if it is so in the underlying leaf diagram L. For such an edge e there is a natural way to define a tree diagram A_e that is a lift of L_e. Namely, if e is bounded by the leaves v and w, the arrows of A_e are the arrows of A with w replaced with v every time it appears. This edge-shrinking process is compatible with the triangle relations. We define a linear map $\tilde{\delta}^i_d \colon \widetilde{\mathcal{A}}^{i-1}_d \to \widetilde{\mathcal{A}}^i_d$ on the generators by

$$\tilde{\delta}^i_d(A) = \sum_{e \text{ admissible}} \chi(\Gamma_v)\chi(\Gamma_w)\varepsilon_L(e) \cdot A_e,$$

where χ is the consistency from Definition 2.10, and Γ_v and Γ_w are the graphs containing the leaves v and w, respectively.

We are now ready for the main theorem of this paper.

Theorem 3.11 (1) *The collection of maps $\tilde{\delta}^i_d$ and sets $\widetilde{\mathcal{A}}^i_d$ forms a graded, finite cochain complex. We denote by $H^i_{d,\mathrm{far}}$ the submodule of those i^{th} homology classes in degree d that have a representative cocycle in $\widetilde{\mathcal{A}}^i_{d,\mathrm{far}}$.*

(2) *(Stokes formula) For any $d \geq 0$, $i \geq 1$, $\alpha \in \widetilde{\mathcal{A}}^{i-1}_{d,\mathrm{far}}$ and $\gamma \in \mathcal{G}_i$,*

$$\langle\langle \tilde{\delta}^i_d(\alpha), \gamma \rangle\rangle = \langle\langle \alpha, \partial_i(\gamma) \rangle\rangle.$$

(3) *There is a natural map*

$$H^i_{d,\mathrm{far}} \to H^i(\mathcal{K} \setminus \Sigma)$$

induced by the pairing $\{\cdot,\cdot\}$. For $i = 0$, the image of this map consists of invariants induced by homogeneous Goussarov–Polyak–Viro formulas [10] *for long virtual knots. For $i = 1$, the image consists of arrow-germ formulas as defined by the author in* [16].

Remark 3.12 The farness constraint could be lightened, by allowing R-III–close diagrams. In the case $i = 0$, this is harmless (there are no additional equations) thanks to Lemma 3.2 of [15], and it yields all GPV invariants [15, Theorem 3.6]. For higher values of i, it would require us to compute the proper ε signs to associate with type 3-3 degeneracies and to consider subgerms whose graphs are not necessarily trees.

One could also think of removing the R-I– and R-II–farness condition; by contrast, this would require one to handle arbitrary geometric strata, resulting in a far more complicated story. For $i = 0$ it is pointless, R-I– and R-II–farness is actually a necessary condition for cocyclicity [15, Lemma 3.4]. For $i = 1$ it brings no new cohomology classes [16, Theorem 2.11].

Conjecture 3.13 *The image of the map $H^i_{d,\mathrm{far}} \to H^i(\mathcal{K} \setminus \Sigma)$ consists of Vassiliev cohomology classes of degree at most d.*

This conjecture holds when $i = 0$, when $i = 1$ and $d = 3$ (the case of the Teiblum–Turchin cocycle; see Turchin [22] and Vassiliev [24]), and also over $\mathbb{Z}_2$ when $d = i + 1$ (the extreme case of diagrams with only one tree).

Proof of Theorem 3.11 (1) This follows from Theorem 1.3 after noticing that the additional contribution $\chi(\Gamma_v)\chi(\Gamma_w)$ always cancels itself out in $\tilde{\delta} \circ \tilde{\delta}$.

(2) For simplicity we omit the subscripts and superscripts in the maps ∂_i and $\tilde{\delta}^i_d$. By linearity we may also assume that α is a tree diagram and γ a germ. Note that α cannot be a subgerm of both a k–splitting and an l–splitting of γ for $k \neq l$, so the proof can be split according to the at most unique value of k such that the right-hand side stands a chance to be non-zero when $\partial(\gamma)$ is restricted to k–splittings. As a last preliminary, note that we prove the formula at the level of $\mathcal{A}^i_{d,\mathrm{far}}$, ie *before the quotient by triangle relations*.

If $k > 2$, then because α is R–far we see that any subgerm of a term in $\partial(\gamma)$ that contributes non-trivially to the right-hand side must exclude every two-leaved graph that resulted from the splitting. Similarly, if $k = 2$, at most one of these graphs may have survived. Also, if none of them has survived, then the subgerm's possible contribution is cancelled out by the corresponding subgerm in the *opposite* 2–splitting (where the sliding branch has been pushed in the opposite direction); indeed, according to parity, the signs of these splittings differ either by $\mathrm{E}(P_1, e(s))$, or by $\mathrm{E}(P_1, e(s))$, $\mathrm{E}(P_2, e(s))$ and $\mathrm{lk}(e)$. Thus we see that for any value of k, we can restrict $\mathrm{I}(\partial(\gamma))$ to certain subgerms such that the corresponding subgerms of γ are *signed tree diagrams*. Note also that these subgerms have a well-defined preferred edge $e(s)$.

We now use a divide and conquer trick. Arrange the non-trivial contributions to the right-hand side according to which edge of α corresponds to $e(s)$. This edge must clearly be admissible in α, so a corresponding arrangement can be realised in the left-hand side. Now it is easy to see that the contributions in each pack are naturally in one-to-one correspondence, and that the signs match.

(3) The map $\alpha \mapsto \{\alpha, \cdot\}$ makes tree diagrams into cochains in $\mathcal{K}$. By Theorem 2.12, Lemma 3.10 and the Stokes formula, it maps cocycles to cocycles and coboundaries to coboundaries, thus inducing a map $H^i_{d,\mathrm{far}} \to H^i(\mathcal{K} \setminus \Sigma)$.

For $i = 0$, the map $\tilde{\delta}^1_d$ is isomorphic to the map d^Λ from [16] restricted to Gauss degree d, and this isomorphism is compatible with the Stokes formulas. There, it is proved that Goussarov–Polyak–Viro invariants are exactly the kernel of a certain map $d^\Lambda \oplus d^\Delta \oplus d^{\mathrm{I}} \oplus d^{\mathrm{II}}$, and our R–farness condition ensures that the diagrams live in the kernel of $d^\Delta \oplus d^{\mathrm{I}} \oplus d^{\mathrm{II}}$.

For $i = 1$, we use the result and terminology of [16, Theorem 2.11]. By our R–farness condition the condition of the theorem is satisfied, and also the *cube equations* associated with –strata are empty. Now it is straightforward to check that the tetrahedron equations associated with –strata yield the kernel of the map $\tilde{\delta}^2_d$ restricted to edges that are bounded by one leaf from the triangle, and the remaining equations from –strata are encoded by the restriction of $\tilde{\delta}^2_d$ to the complementary set of edges. Finally, considering the number of leaves in the polygons, the kernel of $\tilde{\delta}^2_d$ is the intersection of the kernels of these two restrictions. $\square$

4 Examples and comments

An essential aspect of our construction is that it is of a virtual nature. That is, the equations do not care about the fact that the germs at which we evaluate the bracket $\langle\langle \alpha, \cdot \rangle\rangle$ may or may not correspond to classical knots. A major benefit is that it makes the theory simple and computable. Taking care of classicality would be much more complicated: to the best of our knowledge there is no complete characterisation of Gauss diagrams of classical knots that do not require actually trying to draw the knot, although there are some powerful invariants which detect non-classicality in a lot of cases, such as Manturov's Gaussian parity [13], the Miyazawa polynomial (see for instance Kamada [12]) and Dye and Kauffman's arrow polynomial [6].

On the side of drawbacks, the map $H^i_{d,\mathrm{far}} \to H^i(\mathcal{K} \setminus \Sigma)$ is unlikely to surject onto the subgroup of Vassiliev cohomology classes. For instance, the Vassiliev invariant of order 3 given by Polyak and Viro's formula v_3 in [10, Theorem 2] cannot be found

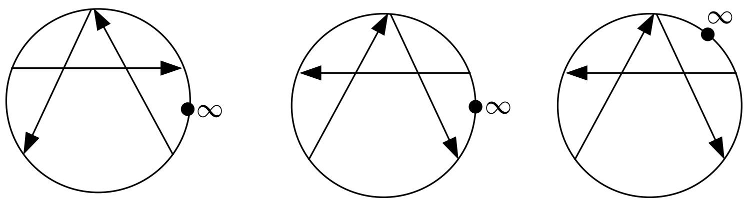
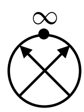
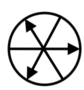
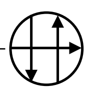
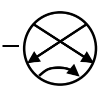
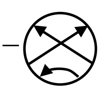


Figure 6: The 1–cocycle $\widetilde{\alpha}_3^1$

in $H^0_{3,\mathrm{far}}$ (a virtual version of v_3 is constructed by Chmutov and Polyak [4], but its nonhomogeneity makes it of a strongly different nature). However, our cochain complex produces a formula for v_3, quite unexpectedly, not from $H^0_{3,\mathrm{far}}$ but from $H^1_{3,\mathrm{far}}$, by integrating a 1–cocycle over the Fox–Hatcher loop (see Section 4.1). More precisely, we have the following as a corollary of Theorem 4.2:

Theorem 4.1 *The tree diagram $\widetilde{\alpha}_3^1$ in Figure 6 is an R–far 1–cocycle. Moreover, the integration of $\widetilde{\alpha}_3^1$ on the Gramain loop and the Fox–Hatcher loop of a knot K yield respectively the Gauss diagram formulas*

$\int_{\mathrm{rot}(K)} \widetilde{\alpha}_3^1 = \Big\langle\!\Big\langle$  $, K \Big\rangle\!\Big\rangle = v_2(K),$

$\int_{\mathrm{FH}(K)} \widetilde{\alpha}_3^1 = \Big\langle\!\Big\langle$  $+$  $-$  $-$  $, K \Big\rangle\!\Big\rangle = 6v_3(K) - w(K)v_2(K),$

where $w(K)$ denotes the blackboard framing of the diagram of K considered. In particular the map $H^1_{3,\mathrm{far}} \to H^1(\mathcal{K} \setminus \Sigma)$ has rank at least 1. If Conjecture 3.13 holds, then this rank is 1 and $\widetilde{\alpha}_3^1$ is a realisation of the Teiblum–Turchin cocycle over the integers.

The first line features Polyak and Viro's formula for v_2 [20], while the formula in the second line is new. As far as we know, this is the first time a Gauss diagram invariant specific to classical knots is found without using Gauss diagram identities (as in Östlund [17]). The only step where we did leave the comfortable field of virtual arguments is when we used the existence of the Fox–Hatcher loop!

Note that the second formula is unbased, which is a general phenomenon when integrating over the Fox–Hatcher loop. It denotes an invariant of closed knots (and hence of long knots since the two theories are equivalent). The evaluation bracket is then defined similarly to the based version, but it counts subdiagrams with multiplicity, which is given by the order of their symmetry group (see [17, Sections 2.2 and 2.4; 14, Section 4.1.2]).

4.1 Formal integration of 1–cocycles

A deep result due to Hatcher [11] states that the connected component of $\mathcal{K} \setminus \Sigma$ corresponding to a non-satellite long knot K has the homotopy type of $\mathbb{S}^1$ if K is a torus knot, and of $\mathbb{S}^1 \times \mathbb{S}^1$ if K is hyperbolic. For those knots there are essentially two interesting elements in $H_1(\mathcal{K}_K)$: the *Gramain loop*, $\mathrm{rot}(K)$, which consists of a rotation of a long knot around its axis, and the *Fox rolling*, or *Hatcher loop*, $\mathrm{FH}(K)$, which consists of sliding the ball at infinity (in $\mathbb{S}^3$) along the knot.

The Gramain loop does not depend on the Reidemeister moves we use to represent it. However, the Fox–Hatcher loop depends on a framing choice: indeed, each time one adds $+1$ to the framing of K, the ball at infinity makes one positive full spin on itself, which amounts to a negative spin of K, hence it adds $-\mathrm{rot}(K)$ to $\mathrm{FH}(K)$.

Let A be a monotonic tree diagram of codimension 1, so that its only polygon with more than two leaves is a triangle T. If the highest (respectively lowest) point of this triangle with respect to the order $<_A$ is also the lowest (resp. highest) of all leaves in A with respect to the $\mathbb{R}$ order, then we define a new diagram $\int_{\mathrm{rot}}^h A$ (resp. $\int_{\mathrm{rot}}^l A$) of codimension 0 by forgetting the arrow containing that point, with sign rule as indicated in Figure 7. Otherwise, we set $\int_{\mathrm{rot}}^h A = 0$ (resp. $\int_{\mathrm{rot}}^l A = 0$). This defines linear maps $\widetilde{\mathcal{A}}_d^1 \to \widetilde{\mathcal{A}}_{d-1}^0$.

With the same notations, let (a,b) and (b,c) denote the two arrows of T. We construct two unbased diagrams by replacing the arrow (a,b) with (a,∞) (resp. (b,c) with (∞,c)) while forgetting the point ∞, and give them signs depending only on the relative position of a, b and c in the cyclic order; see the rule in Figure 8. The difference is denoted by $\int_{\mathrm{FH}} A$ and defines a map $\widetilde{\mathcal{A}}_d^1 \to \widetilde{\mathcal{A}}_d^0$.

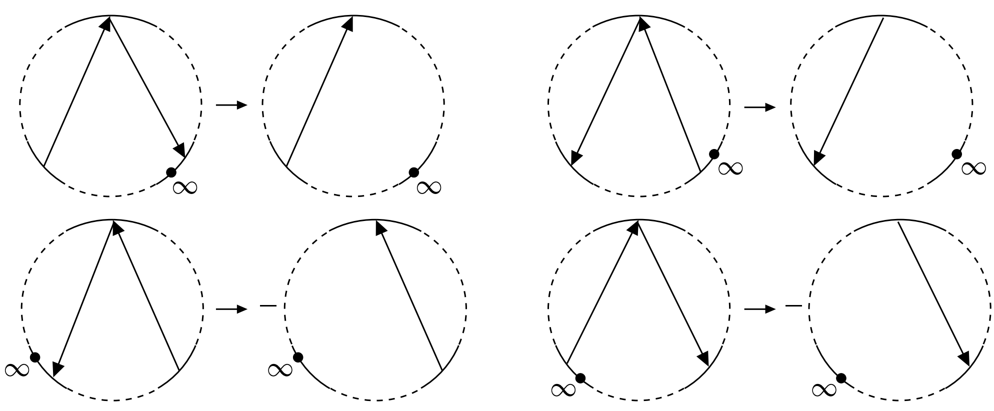


Figure 7: Sign rules for $\int_{\mathrm{rot}}^h$ (on the left) and $\int_{\mathrm{rot}}^l$ (on the right)

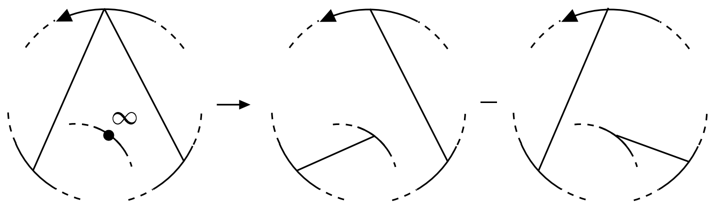


Figure 8: Sign rule for $\int_{\mathrm{FH}}$. It does not depend on the position of the point ∞.

Theorem 4.2 *Let $\alpha \in \widetilde{\mathcal{A}}^1_{d,\mathrm{far}} \cap \operatorname{Ker} \delta^2_d$. Then for any classical knot K,*

(1) $$\int_{\mathrm{rot}(K)} \alpha = \left\langle\!\!\left\langle \int^h_{\mathrm{rot}} \alpha + \int^l_{\mathrm{rot}} \alpha, K \right\rangle\!\!\right\rangle.$$

In particular, the right-hand side defines a finite-type invariant of K of degree at most $d-1$. However, $\int^h_{\mathrm{rot}} \alpha + \int^l_{\mathrm{rot}} \alpha$ might not lie in $\operatorname{Ker} \delta^1_{d-1}$.

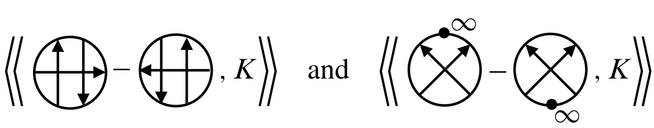
(2) $$\int_{\mathrm{FH}(K)} \alpha = \left\langle\!\!\left\langle \int_{\mathrm{FH}} \alpha, K \right\rangle\!\!\right\rangle.$$

The right-hand side defines a regular *invariant of K. Its value on a diagram of K with trivial blackboard framing defines a finite-type invariant of K of degree at most d. (Recall that here the bracket on the right counts subdiagrams with their potential multiplicity due to symmetry.)*

This theorem can be proved by analysing the presentation of rot from [7, Figure 144], and that of FH given by Fox [9] from the viewpoint of Gauss diagrams, as in the proof of [16, Theorem 3.3]. Reidemeister farness is crucial in the proof, not only for the theory to work properly, but to have good control of the non-trivial contributions to the integrals. For example, the 1–cocycle formula from [16, Theorem 3.2], which allows R-III–close diagrams, is impossible (for us) to integrate directly on the Fox–Hatcher loop, because of uncontrollable contributions.

Gauss diagram identities This theorem can be useful even when applied to a cocycle that is trivial in $H^1(\mathcal{K} \setminus \Sigma)$. Indeed, it may happen that the integration of such a cocycle is not formally zero. When this happens, it means that we have found a Gauss diagram identity, that is, a formula for the trivial invariant. But since there are no such formulas for virtual knots, we have there a non-trivial obstruction to classicality.

Among the low-degree examples, we have thereby a new proof that the Gauss diagram formulas



vanish for classical knots, which was first stated by Polyak and Viro [20] and proved by Östlund [17].

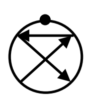
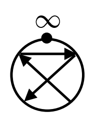
4.2 Higher-degree examples

A number of higher-degree formulas come for free as, in general, $\widetilde{\mathcal{A}}^i_{i+1,\mathrm{far}} \cong H^i_{i+1,\mathrm{far}}$, whose rank grows at least quadratically with i. All of those can be proved to be Vassiliev classes at least over $\mathbb{Z}_2$ using Vassiliev's homological calculus [24]. One could study their non-triviality by using the results of Budney [1] and Budney and Cohen [2], which are the state of the art and an excellent sequel to and completion of Hatcher's work on the topology of spaces of knots. We study here the cocycles in $H^2_{3,\mathrm{far}}$.

Our main motivation for computing higher-degree examples lies in reinterpreting a result of Budney, Conant, Scannell and Sinha [3], which states that it is possible to compute the invariant v_2 by counting an appropriate kind of quadrisecant with appropriate signs.

In the present language, a quadrisecant of a knot is a particular direction of projection for which the knot respects a germ with one polygon and four leaves. Hence, counting quadrisecants with signs is precisely what 2–cocycles in $H^2_{3,\mathrm{far}}$ do. More precisely, given a knot K, consider a sphere in $\mathbb{R}^3$, centred at the origin and with radius large enough to intersect K only in two points where it is arbitrarily close to its axis. Each point in that sphere defines a different direction of projection, except for the two intersection points with the axis of K. So we do not have a 2–cycle, but still a canonical 2–chain, where evaluating our cocycles makes sense since generically the quadrisecants stay far away from the knot axis during an isotopy of K. We call that 2–chain $\mathbb{S}_\infty(K)$.

The module $\widetilde{\mathcal{A}}^2_{3,\mathrm{far}}$ has two generators

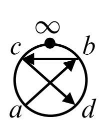
$v^2_3 =$  and $\tilde{v}^2_3 =$ 

and both are cocycles.

Theorem 4.3 *For any knot K,*

$$v_2(K) = \langle\langle v^2_3, \mathbb{S}_\infty(K) \rangle\rangle = \sum w(a,b)w(c,d),$$

where the sum is over all quadrisecants of K of type



and $w(a,b)$ denotes the writhe of the simple crossing between the branches a and b.

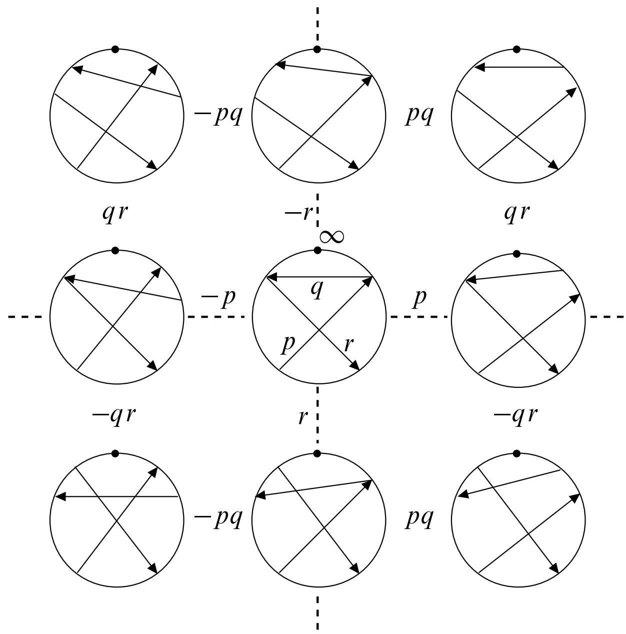


Figure 9: The meridian of a germ with underlying tree diagram v_3^2. The numbers p, q, r are the writhes of the arrows as indicated in the middle diagram. A sign between two diagrams indicates the co-orientation.

One can see that this is a new point of view on [3, Proposition 6.2], with a much simpler formula to think of. Indeed, the quadrisecants counted by v_3^2 are precisely those which "determine the cycle (1342)" in the language of [3, Section 6].

Proof We begin with the second equality. It is proved by analysing the co-orientation defined by v_3^2 and understanding what orientation it defines on $\mathbb{S}_\infty(K)$. The natural orientation of the plane in Figure 9 (ie the co-orientation of the codimension-2 stratum in the middle of the picture), as defined by Theorem 2.12, is counterclockwise if and only if the product of writhes pqr is $+1$.

Now we need to draw the picture of Figure 9 on the sphere $\mathbb{S}_\infty$. For this, observe the following. Choose a point x on $\mathbb{S}_\infty$ which defines a diagram K_x with exactly one generic triple point, say $f(t_1) >_x f(t_2) >_x f(t_3)$; the set of such points is a 1–submanifold $X \subset \mathbb{S}_\infty$. By moving the centre of $\mathbb{S}_\infty$ so that it lies on the line containing the triple point, the derivatives $f'(t_1)$, $f'(t_2)$ and $f'(t_3)$ project to a generic triple of vectors v_1, v_2, $v_3 \in T_x\mathbb{S}_\infty$.

Fact The direction of T_xX lies in the angular region determined by the directions of v_1 and v_3 that does not contain the direction of v_2.

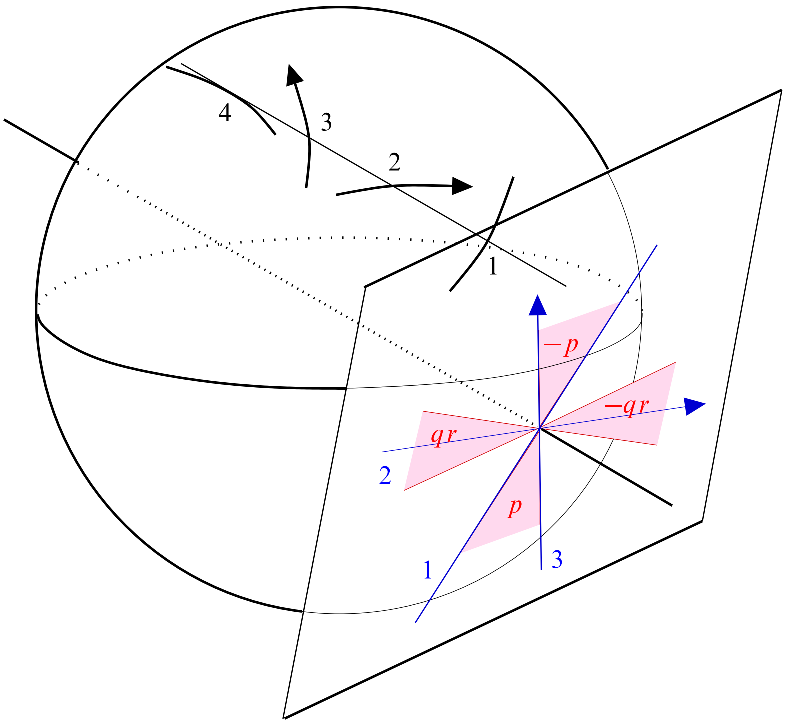


Figure 10: Orientation of $\mathbb{S}_\infty$ induced by v_3^2

To see this, think of the top and bottom branches as locally spiraling around the medium branch; see also Figure 10: the zones labelled p and $-p$ contain the directions from which one sees a triple point between branches 1, 2 and 3.

Figure 10 reads like this. Independently of the direction of the furthest branch (4), we know that the branch of X that slides 4 away and keeps the triple point $\{1,2,3\}$ lies in the region bounded by 1 and 3 that does not contain 2. Also, it appears that the orientation $p/-p$ (defined by the middle horizontal line of Figure 9) depends only on the orientation of the branch 3 as indicated. It is then easy to see that the splitting of the remaining triple point is supported by the direction 2, and that the orientation $qr/-qr$ depends only on the orientation of the branch 2.

To conclude, the relative position of p and qr in Figure 10 is dictated by the sign q (writhe of the crossing between branches 2 and 3). Hence the orientation induced on $\mathbb{S}_\infty$ by v_3^2 is dictated by the sign $pqr \cdot q = pr$, which is the result announced.

Using this formula, it is straightforward to see that $\langle\langle v_3^2, \mathbb{S}_\infty(K)\rangle\rangle$ is a Vassiliev invariant of degree at most 2; therefore it suffices to check the first equality for the trefoil. $\square$

References

[1] **R Budney**, *Topology of knot spaces in dimension* 3, Proc. Lond. Math. Soc. 101 (2010) 477–496 MR2679699

[2] **R Budney**, **F Cohen**, *On the homology of the space of knots*, Geom. Topol. 13 (2009) 99–139 MR2469515

[3] **R Budney**, **J Conant**, **K P Scannell**, **D Sinha**, *New perspectives on self-linking*, Adv. Math. 191 (2005) 78–113 MR2102844

[4] **S Chmutov**, **M Polyak**, *Elementary combinatorics of the HOMFLYPT polynomial*, Int. Math. Res. Not. 2010 (2010) 480–495 MR2587575

[5] **J M S David**, *Projection-generic curves*, J. London Math. Soc. 27 (1983) 552–562 MR697147

[6] **H A Dye**, **L H Kauffman**, *Virtual crossing number and the arrow polynomial*, J. Knot Theory Ramifications 18 (2009) 1335–1357 MR2583800

[7] **T Fiedler**, *Quantum one-cocycles for knots*, preprint (2013) `arXiv:1304.0970`

[8] **T Fiedler**, **V Kurlin**, *A* 1*–parameter approach to links in a solid torus*, J. Math. Soc. Japan 62 (2010) 167–211 MR2648220

[9] **R H Fox**, *Rolling*, Bull. Amer. Math. Soc. 72 (1966) 162–164 MR0184221

[10] **M Goussarov**, **M Polyak**, **O Viro**, *Finite-type invariants of classical and virtual knots*, Topology 39 (2000) 1045–1068 MR1763963

[11] **A Hatcher**, *Spaces of knots*, preprint (1999) `arXiv:9909095`

[12] **N Kamada**, *An index of an enhanced state of a virtual link diagram and Miyazawa polynomials*, Hiroshima Math. J. 37 (2007) 409–429 MR2376727

[13] **V O Manturov**, *Parity in knot theory*, Mat. Sb. 201 (2010) 65–110 MR2681114 In Russian; translated in Sb. Math. 201 (2010) 693–733

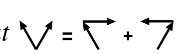
[14] **A Mortier**, *Virtual knot theory on a group*, preprint `arXiv:1403.3419`

[15] **A Mortier**, *Polyak type equations for virtual arrow diagram invariants in the annulus*, J. Knot Theory Ramifications 22 (2013) 1350034, 21 MR3084752

[16] **A Mortier**, *Finite-type* 1*–cocycles of knots and virtual knots given by Polyak–Viro formulas*, J. Knot Theory Ramifications 24 (2015) 1540004

[17] **O P Östlund**, *A combinatorial approach to Vassiliev knot invariants*, project report, Uppsala University (1996)

[18] **M Polyak**, *On the algebra of arrow diagrams*, Lett. Math. Phys. 51 (2000) 275–291 MR1778064

[19] **M Polyak**, *Three stories about* , lecture (2011) Available at `http://drorbn.net/dbnvp/SK11_Polyak.php`

[20] **M Polyak**, **O Viro**, *Gauss diagram formulas for Vassiliev invariants*, Internat. Math. Res. Notices (1994) 445–453 MR1316972

[21] **M Polyak**, **O Viro**, *On the Casson knot invariant*, J. Knot Theory Ramifications 10 (2001) 711–738 MR1839698

[22] **V Turchin**, *Computation of the first nontrivial* 1*–cocyle in the space of long knots*, Mat. Zametki 80 (2006) 105–114 MR2280743

[23] **V A Vassiliev**, *Cohomology of knot spaces*, from: "Theory of singularities and its applications", (V I Arnol′d, editor), Adv. Soviet Math. 1, Amer. Math. Soc. (1990) 23–69 MR1089670

[24] **V A Vassiliev**, *Combinatorial formulas for cohomology of spaces of knots*, from: "Advances in topological quantum field theory", (J M Bryden, editor), NATO Sci. Ser. II Math. Phys. Chem. 179, Kluwer, Dordrecht (2004) 1–21 MR2147414

[25] **C T C Wall**, *Geometric properties of generic differentiable manifolds*, from: "Geometry and topology", (J Palis, M do Carmo, editors), Lecture Notes in Math. 597, Springer, Berlin (1977) 707–774 MR0494233

15 Route des Futaies, 57100 Thionville, France

Osaka City University Advanced Mathematical Institute
OCAMI, 3-3-138 Sugimotocho, Sumiyoshi-ku, Osaka-shi, Osaka 558-8585, Japan

arno.mortier@laposte.net

http://sites.google.com/site/mortier2x0/

Received: 10 September 2014 Revised: 12 March 2015

msp

Algebraic & Geometric Topology 15 (2015) 3467–3483

On the K–theory of subgroups of virtually connected Lie groups

DANIEL KASPROWSKI

We prove that for every finitely generated subgroup G of a virtually connected Lie group which admits a finite-dimensional model for $\underline{E}G$, the assembly map in algebraic K–theory is split injective. We also prove a similar statement for algebraic L–theory which, in particular, implies the generalized integral Novikov conjecture for such groups.

18F25, 19A31, 19B28, 19G24

1 Introduction

For every group G and every ring R there is a functor $\mathbb{K}_R\colon \operatorname{Or} G \to \mathfrak{Spectra}$ from the orbit category of G to the category of spectra, sending G/H to (a spectrum weakly equivalent to) the K–theory spectrum $\mathbb{K}(R[H])$ for every subgroup $H \leq G$. By K–theory we will always mean nonconnective K–theory as constructed by Pedersen and Weibel [26]. For any such functor $F\colon \operatorname{Or} G \to \mathfrak{Spectra}$ a G–homology theory $\mathbb{F}$ can be constructed via

$$\mathbb{F}(X) := \operatorname{Map}_G(\,\cdot\,, X_+) \wedge_{\operatorname{Or} G} F;$$

see Davis and Lück [14]. We will denote its homotopy groups by $H_n^G(\,\cdot\,, F) := \pi_n \mathbb{F}(X)$. Let $\mathcal{F}$ be a family of subgroups of G. The K–theoretic assembly map for $\mathcal{F}$ is the map

$$\alpha_{\mathcal{F}}\colon H_n^G(E_{\mathcal{F}}G; \mathbb{K}_R) \to H_n^G(\mathrm{pt}; \mathbb{K}_R) \cong K_n(R[G]),$$

induced by the map $E_{\mathcal{F}}G \to \mathrm{pt}$. Here $E_{\mathcal{F}}G$ denotes the classifying space for the family $\mathcal{F}$; see Lück [22]. The assembly map is a helpful tool to relate the K–theory of the group ring $R[G]$ to the K–theory of the group rings over $H \in \mathcal{F}$. The assembly map can be defined more generally for any small additive G–category instead of R; see Bartels and Reich [11]. In this article all additive categories will be small.

Analogously, for every additive G–category $\mathcal{A}$ with involution and every family of subgroups $\mathcal{F}$ we can define the L–theoretic assembly map

$$\alpha_{\mathcal{F}}\colon H_n^G(E_{\mathcal{F}}G; \mathbb{L}_{\mathcal{A}}^{\langle -\infty \rangle}) \to H_n^G(\mathrm{pt}; \mathbb{L}_{\mathcal{A}}^{\langle -\infty \rangle}).$$

Published: 12 January 2016 DOI: 10.2140/agt.2015.15.3467

The Farrell–Jones conjecture [15] states that the assembly maps $\alpha_{\mathcal{V}cyc}$ for the family of virtually cyclic subgroups in K– and L–theory are isomorphisms for all additive G–categories $\mathcal{A}$ (with involution) and all $n \in \mathbb{Z}$. The Farrell–Jones conjecture has been proven for a large class of groups, for example hyperbolic and CAT(0)–groups (Bartels and Lück [7; 8], Bartels, Lück and Reich [9; 10] and Wegner [29]), virtually solvable groups (Wegner [30]), and lattices in virtually connected Lie groups (Bartels, Farrell and Lück [4] and Kammeyer, Lück and Rüping [19]). The Farrell–Jones conjecture implies that the assembly maps $\alpha_{\mathcal{F}in}$ for the family of finite subgroups are split injective; see Bartels [2, Theorem 1.3]. The rational split injectivity of the map $\alpha_{\mathcal{F}in}$ in L–theory implies the Novikov conjecture. The integral split injectivity of $\alpha_{\mathcal{F}in}$ is called the *generalized integral Novikov conjecture*; for more details see Section 6. Kasparov proved the Novikov conjecture for all discrete subgroups of virtually connected Lie groups in [20, Theorem 6.9]. More generally, the Novikov conjecture is true for groups which uniformly embed into a Hilbert space; see Skandalis, Tu and Yu [27]. This includes all amenable groups and all groups with finite asymptotic dimension. By Carlsson and Goldfarb [12, Section 3] and Ji [17, Corollary 3.4], discrete subgroups of virtually connected Lie groups have finite asymptotic dimension, giving a second proof that the Novikov conjecture holds for these groups. Here we will show that, in particular, discrete subgroups of virtually connected Lie groups also satisfy the generalized integral Novikov conjecture.

In [21] the author proved the split injectivity of the assembly map for finitely generated subgroups G of $\mathrm{GL}_n(\mathbb{C})$ which have an upper bound on the Hirsch length of the unipotent subgroups. For a definition of the Hirsch length see Section 3. The bound on the Hirsch length exists if and only if G has finite virtual cohomological dimension by Alperin and Shalen [1]. Since G is virtually torsion-free, this is the case if and only if there is a finite-dimensional model for $\underline{E}G$ where we consider G with the discrete topology; see Lück [22, Theorem 3.1]. In this article we want to extend this theorem to subgroups of all virtually connected Lie groups. Note that in the theorem we again consider G with the discrete topology.

Theorem 1.1 *Let G be a finitely generated subgroup of a virtually connected Lie group, and assume there exists a finite-dimensional model for $\underline{E}G$. Then the K–theoretic assembly map*

$$H_n^G(\underline{E}G;\mathbb{K}_{\mathcal{A}}) \to K_n(\mathcal{A}[G])$$

is split injective for every additive G–category $\mathcal{A}$.

A similar version holds for L–theory as well, which implies, in particular, the generalized integral Novikov conjecture for these groups; see Section 6.

If G is a discrete subgroup of a virtually connected Lie group H, and K the maximal compact subgroup of H, then H/K is a finite-dimensional model for $\underline{E}G$; see Lück [23, Theorem 4.4]. In particular, we get the following corollary.

Corollary 1.2 *Let G be a finitely generated discrete subgroup of a virtually connected Lie group. Then the K–theoretic assembly map*

$$H_n^G(\underline{E}G; \mathbb{K}_{\mathcal{A}}) \to K_n(\mathcal{A}[G])$$

is split injective for every additive G–category $\mathcal{A}$.

The condition on the existence of a finite-dimensional model for $\underline{E}G$ can be reformulated in the following way.

Proposition 1.3 *A finitely generated subgroup G of a virtually connected Lie group admits a finite-dimensional model for $\underline{E}G$ if and only if there exists $N \in \mathbb{N}$ such that every finitely generated abelian subgroup of G has rank at most N.*

The rank of an abelian group A is defined as $\mathrm{rk}(A) := \dim_{\mathbb{Q}}(A \otimes_{\mathbb{Z}} \mathbb{Q})$ or, equivalently, as the cardinality of a maximal linearly independent subset of A. The statement that every finitely generated abelian subgroup of G has rank at most N is equivalent to the statement that every abelian subgroup of G has rank at most N. For a proof of the proposition, see Section 3.

In Section 7, we prove that Theorem 1.1 and its L–theoretic analog also hold without the assumption that G is finitely generated.

Acknowledgments: I would like to thank Johannes Ebert for helpful discussions, and Henrik Rüping and the referee for useful comments and suggestions. This work was partially supported by the SFB 878 "Groups, Geometry and Actions" and the Max Planck Society.

2 Lie groups

A Lie group is virtually connected if it has only finitely many connected components. For the rest of this section let H be a virtually connected Lie group with Lie algebra $\mathfrak{h}$ (which we identify with $T_e H$). The Lie group H acts on itself by conjugation;

$$c\colon H \to \mathrm{Aut}(H),\ g \mapsto (h \mapsto ghg^{-1}).$$

Taking the derivative yields a map

$$\mathrm{Ad}\colon H \to \mathrm{Aut}(\mathfrak{h}),\ g \mapsto D_e(c(g)).$$

Since $\operatorname{Aut}(\mathfrak{h})$ is a Lie subgroup of $\mathrm{GL}(\mathfrak{h})$, Ad gives a representation of H. The kernel of the representation Ad is the centralizer $C_H(H_0)$ of the unit component H_0 of H.

By definition of the centralizer, the group $C_H(H_0) \cap H_0$ is abelian, and since H is virtually connected the centralizer $C_H(H_0)$ is, therefore, virtually abelian. For every subgroup G of H we obtain a short exact sequence

$$1 \to C_H(H_0) \cap G \to G \to \mathrm{Ad}(G) \to 1,$$

with virtually abelian kernel and linear quotient. We will use this sequence to extend the results of [21] to general virtually connected Lie groups. Before we can do so, we first need to prove Proposition 1.3, which will be done in the next chapter.

3 A bound on the rank of abelian subgroups

In the proof of Proposition 1.3, a bound on the Hirsch length of the finitely generated nilpotent subgroups is needed. First we review some facts about nilpotent groups to see that this is the same as a bound on the ranks of the finitely generated abelian subgroups.

Let G be a group. Define $G_1 := G$ and, recursively, $G_{n+1} := [G_n, G]$. The series $G = G_1 \geq G_2 \geq \cdots$ is called the *lower central series* of G. A group G is *nilpotent* if there exists $c \in \mathbb{N}$ with $G_{c+1} = 1$. The smallest such c is called the *nilpotency class* of G; we denote it by $c(G)$. The *upper central series* $1 = Z_0(G) \leq Z_1(G) \leq \cdots$ of G is recursively defined by

$$Z_{i+1}(G) := \{g \in G \mid \forall h \in G : [g,h] \in Z_i(G)\}.$$

If G is nilpotent, then $Z_{c(G)}(G) = G$ and the length of the upper and lower central series agree. For any normal subgroup $H \leq G$ the quotient G/H is again nilpotent.

The *Hirsch length* $h(G)$ of G is

$$h(G) := \mathrm{rk}(G_1/G_2) + \cdots + \mathrm{rk}(G_{c-1}/G_c) + \mathrm{rk}(G_c),$$

where $\mathrm{rk}(H)$ denotes the rank of an abelian group H; ie $\mathrm{rk}(H) := \dim_{\mathbb{Q}}(H \otimes_{\mathbb{Z}} \mathbb{Q})$.

Let $n(G)$ denote

$$\max\{\mathrm{rk}(A) \mid A \trianglelefteq G \text{ an abelian normal subgroup}\}.$$

Let H be and G be a group acting on H. G *acts nilpotently* if there is a series

$$1 = H_0 \leq H_1 \leq \cdots \leq H_n = H$$

of G–invariant normal subgroups of H such that the induced action on H_i/H_{i-1} is trivial. In the special case where $H = G$ and the action is by conjugation, G acts nilpotently on itself if and only if G is nilpotent.

Proposition 3.1 *Let G be finitely generated nilpotent. Then $h(G) \le \frac{n(G)(n(G)+1)}{2}$.*

The proposition is proved in Möhres [25, Theorem 2] for torsion-free nilpotent groups instead of finitely generated nilpotent groups. For the convenience of the reader we give a proof. For this we need the following well-known statements about nilpotent groups.

Lemma 3.2 *A subgroup of a finitely generated nilpotent group is finitely generated.*

Proof The statement follows by induction on the nilpotency class. □

Lemma 3.3 [28, Theorem 1.3] *Let G be nilpotent and $N \trianglelefteq G$ a nontrivial normal subgroup. Then $N \cap Z(G)$ is nontrivial, where $Z(G)$ denotes the center of G.*

Lemma 3.4 *Let G be nilpotent and A a maximal abelian normal subgroup. Then $C_G(A) = A$, where $C_G(A)$ is the centralizer of A in G.*

Proof Since $A \trianglelefteq G$ is normal, so is $C_G(A)$. Suppose $A \ne C_G(A)$. Then $C_G(A)/A$ is a nontrivial normal subgroup of G/A, and $H := C_G(A)/A \cap Z(G/A)$ is nontrivial by the previous lemma. Let $C = \langle c\rangle$ be a cyclic subgroup of H. Then $C \trianglelefteq Z(G/A) \trianglelefteq G/A$ and, since C lies in the center, it is a normal subgroup of G/A. Let $c' \in C_G(A)$ be a preimage of c; then the preimage of C is $\langle A, c'\rangle$. This is abelian and normal in G; hence, A was not maximal with this property. □

Lemma 3.5 *Let $\mathrm{Tr}(n,\mathbb{Z}) \le \mathrm{GL}_n(\mathbb{Z})$ denote the subgroup of unitriangular matrices; ie every element of $\mathrm{Tr}(n,\mathbb{Z})$ has 1's on the diagonal and 0's below the diagonal. If $G \le \mathrm{GL}_n(\mathbb{Z})$ acts nilpotently on $\mathbb{Z}^n$, then it is unipotent and conjugate to a subgroup of $\mathrm{Tr}(n,\mathbb{Z})$.*

Proof Since $\mathrm{Tr}(n,\mathbb{Z})$ is unipotent, it suffices to prove that G is conjugate to a subgroup of it. Let

$$0 = H_0 \trianglelefteq H_1 \trianglelefteq H_2 \trianglelefteq \dots \trianglelefteq H_k = \mathbb{Z}^n$$

be a sequence of G–invariant subspaces and let G act trivially on H_i/H_{i-1} for all $i = 1,\dots,k$. The lemma is obvious for $k = 1$, and we will prove it by induction on k. Let $H' := \{z \in \mathbb{Z}^n \mid \exists\ l \in \mathbb{Z} : lz \in H_1\}$. Let $z \in H'$ and $l \in \mathbb{Z}$ with $lz \in H_1$. For every

$g \in G$ we have $lg(z) = g(lz) = lz$ and thus also $g(z) = z$; ie G acts trivially on H'. By construction, $\mathbb{Z}^n/H'$ is torsion-free, and we obtain a splitting $\mathbb{Z}^n \cong H' \oplus \mathbb{Z}^n/H'$. The sequence

$$0 = H'/H' \trianglelefteq H_2 + H'/H' \trianglelefteq \cdots \trianglelefteq H_k + H'/H' = \mathbb{Z}^n/H'$$

consists of G–invariant subspaces, and G acts trivially on the quotients. By induction there is a basis of $\mathbb{Z}^n/H$ such that $G \leq \mathrm{GL}(\mathbb{Z}^n/H)$ is unitriangular. Using this basis together with a basis of H' yields a basis of $\mathbb{Z}^n$ for which G lies in $\mathrm{Tr}(n, \mathbb{Z})$. □

Proof of Proposition 3.1 Let $n := n(G)$ and A be a maximal abelian normal subgroup. Then A again is finitely generated by Lemma 3.2, and $A \cong \mathbb{Z}^n \oplus F$ with F a finite group. The group G acts by conjugation on A and, since $C_G(A) = A$, the induced map $G/A \to \mathrm{Aut}(A)$ is injective. Since F is finite, the projection to $\mathrm{Aut}(\mathbb{Z}^n) = \mathrm{GL}_n(\mathbb{Z})$ has finite kernel. The group G is nilpotent, and thus it acts nilpotently on $\mathbb{Z}^n$ (by conjugation). This implies that the image G/A in $\mathrm{GL}_n(\mathbb{Z})$ is conjugate to a subgroup of the unitriangular matrices $\mathrm{Tr}(n, \mathbb{Z})$. Since $h(\mathrm{Tr}(n, \mathbb{Z})) = n(n-1)/2$, we have

$$\begin{aligned} h(G) &\leq h(A) + h(\ker(\mathrm{Aut}(A) \to \mathrm{GL}_n(\mathbb{Z}))) + h(\mathrm{Tr}(n, \mathbb{Z})) \\ &= n + 0 + \frac{n(n-1)}{2} = \frac{n(n+1)}{2}. \end{aligned}$$ □

A direct corollary of Proposition 3.1 is the following.

Corollary 3.6 *Let G be a group. Then G has a bound on the Hirsch length of its finitely generated nilpotent subgroups if and only if it has a bound on the rank of its finitely generated abelian subgroups.*

Before we can prove Proposition 1.3 we need the following lemma.

Lemma 3.7 *Let A be a (countable) abelian group with finite rank, then there is a finite-dimensional model for $\underline{E}A$.*

Proof Let $\mathrm{rk}\, A = n$. Then there exists a subgroup $B \leq A$ isomorphic to $\mathbb{Z}^n$. The quotient $Q := A/B$ has rank 0 and thus is a torsion group. For $n \in \mathbb{N}$ let $F_n \leq Q$ be finite subgroups with $F_n \leq F_{n+1}$ and $Q = \bigcup_{n \in \mathbb{N}} F_n$. Define a Q–CW-complex X by taking $\coprod_{n \in \mathbb{N}} Q/F_n$ as the zero skeleton and for every $n \in \mathbb{N}$ adding a 1–cell with stabilizer F_n between the 0–cells Q/F_n and Q/F_{n+1}. This defines a 1–dimensional model X for $\underline{E}Q$. Let $p\colon A \to Q$ be the quotient map. For every finite subgroup $F \leq Q$, the preimage $p^{-1}(F)$ is finitely generated abelian of rank n and thus has $\mathbb{R}^n$ as an n–dimensional model for $\underline{E}p^{-1}(Q)$. Therefore, the proof of Lück [22, Theorem 3.1] shows that A has a model for $\underline{E}A$ of dimension $n+1$. □

Let G be a subgroup of $\mathrm{GL}_n(\mathbb{C})$ and assume there exists $N \in \mathbb{N}$ such that the rank of every finitely generated unipotent subgroup of G is at most N. Then, by Alperin and Shalen [1], the virtual cohomological dimension of G is bounded and therefore admits a finite-dimensional model for $\underline{E}G$ by [22, Theorem 6.4]. Using this, we now can prove Proposition 1.3.

Proof of Proposition 1.3 Let G be a subgroup of a virtually connected Lie group H such that there exists a finite dimensional model X for $\underline{E}G$. Then, in particular, X is a model for $\underline{E}A$ for every abelian subgroup $A \leq G$ and $\operatorname{rk} A \leq \dim X$.

For the other direction, let G be a finitely generated subgroup of a virtually connected Lie group H such that there exists a bound on the rank of the finitely generated abelian subgroups of G. Then, by Corollary 3.6, G has also a bound on the Hirsch length of its finitely generated nilpotent subgroups. Let $G_0 := G \cap H_0$, and consider the extension

$$1 \to C_H(H_0) \cap G_0 \to G_0 \to \mathrm{Ad}(G_0) \to 1$$

from Section 2. Since $C_H(H_0) \cap G_0$ is contained in the center of G_0, $\mathrm{Ad}(G_0)$ also has a bound on the Hirsch length of its finitely generated nilpotent subgroups and, thus, on the finitely generated unipotent subgroups. By the above it admits a finite dimensional model for $\underline{E}\,\mathrm{Ad}(G_0)$. And since also $K := C_H(H_0) \cap G_0$ has finite rank, there is a finite dimensional model for $\underline{E}K$ by Lemma 3.7. Consider the extensions

$$1 \to K \to G_0 \to \mathrm{Ad}(G_0) \to 1,$$
$$1 \to G_0 \to G \to F \to 1,$$

with F finite. The group G_0 is finitely generated since finite index subgroups of finitely generated groups are again finitely generated. Thus $\mathrm{Ad}(G_0)$ is virtually torsion-free by Selberg's lemma, and we can use [22, Theorem 3.1] to obtain a finite dimensional model for $\underline{E}G$ from these sequences. □

Remark 3.8 Using the results of the author from [21], the short exact sequence

$$1 \to C_H(H_0) \cap G \to G \to \mathrm{Ad}(G) \to 1$$

implies that G has fqFDC, which also is defined in [21]. In particular, if G has a bound on the order of the finite subgroups, then the main result of [21] directly implies the split injectivity of the K–theoretic assembly map and a similar result in L–theory. Since we do not know if this always holds, we use a different approach using inheritance properties; see Sections 4 and 5.

4 Inheritance properties

To use the short exact sequence from Section 2 we want to show the following inheritance property.

Proposition 4.1 *Assume there is a short exact sequence of groups*

$$1 \to J \to G \xrightarrow{\phi} Q \to 1$$

such that for every virtually cyclic subgroup $V \leq Q$ the preimage $\phi^{-1}(V)$ satisfies the Farrell–Jones conjecture. Furthermore, assume that the assembly map

$$H_n^G(\underline{E}Q;\mathbb{K}_{\mathcal{B}}) \to K_n(\mathcal{B}[Q])$$

is split injective for every $n \in \mathbb{Z}$ and every additive Q–category $\mathcal{B}$. Then the K–theoretic assembly map

$$H_n^G(\underline{E}G;\mathbb{K}_{\mathcal{A}}) \to K_n(\mathcal{A}[G])$$

is split injective for every $n \in \mathbb{Z}$ and every additive G–category $\mathcal{A}$.

Proof Let $\mathcal{A}$ be an additive G–category. The fact that $\phi^{-1}(V)$ satisfies the Farrell–Jones conjecture for every virtually cyclic subgroup $V \leq Q$ implies that the natural map $H_n^G(E_{\mathcal{V}cyc}G;\mathbb{K}_{\mathcal{A}}) \to H_n^G(E_{\phi^*\mathcal{V}cyc}G;\mathbb{K}_{\mathcal{A}})$ is an isomorphism, by Bartels and Lück [6, Lemma 2.2], where $\phi^*\mathcal{V}cyc := \{K \leq G \mid \phi(K) \in \mathcal{V}cyc\}$. Here we used that the projection $E_{\mathcal{V}cyc}G \times E_{\phi^*\mathcal{V}cyc}G \to E_{\phi^*\mathcal{V}cyc}G$ is a model for the natural map $E_{\mathcal{V}cyc}G \to E_{\phi^*\mathcal{V}cyc}G$. Furthermore, the natural map $H_n^G(\underline{E}G;\mathbb{K}_{\mathcal{A}}) \to H_n^G(E_{\mathcal{V}cyc}G;\mathbb{K}_{\mathcal{A}})$ is split injective by Bartels [2]. Now the commutative diagram

$$\begin{tikzcd}
H_n^G(\underline{E}G;\mathbb{K}_{\mathcal{A}}) \arrow[r] \arrow[d, hook] & H_n^G(E_{\phi^*\mathcal{F}in}G;\mathbb{K}_{\mathcal{A}}) \arrow[d] \\
H_n^G(E_{\mathcal{V}cyc}G;\mathbb{K}_{\mathcal{A}}) \arrow[r, "\cong"] & H_n^G(E_{\phi^*\mathcal{V}cyc}G;\mathbb{K}_{\mathcal{A}})
\end{tikzcd}$$

implies that the map $H_n^G(\underline{E}G;\mathbb{K}_{\mathcal{A}}) \to H_n^G(E_{\phi^*\mathcal{F}in}G;\mathbb{K}_{\mathcal{A}})$ is split injective, where $\phi^*\mathcal{F}in := \{K \leq G \mid \phi(K) \in \mathcal{F}in\}$. By Bartels and Reich [11, Corollary 4.3] the split injectivity for Q implies that the assembly map $H_n^G(E_{\phi^*\mathcal{F}in}G;\mathbb{K}_{\mathcal{A}}) \to K_n(\mathcal{A}[G])$ is split injective. Combining these results yields the proposition. □

To apply the above proposition for the short exact sequence from the previous section, we need the following.

Lemma 4.2 *The class of virtually solvable groups is closed under group extensions.*

The idea of the proof is taken from math.stackexchange.com; see [13].

Proof Let
$$1 \to N \to G \xrightarrow{p} Q \to 1$$
be a short exact sequence, and let N and Q be virtually solvable. Let $Q' \leq Q$ be a solvable subgroup with $[Q : Q'] < \infty$; then $[G : p^{-1}(Q')] < \infty$. Thus we can assume that Q is solvable. We will first consider the case that N is finite. Since N is normal in G, G acts on N by conjugation, which induces a map $c\colon G \to \mathrm{Aut}(N)$. The centralizer $C_G(N)$ of N in G is the kernel of c. Since the class of solvable groups is closed under extension, and $C_G(N) \cap N$ is abelian, the exact sequence
$$1 \to C_G(N) \cap N \to C_G(N) \to p(C_G(N)) \to 1$$
shows that $C_G(N)$ is solvable. The group N is finite; thus $C_G(N)$ has finite index in G.

Now let N be any virtually solvable group. And let $\mathcal{S}$ be the set of all normal, solvable, finite-index subgroups of N, ordered by inclusion. This is not empty, and we can choose K to be a maximal element of S. For every $g \in G$ also $gKg^{-1}K$ is a solvable, normal, finite-index subgroup of N. Since K was maximal, it therefore has to be normal in G. From the short exact sequence
$$1 \to N/K \to G/K \to Q \to 1,$$
it follows from the first case that G/K is virtually solvable. Since K is solvable, the sequence
$$1 \to K \to G \to G/K \to 1$$
implies that G is virtually solvable. □

5 Proof of Theorem 1.1

For this section let H be a virtually connected Lie group and $G \leq H$ a finitely generated subgroup such that there exists a finite dimensional model for $\underline{E}G$. The proof of Theorem 1.1 follows easily from the statements of the previous section.

Proof of Theorem 1.1 Let $\Gamma := \mathrm{Ad}(G)$ be the image of G under $\mathrm{Ad}\colon H \to \mathrm{GL}(\mathfrak{h})$. Since $C_H(H_0) \cap G \cap H_0$ is contained in the center of G, the preimage of any unipotent subgroup U of $\mathrm{Ad}(G \cap H_0)$ is a nilpotent subgroup of $G \cap H_0$. By Corollary 3.6 and

Proposition 1.3 there is a bound on the Hirsch length of the nilpotent subgroups of $G \cap H_0$ and, in particular, there is a bound on the Hirsch length of U. Since $G \cap H_0$ has finite index in G, this implies that there also is a bound on the Hirsch length of the unipotent subgroups of Γ. Now we can apply the following:

[21, Corollary 3] *Let F be a field of characteristic zero, and let Γ be a finitely generated subgroup of $\mathrm{GL}_n(F)$ with a global upper bound on the Hirsch rank of its unipotent subgroups. Then the K–theoretic assembly map*

$$H_*^\Gamma(\underline{E}G; \mathbb{K}_{\mathcal{A}}) \to H_*^\Gamma(pt; \mathbb{K}_{\mathcal{A}}) \cong K_*(\mathcal{A}[\Gamma])$$

is split injective for every additive Γ–category $\mathcal{A}$.

Note that [21, Corollary 3] is stated only for rings instead of additive Γ–categories, but by [21, Theorem 8.1] it is true for any additive Γ–category.

Furthermore, by Wegner [30], every virtually solvable group satisfies the Farrell–Jones conjecture. Using this and Lemma 4.2, we see that the sequence

$$1 \to C_H(H_0) \cap G \to G \to \mathrm{Ad}(G) \to 1$$

satisfies the conditions of Proposition 4.1. Therefore, the assembly map

$$H_*^G(\underline{E}G; \mathbb{K}_{\mathcal{A}}) \to K_*(\mathcal{A}[G])$$

is split injective for every additive G–category $\mathcal{A}$. □

6 L–theory

Most of the statements from the previous sections also hold for L–theory. For the rest of the section let G be a finitely generated subgroup of a virtually connected Lie group H with a finite dimensional model for $\underline{E}G$, and let Q be the image of G under Ad: $H \to \mathrm{GL}(\mathfrak{h})$. Furthermore, let ϕ denote $\mathrm{Ad}|_G$, and let $\mathcal{A}$ be an additive G–category with involution. As above we obtain the commutative diagram

$$\begin{tikzcd}
H_n^G(\underline{E}G; \mathbb{L}_{\mathcal{A}}^{\langle -\infty \rangle}) \arrow[r] \arrow[d] & H_n^G(E_{\phi^*\mathcal{F}in}G; \mathbb{L}_{\mathcal{A}}^{\langle -\infty \rangle}) \arrow[d] \\
H_n^G(E_{\mathcal{V}cyc}G; \mathbb{L}_{\mathcal{A}}^{\langle -\infty \rangle}) \arrow[r, "\cong"] & H_n^G(E_{\phi^*\mathcal{V}cyc}G; \mathbb{L}_{\mathcal{A}}^{\langle -\infty \rangle}),
\end{tikzcd}$$

and the lower horizontal map is still an isomorphism by Bartels and Lück [6, Lemma 2.2] and Wegner [30]. But for the vertical map on the left to be injective we need that for every virtually cyclic subgroup $V \subseteq G$ there is an $i_0 \in \mathbb{N}$ such that for every $i \geq i_0$ we have $K_{-i}(\mathcal{A}[V]) = 0$; see Bartels [2]. Then it remains to show that

$$H_n^G(E_{\phi^*\mathcal{F}in}G; \mathbb{L}_{\mathcal{A}}^{\langle -\infty \rangle}) \to L_n^{\langle -\infty \rangle}(\mathcal{A}[G])$$

is split injective. By Bartels and Reich [11, Proposition 4.2 and Corollary 4.3], this follows if

$$H_n^Q(\underline{E}G; \mathbb{L}_{\mathrm{ind}_\phi \mathcal{A}}^{\langle -\infty \rangle}) \to L_n^{\langle -\infty \rangle}((\mathrm{ind}_\phi \mathcal{A})[Q])$$

is split injective. See [11] for the definition of $\mathrm{ind}_\phi \mathcal{A}$. To apply [21, Theorem 9.1] as above, we need the further assumption that for every finite subgroup $H \leq Q$ there is an $i_0 \in \mathbb{N}$ such that for every $i \geq i_0$ we have

$$0 = K_{-i}((\mathrm{ind}_\phi \mathcal{A})[H]) \cong K_{-i}(\mathcal{A}[\phi^{-1}(H)]).$$

Since $\phi^{-1}(H)$ is virtually abelian, we obtain the following version of the main theorem for L–theory.

Theorem 6.1 *Let G be a finitely generated subgroup of a virtually connected Lie group, and assume there exists an $N \in \mathbb{N}$ such that every finitely generated abelian subgroup of G has rank at most N. Let $\mathcal{A}$ be an additive G–category with involution. Assume further that for every virtually abelian subgroup H of G there is an $i_0 \in \mathbb{N}$ such that for every $i \geq i_0$ we have $K_{-i}(\mathcal{A}[H]) = 0$; then the L–theoretic assembly map*

$$H_n^G(\underline{E}G; \mathbb{L}_{\mathcal{A}}^{\langle -\infty \rangle}) \to L_n^{\langle -\infty \rangle}(\mathcal{A}([G])$$

is split injective.

For torsion-free groups G the *integral Novikov conjecture* states that the assembly map

$$H_n^G(EG; \mathbb{L}_{\mathbb{Z}}^{\langle -\infty \rangle}) \to L_n^{\langle -\infty \rangle}(\mathbb{Z}[G])$$

is injective. It is known that the integral Novikov conjecture is false for groups containing torsion. Following Ji [18], we say that G satisfies the *generalized integral Novikov conjecture* if the assembly maps

$$H_n^G(\underline{E}G; \mathbb{L}_{\mathbb{Z}}^{\langle -\infty \rangle}) \to L_n^{\langle -\infty \rangle}(\mathbb{Z}[G]), \quad H_n^G(\underline{\mathrm{E}}G; \mathbb{K}_{\mathbb{Z}}) \to K_n(\mathbb{Z}[G])$$

are injective. By Lück and Reich [24, Propostion 2.20], the relative rational assembly map

$$H_n^G(EG; \mathbb{L}_{\mathbb{Z}}^{\langle -\infty \rangle}) \otimes_{\mathbb{Z}} \mathbb{Q} \to H_n^G(\underline{E}G; \mathbb{L}_{\mathbb{Z}}^{\langle -\infty \rangle}) \otimes_{\mathbb{Z}} \mathbb{Q}$$

is injective. Observe that, since the $\mathbb{Z}/2$–Tate cohomology groups vanish rationally, there is no difference between the various decorations in L–theory as can be seen using the Rothenberg sequence. Therefore, by [24, Proposition 1.53], the injectivity of the rational assembly map

$$H_n^G(EG;\mathbb{L}_{\mathbb{Z}}^{\langle -\infty\rangle})\otimes_{\mathbb{Z}}\mathbb{Q}\to L_n^{\langle -\infty\rangle}(\mathbb{Z}[G])\otimes_{\mathbb{Z}}\mathbb{Q}$$

implies the Novikov conjecture about the homotopy invariance of higher signatures. In particular, the generalized integral Novikov conjecture implies the (classical) Novikov conjecture.

We will show that $K_{-n}(\mathbb{Z}[G])=0$ for $n>1$ and any virtually abelian group A. Therefore, Theorem 6.1 implies the generalized integral Novikov conjecture for the groups G appearing in the theorem; ie we get the following corollary.

Corollary 6.2 *Let G be a finitely generated subgroup of a virtually connected Lie group, and assume there exists an $N\in\mathbb{N}$ such that every finitely generated abelian subgroup of G has rank at most N. Then G satisfies the generalized integral Novikov conjecture.*

By Farrell and Jones [16, Theorem 2.1], for every virtually cyclic group V and $n>1$,

$$K_{-n}(\mathbb{Z}[V])=0.$$

Let G be a group and let X be a finite G–CW-complex with virtually cyclic stabilizers. By induction on the dimension of X we prove that

$$H_{-n}^G(X;\mathbb{K}_{\mathbb{Z}})=0$$

for every $n>1$. For $\dim X=0$, we have

$$H_{-n}^G(X;\mathbb{K}_{\mathbb{Z}})\cong\bigoplus_{x\in X}\mathbb{K}_{-n}(\mathbb{Z}[G_x])=0,$$

where the stabilizers G_x are virtually cyclic by assumption. Assume the above holds for m and let $\dim X=m+1$. Then we have the exact sequence

$$0=H_{-n}^G(X^{(m)};\mathbb{K}_{\mathbb{Z}})\to H_{-n}^G(X;\mathbb{K}_{\mathbb{Z}})\to H_{-n}^G(X,X^{(m)};\mathbb{K}_{\mathbb{Z}}),$$

and

$$H_{-n}^G(X,X^{(m)};\mathbb{K}_{\mathbb{Z}})\cong\bigoplus_{c\in C_m}\mathbb{K}_{-n-m-1}(\mathbb{Z}[G_c])=0,$$

where C_m denotes the set of m–cells of X and G_c the (virtually cyclic) stabilizer of the cell c. Since every virtually abelian group A satisfies the Farrell–Jones conjecture, we have

$$K_{-n}(\mathbb{Z}[A]) \cong H^{A}_{-n}(X;\mathbb{K}_{\mathbb{Z}}) \cong \operatorname*{colim}_{K} H^{A}_{-n}(AK;\mathbb{K}_{\mathbb{Z}}) = 0,$$

where X is an A–CW-complex model for $E_{\mathcal{V}cyc}A$, and the colimit is taken over all finite subcomplexes $K \subseteq X$.

7 Inheritance under colimits

In this section we want to show that Theorem 1.1 and Theorem 6.1 hold without the assumption that G is finitely generated.

By Bartels, Echterhoff and Lück [3, Lemma 2.4 and Lemma 6.2] for every system G_α of finitely generated subgroups of G such that $\operatorname{colim}_\alpha G_\alpha \cong G$, the assembly map

$$H^G_n(\underline{E}G;\mathbb{K}_{\mathcal{A}}) \to K_n(\mathcal{A}[G])$$

is the colimit of the assembly maps

$$H^{G_\alpha}_n(\underline{E}G_\alpha;\mathbb{K}_{\mathcal{A}}) \to K_n(\mathcal{A}[G_\alpha]),$$

for any additive G–category $\mathcal{A}$. The same statement holds in L–theory for any additive G–category with involution. Note that the statement in [3] is formulated for rings with G–action instead of additive G–categories, but the statement for G–categories holds in the same way. Furthermore, a finite-dimensional model for $\underline{E}G$ gives a finite-dimensional model for $\underline{E}G_\alpha$ by restricting the action to G_α. So taking the colimit over all finitely generated subgroups proves that injectivity holds without the assumption that G is finitely generated. For the construction of a splitting we need to see that the splittings for the finitely generated subgroups are natural with respect to the structure maps of the colimit. In the proof of Theorem 1.1 and Theorem 6.1 the assumption that G is finitely generated is only needed to apply [21, Corollary 3] and its L–theoretic analog, respectively. So it suffices to see that the splittings constructed in [21] are natural with respect to the structure maps of the colimit.

We will use the definitions of controlled categories and bounded mapping spaces from [21, Sections 5 and 7]. In the following let X denote a finite dimensional simplicial model for $\underline{E}G$. By Bartels, Farrell, Jones and Reich [5, Section 6] the assembly map

$$H^G_n(\underline{E}G;\mathbb{K}_{\mathcal{A}}) \to K_n(\mathcal{A}[G])$$

can be identified with the map

$$\operatorname*{colim}_{K\subseteq X \text{ fin.}} \pi_{n+1}(\mathbb{K}\mathcal{A}_G(GK)^\infty)^G \to \operatorname*{colim}_{K\subseteq X \text{ fin.}} \pi_n(\mathbb{K}\mathcal{A}_G(GK)_0)^G.$$

Now consider the diagram

$$\begin{tikzcd}
\operatorname*{colim}_{K\subseteq X \text{ fin.}} \pi_{n+1}(\mathbb{K}\mathcal{A}_G(GK)^\infty)^G \arrow[r] \arrow[d, "f"] & \operatorname*{colim}_{K\subseteq X \text{ fin.}} \pi_n(\mathbb{K}\mathcal{A}_G(GK)_0)^G \arrow[d, "j"] \\
\operatorname*{colim}_{K\subseteq X \text{ fin.}} \pi_{n+1}\operatorname{Map}^{bd}_G(X,\mathbb{K}\mathcal{A}_G(GK)^\infty) \arrow[r, "h"'] & \operatorname*{colim}_{K\subseteq X \text{ fin.}} \pi_n\operatorname{Map}^{bd}_G(X,\mathbb{K}\mathcal{A}_G(GK)_0).
\end{tikzcd}$$

By [21, Remark 7.7] the map f is an isomorphism and the map h is an isomorphism in the situation of [21, Corollary 3].

Let $\Gamma \to \Lambda$ be an injective group homomorphism. For every Λ–set J and every subcomplex $K \subseteq X$ we can define a map

$$\Big(\prod_J^{bd} \mathcal{A}_\Gamma(\Gamma K)^\infty\Big)^\Gamma \to \Big(\prod_J^{bd} \mathcal{A}_\Lambda(\Lambda K)^\infty\Big)^\Lambda$$

as follows. A controlled module $(M_j) \in (\prod_J^{bd} \mathcal{A}_\Gamma(\Gamma K))^\Gamma$ is sent to $(M'_j)_j$ with $(M'_j)_{h',x,t} := \bigoplus_{[h]\in\Lambda/\Gamma}(M_{h^{-1}j})_{h^{-1}h',h^{-1}x,t}$ and analogously on morphisms. This map is well defined since (M_j) is Γ–invariant. The above maps induce a map

$$\operatorname{Map}^{bd}_\Gamma(X,\mathbb{K}\mathcal{A}_\Gamma(\Gamma K)) \to \operatorname{Map}^{bd}_\Lambda(X,\mathbb{K}\mathcal{A}_\Lambda(\Lambda K))$$

for every finite subcomplex $K \subseteq X$, and in the special case where $J = \{\mathrm{pt}\}$ we obtain a map

$$(\mathbb{K}\mathcal{A}_\Gamma(\Gamma K)^\infty)^\Gamma \to (\mathbb{K}\mathcal{A}_\Lambda(\Lambda K)^\infty)^\Lambda.$$

The same maps can be constructed with $\mathcal{A}_\Gamma(\Gamma K)^\infty$ and $\mathcal{A}_\Lambda(\Lambda K)^\infty$ replaced by $\mathcal{A}_\Gamma(\Gamma K)_0$ and $\mathcal{A}_\Lambda(\Lambda K)_0$, respectively. So they induce maps from the above diagram for Γ to the same diagram for Λ. We will omit the technical proofs that the maps of the diagram are natural with respect to these maps and that under the identification with the assembly map they correspond to the structure maps of the colimit from [5]. This shows that the splitting $f^{-1}\circ h^{-1}\circ j$ is natural with respect to the structure maps of the colimit.

Now let us consider the L–theoretic version. For [21, Remark 7.7] it was used that the category

$$\Big(\prod_{j\in J} \mathbb{K}\mathcal{A}_G(GK)^\infty\Big)^G \simeq \prod_{[j]\in G\backslash J} \mathbb{K}\mathcal{A}_G^{G_j}(GK)^\infty$$

is weakly equivalent to

$$\Big(\mathbb{K}\prod_{j\in J}\mathcal{A}_G(GK)^\infty\Big)^G \simeq \mathbb{K}\prod_{[j]\in G\backslash J}\mathcal{A}_G^{G_j}(GK)^\infty$$

for every G–set J with finite stabilizers and every finite subcomplex $K\subseteq X$, where G_j is the stabilizer of $j\in J$. Let $H\le G$ be finite; then

$$K_n(\mathcal{A}_G^{G_j}(G/H)^\infty)\cong\prod_{G_j\backslash G/H}K_n(\mathcal{A}_G^{G_j}(G_j/(G_j\cap H))^\infty)\cong\prod_{G_j\backslash G/H}K_{n-1}(\mathcal{A}[G_j\cap H]).$$

If for each finite subgroup $H\le G$ there exists $N\in\mathbb{N}$ such that for each $n\ge N$ the groups $K_{-n}\mathcal{A}[H]$ vanish, then by induction on the cells this implies that for every finite subcomplex $K\subseteq X$ there exists $N\in\mathbb{N}$ such that for $n\ge N$ the groups $K_{-n}(\mathcal{A}_G^{G_j}(GK)^\infty)$ vanish. Therefore, under this assumption, L–theory commutes with the above product, and we get that the map

$$\phi\colon \mathrm{Map}_G^{bd}(X,\mathbb{L}\mathcal{A}_G(GK)^\infty)\to\mathrm{Map}_G(X,\mathbb{L}\mathcal{A}_G(GK)^\infty)$$

is an isomorphism. Also, under the above assumption,

$$\psi\colon (\mathbb{L}\mathcal{A}_G(GK)^\infty)^G\to\mathrm{Map}_G(X,\mathbb{L}\mathcal{A}_G(GK)^\infty)$$

is an isomorphism; see [21, Section 9]. Since ψ factors over ϕ, the map

$$(\mathbb{L}\mathcal{A}_G(GK)^\infty)^G\to\mathrm{Map}_G^{bd}(X,\mathbb{L}\mathcal{A}_G(GK)^\infty)$$

is an isomorphism as well. Therefore, we obtain the naturality of the splitting as in the case for K–theory.

References

[1] **R C Alperin**, **P B Shalen**, *Linear groups of finite cohomological dimension*, Bull. Amer. Math. Soc. 4 (1981) 339–341 MR609046

[2] **A C Bartels**, *On the domain of the assembly map in algebraic K–theory*, Algebr. Geom. Topol. 3 (2003) 1037–1050 MR2012963

[3] **A Bartels**, **S Echterhoff**, **W Lück**, *Inheritance of isomorphism conjectures under colimits*, from: "K–theory and noncommutative geometry", (G Cortiñas, J Cuntz, M Karoubi, R Nest, C A Weibel, editors), Eur. Math. Soc., Zürich (2008) 41–70 MR2513332

[4] **A Bartels**, **F T Farrell**, **W Lück**, *The Farrell–Jones conjecture for cocompact lattices in virtually connected Lie groups*, J. Amer. Math. Soc. 27 (2014) 339–388 MR3164984

[5] **A Bartels**, **T Farrell**, **L Jones**, **H Reich**, *On the isomorphism conjecture in algebraic K–theory*, Topology 43 (2004) 157–213 MR2030590

[6] **A Bartels**, **W Lück**, *Isomorphism conjecture for homotopy K–theory and groups acting on trees*, J. Pure Appl. Algebra 205 (2006) 660–696 MR2210223

[7] **A Bartels**, **W Lück**, *The Borel conjecture for hyperbolic and* $\mathrm{CAT}(0)$*–groups*, Ann. of Math. 175 (2012) 631–689 MR2993750

[8] **A Bartels**, **W Lück**, *Geodesic flow for* $\mathrm{CAT}(0)$*–groups*, Geom. Topol. 16 (2012) 1345–1391 MR2967054

[9] **A Bartels**, **W Lück**, **H Reich**, *Equivariant covers for hyperbolic groups*, Geom. Topol. 12 (2008) 1799–1882 MR2421141

[10] **A Bartels**, **W Lück**, **H Reich**, *The K–theoretic Farrell–Jones conjecture for hyperbolic groups*, Invent. Math. 172 (2008) 29–70 MR2385666

[11] **A Bartels**, **H Reich**, *Coefficients for the Farrell–Jones conjecture*, Adv. Math. 209 (2007) 337–362 MR2294225

[12] **G Carlsson**, **B Goldfarb**, *On homological coherence of discrete groups*, J. Algebra 276 (2004) 502–514 MR2058455

[13] **Y de Cornulier**, *Answer to "If N and G/N are virtually solvable, then G is virtually solvable?"* (2013) Available at `http://math.stackexchange.com/q/437063`

[14] **J F Davis**, **W Lück**, *Spaces over a category and assembly maps in isomorphism conjectures in K– and L–theory*, K–Theory 15 (1998) 201–252 MR1659969

[15] **F T Farrell**, **L E Jones**, *Isomorphism conjectures in algebraic K–theory*, J. Amer. Math. Soc. 6 (1993) 249–297 MR1179537

[16] **F T Farrell**, **L E Jones**, *The lower algebraic K–theory of virtually infinite cyclic groups*, K–Theory 9 (1995) 13–30 MR1340838

[17] **L Ji**, *Asymptotic dimension and the integral K–theoretic Novikov conjecture for arithmetic groups*, J. Differential Geom. 68 (2004) 535–544 MR2144540

[18] **L Ji**, *Integral Novikov conjectures and arithmetic groups containing torsion elements*, Comm. Anal. Geom. 15 (2007) 509–533 MR2379803

[19] **H Kammeyer**, **W Lück**, **H Rüping**, *The Farrell–Jones conjecture for arbitrary lattices in virtually connected Lie groups*, preprint (2014) `arXiv:1401.0876v1` To appear in Geom. Topol.

[20] **G G Kasparov**, *Equivariant KK–theory and the Novikov conjecture*, Invent. Math. 91 (1988) 147–201 MR918241

[21] **D Kasprowski**, *On the K–theory of groups with finite decomposition complexity*, Proc. Lond. Math. Soc. 110 (2015) 565–592 MR3342098

[22] **W Lück**, *The type of the classifying space for a family of subgroups*, J. Pure Appl. Algebra 149 (2000) 177–203 MR1757730

[23] **W Lück**, *Survey on classifying spaces for families of subgroups*, from: "Infinite groups: Geometric, combinatorial and dynamical aspects", (L Bartholdi, T Ceccherini-Silberstein, T Smirnova-Nagnibeda, A Zuk, editors), Progr. Math. 248, Birkhäuser, Basel (2005) 269–322 MR2195456

[24] **W Lück**, **H Reich**, *The Baum–Connes and the Farrell–Jones conjectures in K– and L–theory*, from: "Handbook of K–theory, Vol. 1, 2", (E M Friedlander, D R Grayson, editors), Springer, Berlin (2005) 703–842 MR2181833

[25] **W Möhres**, *Torsion-free nilpotent groups with bounded ranks of the abelian subgroups*, from: "Group theory", (O H Kegel, F Menegazzo, G Zacher, editors), Lecture Notes in Math. 1281, Springer, Berlin (1987) 115–117 MR921696

[26] **E K Pedersen**, **C A Weibel**, *A nonconnective delooping of algebraic K–theory*, from: "Algebraic and geometric topology", (A Ranicki, N Levitt, F Quinn, editors), Lecture Notes in Math. 1126, Springer, Berlin (1985) 166–181 MR802790

[27] **G Skandalis**, **J L Tu**, **G Yu**, *The coarse Baum–Connes conjecture and groupoids*, Topology 41 (2002) 807–834 MR1905840

[28] **R B Warfield, Jr**, *Nilpotent groups*, Lecture Notes in Mathematics 513, Springer, Berlin (1976) MR0409661

[29] **C Wegner**, *The K–theoretic Farrell–Jones conjecture for CAT(0)–groups*, Proc. Amer. Math. Soc. 140 (2012) 779–793

[30] **C Wegner**, *The Farrell–Jones conjecture for virtually solvable groups*, preprint (2013) `arXiv:1308.2432v2` To appear in J. Topol. Available at `http://dx.doi.org/10.1112/jtopol/jtv026`

Max-Planck-Institut für Mathematik
Vivatsgasse 7, D-53111 Bonn, Germany

`kasprowski@mpim-bonn.mpg.de`

`http://guests.mpim-bonn.mpg.de/kasprowski/`

Received: 24 September 2014 Revised: 9 February 2015

Algebraic & Geometric Topology 15 (2015) 3485–3534

McCool groups of toral relatively hyperbolic groups

VINCENT GUIRARDEL
GILBERT LEVITT

The outer automorphism group $\mathrm{Out}(G)$ of a group G acts on the set of conjugacy classes of elements of G. McCool proved that the stabilizer $\mathrm{Mc}(\mathcal{C})$ of a finite set of conjugacy classes is finitely presented when G is free. More generally, we consider the group $\mathrm{Mc}(\mathcal{H})$ of outer automorphisms Φ of G acting trivially on a family of subgroups H_i, in the sense that Φ has representatives α_i that are equal to the identity on H_i.

When G is a toral relatively hyperbolic group, we show that these two definitions lead to the same subgroups of $\mathrm{Out}(G)$, which we call "McCool groups" of G. We prove that such McCool groups are of type VF (some finite-index subgroup has a finite classifying space). Being of type VF also holds for the group of automorphisms of G preserving a splitting of G over abelian groups.

We show that McCool groups satisfy a uniform chain condition: there is a bound, depending only on G, for the length of a strictly decreasing sequence of McCool groups of G. Similarly, fixed subgroups of automorphisms of G satisfy a uniform chain condition.

20F28; 20F65, 20F67

1 Introduction

Mapping class groups of punctured surfaces may be viewed as subgroups of $\mathrm{Out}(F_n)$ for some n (with F_n denoting the free group of rank n). Indeed, they consist of automorphisms of F_n fixing conjugacy classes corresponding to punctures. More generally, the group of automorphisms of F_n fixing a finite number of conjugacy classes was studied by McCool [30], who proved in particular that such groups are finitely presented. We therefore define:

Definition 1.1 Let G be a group. Let $\mathcal{C}$ be a set of conjugacy classes $[c_i]$ of elements of G. We denote by $\mathrm{Mc}(\mathcal{C})$ the subgroup of $\mathrm{Out}(G)$ consisting of outer automorphisms fixing each $[c_i]$. If $\mathcal{C}$ is finite, we say that $\mathrm{Mc}(\mathcal{C})$ is an *elementary McCool group* of G (or of $\mathrm{Out}(G)$).

Published: 12 January 2016 DOI: 10.2140/agt.2015.15.3485

Work on automorphisms suggests a more general definition:

Definition 1.2 Let G be a group. Let $\mathcal{H} = \{H_i\}$ be an arbitrary family of subgroups of G. We say that $\varphi \in \mathrm{Aut}(G)$ and its image $\Phi \in \mathrm{Out}(G)$ *act trivially on* $\mathcal{H}$ if φ acts on each H_i as conjugation by some $g_i \in G$. Note that Φ acts trivially if and only if it has representatives $\varphi_i \in \mathrm{Aut}(G)$ with φ_i equal to the identity on H_i.

We denote by $\mathrm{Mc}(\mathcal{H})$ or $\mathrm{Mc}_G(\mathcal{H})$ the subgroup of $\mathrm{Out}(G)$ consisting of all Φ acting trivially on $\mathcal{H}$.

If $\mathcal{H}$ is a finite family of finitely generated subgroups, we say that $\mathrm{Mc}(\mathcal{H})$ is a *McCool group* of G (or of $\mathrm{Out}(G)$).

Elementary McCool groups correspond to McCool groups with $\mathcal{H}$ a finite family of cyclic groups. $\mathrm{Mc}(\mathcal{H})$ does not change if we replace the H_i by conjugate subgroups, so it is really associated to a family of conjugacy classes of subgroups.

For a topological analogy, one may think of $\mathrm{Mc}(\mathcal{H})$ as the group of automorphisms of $G = \pi_1(X)$ induced by homeomorphisms of X equal to the identity on subspaces Y_i with $\pi_1(Y_i) = H_i$.

McCool groups are relevant for automorphisms for the following reason (see Guirardel and Levitt [25]). Consider a splitting of a group $\widehat{G}$ as a graph of groups in which G is a vertex group and the H_i are the incident edge groups. Then any element of $\mathrm{Mc}_G(\mathcal{H})$ extends "by the identity" to an automorphism of $\widehat{G}$. Topologically, if X is a vertex space in a graph of spaces $\widehat{X}$ and edge spaces are attached to subspaces $Y_i \subset X$, then any homeomorphism of X equal to the identity on the Y_i extends to $\widehat{X}$ by the identity.

In this paper we will consider McCool groups when G is a *toral relatively hyperbolic group*: G is torsion-free and hyperbolic relative to a finite set of finitely generated abelian subgroups. This includes in particular torsion-free hyperbolic groups, limit groups and groups acting freely on $\mathbb{R}^n$–trees.

We will show (Corollary 1.6) that in this case any $\mathrm{Mc}(\mathcal{H})$ is an elementary McCool group $\mathrm{Mc}(\mathcal{C})$; in other words, it is equivalent for a subgroup of $\mathrm{Out}(G)$ to be an elementary McCool group $\mathrm{Mc}(\mathcal{C})$, or to be a McCool group $\mathrm{Mc}(\mathcal{H})$ with $\mathcal{H}$ a finite family of finitely generated groups, or to be $\mathrm{Mc}(\mathcal{H})$ with $\mathcal{H}$ arbitrary. We will not always make the distinction in the statements given below.

It was proved by McCool [30] that (elementary) McCool groups of a free group are finitely presented. Culler and Vogtmann [9, Corollary 6.1.4] proved that they are of *type* VF: they have a finite-index subgroup with a finite classifying space (ie there exists a classifying space which is a finite complex). We proved in [25] that $\mathrm{Out}(G)$

is of type VF if G is toral relatively hyperbolic (in particular, $\mathrm{Out}(G)$ is virtually torsion-free). Our first main results extend this to certain naturally defined subgroups of $\mathrm{Out}(G)$.

Theorem 1.3 *If G is a toral relatively hyperbolic group, then any McCool group $\mathrm{Mc}(\mathcal{H}) \subset \mathrm{Out}(G)$ is of type* VF.

Theorem 1.4 *If G is a toral relatively hyperbolic group and T is a simplicial tree on which G acts with abelian edge stabilizers, then the group of automorphisms $\mathrm{Out}(T) \subset \mathrm{Out}(G)$ leaving T invariant is of type* VF.

Our most general result in this direction (Corollary 6.3) combines these two theorems; it implies in particular that $\mathrm{Mc}(\mathcal{H}) \cap \mathrm{Out}(T)$ *is of type* VF *if T is as above and $\mathcal{H}$ is any family of subgroups each of which fixes a point in T.*

Remark Some of these results may be extended to groups which are hyperbolic relative to virtually polycyclic subgroups, but with the weaker conclusion that the automorphism groups are of type F_∞ (see Guirardel and Levitt [17]). On the other hand, one can show that, if there exists a hyperbolic group which is not residually finite, then there exists a hyperbolic group with $\mathrm{Out}(G)$ not virtually torsion-free (hence not VF).

Our second main result is the following:

Theorem 1.5 *Let G be a toral relatively hyperbolic group. McCool groups of G satisfy a uniform chain condition: there exists $C = C(G)$ such that, if*

$$\mathrm{Mc}(\mathcal{H}_0) \supsetneq \mathrm{Mc}(\mathcal{H}_1) \supsetneq \cdots \supsetneq \mathrm{Mc}(\mathcal{H}_p)$$

is a strictly decreasing chain of McCool groups in $\mathrm{Out}(G)$, *then* $p \leq C$.

This is based, among other things, on the vertex finiteness we proved in [24]: if G is toral relatively hyperbolic, then all vertex groups occurring in splittings of G over abelian groups lie in finitely many isomorphism classes.

The chain condition, proved in Section 5 for McCool groups $\mathrm{Mc}(\mathcal{H})$ with $\mathcal{H}$ a finite family of finitely generated groups, implies:

Corollary 1.6 *Let G be a toral relatively hyperbolic group. If $\mathcal{H}$ is a (possibly infinite) family of (possibly infinitely generated) subgroups $H_i \subset G$, there exists a finite set of conjugacy classes $\mathcal{C}$ such that* $\mathrm{Mc}(\mathcal{H}) = \mathrm{Mc}(\mathcal{C})$. *In particular, any* $\mathrm{Mc}(\mathcal{H})$ *is a McCool group and any McCool group is an elementary McCool group* $\mathrm{Mc}(\mathcal{C})$.

The chain condition also implies that no McCool group $\mathrm{Mc}(\mathcal{H}) \subset \mathrm{Out}(G)$ is conjugate to a proper subgroup. Note, however, that McCool groups may fail to be co-Hopfian (they may be isomorphic to proper subgroups). To illustrate the variety of McCool groups, we show:

Proposition 1.7 $\mathrm{Out}(F_n)$ *contains infinitely many non-isomorphic McCool groups if* $n \geq 4$*; it contains infinitely many non-conjugate McCool groups if* $n \geq 3$.

It may be shown that the bounds on n are sharp (see the appendix). We will also show in the appendix that, if G is a torsion-free, *one-ended* hyperbolic group, then $\mathrm{Out}(G)$ only contains finitely many McCool groups up to conjugacy.

Say that $J \subset G$ is *a fixed subgroup* if there is a family of automorphisms $\alpha_i \in \mathrm{Aut}(G)$ such that $J = \bigcap_i \mathrm{Fix}\,\alpha_i$, with $\mathrm{Fix}\,\alpha = \{g \in G \mid \alpha(g) = g\}$. The chain condition also implies:

Theorem 1.8 *Let* G *be a toral relatively hyperbolic group. There is a constant* $c = c(G)$ *such that, if* $J_0 \subsetneq J_1 \subsetneq \cdots \subsetneq J_p$ *is a strictly ascending chain of fixed subgroups, then* $p \leq c$.

This was proved by Martino and Ventura [29] for G free, with $c(F_n) = 2n$. In [18], we will apply Theorems 1.3 and 1.8 to the study of stabilizers for the action of $\mathrm{Out}(G)$ on spaces of $\mathbb{R}$–trees.

As explained above, one does not get new groups by allowing the set $\mathcal{C}$ in Definition 1.1 to be infinite or by considering arbitrary subgroups as in Definition 1.2. The following definition provides a genuine generalization.

Definition 1.9 Let G be a group, and $\mathcal{C}$ a finite set of conjugacy classes $[c_i]$. We write $\mathcal{C}^{-1}$ for the set of classes $[c_i^{-1}]$. Let $\widehat{\mathrm{Mc}}(\mathcal{C})$ be the subgroup of $\mathrm{Out}(G)$ consisting of automorphisms leaving $\mathcal{C} \cup \mathcal{C}^{-1}$ globally invariant; it contains $\mathrm{Mc}(\mathcal{C})$ as a normal subgroup of finite index. We say that $\widehat{\mathrm{Mc}}(\mathcal{C})$ is an *extended elementary McCool group* of G.

More generally, if $\mathcal{H}$ is a finite family of subgroups, one can define finite extensions of $\mathrm{Mc}(\mathcal{H})$ by allowing the H_i to be permuted or the action on H_i to be only "almost" trivial.

Proposition 1.10 *Given a toral relatively hyperbolic group* G*, there exists a number* C *such that, if a subgroup* $\widehat{M} \subset \mathrm{Out}(G)$ *contains a group* $\mathrm{Mc}(\mathcal{H})$ *with finite index, then the index* $[\widehat{M} : \mathrm{Mc}(\mathcal{H})]$ *is bounded by* C.

In particular, for $\mathcal{C}$ *finite, the index of* $\mathrm{Mc}(\mathcal{C})$ *in* $\widehat{\mathrm{Mc}}(\mathcal{C})$ *is bounded by a constant depending only on* G.

It follows that extended elementary McCool groups satisfy a uniform chain condition as in Theorem 1.5 (see Corollary 6.4). We also have:

Corollary 1.11 *Let G be a toral relatively hyperbolic group. Let A be any subgroup of $\mathrm{Out}(G)$ and let $\mathcal{C}_A$ be the (possibly infinite) set of conjugacy classes of G whose A–orbit is finite. The image of A in the group of permutations of $\mathcal{C}_A$ is finite and its order is bounded by a constant depending only on G. In other words, there is a subgroup $A_0 \subset A$ of bounded finite index such that every conjugacy class in G is fixed by A_0 or has infinite orbit under A_0.*

When G is free, one may take for A_0 the intersection of A with a fixed finite-index subgroup of $\mathrm{Out}(G)$ (independent of A); see Handel and Mosher [26].

One may also consider subgroups of $\mathrm{Aut}(G)$.

Definition 1.12 Let $\mathcal{H}$ be a family of (conjugacy classes of) subgroups, and $H_0 < G$ another subgroup. Let $\mathrm{Ac}(\mathcal{H}, H_0) \subset \mathrm{Aut}(G)$ be the group of automorphisms acting trivially on $\mathcal{H}$ (in the sense of Definition 1.2) and fixing the elements of H_0.

Proposition 1.13 *If G is a non-abelian, toral relatively hyperbolic group, then the group $\mathrm{Ac}(\mathcal{H}, H_0)$ is an extension*

$$1 \longrightarrow K \longrightarrow \mathrm{Ac}(\mathcal{H}, H_0) \longrightarrow \mathrm{Mc}(\mathcal{H}') \longrightarrow 1,$$

where $\mathrm{Mc}(\mathcal{H}') \subset \mathrm{Out}(G)$ is a McCool group and K is the centralizer of H_0 (isomorphic to G or to $\mathbb{Z}^n$ for some $n \geq 0$).

Corollary 1.14 *Theorems 1.3 and 1.5 also hold in $\mathrm{Aut}(G)$: groups of the form $\mathrm{Ac}(\mathcal{H}, H_0)$ are of type VF and satisfy a uniform chain condition.*

Theorems 1.3 and 1.4 are proved in Section 3 and Theorem 1.5 is proved in Section 5. All other results are proved in Section 6.

Acknowledgements Guirardel acknowledges support from ANR-11-BS01-013, the Institut Universitaire de France and the Lebesgue Center of Mathematics. Levitt acknowledges support from ANR-10-BLAN-116-03.

2 Preliminaries

In this paper, G will always denote a toral relatively hyperbolic group. Any non-trivial abelian subgroup A of G is contained in a unique maximal abelian subgroup. The maximal abelian subgroups are malnormal (G is CSA), finitely generated and there are finitely many non-cyclic ones up to conjugacy. Two subgroups of A which are conjugate in G are equal.

The center of a group H will be denoted by $Z(H)$. We write $N_K(H)$ for the normalizer of a group H in a group K, with $N(H) = N_G(H)$. Centralizers are called $Z_K(H)$.

We say that $\Phi \in \mathrm{Out}(G)$ preserves a subgroup H, or leaves H invariant, if its representatives $\varphi \in \mathrm{Aut}(G)$ map H to a conjugate. If $\varphi \in \mathrm{Aut}(G)$ equals the identity on H, we say that it fixes H.

Definition 2.1 If $\mathcal{H}$ is a family of subgroups, we let $\mathrm{Out}(G;\mathcal{H}) \subset \mathrm{Out}(G)$ be the group of automorphisms preserving each $H \in \mathcal{H}$, and $\widehat{\mathrm{Out}}(G;\mathcal{H})$ the group of automorphisms preserving $\mathcal{H}$ globally (possibly permuting groups in $\mathcal{H}$).

We denote by

$$\mathrm{Out}(G;\mathcal{H}^{(\mathrm{t})}) = \mathrm{Mc}(\mathcal{H}) \subset \mathrm{Out}(G)$$

the group of automorphisms acting trivially on groups in $\mathcal{H}$ (as in Definition 1.2).

We write

$$\begin{aligned} \mathrm{Out}(G;\mathcal{H}^{(\mathrm{t})},\mathcal{K}) &:= \mathrm{Out}(G;\mathcal{H}^{(\mathrm{t})}) \cap \mathrm{Out}(G;\mathcal{K}), \\ \mathrm{Out}(G;\mathcal{H},\mathcal{K}) &:= \mathrm{Out}(G;\mathcal{H} \cup \mathcal{K}). \end{aligned}$$

Remark $\mathrm{Out}(G;\mathcal{H}^{(\mathrm{t})})$ and $\mathrm{Mc}(\mathcal{H})$ denote the same group. The notation $\mathrm{Out}(G;\mathcal{H}^{(\mathrm{t})})$ is more flexible and will be convenient in Section 3.

We will often view a set of conjugacy classes $\mathcal{C} = \{[c_i]\}$ as a family of cyclic subgroups $\mathcal{H} = \{\langle c_i\rangle\}$ since $\mathrm{Mc}(\mathcal{C}) = \mathrm{Mc}(\mathcal{H})$. Note that $\mathrm{Out}(G;\mathcal{H})$ is larger than $\mathrm{Mc}(\mathcal{C}) = \mathrm{Mc}(\mathcal{H})$ since c_i may sent to a conjugate of c_i^{-1}.

For example, suppose that $H < G = \mathbb{Z}^n$ is the subgroup generated by the first k basis elements and $\mathcal{H} = \{H\}$. Then $\mathrm{Out}(G) = \mathrm{GL}(n,\mathbb{Z})$, the group $\mathrm{Out}(G;\mathcal{H})$ consists of block triangular matrices, and $\mathrm{Out}(G;\mathcal{H}^{(\mathrm{t})}) = \mathrm{Mc}(\mathcal{H})$ is the group of matrices fixing the first k basis vectors.

There are inclusions $\mathrm{Out}(G;\mathcal{H}^{(\mathrm{t})}) \subset \mathrm{Out}(G;\mathcal{H}) \subset \widehat{\mathrm{Out}}(G;\mathcal{H})$. Note that $\mathrm{Out}(G;\mathcal{H}^{(\mathrm{t})})$ has finite index in $\mathrm{Out}(G;\mathcal{H})$ and $\widehat{\mathrm{Out}}(G;\mathcal{H})$ if $\mathcal{H}$ is a finite family of cyclic groups.

Given a family $\mathcal{H}$ and a subgroup J, we denote by $\mathcal{H}_{|J}$ the J–conjugacy classes of subgroups of J conjugate to a group of $\mathcal{H}$. We view $\mathcal{H}_{|J}$ as a family of subgroups of J, each defined up to conjugacy in J. In the next subsection we will define a closely related notion $\mathcal{H}_{\|J}$ when $J = G_v$ is a vertex stabilizer in a tree.

If $\mathcal{C}$ is a set of conjugacy classes $[c_i]$, viewed as a set of cyclic subgroups, $\mathcal{C}_{|J}$ is the set of J–conjugacy classes of elements of J representing elements in $\mathcal{C}$.

Now suppose that subgroups of J which are conjugate in G are conjugate in J; this holds for instance if J is malnormal (in particular if J is a free factor) and also if J is abelian. In this case we may view $\mathcal{H}_{|J}$ as a subset of $\mathcal{H}$; it is finite if $\mathcal{H}$ is.

2.1 Trees and splittings

A tree will be a simplicial tree T with an action of G without inversions. A tree T is *relative to* $\mathcal{H}$ (resp. $\mathcal{C}$) if any group in $\mathcal{H}$ (resp. any element representing a class in $\mathcal{C}$) fixes a point in T.

Two trees are considered to be the same if there is a G–equivariant isomorphism between them. In this paper, all trees will have abelian edge stabilizers.

Unless mentioned otherwise, we assume that the action is *minimal* (there is no proper invariant subtree). We usually assume that there is *no redundant vertex* (if $T \setminus \{x\}$ has two components, some $g \in G$ interchanges them). If a finitely generated subgroup $H \subset G$ acts on T with no global fixed point, there is a smallest H–invariant subtree, called the *minimal subtree* of H.

The tree T is *trivial* if there is a global fixed point (minimality then implies that T is a point). An element or a subgroup of G is *elliptic* if it fixes a point in T. Conjugates of elliptic subgroups are elliptic, so we also consider elliptic conjugacy classes.

An action of G on a tree T gives rise to a splitting of G, ie a decomposition of G as the fundamental group of the quotient graph of groups $\Gamma = T/G$. Conversely, T is the Bass–Serre tree of Γ. All definitions given here apply to both splittings and trees. In particular, a splitting is relative to $\mathcal{H}$ if every $H \in \mathcal{H}$ has a conjugate contained in a vertex group.

Minimality implies that the graph Γ is finite. There is a one-to-one correspondence between vertices (resp. edges) of Γ and G–orbits of vertices (resp. edges) of T. We denote by V the set of vertices of Γ and by G_v the group carried by a vertex $v \in V$. We also view v as a vertex of T with stabilizer G_v. Similarly, we denote by e an edge of Γ or T, by G_e the corresponding group (always abelian in this paper) and by E the set of non-oriented edges of Γ.

Edge groups being abelian, hence relatively quasiconvex, every vertex group G_v is toral relatively hyperbolic (see for instance [25]).

The edge groups carried by edges of Γ incident to a given vertex v will be called the *incident edge groups* of G_v. We denote by Inc_v the family of incident edge groups (we view it as a finite family of subgroups of G_v, each well defined up to conjugacy).

If $\mathcal{H}$ is a finite family of subgroups of G and v is a vertex stabilizer of T, we denote by $\mathcal{H}_{\|G_v}$ the family of subgroups $H \subset G_v$ which are conjugate to a group of $\mathcal{H}$ and fix no other point in T. Two such groups are conjugate in G_v if they are conjugate in G (see [25, Lemma 2.2], where the notation $\mathcal{H}_{|G_v}$ is used instead), so we may also view $\mathcal{H}_{\|G_v}$ as a subset of $\mathcal{H}$ (it contains some of the groups of $\mathcal{H}$ having a conjugate in G_v), or as a finite family of subgroups of G_v, each well-defined up to conjugacy ($\mathcal{H}_{\|G_v}$ may be smaller than $\mathcal{H}_{|G_v}$ because we do not include subgroups of edge groups).

Any splitting of G_v relative to Inc_v extends to a splitting of G. If T is relative to $\mathcal{H}$, any splitting of G_v relative to $\mathrm{Inc}_v \cup \mathcal{H}_{\|G_v}$ is relative to $\mathcal{H}_{|G_v}$ and extends to a splitting of G relative to $\mathcal{H}$.

If $\mathcal{C}$ is a set of conjugacy classes, we view $\mathcal{C}_{\|G_v}$ as the subset of $\mathcal{C}$ consisting of classes having a representative that fixes v and no other vertex. In particular, $\mathcal{C}_{\|G_v}$ is finite if $\mathcal{C}$ is.

A tree T' is a *collapse* of T if it is obtained from T by collapsing each edge in a certain G–invariant collection to a point; conversely, we say that T *refines* T'. In terms of graphs of groups, one passes from $\Gamma = T/G$ to $\Gamma' = T'/G$ by collapsing edges; for each vertex $v' \in \Gamma'$, the vertex group $G_{v'}$ is the fundamental group of the graph of groups $\Gamma_{v'}$ occurring as the preimage of v' in Γ.

All maps between trees will be G–equivariant. Given two trees T and T', we say that T *dominates* T' if there is a map $f\colon T \to T'$ or, equivalently, if every subgroup which is elliptic in T is also elliptic in T'; in particular, T dominates any collapse T'. We sometimes say that f is a *domination map.* Minimality implies that it is onto.

Two trees belong to the same *deformation space* if they dominate each other. In other words, a deformation space $\mathcal{D}$ is the set of all trees having a given family of subgroups as their elliptic subgroups. We say that $\mathcal{D}$ dominates $\mathcal{D}'$ if trees in $\mathcal{D}$ dominate those in $\mathcal{D}'$.

2.2 JSJ decompositions [21; 22]

Let $\mathcal{H}$ be a family of subgroups of G. Recall that a tree T is *relative* to $\mathcal{H}$ if all groups of $\mathcal{H}$ are elliptic in T.

We denote by $\mathcal{H}^{+\mathrm{ab}}$ the family obtained by adding to $\mathcal{H}$ all non-cyclic abelian subgroups of G.

The group G is *freely indecomposable* relative to $\mathcal{H}$ if it does not split over the trivial group relative to $\mathcal{H}$; equivalently, G cannot be written non-trivially as $A*B$ with every group of $\mathcal{H}$ contained in a conjugate of A or B (if $\mathcal{H}$ is trivial, we also require $G \neq \mathbb{Z}$, as we consider $\mathbb{Z}$ as freely decomposable). Non-cyclic abelian groups being one-ended, being freely indecomposable relative to $\mathcal{H}$ is the same as being so relative to $\mathcal{H}^{+\mathrm{ab}}$.

Let $\mathcal{A}$ be another family of subgroups (in this paper, $\mathcal{A}$ consists of the trivial group or is the family of all abelian subgroups). Once $\mathcal{H}$ and $\mathcal{A}$ are fixed, we only consider trees relative to $\mathcal{H}$, with edge stabilizers in $\mathcal{A}$. We also assume that trees are minimal.

A tree T (with edge stabilizers in $\mathcal{A}$, relative to $\mathcal{H}$) is *universally elliptic* (with respect to $\mathcal{H}$) if its edge stabilizers are elliptic in every tree. It is a *JSJ tree* if, moreover, it dominates every universally elliptic tree. The set of JSJ trees is called the *JSJ deformation space* (over $\mathcal{A}$ relative to $\mathcal{H}$). All JSJ trees have the same vertex stabilizers, provided one restricts to stabilizers not in $\mathcal{A}$.

When $\mathcal{A}$ consists of the trivial group, the JSJ deformation space is called the *Grushko deformation space* (relative to $\mathcal{H}$). The group G has a relative Grushko decomposition $G = G_1 * \cdots * G_n * F_p$, with F_p free, every $H \in \mathcal{H}$ contained in some G_i (up to conjugacy) and G_i freely indecomposable relative to $\mathcal{H}_{|G_i}$. Vertex stabilizers of the relative Grushko deformation space $\mathcal{D}$ are precisely conjugates of the G_i. The deformation space is trivial (it only contains the trivial tree) if and only if G is freely indecomposable relative to $\mathcal{H}$. Writing $\mathcal{G} = \{G_1, \dots, G_n\}$, note that $\mathrm{Out}(G; \mathcal{H} \cup \mathcal{G})$ has finite index in $\mathrm{Out}(G; \mathcal{H})$, because automorphisms in $\mathrm{Out}(G; \mathcal{H})$ leave $\mathcal{D}$ invariant and therefore permute the G_i (up to conjugacy).

Now suppose that $\mathcal{A}$ consists of all abelian subgroups and G is freely indecomposable relative to a family $\mathcal{H}$. Then [22, Theorem 11.1] the JSJ deformation space relative to $\mathcal{H}^{+\mathrm{ab}}$ contains a preferred tree T_{can}; this tree is invariant under $\widehat{\mathrm{Out}}(G; \mathcal{H})$ (the group of automorphisms preserving $\mathcal{H}$).

It is obtained as a *tree of cylinders*. We describe this construction in the case that will be needed here (see [23, Proposition 6.3] for details). Let T be any tree with non-trivial abelian edge stabilizers, relative to all non-cyclic abelian subgroups. Say that two edges e and e' belong to the same cylinder if their stabilizers commute. Cylinders are subtrees intersecting in at most one point.

The tree of cylinders T_c is defined as follows. It is bipartite, with vertex set $\mathcal{V}_0 \cup \mathcal{V}_1$. Vertices in $\mathcal{V}_0$ are vertices of T belonging to at least two cylinders. Vertices in $\mathcal{V}_1$ are cylinders of T. A vertex $v \in \mathcal{V}_0$ is joined to a vertex $Y \in \mathcal{V}_1$ if v (viewed as a vertex of T) belongs to Y (viewed as a subtree of T). Equivalently, one obtains T_c from T by replacing each cylinder Y by the cone on its boundary (points of Y belonging to at least one other cylinder).

The tree T_c only depends on the deformation space $\mathcal{D}$ containing T and it belongs to $\mathcal{D}$. Like T, it has non-trivial abelian edge stabilizers and is relative to all non-cyclic abelian subgroups. It is minimal if T is minimal, but vertices in $\mathcal{V}_1$ may be redundant vertices.

The stabilizer of a vertex $v_1 \in \mathcal{V}_1$ is a maximal abelian subgroup. The stabilizer of a vertex in $\mathcal{V}_0$ is non-abelian and is the stabilizer of a vertex of T. The stabilizer of an edge v_0v_1 with $v_i \in \mathcal{V}_i$ is an infinite abelian subgroup; it is a maximal abelian subgroup of G_{v_0} (but it is not always maximal abelian in G_{v_1}).

The $\widehat{\mathrm{Out}}(G;\mathcal{H})$–invariant tree T_{can} mentioned above is the tree of cylinders of JSJ trees relative to $\mathcal{H}^{+\mathrm{ab}}$. It is a JSJ tree and the tree of cylinders of T_{can} is T_{can} itself.

Let $\Gamma_{\mathrm{can}} = T_{\mathrm{can}}/G$ be the quotient graph of groups and let $v \in \mathcal{V}_0/G$ be a vertex with G_v non-abelian. If G_v does not split over an abelian group relative to incident edge groups and to $\mathcal{H}_{\|G_v}$, it is universally elliptic (with respect to both $\mathcal{H}$ and $\mathcal{H}^{+\mathrm{ab}}$) and we say that G_v (or v) is *rigid*; otherwise, it is *flexible*.

A key fact here is that every flexible vertex v of Γ_{can} is *quadratically hanging (QH)*. The group G_v is the fundamental group of a compact (possibly non-orientable) surface Σ, and incident edge groups are boundary subgroups of $\pi_1(\Sigma)$ (ie fundamental groups of boundary components of Σ); in particular, incident edge groups are cyclic. At most one incident edge group is attached to a given boundary component (groups carried by distinct incident edges are non-conjugate in G_v). If H is conjugate to a group of $\mathcal{H}$, then $H \cap G_v$ is contained in a boundary subgroup. Conversely, every boundary subgroup is an incident edge group or has a finite-index subgroup which is conjugate to a group of $\mathcal{H}$.

As Szepietowski [34] does, we denote by $\mathcal{PM}^+(\Sigma)$ the group of isotopy classes of homeomorphisms of Σ mapping each boundary component to itself in an orientation-preserving way. We view $\mathcal{PM}^+(\Sigma)$ as a subgroup of $\mathrm{Out}(\pi_1(\Sigma)) = \mathrm{Out}(G_v)$; indeed, $\mathcal{PM}^+(\Sigma) = \mathrm{Out}(G_v; \mathrm{Inc}_v^{(t)}, \mathcal{H}_{\|G_v}^{(t)})$.

2.3 Automorphisms of trees

There is a natural action of $\mathrm{Out}(G)$ on the set of trees, given by precomposing the action on T with an automorphism of G. We denote by $\mathrm{Out}(T)$ the stabilizer of a tree T. We write $\mathrm{Out}(T,\mathcal{H})$ for $\mathrm{Out}(T) \cap \mathrm{Out}(G;\mathcal{H})$, and so on.

If T is a point, $\mathrm{Out}(T)=\mathrm{Out}(G)$. If G is abelian and T is not a point, then T is a line on which G acts by integral translations and $\mathrm{Out}(T)$ is the group of automorphisms of G preserving the kernel of the action.

We now study $\mathrm{Out}(T)$ in the general case, following Levitt [27].

We always assume that edge stabilizers are abelian. This implies that all vertex or edge stabilizers H have the property that the normalizer $N(H)$ acts on H by inner automorphisms; indeed, $N(H)$ is abelian if H is abelian and is equal to H if H is not abelian.

One first considers the action of $\mathrm{Out}(T)$ on the finite graph $\Gamma = T/G$. We always denote by $\mathrm{Out}^0(T)$ the finite-index subgroup consisting of automorphisms acting trivially.

We study it through the natural map

$$\rho=\prod_{v\in V}\rho_v\colon \mathrm{Out}^0(T)\longrightarrow \prod_{v\in V}\mathrm{Out}(G_v)$$

recording the action of automorphisms on vertex groups (see [27, Section 2]); recall that V is the vertex set of Γ. Since $N(G_v)$ acts on G_v by inner automorphisms, $\rho_v(\Phi)$ is simply defined as the class of $\alpha_{|G_v}$, where $\alpha\in\mathrm{Aut}(G)$ is any representative of $\Phi\in\mathrm{Out}^0(T)$ leaving G_v invariant.

The image of ρ is contained in $\prod_{v\in V}\mathrm{Out}(G_v;\mathrm{Inc}_v)$ (the family of incident edge groups at a given v is preserved). It contains the subgroup $\prod_{v\in V}\mathrm{Out}(G_v;\mathrm{Inc}_v^{(t)})$ because automorphisms of G_v acting trivially on incident edge groups extend "by the identity" to automorphisms of G preserving T.

The kernel of ρ is the *group of twists* $\mathcal{T}$, a finitely generated abelian group when no edge group is trivial (bitwists as defined in [27] belong to $\mathcal{T}$ because the normalizer of an abelian subgroup is its centralizer). We therefore have an exact sequence

$$1\longrightarrow \mathcal{T}\longrightarrow \mathrm{Out}^0(T)\xrightarrow{\rho}\prod_{v\in V}\mathrm{Out}(G_v;\mathrm{Inc}_v).$$

Now suppose that T is relative to families $\mathcal{H}$ and $\mathcal{K}$ (ie each H_i and K_j fixes a point in T). A trivial but important remark is that $\mathcal{T}\subset\mathrm{Out}(G;\mathcal{H}^{(t)},\mathcal{K}^{(t)})$. As pointed out in [25, Lemma 2.10], we have

$$\prod_{v\in V}\mathrm{Out}(G_v;\mathrm{Inc}_v^{(t)},\mathcal{H}_{\|G_v}^{(t)},\mathcal{K}_{\|G_v})\subset\rho\big(\mathrm{Out}^0(T)\cap\mathrm{Out}(G;\mathcal{H}^{(t)},\mathcal{K})\big)$$
$$\subset\prod_{v\in V}\mathrm{Out}(G_v;\mathrm{Inc}_v,\mathcal{H}_{\|G_v}^{(t)},\mathcal{K}_{\|G_v})$$

(see Section 2.1 for the definition of $\mathcal{H}_{\|G_v}$; groups of $\mathcal{H}_{\|G_v}$ that are conjugate in G are necessarily conjugate in G_v).

The fact noted above that the image of $\mathrm{Out}^0(T)$ by ρ contains $\prod_{v\in V}\mathrm{Out}(G_v;\mathrm{Inc}_v^{(\mathrm{t})})$ expresses that automorphisms $\Phi_v\in\mathrm{Out}(G_v)$ acting trivially on incident edge groups may be combined into a global $\Phi\in\mathrm{Out}(G)$. In Section 3.2.4 we will need a more general result, where we only assume that the Φ_v have compatible actions on edge groups.

Given an edge e of Γ, there is a natural map $\rho_e\colon \mathrm{Out}^0(T)\to\mathrm{Out}(G_e)$, defined in the same way as ρ_v above. If v is an endpoint of e, the inclusion of G_e into G_v induces a homomorphism $\rho_{v,e}\colon \mathrm{Out}(G_v;\mathrm{Inc}_v)\to\mathrm{Out}(G_e)$ with $\rho_e=\rho_{v,e}\circ\rho_v$ (it is well-defined because the normalizer $N_{G_v}(G_e)$ acts on G_e by inner automorphisms).

Lemma 2.2 *Consider a family of automorphisms $\Phi_v\in\mathrm{Out}(G_v;\mathrm{Inc}_v)$ such that, if $e=vw$ is any edge of Γ, then $\rho_{v,e}(\Phi_v)=\rho_{w,e}(\Phi_w)$. There exists $\Phi\in\mathrm{Out}^0(T)$ such that $\rho_v(\Phi)=\Phi_v$ for every v.*

We leave the proof to the reader. The lemma applies to any graph of groups such that, for every vertex or edge group H, the normalizer $N(H)$ acts on H by inner automorphisms. Φ is not unique: it may be composed with any element of $\mathcal{T}$.

In Section 3.2.4 we will have a family of automorphisms $\Phi_e\in\mathrm{Out}(G_e)$ and we will want $\Phi\in\mathrm{Out}^0(T)$ such that $\rho_e(\Phi)=\Phi_e$ for every e. By the lemma, it suffices to find automorphisms $\Phi_v\in\mathrm{Out}(G_v;\mathrm{Inc}_v)$ inducing the Φ_e.

2.4 Rigid vertices

We now specialize to the case when $T=T_{\mathrm{can}}$ is the canonical JSJ decomposition relative to $\mathcal{H}^{+\mathrm{ab}}$ discussed in Section 2.2.

If v is a QH vertex, the image of $\mathrm{Out}^0(T)\cap\mathrm{Out}(G;\mathcal{H}^{(\mathrm{t})})$ in $\mathrm{Out}(G_v)$ contains $\mathcal{PM}^+(\Sigma)=\mathrm{Out}(G_v;\mathrm{Inc}_v^{(t)},\mathcal{H}_{\|G_v}^{(t)})$ with finite index (see [25, Proposition 4.7]).

If v is a rigid vertex, then G_v does not split over an abelian group relative to $\mathrm{Inc}_v\cup\mathcal{H}_{\|G_v}$. By the Bestvina–Paulin method and Rips theory, one deduces that the image of $\mathrm{Out}^0(T)\cap\mathrm{Out}(G;\mathcal{H}^{(\mathrm{t})})$ in $\mathrm{Out}(G_v)$ is finite if $\mathcal{H}$ is a finite family of finitely generated subgroups (see [25, Theorem 3.9 and Proposition 4.7]).

Lemma 2.3 *Let $\mathcal{H}$ and $\mathcal{K}$ be finite families of finitely generated subgroups, with each group in $\mathcal{K}$ abelian. Assume that G is one-ended relative to $\mathcal{H}\cup\mathcal{K}$ and let T_{can} be the canonical JSJ tree relative to $(\mathcal{H}\cup\mathcal{K})^{+\mathrm{ab}}$.*

The image of

$$\mathrm{Out}^0(T) \cap \mathrm{Out}(G; \mathcal{H}^{(t)}, \mathcal{K})$$

by $\rho_v\colon \mathrm{Out}^0(T) \to \mathrm{Out}(G_v)$ *is finite if* v *is a rigid vertex of* T_{can}. *Its image by* $\rho_e\colon \mathrm{Out}^0(T) \to \mathrm{Out}(G_e)$ *is finite if* e *is any edge.*

Proof Define $\mathcal{K}_{\mathbb{Z}}$ by removing all non-cyclic groups from $\mathcal{K}$. Being freely indecomposable relative to $\mathcal{H} \cup \mathcal{K}$ is the same as being freely indecomposable relative to $\mathcal{H} \cup \mathcal{K}_{\mathbb{Z}}$, and a tree is relative to $(\mathcal{H} \cup \mathcal{K})^{+\mathrm{ab}}$ if and only if it is relative to $(\mathcal{H} \cup \mathcal{K}_{\mathbb{Z}})^{+\mathrm{ab}}$. We may therefore view T_{can} as the canonical JSJ tree relative to $(\mathcal{H} \cup \mathcal{K}_{\mathbb{Z}})^{+\mathrm{ab}}$.

Let v be a rigid vertex. The group $\mathrm{Out}(G; \mathcal{H}^{(t)}, \mathcal{K})$ is contained in $\mathrm{Out}(G; \mathcal{H}^{(t)}, \mathcal{K}_{\mathbb{Z}})$, which contains $\mathrm{Out}(G; \mathcal{H}^{(t)}, \mathcal{K}_{\mathbb{Z}}^{(t)})$ with finite index. As explained above, the image of $\mathrm{Out}^0(T) \cap \mathrm{Out}(G; \mathcal{H}^{(t)}, \mathcal{K}_{\mathbb{Z}}^{(t)})$ in $\mathrm{Out}(G_v)$ is finite [25, Proposition 4.7]. The first assertion of the lemma follows.

Since T_{can} is bipartite, every edge e is incident to a vertex v which is QH or rigid. In the first case G_e is cyclic, so there is nothing to prove. In the second case the map $\rho_e\colon \mathrm{Out}^0(T) \to \mathrm{Out}(G_e)$ factors through $\mathrm{Out}(G_v)$ and the second assertion follows from the first. $\square$

3 Finite classifying space

In this section, we prove that McCool groups of a toral relatively hyperbolic group have type VF (Theorem 1.3) and that so does the stabilizer of a splitting (Theorem 1.4). In the course of the proof, we will describe the automorphisms of a given maximal abelian subgroup which are restrictions of an automorphism of G belonging to a given McCool group (Proposition 3.10).

We start by recalling some standard facts about groups of type VF.

A group has type F if it has a finite classifying space and type VF if some finite-index subgroup is of type F. A key tool for proving that groups have type F is the following statement:

Theorem 3.1 (See for instance Geoghegan [15, Theorem 7.3.4]) *Suppose that* G *acts simplicially and cocompactly on a contractible simplicial complex* X. *If all point stabilizers have type* F, *so does* G. *In particular, being of type* F *is stable under extensions.*

If G has a finite-index subgroup acting as in the theorem, then G has type VF. In particular:

Corollary 3.2 *Given an exact sequence* $1 \to N \to G \to Q \to 1$, *suppose that* Q *has type* VF *and* G *has a finite-index subgroup* $G_0 < G$ *such that* $G_0 \cap N$ *has type* F. *Then* G *has type* VF.

Remark 3.3 Suppose that G acts on X as in Theorem 3.1. If point stabilizers are only of type VF, one cannot claim that G has type VF, even if G is torsion-free. This subtlety was overlooked in [20, Theorem 5.2] (we will give a corrected statement in Corollary 3.8) and it introduces technical complications (which would not occur if we only wanted to prove that the groups under consideration have type F_∞). In particular, to study the stabilizer of a tree with non-cyclic edge stabilizers in Section 3.2.3, we have to prove more precise versions of certain results (such as the "moreover" in Theorem 3.4).

3.1 McCool groups are VF

In this subsection we prove the following strengthening of Theorem 1.3:

Theorem 3.4 *Let* G *be a toral relatively hyperbolic group. Let* $\mathcal{H}$ *and* $\mathcal{K}$ *be two finite families of finitely generated subgroups, with each group in* $\mathcal{K}$ *abelian. Then* $\mathrm{Out}(G;\mathcal{H}^{(t)},\mathcal{K})$ *is of type* VF.

Moreover, if groups in $\mathcal{H}$ *are also abelian, then there exists a finite-index subgroup* $\mathrm{Out}^1(G;\mathcal{H},\mathcal{K}) \subset \mathrm{Out}(G;\mathcal{H},\mathcal{K})$ *such that* $\mathrm{Out}^1(G;\mathcal{H},\mathcal{K}) \cap \mathrm{Out}(G;\mathcal{H}^{(t)},\mathcal{K})$ *is of type* F.

Recall (Definition 2.1) that $\mathrm{Out}(G;\mathcal{H}^{(t)},\mathcal{K})$ consists of classes of automorphisms acting trivially on each group $H_i \in \mathcal{H}$ (ie as conjugation by some $g_i \in G$) and leaving each $K_j \in \mathcal{K}$ invariant up to conjugacy.

It will follow from Corollary 1.6 that the main assertion of Theorem 3.4 holds if $\mathcal{H}$ is an arbitrary family of subgroups (see Corollary 6.3), but finiteness is needed at this point in order to apply Lemma 2.3.

Convention 3.5 In this subsection, a superscript 1, as in $\mathrm{Out}^1(G;\mathcal{H},\mathcal{K})$, always indicates a subgroup of finite index. The superscript 0 refers to a trivial action on a quotient graph of groups (see Section 2.3).

3.1.1 The abelian case

The following lemma deals with the case when $G = \mathbb{Z}^n$.

Lemma 3.6 *Let* $\mathcal{H}$ *and* $\mathcal{K}$ *be finite families of subgroups of* $\mathbb{Z}^n$. *Consider the subgroup* $A = \mathrm{Out}(\mathbb{Z}^n;\mathcal{H}^{(t)},\mathcal{K})$ *of* $\mathrm{GL}(n,\mathbb{Z})$ *consisting of matrices acting as the identity on groups* $H_i \in \mathcal{H}$ *and leaving each* $K_j \in \mathcal{K}$ *invariant. Then* A *is of type* VF. *More precisely, every torsion-free subgroup of finite index* $A' \subset A$ *is of type* F.

Recall that $\mathrm{GL}(n,\mathbb{Z})$ is virtually torsion-free, so groups such as A' exist.

Proof The set of endomorphisms of $\mathbb{Z}^n$ acting as the identity on H_i and preserving K_j is a linear subspace defined by linear equations with rational coefficients. It follows that the groups A and A' are arithmetic: they are commensurable with a subgroup of $\mathrm{GL}(n,\mathbb{Z})$ defined by $\mathbb{Q}$–linear equations. By Borel and Serre [7], every torsion-free arithmetic subgroup of $\mathrm{GL}(n,\mathbb{Q})$ is of type F. □

To deduce Theorem 3.4 when G is abelian, we simply define $\mathrm{Out}^1(G;\mathcal{H},\mathcal{K})$ as any torsion-free, finite-index subgroup of $\mathrm{Out}(G;\mathcal{H},\mathcal{K})$.

If G is not abelian, we shall distinguish two cases.

3.1.2 The one-ended case We first assume that G is freely indecomposable relative to $\mathcal{H}\cup\mathcal{K}$: one cannot write $G=A*B$ with each group of $\mathcal{H}\cup\mathcal{K}$ contained in a conjugate of A or B. We then consider the canonical tree T_{can} as in Section 2.2 (it is a JSJ tree relative to $\mathcal{H}$, $\mathcal{K}$ and to non-cyclic abelian subgroups). It is invariant under $\mathrm{Out}(G;\mathcal{H},\mathcal{K})$, so $\mathrm{Out}(G;\mathcal{H},\mathcal{K})\subset\mathrm{Out}(T_{\mathrm{can}})$.

We write $\mathrm{Out}^0(T_{\mathrm{can}})$ for the finite-index subgroup consisting of automorphisms acting trivially on the finite graph $\Gamma_{\mathrm{can}}=T_{\mathrm{can}}/G$ and

$$\mathrm{Out}^0(G;\mathcal{H},\mathcal{K})=\mathrm{Out}(G;\mathcal{H},\mathcal{K})\cap\mathrm{Out}^0(T_{\mathrm{can}}),$$

which has finite index in $\mathrm{Out}(G;\mathcal{H},\mathcal{K})$.

Recall that non-abelian vertex stabilizers G_v of T_{can} (or vertex groups of Γ_{can}) are rigid or QH. Also recall from Section 2.3 that, for each vertex v, there is a map $\rho_v\colon \mathrm{Out}^0(T_{\mathrm{can}})\to\mathrm{Out}(G_v;\mathrm{Inc}_v)$, with Inc_v the family of incident edge groups (see Section 2.1).

We define a subgroup $\mathrm{Out}^r(G;\mathcal{H},\mathcal{K})\subset\mathrm{Out}(G;\mathcal{H},\mathcal{K})$ by restricting to automorphisms $\Phi\in\mathrm{Out}^0(G;\mathcal{H},\mathcal{K})$ and imposing conditions on the image of Φ by the maps ρ_v:

- If G_v is rigid, we ask that $\rho_v(\Phi)$ be trivial.

- If G_v is abelian, we fix a torsion-free subgroup of finite index $\mathrm{Out}^1(G_v)\subset\mathrm{Out}(G_v)$ and we ask that $\rho_v(\Phi)$ belong to $\mathrm{Out}^1(G_v)$.

- If G_v is QH, it is the fundamental group of a compact surface Σ. Each boundary component is associated to an incident edge or a group in $\mathcal{H}\cup\mathcal{K}$ (see Section 2.2), so $\rho_v(\Phi)$ preserves the peripheral structure of $\pi_1(\Sigma)$ and may therefore be represented by a homeomorphism of Σ. Since groups in $\mathcal{H}\cup\mathcal{K}$, and their conjugates, only intersect G_v

along boundary subgroups, the image of $\mathrm{Out}^0(G;\mathcal{H},\mathcal{K})$ by ρ_v contains the mapping class group

$$\mathcal{PM}^+(\Sigma)=\mathrm{Out}(G_v;\mathrm{Inc}_v^{(t)},\mathcal{H}_{\|G_v}^{(t)},\mathcal{K}_{\|G_v}^{(t)})$$

(see Section 2.2); the index is finite. We fix a finite-index subgroup $\mathcal{PM}^{+,1}(\Sigma)$ of type F and we require $\rho_v(\Phi)\in\mathcal{PM}^{+,1}(\Sigma)$. In particular, Φ acts trivially on all boundary subgroups of Σ.

Let $\mathrm{Out}^r(G;\mathcal{H},\mathcal{K})$ consist of automorphisms $\Phi\in\mathrm{Out}^0(G;\mathcal{H},\mathcal{K})$ whose images $\rho_v(\Phi)$ satisfy the above conditions. These automorphisms act trivially on edge stabilizers.

It follows from Lemma 2.3 that $\mathrm{Out}^r(G;\mathcal{H},\mathcal{K})\cap\mathrm{Out}(G;\mathcal{H}^{(t)},\mathcal{K})$ always has finite index in $\mathrm{Out}(G;\mathcal{H}^{(t)},\mathcal{K})$. If groups in $\mathcal{H}$ are abelian, then $\mathrm{Out}^r(G;\mathcal{H},\mathcal{K})$ has finite index in $\mathrm{Out}(G;\mathcal{H},\mathcal{K})$. It therefore suffices to prove that

$$O:=\mathrm{Out}^r(G;\mathcal{H},\mathcal{K})\cap\mathrm{Out}(G;\mathcal{H}^{(t)},\mathcal{K})$$

is of type F (this argument, based on Lemma 2.3, is the only place where we use the assumptions on $\mathcal{H}$ and $\mathcal{K}$).

Every edge of T_{can} has an endpoint v with G_v rigid or QH, so elements of O act trivially on edge stabilizers of T_{can}. Consider an abelian vertex stabilizer G_v. Elements in $\rho_v(O)$ are the identity on incident edge groups and groups in $\mathcal{H}_{\|G_v}$, and leave groups in $\mathcal{K}_{\|G_v}$ invariant. By Lemma 3.6 these conditions define a group $B_v\subset\mathrm{Out}(G_v)$ which is of type VF and $C_v:=B_v\cap\mathrm{Out}^1(G_v)$ is a group of type F containing $\rho_v(O)$.

Recall from Section 2.3 the exact sequence

$$1\longrightarrow\mathcal{T}\longrightarrow\mathrm{Out}^0(T_{\mathrm{can}})\xrightarrow{\rho}\prod_{v\in V}\mathrm{Out}(G_v;\mathrm{Inc}_v).$$

We claim that the image of O by ρ is a direct product $\prod_{v\in V}C_v$, with C_v as above if G_v is abelian, $C_v=\mathcal{PM}^{+,1}(\Sigma)$ if v is QH, and C_v trivial if v is rigid. The image is contained in the product. Conversely, given a family $(\Phi_v)_{v\in V}$, with $\Phi_v\in C_v$, the automorphisms Φ_v act trivially on incident edge groups, so there is $\Phi\in\mathrm{Out}^0(T_{\mathrm{can}})$ with $\rho_v(\Phi)=\Phi_v$. Since Φ_v acts trivially on $\mathrm{Inc}_v\cup\mathcal{H}_{\|G_v}$ and preserves $\mathcal{K}_{\|G_v}$, this automorphism is in O. This proves the claim.

It follows that $\rho(O)$ is of type F. The group of twists $\mathcal{T}$ is contained in O, because twists act trivially on vertex groups and T is relative to $\mathcal{H}\cup\mathcal{K}$, so we can conclude that O is of type F by Theorem 3.1 if we know that $\mathcal{T}$ is of type F. The group $\mathcal{T}$ is a finitely generated abelian group. It is torsion-free, hence of type F, as shown in [25, Section 4] (alternatively, one can replace $\mathrm{Out}^r(G;\mathcal{H},\mathcal{K})$ by its intersection with a torsion-free, finite-index subgroup of $\mathrm{Out}(G)$, which exists by [25, Corollary 4.4]).

This proves Theorem 3.4 in the freely indecomposable case. To prove it in general, we need to study automorphisms of free products.

3.1.3 Automorphisms of free products In this subsection, G does not have to be relatively hyperbolic.

Let $\mathcal{G} = \{G_i\}$ be a family of subgroups of G. We have defined $\mathrm{Out}(G;\mathcal{G})$ as automorphisms leaving the conjugacy class of each G_i invariant and $\mathrm{Out}(G;\mathcal{G}^{(t)})$ as automorphisms acting trivially on each G_i.

More generally, consider a group of automorphisms $\mathcal{Q}_i \subset \mathrm{Out}(G_i)$ and $\mathcal{Q} = \{\mathcal{Q}_i\}$. We would like to define $\mathrm{Out}(G;\mathcal{G}^{(\mathcal{Q})}) \subset \mathrm{Out}(G;\mathcal{G})$ as the automorphisms Φ acting on each G_i as an element of $\mathcal{Q}_i$. To be precise, given $\Phi \in \mathrm{Out}(G;\mathcal{G})$, choose representatives φ_i of Φ in $\mathrm{Aut}(G)$ with $\varphi_i(G_i) = G_i$. We say that Φ belongs to $\mathrm{Out}(G;\mathcal{G}^{(\mathcal{Q})})$ if every φ_i represents an element of $\mathcal{Q}_i$. This is well-defined (independent of the chosen φ_i) if each G_i is a free factor (more generally, if the normalizer of G_i acts on G_i by inner automorphisms).

The goal of this subsection is to show:

Proposition 3.7 *Let $G = G_1 * \cdots * G_n * F_p$, with F_p free of rank p, and let $\mathcal{G} = \{G_i\}$. Assume that all groups G_i and $G_i/Z(G_i)$ have type* F.

Let $\mathcal{Q} = \{\mathcal{Q}_i\}$ be a family of subgroups $\mathcal{Q}_i \subset \mathrm{Out}(G_i)$. If every $\mathcal{Q}_i$ is of type VF*, then $\mathrm{Out}(G;\mathcal{G}^{(\mathcal{Q})})$ has type* VF.

More precisely, there exists a finite-index subgroup $\mathrm{Out}^1(G;\mathcal{G}) \subset \mathrm{Out}(G;\mathcal{G})$, independent of $\mathcal{Q}$, such that, if every $\mathcal{Q}_i$ is of type F*, then $\mathrm{Out}^1(G;\mathcal{G}) \cap \mathrm{Out}(G;\mathcal{G}^{(\mathcal{Q})})$ has type* F.

The "more precise" assertion implies the first one, since $\mathrm{Out}(G;\mathcal{G}^{(\mathcal{Q}')})$ has finite index in $\mathrm{Out}(G;\mathcal{G}^{(\mathcal{Q})})$ if every $\mathcal{Q}_i'$ is a finite-index subgroup of $\mathcal{Q}_i$.

Assume that G_i and $G_i/Z(G_i)$ have type F. The proposition says in particular that the Fouxe-Rabinovitch group $\mathrm{Out}(G;\mathcal{G}^{(t)})$ is of type VF, and that $\mathrm{Out}(G;\mathcal{G})$ is of type VF if every $\mathrm{Out}(G_i)$ is. If we consider the Grushko decomposition of G, then $\mathrm{Out}(G;\mathcal{G})$ has finite index in $\mathrm{Out}(G)$ and we get:

Corollary 3.8 (Correcting [20, Theorem 5.2]) *Let $G = G_1 * \cdots * G_n * F_p$, with F_p free and G_i non-trivial, not isomorphic to $\mathbb{Z}$ and not a free product. If every G_i and $G_i/Z(G_i)$ has type* F *and every $\mathrm{Out}(G_i)$ has type* VF*, then $\mathrm{Out}(G)$ has type* VF.

Proof of Proposition 3.7 We prove the "more precise" assertion, so we assume that $\mathcal{Q}_i \subset \mathrm{Out}(G_i)$ has type F. We shall apply Theorem 3.1 to the action of $\mathrm{Out}(G;\mathcal{G}^{(\mathcal{Q})})$ on the outer space defined in [20]. We let $\mathcal{D}$ be the Grushko deformation space relative to $\mathcal{G}$, ie the JSJ deformation space of G over the trivial group relative to $\mathcal{G}$ (see Section 2.2). Trees in $\mathcal{D}$ have trivial edge stabilizers and non-trivial vertex stabilizers are conjugates of the G_i.

Like ordinary outer space [9], the projectivization $\widehat{\mathcal{D}}$ of $\mathcal{D}$ is a complex consisting of simplices with missing faces and the spine of $\widehat{\mathcal{D}}$ is a simplicial complex. It is contractible for the weak topology [19].

The group $\mathrm{Out}(G;\mathcal{G})$ acts on $\mathcal{D}$, hence on the spine, and the action of the Fouxe-Rabinovitch group $\mathrm{Out}(G;\mathcal{G}^{(\mathrm{t})}) \subset \mathrm{Out}(G;\mathcal{G}^{(\mathcal{Q})})$ is cocompact because there are finitely many possibilities for the quotient graph T/G for $T \in \mathcal{D}$. In order to apply Theorem 3.1, we just need to show that stabilizers are of type F.

$\mathrm{Out}(G;\mathcal{G})$ also acts on the free group (isomorphic to F_p) obtained from G by killing all the G_i (it may be viewed as the topological fundamental group of $\Gamma = T/G$ for any $T \in \mathcal{D}$). In other words, there is a natural map $\mathrm{Out}(G;\mathcal{G}) \to \mathrm{Out}(F_p)$. We fix a torsion-free, finite-index subgroup $\mathrm{Out}^1(F_p) \subset \mathrm{Out}(F_p)$ and we define $\mathrm{Out}^1(G;\mathcal{G}) \subset \mathrm{Out}(G;\mathcal{G})$ as the pullback of $\mathrm{Out}^1(F_p)$.

Given $T \in \mathcal{D}$, we let S be its stabilizer for the action of $\mathrm{Out}^1(G;\mathcal{G})$ and $S_{\mathcal{Q}}$ its stabilizer for the action of $\mathrm{Out}^1(G;\mathcal{G}) \cap \mathrm{Out}(G;\mathcal{G}^{(\mathcal{Q})})$. We complete the proof by showing that $S_{\mathcal{Q}}$ has type F.

We first claim that S equals $\mathrm{Out}^0(T)$, the group of automorphisms of G leaving T invariant and acting trivially on $\Gamma = T/G$. Clearly $\mathrm{Out}^0(T) \subset S$. Conversely, we have to show that any $\Phi \in S$ acts as the identity on Γ. First, Φ fixes all vertices of Γ carrying a non-trivial group G_v, because G_v is a G_i (up to conjugacy) and the G_i are not permuted. In particular, by minimality of T, all terminal vertices of Γ are fixed. Also, by our definition of $\mathrm{Out}^1(G;\mathcal{G})$, the image of Φ in $\mathrm{Out}(\pi_1(\Gamma))$ is trivial or has infinite order. The claim follows because any non-trivial symmetry of Γ fixing all terminal vertices maps to a non-trivial element of finite order in $\mathrm{Out}(\pi_1(\Gamma))$ if Γ is not a circle.

The map ρ (see Section 2.3) maps S onto $\prod_i \mathrm{Out}(G_i)$, and the image of $S_{\mathcal{Q}}$ is $\prod_i \mathcal{Q}_i$, a group of type F. The kernel is the group of twists $\mathcal{T}$, which is contained in $S_{\mathcal{Q}}$, so it suffices to check that $\mathcal{T}$ has type F. Since edge stabilizers are trivial, $\mathcal{T}$ is a direct product $\prod_i K_i$, with $K_i = G_i^{n_i}/Z(G_i)$; here n_i is the valence of the vertex carrying G_i in Γ and the center $Z(G_i)$ is embedded diagonally (see [27]). There are exact sequences

$$1 \longrightarrow G_i^{n_i-1} \longrightarrow G_i^{n_i}/Z(G_i) \longrightarrow G_i/Z(G_i) \longrightarrow 1,$$

so the assumptions of the proposition ensure that $\mathcal{T}$ is of type F. $\square$

3.1.4 The infinitely ended case We can now prove Theorem 3.4 in full generality. We let $G = G_1 * \cdots * G_n * F_p$ be the Grushko decomposition of G relative to $\mathcal{H} \cup \mathcal{K}$ (see Section 2.2) and $\mathcal{G} = \{G_i\}$. Each G_i is toral relatively hyperbolic, so has type F by Dahmani [10]. Its center is trivial if G_i is nonabelian, so $G_i/Z(G_i)$ also has type F. This will allow us to use Proposition 3.7.

Lemma 3.9 *Let $\mathcal{Q} = \{\mathcal{Q}_i\}$ with $\mathcal{Q}_i = \mathrm{Out}(G_i; \mathcal{H}^{(\mathrm{t})}_{|G_i}, \mathcal{K}_{|G_i})$ and let $\mathcal{R} = \{\mathcal{R}_i\}$ with $\mathcal{R}_i = \mathrm{Out}(G_i; \mathcal{H}_{|G_i}, \mathcal{K}_{|G_i})$. Then*

$$\mathrm{Out}(G; \mathcal{G}^{(\mathcal{Q})}) = \mathrm{Out}(G; \mathcal{H}^{(\mathrm{t})}, \mathcal{K}) \cap \mathrm{Out}(G; \mathcal{G}),$$
$$\mathrm{Out}(G; \mathcal{G}^{(\mathcal{R})}) = \mathrm{Out}(G; \mathcal{H}, \mathcal{K}) \cap \mathrm{Out}(G; \mathcal{G}).$$

Moreover, $\mathrm{Out}(G; \mathcal{G}^{(\mathcal{Q})})$ has finite index in $\mathrm{Out}(G; \mathcal{H}^{(\mathrm{t})}, \mathcal{K})$ and $\mathrm{Out}(G; \mathcal{G}^{(\mathcal{R})})$ has finite index in $\mathrm{Out}(G; \mathcal{H}, \mathcal{K})$.

Proof If Φ belongs to $\mathrm{Out}(G; \mathcal{G}^{(\mathcal{Q})})$, it belongs to $\mathrm{Out}(G; \mathcal{H}^{(\mathrm{t})}, \mathcal{K})$, because every group in $\mathcal{H} \cup \mathcal{K}$ has a conjugate contained in some G_i. Conversely, automorphisms in $\mathrm{Out}(G; \mathcal{H}^{(\mathrm{t})}, \mathcal{K})$ preserve the Grushko deformation space relative to $\mathcal{H} \cup \mathcal{K}$ and therefore permute the G_i, so $\mathrm{Out}(G; \mathcal{G}) \cap \mathrm{Out}(G; \mathcal{H}^{(\mathrm{t})}, \mathcal{K})$ has finite index in $\mathrm{Out}(G; \mathcal{H}^{(\mathrm{t})}, \mathcal{K})$. If $\varphi \in \mathrm{Aut}(G)$ leaves G_i invariant and maps a non-trivial $H \subset G_i$ to a conjugate gHg^{-1}, then $g \in G_i$ because G_i is a free factor. This shows

$$\mathrm{Out}(G; \mathcal{H}^{(\mathrm{t})}, \mathcal{K}) \cap \mathrm{Out}(G; \mathcal{G}) \subset \mathrm{Out}(G; \mathcal{G}^{(\mathcal{Q})}),$$

completing the proof for $\mathrm{Out}(G; \mathcal{G}^{(\mathcal{Q})})$. The proof for $\mathrm{Out}(G; \mathcal{G}^{(\mathcal{R})})$ is similar. $\square$

The first assertion of Theorem 3.4 now follows immediately from the one-ended case together with Proposition 3.7, since $\mathrm{Out}(G; \mathcal{H}^{(\mathrm{t})}, \mathcal{K})$ contains $\mathrm{Out}(G; \mathcal{G}^{(\mathcal{Q})})$ with finite index. There remains to prove the "moreover".

Each G_i is freely indecomposable relative to $\mathcal{H}_{|G_i} \cup \mathcal{K}_{|G_i}$, so we may apply the "moreover" of Theorem 3.4 to G_i. We get a finite-index subgroup $\mathcal{R}_i^1 \subset \mathcal{R}_i$ such that $\mathcal{Q}_i^1 := \mathcal{R}_i^1 \cap \mathcal{Q}_i$ has type F. Let $\mathcal{R}^1 = \{\mathcal{R}_i^1\}$ and $\mathcal{Q}^1 = \{\mathcal{Q}_i^1\}$.

By Proposition 3.7, there is a finite-index subgroup $\mathrm{Out}^1(G; \mathcal{G}) \subset \mathrm{Out}(G; \mathcal{G})$ such that $\mathrm{Out}^1(G; \mathcal{G}) \cap \mathrm{Out}(G; \mathcal{G}^{(\mathcal{Q}^1)})$ has type F. Now write

$$\mathrm{Out}^1(G; \mathcal{G}) \cap \mathrm{Out}(G; \mathcal{G}^{(\mathcal{Q}^1)}) = \mathrm{Out}^1(G; \mathcal{G}) \cap \mathrm{Out}(G; \mathcal{G}^{(\mathcal{R}^1)}) \cap \mathrm{Out}(G; \mathcal{G}^{(\mathcal{Q})}).$$

By Lemma 3.9, we may replace the last term $\mathrm{Out}(G; \mathcal{G}^{(\mathcal{Q})})$ by $\mathrm{Out}(G; \mathcal{H}^{(\mathrm{t})}, \mathcal{K})$. Defining

$$\mathrm{Out}^1(G; \mathcal{H}, \mathcal{K}) := \mathrm{Out}^1(G; \mathcal{G}) \cap \mathrm{Out}(G; \mathcal{G}^{(\mathcal{R}^1)}),$$

we have shown that $\mathrm{Out}^1(G;\mathcal{H},\mathcal{K})\cap\mathrm{Out}(G;\mathcal{H}^{(\mathrm{t})},\mathcal{K})$ has type F. There remains to check that $\mathrm{Out}^1(G;\mathcal{H},\mathcal{K})$ is a finite-index subgroup of $\mathrm{Out}(G;\mathcal{H},\mathcal{K})$.

Since $\mathrm{Out}^1(G;\mathcal{G})$ has finite index in $\mathrm{Out}(G;\mathcal{G})$ and $\mathcal{R}_i^1$ is a finite-index subgroup of $\mathcal{R}_i$, the group $\mathrm{Out}^1(G;\mathcal{H},\mathcal{K})$ has finite index in $\mathrm{Out}(G;\mathcal{G})\cap\mathrm{Out}(G;\mathcal{G}^{(\mathcal{R})})$, which equals $\mathrm{Out}(G;\mathcal{G}^{(\mathcal{R})})$ and has finite index in $\mathrm{Out}(G;\mathcal{H},\mathcal{K})$ by Lemma 3.9.

This completes the proof of Theorem 3.4.

3.1.5 The action on abelian groups We study the action of $\mathrm{Out}(G)$ on abelian subgroups. The result of this subsection (Proposition 3.10) will be needed in Section 3.2.4.

A toral relatively hyperbolic group has finitely many conjugacy classes of non-cyclic maximal abelian subgroups. Fix a representative A_j in each class. Automorphisms of G preserve the set of A_j (up to conjugacy), so some finite-index subgroup of $\mathrm{Out}(G)$ maps to $\prod_j \mathrm{Out}(A_j)$. We shall show in particular that the image of a suitable finite-index subgroup $\mathrm{Out}'(G)\subset\mathrm{Out}(G)$ is a product of McCool groups $\prod_j \mathrm{Out}(A_j;\{F_j\}^{(\mathrm{t})})\subset\prod_j \mathrm{Out}(A_j)$.

This product structure expresses the fact that automorphisms of non-conjugate maximal non-cyclic abelian subgroups do not interact. Indeed, consider a family of elements $\Phi_j\in\mathrm{Out}(A_j)$ and suppose that each Φ_j, taken individually, extends to an automorphism $\widehat{\Phi}_j\in\mathrm{Out}'(G)$; then there is $\Phi\in\mathrm{Out}'(G)$ inducing all Φ_j simultaneously.

In fact, we will work with two (possibly empty) finite families $\mathcal{H}$ and $\mathcal{K}$ of abelian subgroups and we will restrict to $\mathrm{Out}(G;\mathcal{H}^{(\mathrm{t})},\mathcal{K})$. We shall therefore define a finite-index subgroup $\mathrm{Out}'(G;\mathcal{H}^{(\mathrm{t})},\mathcal{K})\subset\mathrm{Out}(G;\mathcal{H}^{(\mathrm{t})},\mathcal{K})$.

First assume that G is freely indecomposable relative to $\mathcal{H}\cup\mathcal{K}$. As in Section 3.1.2, we consider the canonical JSJ tree T_{can}, we restrict to automorphisms $\Phi\in\mathrm{Out}(G;\mathcal{H}^{(\mathrm{t})},\mathcal{K})$ acting trivially on $\Gamma_{\mathrm{can}}=T_{\mathrm{can}}/G$ and we define $\mathrm{Out}'(G;\mathcal{H}^{(\mathrm{t})},\mathcal{K})$ by imposing conditions on the action on non-abelian vertex groups G_v: if G_v is QH, the action should be trivial on all boundary subgroups of Σ (ie $\rho_v(\Phi)\in\mathcal{PM}^+(\Sigma)$); if G_v is rigid, then $\rho_v(\Phi)$ should be trivial. We have explained in Section 3.1.2 why this defines a subgroup of finite index $\mathrm{Out}'(G;\mathcal{H}^{(\mathrm{t})},\mathcal{K})$ in $\mathrm{Out}(G;\mathcal{H}^{(\mathrm{t})},\mathcal{K})$. Note that $\mathrm{Out}'(G;\mathcal{H}^{(\mathrm{t})},\mathcal{K})$ acts trivially on edge groups of T_{can}.

If G is not freely indecomposable relative to $\mathcal{H}\cup\mathcal{K}$, let $G=G_1*\cdots*G_n*F_p$ be the relative Grushko decomposition. To define $\mathrm{Out}'(G;\mathcal{H}^{(\mathrm{t})},\mathcal{K})$, we require that Φ maps G_i to G_i (up to conjugacy) and the induced automorphism belongs to $\mathrm{Out}'(G_i;\mathcal{H}^{(\mathrm{t})}_{|G_i},\mathcal{K}_{|G_i})$ as defined above.

Elements of $\mathrm{Out}'(G;\mathcal{H}^{(t)},\mathcal{K})$ leave every A_j invariant (up to conjugacy) and we denote by

$$\theta\colon \mathrm{Out}'(G;\mathcal{H}^{(t)},\mathcal{K}) \longrightarrow \prod_j \mathrm{Out}(A_j)$$

the natural map.

We can now state:

Proposition 3.10 *Let $\mathcal{H}$ and $\mathcal{K}$ be two finite families of abelian subgroups and let $\mathrm{Out}'(G;\mathcal{H}^{(t)},\mathcal{K})$ be the finite-index subgroup of $\mathrm{Out}(G;\mathcal{H}^{(t)},\mathcal{K})$ defined above.*

There are subgroups $F_j \subset A_j$ such that the image of $\theta\colon \mathrm{Out}'(G;\mathcal{H}^{(t)},\mathcal{K}) \to \prod_j \mathrm{Out}(A_j)$ equals $\prod_j \mathrm{Out}(A_j;\{F_j\}^{(t)},\mathcal{K}_{|A_j})$.

Recall that the A_j are representatives of conjugacy classes of non-cyclic maximal abelian subgroups.

Proof The A_j are contained (up to conjugacy) in factors G_i of the Grushko decomposition relative to $\mathcal{H}\cup\mathcal{K}$ and the G_i are invariant under $\mathrm{Out}'(G;\mathcal{H}^{(t)},\mathcal{K})$. Since any family of automorphisms $\Phi_i \in \mathrm{Out}'(G_i;\mathcal{H}^{(t)}_{|G_i},\mathcal{K}_{|G_i})$ extends to an automorphism $\Phi \in \mathrm{Out}'(G;\mathcal{H}^{(t)},\mathcal{K})$, we may assume that G is freely indecomposable relative to $\mathcal{H}\cup\mathcal{K}$.

Let T_{can} be as above. If A_j is contained in a rigid vertex stabilizer, then $\mathrm{Out}'(G;\mathcal{H}^{(t)},\mathcal{K})$ acts trivially on A_j and we define $F_j = A_j$. If not, A_j is a vertex stabilizer G_v. Vertex stabilizers adjacent to v are rigid or QH and, because of the way we defined it, $\mathrm{Out}'(G;\mathcal{H}^{(t)},\mathcal{K})$ leaves A_j invariant and acts trivially on incident edge groups. It also acts trivially on the groups belonging to $\mathcal{H}_{|A_j}$.

Defining F_j as the subgroup of A_j generated by incident edge groups and groups in $\mathcal{H}_{|A_j}$, we have proved that the image of θ is contained in $\prod_j \mathrm{Out}(A_j;\{F_j\}^{(t)},\mathcal{K}_{|A_j})$. Conversely, choose a family $\Phi_j \in \mathrm{Out}(A_j;\{F_j\}^{(t)},\mathcal{K}_{|A_j})$. As explained in Section 2.3, there exists $\Phi \in \mathrm{Out}^0(T_{\mathrm{can}})$ acting trivially on cyclic, rigid and QH vertex stabilizers and inducing Φ_j on A_j. We check that Φ acts trivially on any $H \in \mathcal{H}$. Such a group H fixes a vertex $v \in T_{\mathrm{can}}$. If G_v is cyclic, rigid or QH, the action of Φ on H is trivial. If not, G_v is (conjugate to) an A_j and the action is trivial because $H \subset F_j$. A similar argument shows that Φ preserves $\mathcal{K}$ up to conjugacy, so $\Phi \in \mathrm{Out}(G;\mathcal{H}^{(t)},\mathcal{K})$. Since Φ acts trivially on rigid and QH vertex stabilizers, $\Phi \in \mathrm{Out}'(G;\mathcal{H}^{(t)},\mathcal{K})$. □

3.2 Automorphisms preserving a tree

We now study the stabilizer of a tree. The following theorem clearly implies Theorem 1.4.

Theorem 3.11 *Let G be a toral relatively hyperbolic group. Let T be a simplicial tree on which G acts with abelian edge stabilizers. Let $\mathcal{K}$ be a finite family of abelian subgroups of G, each of which fixes a point in T. Then $\mathrm{Out}(T,\mathcal{K}) = \mathrm{Out}(T) \cap \mathrm{Out}(G;\mathcal{K})$ is of type* VF.

The group $\mathrm{Out}(T,\mathcal{K})$ is the subgroup of $\mathrm{Out}(G)$ consisting of automorphisms leaving T invariant and mapping each group of $\mathcal{K}$ to a conjugate (in an arbitrary way). The tree T is assumed to be minimal, but it may be a point, it may have trivial edge stabilizers, and non-cyclic abelian subgroups need not be elliptic.

Theorem 3.4 proves Theorem 3.11 when T is a point. Also note that, if G is abelian and T is not a point, then T is a line on which G acts by integral translations and $\mathrm{Out}(T,\mathcal{K})$ is of type VF because it equals $\mathrm{Out}(G;\mathcal{K}\cup\{N\})$, with N the kernel of the action of G on T.

Thus, we assume from now on that G is not abelian. We will prove Theorem 3.11 when T has cyclic edge stabilizers before treating the general case. This special case is much easier because $\mathrm{Out}(G_e)$ is finite for every edge stabilizer G_e and we may apply [27, Proposition 2.3].

3.2.1 Cyclic edge stabilizers In this subsection we prove Theorem 3.11 when all edge stabilizers G_e of T are cyclic (possibly trivial); this happens in particular if G is hyperbolic.

As in Section 2.3, we consider the exact sequence

$$1 \longrightarrow \mathcal{T} \longrightarrow \mathrm{Out}^0(T) \xrightarrow{\rho} \prod_{v\in V} \mathrm{Out}(G_v;\mathrm{Inc}_v).$$

The image of ρ contains $\prod_{v\in V}\mathrm{Out}(G_v;\mathrm{Inc}_v^{(t)})$ and the index is finite because all groups $\mathrm{Out}(G_e)$ are finite (see [27], where $\mathrm{Out}(G_v;\mathrm{Inc}_v^{(t)})$ is denoted by $PMCG(G_v)$). The preimage of $\prod_{v\in V}\mathrm{Out}(G_v;\mathrm{Inc}_v^{(t)})$ is thus a finite index subgroup $\mathrm{Out}^1(T)\subset\mathrm{Out}(T)$.

We want to prove that $\mathrm{Out}(T,\mathcal{K})$ is of type VF, so we restrict the preceding discussion to $\mathrm{Out}(T,\mathcal{K})$. Let

$$\mathrm{Out}^1(T,\mathcal{K}) = \mathrm{Out}^1(T)\cap\mathrm{Out}(G;\mathcal{K}),$$

a finite-index subgroup. We show that $\mathrm{Out}^1(T,\mathcal{K})$ is of type VF (this will not use the assumption that edge stabilizers are cyclic).

The image of $\mathrm{Out}^1(T,\mathcal{K})$ by ρ is contained in $\prod_{v\in V}\mathrm{Out}(G_v;\mathrm{Inc}_v^{(t)},\mathcal{K}_{\|G_v})$, with $\mathcal{K}_{\|G_v}$ as in Section 2.1 and, arguing as in Section 2.3, one sees that equality holds. On the other hand, $\mathrm{Out}^1(T,\mathcal{K})$ contains $\mathcal{T}$ because twists act trivially on vertex stabilizers, hence on $\mathcal{K}$ since groups of $\mathcal{K}$ are elliptic in T. We therefore have an exact sequence

$$1\longrightarrow\mathcal{T}\longrightarrow\mathrm{Out}^1(T,\mathcal{K})\longrightarrow\prod_{v\in V}\mathrm{Out}(G_v;\mathrm{Inc}_v^{(t)},\mathcal{K}_{\|G_v})\longrightarrow 1.$$

Vertex stabilizers are toral relatively hyperbolic, so the product is of type VF by Theorem 3.4 applied to the G_v. We conclude the proof by showing that $\mathcal{T}$ is of type F. This will imply that $\mathrm{Out}^1(T,\mathcal{K})$, and hence $\mathrm{Out}(T,\mathcal{K})$, is VF.

We claim that $\mathcal{T}$ is isomorphic to the direct product of a finitely generated abelian group and a finite number of copies of non-abelian vertex groups G_v. We use the presentation of $\mathcal{T}$ given in [27, Proposition 3.1]. It says that $\mathcal{T}$ can be written as a quotient

$$\mathcal{T}=\prod_{e,v}Z_{G_v}(G_e)/\langle\mathcal{R}_V,\mathcal{R}_E\rangle,$$

the product being taken over all pairs (e,v) where e is an edge incident to v; here $\mathcal{R}_E=\prod_e Z(G_e)$ is the group of edge relations and $\mathcal{R}_V=\prod_v Z(G_v)$ is the group of vertex relations, both embedded naturally in $\prod_{e,v}Z_{G_v}(G_e)$. Every group $Z_{G_v}(G_e)$ is abelian, unless G_e is trivial and G_v is non-abelian. In this case $Z_{G_v}(G_e)=G_v$ and it is not affected by the edge and vertex relations since both $Z(G_v)$ and $Z(G_e)$ are trivial. Our claim follows.

It follows that $\mathcal{T}$ is of type F provided that it is torsion-free. One may show that this is always the case, but it is simpler to replace $\mathrm{Out}^1(T,\mathcal{K})$ by its intersection with a torsion-free, finite-index subgroup of $\mathrm{Out}(G)$.

3.2.2 Changing T We shall now prove Theorem 3.11 in the general case.

The first step, carried out in this subsection, is to replace T by a better tree $\widehat{T}$ (satisfying the second assertion of the lemma below). When all edge stabilizers are non-trivial, $\widehat{T}$ may be viewed as the smallest common refinement (called lcm in [22]) of T and its tree of cylinders (see Section 2.2). Here is the construction of $\widehat{T}$.

Consider edges of T with non-trivial stabilizer. We say that two such edges belong to the same cylinder if their stabilizers commute. Cylinders are subtrees and meet in at most one point. A vertex v with all incident edge groups trivial belongs to no cylinder. Otherwise v belongs to one cylinder if G_v is abelian and to infinitely many cylinders if G_v is not abelian. To define $\widehat{T}$, we shall refine T at vertices x belonging to infinitely many cylinders.

Given such an x, let S_x be the set of cylinders Y such that $x \in Y$. We replace x by the cone T_x on S_x: there is a central vertex, again denoted by x, and vertices (x, s_Y) for $Y \in S_x$, with an edge between x and (x, s_Y). Edges e of T incident to x are attached to T_x as follows: if the stabilizer of e is trivial, we attach it to the central vertex x; if not, e is contained in a cylinder Y and we attach e to the vertex (x, s_Y), noting that G_e leaves Y invariant.

Performing this operation at each x belonging to infinitely many cylinders yields a tree $\widehat{T}$. The construction being canonical, there is a natural action of G on $\widehat{T}$, and $\mathrm{Out}(T) \subset \mathrm{Out}(\widehat{T})$.

Lemma 3.12 (1) *Edge stabilizers of $\widehat{T}$ are abelian, $\widehat{T}$ is dominated by T, and $\mathrm{Out}(\widehat{T}) = \mathrm{Out}(T)$.*

(2) *Let G_v be a non-abelian vertex stabilizer of $\widehat{T}$. Non-trivial incident edge stabilizers G_e are maximal abelian subgroups of G_v. If e_1 and e_2 are edges of $\widehat{T}$ incident to v with G_{e_1} and G_{e_2} equal and non-trivial, then $e_1 = e_2$.*

Proof Let Y be a cylinder in S_x (viewed as a subtree of T). The setwise stabilizer G_Y of Y is the maximal abelian subgroup of G containing stabilizers of edges of Y. The stabilizer of the vertex (x, s_Y) of $\widehat{T}$, and also of the edge between (x, s_Y) and x, is $G_x \cap G_Y$; it is non-trivial (it contains the stabilizer of edges of Y incident to x) and is a maximal abelian subgroup of G_x. This proves that edge stabilizers of $\widehat{T}$ are abelian, since the other edges have the same stabilizer as in T.

Every vertex stabilizer of T is also a vertex stabilizer of $\widehat{T}$, so T dominates $\widehat{T}$. Edges of $\widehat{T}$ which are not edges of T (those between (x, s_Y) and x) are characterized as those having non-trivial stabilizer and having an endpoint v with G_v non-abelian. One recovers T from $\widehat{T}$ by collapsing these edges, so $\mathrm{Out}(\widehat{T}) \subset \mathrm{Out}(T)$.

Consider two edges e_1 and e_2 incident to v in $\widehat{T}$, with the same non-trivial stabilizer. They join v to vertices (v, s_{Y_i}) and we have seen that $G_{e_1} = G_{e_2}$ is maximal abelian in G_v. The groups G_{Y_1} and G_{Y_2} are equal because they both contain $G_{e_1} = G_{e_2}$. Edges of Y_i have stabilizers contained in G_{Y_i}, so have commuting stabilizers. Thus $Y_1 = Y_2$, so $e_1 = e_2$. □

Remark 3.13 If G_{e_1} and G_{e_2} are conjugate in G_v, rather than equal, we conclude that e_1 and e_2 belong to the same G_v–orbit. On the other hand, edges belonging to different G_v–orbits may have stabilizers which are conjugate in G (but not in G_v).

3.2.3 The action on edge groups In Section 3.2.1 we could neglect the action of $\mathrm{Out}^0(T)$ on edge groups because all groups $\mathrm{Out}(G_e)$ were finite. We now allow edge stabilizers of arbitrary rank, so we must take these actions into account. We write $\mathrm{Out}^0(T,\mathcal{K}) = \mathrm{Out}^0(T) \cap \mathrm{Out}(G;\mathcal{K})$.

Recall that, for each edge e of $\Gamma = T/G$, there is a natural map $\rho_e\colon \mathrm{Out}^0(T) \to \mathrm{Out}(G_e)$ (see Section 2.3). The collection of all these maps defines a map

$$\psi\colon \mathrm{Out}^0(T,\mathcal{K}) \longrightarrow \prod_{e\in E} \mathrm{Out}(G_e),$$

the product being over all non-oriented edges of Γ. We denote by Q the image of $\mathrm{Out}^0(T,\mathcal{K})$ under ψ, so that we have the exact sequence

$$1 \longrightarrow \ker\psi \longrightarrow \mathrm{Out}^0(T,\mathcal{K}) \longrightarrow Q \longrightarrow 1.$$

Lemma 3.14 *If T satisfies the second assertion of Lemma 3.12, then the group Q is of type* VF.

This lemma will be proved in the next subsection. We first explain how to deduce Theorem 3.11 from it. The first assertion of Lemma 3.12 implies that the theorem holds for T if it holds for $\widehat{T}$, so we may assume that T satisfies the second assertion of Lemma 3.12.

The kernel of ψ is the group discussed in Section 3.2.1 under the name $\mathrm{Out}^1(T,\mathcal{K})$, but now (contrary to Convention 3.5) $\mathrm{Out}^1(T,\mathcal{K})$ may be of infinite index in $\mathrm{Out}(T,\mathcal{K})$; indeed, $\mathrm{Out}(T,\mathcal{K})$ is virtually an extension of $\mathrm{Out}^1(T,\mathcal{K})$ by Q. To avoid confusion, we use the notation $\ker\psi$ rather than $\mathrm{Out}^1(T,\mathcal{K})$.

We proved in Section 3.2.1 that $\ker\psi$ is of type VF and, by the lemma, Q is of type VF, but this is not quite sufficient (see Remark 3.3). We shall now construct a finite-index subgroup $\mathrm{Out}^2(T,\mathcal{K}) \subset \mathrm{Out}^0(T,\mathcal{K})$ such that $\ker\psi \cap \mathrm{Out}^2(T,\mathcal{K})$ has type F. Applying Corollary 3.2 to $\mathrm{Out}^0(T,\mathcal{K})$ then completes the proof of Theorem 3.11.

We argue as in Section 3.2.1. Recall from Section 2.3 the exact sequence

$$1 \longrightarrow \mathcal{T} \longrightarrow \mathrm{Out}^0(T,\mathcal{K}) \xrightarrow{\rho} \prod_{v\in V} \mathrm{Out}(G_v;\mathrm{Inc}_v,\mathcal{K}_{\|G_v})$$

whose restriction to $\ker\psi$ is the exact sequence

$$1 \longrightarrow \mathcal{T} \longrightarrow \ker\psi \xrightarrow{\rho} \prod_{v\in V} \mathrm{Out}(G_v;\mathrm{Inc}_v^{(t)},\mathcal{K}_{\|G_v}) \longrightarrow 1.$$

Using the "more precise" statement of Theorem 3.4 we get, for each $v \in V$, a finite-index subgroup $\mathrm{Out}^1(G_v; \mathrm{Inc}_v, \mathcal{K}_{\| G_v}) \subset \mathrm{Out}(G_v; \mathrm{Inc}_v, \mathcal{K}_{\| G_v})$ such that

$$\mathrm{Out}^1(G_v; \mathrm{Inc}_v, \mathcal{K}_{\| G_v}) \cap \mathrm{Out}(G_v; \mathrm{Inc}_v^{(t)}, \mathcal{K}_{\| G_v})$$

is of type F. Define the finite-index subgroup $\mathrm{Out}^2(T, \mathcal{K}) \subset \mathrm{Out}^0(T, \mathcal{K})$ as the preimage of $\prod_{v \in V} \mathrm{Out}^1(G_v; \mathrm{Inc}_v, \mathcal{K}_{\| G_v})$ under ρ intersected with a torsion-free, finite-index subgroup of $\mathrm{Out}(G)$.

Restricting the exact sequence above, we get an exact sequence

$$1 \longrightarrow \mathcal{T}' \longrightarrow \ker \psi \cap \mathrm{Out}^2(T, \mathcal{K}) \xrightarrow{\rho} L \longrightarrow 1,$$

where L has finite index in the product of the groups

$$\mathrm{Out}^1(G_v; \mathrm{Inc}_v, \mathcal{K}_{\| G_v}) \cap \mathrm{Out}(G_v; \mathrm{Inc}_v^{(t)}, \mathcal{K}_{\| G_v}),$$

hence has type F. The group $\mathcal{T}'$ is a torsion-free, finite-index subgroup of $\mathcal{T}$, so has type F as in Section 3.2.1. We conclude that $\ker \psi \cap \mathrm{Out}^2(T, \mathcal{K})$ has type F. As explained above, this completes the proof of Theorem 3.11 (assuming Lemma 3.14).

3.2.4 Proof of Lemma 3.14 There remains to prove Lemma 3.14. We let E_j be representatives of conjugacy classes of maximal abelian subgroups containing a non-trivial edge stabilizer. Note that E_j is allowed to be cyclic and maximal abelian subgroups of G containing no non-trivial G_e are not included.

Inside each E_j we let B_j be the smallest direct factor containing all edge groups included in E_j (it equals E_j if E_j is cyclic). It is elliptic in T, because it is an abelian group generated (virtually) by elliptic subgroups.

Each automorphism $\Phi \in \mathrm{Out}^0(T, \mathcal{K})$ induces an automorphism of E_j, which preserves B_j and all the edge groups it contains. This defines a map

$$\psi' \colon \mathrm{Out}^0(T, \mathcal{K}) \longrightarrow \prod_j \mathrm{Out}(B_j)$$

having the same kernel as the map $\psi \colon \mathrm{Out}^0(T, \mathcal{K}) \to \prod_{e \in E} \mathrm{Out}(G_e)$ defined in Section 3.2.3. Thus, it suffices to prove that the image of $\mathrm{Out}^0(T, \mathcal{K})$ by ψ' is of type VF. We do so by finding a finite-index subgroup $\mathrm{Out}^1(T, \mathcal{K})$ (not the same as in Section 3.2.1) whose image is a product $\prod_j Q_j$ with each Q_j of type VF.

Consider a non-abelian vertex group G_v. Define $\mathrm{Inc}_{v,\mathbb{Z}} \subset \mathrm{Inc}_v$ by keeping only the incident edge groups which are infinite cyclic, and denote by $E_{\mathrm{nc}}(v)$ the set of edges e of Γ with origin v and G_e non-cyclic (if e is a loop, we subdivide it so that it counts twice in $E_{\mathrm{nc}}(v)$). By Lemma 3.12 and Remark 3.13, the edge groups G_e for $e \in E_{\mathrm{nc}}(v)$ are non-conjugate maximal abelian subgroups of G_v.

We apply Proposition 3.10, describing the action on non-cyclic maximal abelian subgroups, to $\mathrm{Out}(G_v; \mathrm{Inc}_{v,\mathbb{Z}}^{(t)}, \mathcal{K}_{\|G_v})$. We get a subgroup $\mathrm{Out}'(G_v; \mathrm{Inc}_{v,\mathbb{Z}}^{(t)}, \mathcal{K}_{\|G_v})$ of finite index and a subgroup $F_e^v \subset G_e$ for each edge $e \in E_{\mathrm{nc}}(v)$ such that the image of $\mathrm{Out}'(G_v; \mathrm{Inc}_{v,\mathbb{Z}}^{(t)}, \mathcal{K}_{\|G_v})$ in $\prod_{e \in E_{\mathrm{nc}}(v)} \mathrm{Out}(G_e)$ is $\prod_{e \in E_{\mathrm{nc}}(v)} \mathrm{Out}(G_e; \{F_e^v\}^{(t)}, \mathcal{K}_{|G_e})$.

We let $\mathrm{Out}^1(T, \mathcal{K}) \subset \mathrm{Out}^0(T, \mathcal{K})$ be the subgroup consisting of automorphisms acting trivially on cyclic edge stabilizers and acting on non-abelian vertex stabilizers as an element of $\mathrm{Out}'(G_v; \mathrm{Inc}_{v,\mathbb{Z}}^{(t)}, \mathcal{K}_{\|G_v})$. It has finite index because

$$\mathrm{Out}(G_v; \mathrm{Inc}_{v,\mathbb{Z}}^{(t)}, \mathcal{K}_{\|G_v}) \subset \rho_v(\mathrm{Out}^0(T, \mathcal{K})) \subset \mathrm{Out}(G_v; \mathrm{Inc}_{v,\mathbb{Z}}, \mathcal{K}_{\|G_v}),$$

with all indices finite.

We now define $Q_j \subset \mathrm{Out}(B_j)$ as consisting of automorphisms Φ_j such that

(1) if G_e is a cyclic edge stabilizer contained in B_j, then Φ_j acts trivially on G_e;

(2) if B_j contains a non-cyclic G_e and v is an endpoint of e with G_v non-abelian, then Φ_j acts trivially on F_e^v;

(3) non-cyclic edge stabilizers and abelian vertex stabilizers contained in B_j are Φ_j–invariant;

(4) Φ_j extends to an automorphism of E_j leaving $\mathcal{K}_{|E_j}$ invariant; in particular, subgroups of B_j conjugate to a group of $\mathcal{K}$ are Φ_j–invariant.

This definition was designed so that the image of $\mathrm{Out}^1(T, \mathcal{K})$ by ψ' is contained in $\prod_j Q_j$. We claim that equality holds:

Lemma 3.15 *The image of* $\mathrm{Out}^1(T, \mathcal{K})$ *by* ψ' *equals* $\prod_j Q_j$.

Proof We fix automorphisms $\Phi_j \in Q_j \subset \mathrm{Out}(B_j)$ and we have to construct an automorphism $\Phi \in \mathrm{Out}^1(T, \mathcal{K})$. By (1) and (3) above, the Φ_j induce automorphisms Φ_e of edge stabilizers (each non-trivial edge group G_e lies in a unique E_j, so there is no ambiguity in the definition of Φ_e). As explained after Lemma 2.2, it suffices to find automorphisms Φ_v of vertex groups inducing the Φ_e. We distinguish several cases.

If G_v is contained in some B_j (up to conjugacy), it is Φ_j–invariant by (3), so we let Φ_v be the restriction.

If G_v is abelian but not contained in any B_j, we may assume that some incident G_e is non-cyclic (otherwise we let Φ_v be the identity). This G_e is contained in some B_j, and $G_v \subset E_j$. In fact, $G_v = E_j$: since G_v is not contained in B_j, it fixes only v, and E_j fixes v because it commutes with G_v. We may thus extend Φ_j to G_v using (4).

If G_v is not abelian, we construct Φ_v in $\mathrm{Out}'(G_v; \mathrm{Inc}_{v,\mathbb{Z}}^{(t)}, \mathcal{K}_{\|G_v})$ as follows. If $e \in E_{\mathrm{nc}}(v)$, the automorphism Φ_e acts trivially on F_e^v by (2), and preserves $\mathcal{K}_{|G_e}$ by (4). Thus, the collection of automorphisms Φ_e lies in $\prod_{e \in E_{\mathrm{nc}}(v)} \mathrm{Out}(G_e; \{F_e^v\}^{(t)}, \mathcal{K}_{|G_e})$. Proposition 3.10 guarantees that $\mathrm{Out}'(G_v; \mathrm{Inc}_{v,\mathbb{Z}}^{(t)}, \mathcal{K}_{\|G_v})$ contains an automorphism Φ_v inducing Φ_e for all $e \in E_{\mathrm{nc}}(v)$ (and acting trivially on all cyclic incident edge groups).

We have now constructed automorphisms $\Phi_v \in \mathrm{Out}(G_v)$ inducing the Φ_e, so Lemma 2.2 provides an automorphism $\Phi \in \mathrm{Out}^0(T)$ whose image in $\prod_j \mathrm{Out}(B_j)$ is the product of the Φ_j because B_j is virtually generated by edge stabilizers. We show $\Phi \in \mathrm{Out}^1(T, \mathcal{K})$. By construction it acts trivially on cyclic edge groups and acts on non-abelian vertex stabilizers as an element of $\mathrm{Out}'(G_v; \mathrm{Inc}_{v,\mathbb{Z}}^{(t)}, \mathcal{K}_{\|G_v})$. We just have to check that Φ leaves any $K \in \mathcal{K}$ invariant.

The group K is contained in some G_v. If K is contained in some B_j, it is Φ–invariant by (4). Otherwise, K fixes no edge. If G_v is abelian, we have seen that either all incident edge groups are cyclic (and Φ_v is the identity) or G_v equals some E_j and our choice of Φ_v using (4) guarantees that K is invariant. If G_v is not abelian, then K belongs to $\mathcal{K}_{\|G_v}$ because it fixes no edge. It is invariant because we chose $\Phi_v \in \mathrm{Out}'(G_v; \mathrm{Inc}_{v,\mathbb{Z}}^{(t)}, \mathcal{K}_{\|G_v})$. □

We have seen that the group Q of Lemma 3.14 is isomorphic to the image of $\mathrm{Out}^0(T, \mathcal{K})$ under ψ', hence contains $\prod_j Q_j$ with finite index. To show that Q is of type VF, there remains to show that each Q_j is of type VF.

We defined Q_j inside $\mathrm{Out}(B_j)$ by four conditions. As in Lemma 3.6, the first three define an arithmetic group. To deal with the fourth one, we consider the group $\widetilde{Q}_j$ consisting of automorphisms of E_j that leave B_j and $\mathcal{K}_{|E_j}$ invariant with the restriction to B_j satisfying the first three conditions. This is an arithmetic group. It consists of block-triangular matrices and one obtains Q_j by considering the upper-left blocks of matrices in $\widetilde{Q}_j$. It follows that K_j is arithmetic, as the image of an arithmetic group by a rational homomorphism [6, Theorem 6], hence of type VF by Lemma 3.6.

This completes the proof of Lemma 3.14, and hence of Theorem 3.11.

4 A finiteness result for trees

The goal of this subsection is Proposition 4.8, which gives a uniform bound for the size of certain sets of relative JSJ decompositions of G. This an essential ingredient in the proof of the chain condition for McCool groups. We will have to restrict to root-closed (RC) trees, which are introduced in Definitions 4.3 and 4.7 (they are closely related to the primary splittings of Dahmani and Groves [11]).

Definition 4.1 Let H be a subgroup of a group G. Its *root closure* $e(H,G)$, or simply $e(H)$, is the set of elements of G having a power in H. If $e(H)=H$, we say that H is *root-closed*.

If G is toral relatively hyperbolic and H is abelian, $e(H)$ is a direct factor of the maximal abelian subgroup containing H, and H has finite index in $e(H)$. Also note that, given $h \in G$ and $n \geq 2$, there exists at most one element g such that $g^n = h$.

The following fact is completely general:

Lemma 4.2 *Let T be a tree with an action of an arbitrary group. The following are equivalent:*

- *Vertex stabilizers of T are root-closed.*
- *Edge stabilizers of T are root-closed.*

Proof If g^n fixes an edge $e=vw$, it fixes v and w. If vertex stabilizers are root-closed, g fixes v and w, hence fixes e, so edge stabilizers are root-closed.

Conversely, if g^n fixes a vertex v, then g is elliptic, hence fixes a vertex w. Edges between v and w (if any) are fixed by g^n, hence by g if edge stabilizers are root-closed. Thus g fixes v. □

We now go back to a toral relatively hyperbolic group G.

Definition 4.3 A tree T is an *RC tree* if

- all non-cyclic abelian subgroups fix a point in T;
- edge stabilizers of T are abelian and root-closed.

When G is hyperbolic, RC trees are the $\mathcal{Z}_{\max}$–trees of Dahmani and Guirardel [12]: non-trivial edge stabilizers are maximal cyclic subgroups.

Lemma 4.4 (1) *Let T be an RC tree with all edge stabilizers non-trivial. Its tree of cylinders T_c (see Section 2.2) is an RC tree belonging to the same deformation space as T.*

(2) *If T_1 and T_2 are RC trees relative to some family $\mathcal{H}$ and edge stabilizers of T_1 are elliptic in T_2, there is an RC tree $\widehat{T}_1$ relative to $\mathcal{H}$ which refines T_1 and dominates T_2. Moreover, the stabilizer of any edge of $\widehat{T}_1$ fixes an edge in T_1 or in T_2.*

Proof Non-triviality of edge stabilizers ensures that T_c is defined. The vertex stabilizers of T_c are vertex stabilizers of T or maximal abelian subgroups, so are root-closed. The deformation space does not change because T is relative to non-cyclic abelian subgroups (see [23, Proposition 6.3]). This proves (1).

We define a refinement $\widehat{T}_1$ of T_1 dominating T_2 as in [21, Lemma 3.2], by blowing up each vertex v of T_1 into a G_v–invariant subtree of T_2. We just have to check that its edge stabilizers are root-closed. As in the proof of [12, Lemma 4.9], an edge stabilizer of $\widehat{T}_1$ is an edge stabilizer of T_1 or is the intersection of a vertex stabilizer of T_1 with an edge stabilizer of T_2, so is root-closed. □

Proposition 4.5 *Let G be toral relatively hyperbolic. In each of the following two cases, there is a bound for the number of orbits of edges of a minimal tree T with abelian edge stabilizers:*

(1) *T is bipartite: each edge has exactly one endpoint with abelian stabilizer (redundant vertices are allowed).*

(2) *T is an RC tree with no redundant vertex.*

Here and below, the bound has to depend only on G (it is independent of the trees under consideration).

Case 1 applies in particular to trees of cylinders.

Proof We cannot apply Bestvina and Feighn's accessibility theorem [3] directly because T does not have to be reduced in the sense of [3]: $\Gamma = T/G$ may have a vertex v of valence 2 such that an incident edge carries the same group as v. We say that such a v is a non-reduced vertex. The assumptions rule out the possibility that Γ contains long segments consisting of non-reduced vertices (as in the example at the top of [3, page 450]).

If T is bipartite, consider all non-reduced vertices of Γ and collapse exactly one of the incident edges. This yields a reduced graph of groups, and at most half of the edges of Γ are collapsed, so [3] gives a bound.

If T is an RC tree with no redundant vertex, every non-reduced vertex v of $\Gamma = T/G$ has exactly two adjacent edges e_v and f_v, whose groups satisfy $G_{e_v} \subsetneq G_v = G_{f_v}$. Among all edges incident to a non-reduced vertex, consider the set E_m consisting of those with G_e of minimal rank. No two edges of E_m are adjacent at a non-reduced vertex, because T is an RC tree. Now collapse the edges in E_m.

If $I = e_1 \cup e_2 \cup \cdots \cup e_k$ is a maximal segment in the complement of the set of vertices of Γ having degree 3 or carrying a non-abelian group, we never collapse adjacent edges e_i and e_{i+1} (and we do not collapse e_1 if $k = 1$; we may collapse e_1 and e_3 if $k = 3$). It follows that at least one third of the edges of Γ remain after the collapse.

Repeat the process. Denote by M the maximal rank of abelian subgroups of G. After at most M steps one obtains a graph of groups which is reduced in the sense of [3], hence has at most N edges for some fixed N. The number of edges of Γ is bounded by $3^M N$. □

Proposition 4.6 *Given a toral relatively hyperbolic group G, there exists a number M such that, if $T_1 \to T_2 \to \cdots \to T_p$ is a sequence of maps between RC trees belonging to distinct deformation spaces, then $p \leq M$.*

Proof There are two steps:

- The first step is to reduce to the case when no edge stabilizer is trivial. Consider the tree $\overline{T}_i$ (possibly a point) obtained from T_i by collapsing all edges with non-trivial stabilizer. A map $T_i \to T_{i+1}$ cannot send an arc with non-trivial stabilizer to the interior of an edge with trivial stabilizer, so $\overline{T}_i$ dominates $\overline{T}_{i+1}$. Vertex stabilizers of $\overline{T}_i$ are free factors; there are finitely many possibilities for their isomorphism type.

Using Scott's complexity, it is shown in [16, Section 2.2] that the number of times that the deformation space $\mathcal{D}_i$ of $\overline{T}_i$ differs from that of $\overline{T}_{i+1}$ is uniformly bounded. We may therefore assume that $\mathcal{D} = \mathcal{D}_i$ is independent of i.

Let $H_1, \dots, H_k$ be representatives of conjugacy classes of non-trivial vertex stabilizers of trees in $\mathcal{D}$. They are free factors of G, hence toral relatively hyperbolic, and k is bounded.

Consider the action of H_j on its minimal subtree $T_i^j \subset T_i$ (we let T_i^j be any fixed point if the action is trivial). It is an RC tree and no edge stabilizer is trivial. The deformation space of T_i is completely determined by $\mathcal{D}$ and the deformation spaces $\mathcal{D}_i^j$ of the trees T_i^j (viewed as trees with an action of H_j). It therefore suffices to bound (by a constant depending only on H_j) the number of times that $\mathcal{D}_i^j$ changes in a sequence $T_1^j \to T_2^j \to \cdots \to T_p^j$, so we may continue the proof under the additional assumption that the T_i have non-trivial edge stabilizers.

- Now that edge stabilizers are non-trivial, the tree of cylinders of T_i is defined. By the first assertion of Lemma 4.4, we may assume that it equals T_i.

Since all trees are trees of cylinders, we may assume, by [23, Proposition 4.11], that all domination maps $T_i \to T_{i+1}$ send vertex to vertex and map an edge to either a point or an edge. Such a map may collapse an edge to a point, or identify edges belonging to different orbits, or identify edges in the same orbit. The first two phenomena are easy to control, since they decrease the number of orbits of edges; controlling the third one requires more care (and restricting to RC trees).

We associate a complexity $(n, -s)$ to each T_i, with n the number of edges of T_i/G and s the sum of the ranks of its edge groups; complexities are ordered lexicographically. We claim that the complexity of T_{i+1} is strictly smaller than that of T_i. This gives the required uniform bound on p, since n (hence also s) is bounded by the first case of Proposition 4.5.

Let $f_i\colon T_i \to T_{i+1}$ be a domination map as above. Complexity clearly cannot increase when passing from T_i to T_{i+1}. If n does not decrease, no edge of T_i is collapsed in T_{i+1}. Since T_i and T_{i+1} belong to distinct deformation spaces, there exist distinct edges e and e' identified by f_i. They have to belong to the same orbit (otherwise n decreases), so $e' = ge$ for some $g \in G$. The group $\langle g, G_e\rangle$ fixes the edge $f_i(e) = f_i(e')$ of T_{i+1}, so is abelian. It has rank bigger than the rank of G_e because G_e is root-closed and $g \notin G_e$. Thus s increases, and the complexity decreases. □

Let $\mathcal{A}$ be the family of all abelian subgroups. Let $\mathcal{H}$ be a family of subgroups of G. A JSJ tree (over $\mathcal{A}$) relative to $\mathcal{H}$ may be defined as a tree T such that T is relative to $\mathcal{H}$, edge stabilizers of T are elliptic in every tree which is relative to $\mathcal{H}$, and T dominates every tree satisfying the previous conditions (all trees are assumed to have abelian edge stabilizers). This motivates the following definition, where we require that T be an RC tree (compare [12, Section 4.4]). Recall that $\mathcal{H}^{+\mathrm{ab}}$ is obtained by adding all non-cyclic abelian subgroups to $\mathcal{H}$.

Definition 4.7 Let G be a toral relatively hyperbolic group and $\mathcal{H}$ a family of subgroups. A tree T is an *RC JSJ tree relative to $\mathcal{H}^{+\mathrm{ab}}$* if

(1) T is relative to $\mathcal{H}^{+\mathrm{ab}}$ and is an RC tree;
(2) edge stabilizers of T are elliptic in every (not necessarily RC) tree with abelian edge stabilizers which is relative to $\mathcal{H}^{+\mathrm{ab}}$;
(3) T dominates every tree satisfying (1) and (2).

We will construct RC JSJ trees in Section 5. Note that non-cyclic edge stabilizers always satisfy (2).

Proposition 4.8 *Let G be a toral relatively hyperbolic group. Let $\mathcal{H}_1 \subset \cdots \subset \mathcal{H}_i \subset \cdots$ be an increasing sequence (finite or infinite) of families of subgroups with G freely indecomposable relative to $\mathcal{H}_1$. For each i, let U_i be an RC JSJ tree relative to $\mathcal{H}_i^{+\mathrm{ab}}$. There exists a number q, depending only on G, such that the trees U_i belong to at most q distinct deformation spaces.*

Proof Let U_i be as in the proposition. Note that U_i satisfies condition (1) of Definition 4.7 with respect to $\mathcal{H}_j^{+\mathrm{ab}}$ if $j \le i$ and condition (2) with respect to $\mathcal{H}_j^{+\mathrm{ab}}$ if $j \ge i$. But cyclic edge stabilizers of U_i do not necessarily satisfy (2) with respect to $\mathcal{H}_j^{+\mathrm{ab}}$ if $j < i$.

In general, there is no domination map $U_i \to U_{i+1}$, so we cannot apply Proposition 4.6 directly. The easy case is when, for each i, every cyclic edge stabilizer of U_{i+1} is contained in an edge stabilizer of U_i. Indeed, this implies that U_{i+1} satisfies condition (2) with respect to $\mathcal{H}_i^{+\mathrm{ab}}$ (not just to $\mathcal{H}_{i+1}^{+\mathrm{ab}}$). By condition (3), U_i dominates U_{i+1}, so Proposition 4.6 applies.

Next, assume that there is an RC tree T relative to $\mathcal{H}_1$ such that, for all i, there is a domination map $T \to U_i$ that collapses no edge. Each cyclic edge stabilizer G_e of U_{i+1} contains an edge stabilizer $G_{e'}$ of T (take for e' any edge whose image contains a subarc of e). Since G is freely indecomposable relative to $\mathcal{H}_1$ and T is relative to $\mathcal{H}_1$, one has $G_{e'} \neq 1$, and $G_{e'} = G_e$ because $G_{e'}$ is root-closed. Since the map $T \to U_i$ collapses no edge, G_e fixes an edge in U_i and we conclude as above.

We now construct such a tree T. By condition (2) of Definition 4.7, edge stabilizers of U_1 are elliptic in U_2, so by Lemma 4.4 there is an RC tree T_1 relative to $\mathcal{H}_1$ which refines U_1 and dominates U_2; we remove redundant vertices of T_1 if needed. Edge stabilizers of T_1 fix an edge in U_1 or U_2, so are elliptic in U_3 and one may iterate. One obtains RC trees T_i relative to $\mathcal{H}_1$ such that T_i refines T_{i-1} and dominates U_{i+1}. By Proposition 4.5, all trees T_i for i large enough are equal to a fixed RC tree T. We have no control over how large i has to be, but we have a uniform bound for the number of orbits of edges of T.

By construction, there are domination maps $f_i\colon T \to U_i$, but f_i may collapse some G–invariant set of edges. There are only a bounded number of possibilities for the set E_i of edges of T that are collapsed by f_i, so we may assume that $E = E_i$ is independent of i. Collapsing all edges of E then gives a tree T as wanted. □

5 The chain condition

We prove Theorem 1.5. In this section we only consider groups of the form $\mathrm{Out}(G; \mathcal{H}^{(t)})$, so we use the simpler notation $\mathrm{Mc}(\mathcal{H})$. Since we do not yet know that every $\mathrm{Mc}(\mathcal{H})$ is a McCool group, we assume that every $\mathcal{H}_i$ is a finite set of finitely generated subgroups (this is needed to apply Lemma 2.3).

Since $\mathrm{Mc}(\mathcal{H}') = \mathrm{Mc}(\mathcal{H} \cup \mathcal{H}')$ if $\mathrm{Mc}(\mathcal{H}) \supset \mathrm{Mc}(\mathcal{H}')$, we may assume $\mathcal{H}_i \subset \mathcal{H}_{i+1}$. We will use the following procedure several times. We associate an invariant to each family $\mathcal{H}_i$ and we show that, as i varies, the number of distinct values of the invariant is bounded (by which we mean that there is a bound depending only on G). We then continue the proof under the additional assumption that the value of the invariant is independent of i.

- The first invariant is the Grushko deformation space $\mathcal{D}_i$ relative to $\mathcal{H}_i$ (see Section 2.2). The assumption $\mathcal{H}_i \subset \mathcal{H}_{i+1}$ implies that $\mathcal{D}_i$ dominates $\mathcal{D}_{i+1}$. As in the proof of Proposition 4.6, it follows from [16] that the number of times that $\mathcal{D}_i$ changes is bounded. We may therefore assume that $\mathcal{D}_i$ is constant.

Let $G_1, \dots, G_n$ be the free factors in a Grushko decomposition $G = G_1 * \cdots * G_n * F_p$ relative to $\mathcal{H}_i$ (they do not depend on i up to conjugation since $\mathcal{D}_i$ is constant). The subgroup of $\mathrm{Mc}(\mathcal{H}_i)$ consisting of automorphisms sending each factor G_j to a conjugate has bounded index and it is determined by the McCool groups $\mathrm{Mc}_{G_j}(\mathcal{H}_{i|G_j})$, so we are reduced to the case when G is freely indecomposable relative to $\mathcal{H}_i$.

- We then consider the canonical JSJ tree T_i (over abelian subgroups) relative to $\mathcal{H}_i^{+\mathrm{ab}}$, ie to $\mathcal{H}_i$ and all non-cyclic abelian subgroups (see Section 2.2); it is $\mathrm{Mc}(\mathcal{H}_i)$–invariant. We cannot use Proposition 4.8 to say that the number of distinct T_i is bounded, because they are not RC trees, so we shall now replace T_i by an RC JSJ tree U_i.

Any edge e of T_i joins a vertex v_1 whose stabilizer is a maximal abelian subgroup to a vertex v_0 with non-abelian stabilizer. The group G_e is a maximal abelian subgroup of G_{v_0}, but not necessarily of G_{v_1}. Let $\bar{G}_e$ be the root-closure of G_e in G_{v_1} (hence also in G). As in [12, Section 4.3], we can fold all edges in the $\bar{G}_e$–orbit of e together. Doing this for all edges of T_i yields an RC tree U_i which is $\mathrm{Mc}(\mathcal{H}_i)$–invariant.

This construction may also be described in terms of graphs of groups, as follows. We now view $e = v_0v_1$ as an edge of T_i/G. Subdivide it by adding a midpoint u carrying $\bar{G}_e$. This creates two edges v_0u and uv_1, carrying G_e and $\bar{G}_e$, respectively. Do this for every edge e of T_i/G. Collapsing all edges uv_1 yields T_i/G, whereas collapsing all edges v_0u yields U_i/G.

The quotient graph U_i/G is the same as T_i/G, but labels are different. Edge groups are replaced by their root-closure and non-abelian vertex groups have gotten bigger (roots have been adjoined: each fold replaces some G_{v_0} by $G_{v_0} *_{G_e} \bar{G}_e$). Just like T_i, the tree U_i is equal to its tree of cylinders because folding only occurs within cylinders; in particular, U_i is determined by its deformation space.

Note that U_i may have redundant vertices and is not necessarily minimal (this happens if T_i/G has a terminal vertex carrying an abelian group, and the incident edge group has finite index). In this case we replace U_i by its minimal subtree.

We claim that U_i is an RC JSJ tree relative to $\mathcal{H}_i^{+\mathrm{ab}}$, in the sense of Definition 4.7. It satisfies conditions (1) and (2) since its edge stabilizers are finite extensions of edge stabilizers of T_i. Any tree satisfying these two conditions is dominated by T_i because T_i is a JSJ tree. But any RC tree dominated by T_i is also dominated by U_i (with notations as above, e and ge must have the same image if $g \in \overline{G}_e$).

- Proposition 4.8 lets us assume that U_i is a fixed tree U. It is invariant under every $\mathrm{Mc}(\mathcal{H}_i)$. We let $\mathrm{Out}^0(U)$ be the finite-index subgroup of $\mathrm{Out}(U)$ consisting of automorphisms preserving U and acting trivially on $\Gamma = U/G$. The number of edges of Γ is uniformly bounded, by Proposition 4.5, so the index of $\mathrm{Out}^0(U)$ in $\mathrm{Out}(U)$ is bounded and it is enough to prove the chain condition for $\mathrm{Mc}^0(\mathcal{H}_i) := \mathrm{Mc}(\mathcal{H}_i) \cap \mathrm{Out}^0(U)$.

Let V be the set of vertices of Γ. As recalled in Section 2.3, there are maps $\rho_v\colon \mathrm{Out}^0(U) \to \mathrm{Out}(G_v)$ and a product map $\rho\colon \mathrm{Out}^0(U) \to \prod_{v\in V} \mathrm{Out}(G_v)$. Since U is relative to $\mathcal{H}_i$, the group of twists $\mathcal{T} = \ker\rho$ is contained in $\mathrm{Mc}^0(\mathcal{H}_i)$.

Lemma 5.1 *There exist subgroups* $\mathrm{Out}^1(G_v) \subset \mathrm{Out}(G_v)$, *independent of* i, *such that*

(1) $\prod_{v\in V} \mathrm{Out}^1(G_v)$ *is contained in* $\rho(\mathrm{Mc}^0(\mathcal{H}_i))$ *for every* i;

(2) *the index of* $\mathrm{Out}^1(G_v)$ *in* $\rho_v(\mathrm{Mc}^0(\mathcal{H}_i))$ *is uniformly bounded.*

This lemma implies Theorem 1.5 because $\mathrm{Mc}^0(\mathcal{H}_i)$ contains $\rho^{-1}(\prod_{v\in V} \mathrm{Out}^1(G_v))$ with bounded index.

Proof of Lemma 5.1 Let $\mathcal{H}_{i,v} := (\mathcal{H}_i)_{\|G_v}$ be the set of (conjugacy classes of) subgroups of G_v which are conjugate to an element of $\mathcal{H}_i$ and which fix no other point in T (see Section 2.1). Since two such subgroups are conjugate in G_v if and only if they are conjugate in G, we may view $\mathcal{H}_{i,v}$ as a subset of $\mathcal{H}_i$.

Since, as explained in Section 2.3, $\rho(\mathrm{Mc}^0(\mathcal{H}_i))$ contains $\prod_{v\in V} \mathrm{Mc}(\mathrm{Inc}_v \cup \mathcal{H}_{i,v})$, it suffices to fix $v \in V$ and to construct $\mathrm{Out}^1(G_v)$ with $\mathrm{Out}^1(G_v) \subset \mathrm{Mc}(\mathrm{Inc}_v \cup \mathcal{H}_{i,v})$ and the index of $\mathrm{Out}^1(G_v)$ in $\rho_v(\mathrm{Mc}^0(\mathcal{H}_i))$ uniformly bounded. We distinguish several cases:

- First suppose that $G_v \simeq \mathbb{Z}^k$ is abelian, so $\mathrm{Out}(G_v) = \mathrm{Aut}(G_v) = \mathrm{GL}(k,\mathbb{Z})$. Let A_i be the root-closure of the subgroup of G_v generated by incident edge groups and subgroups in $\mathcal{H}_{i,v}$. It is a direct factor and increases with i, so we may assume that it is independent of i. We define $\mathrm{Out}^1(G_v) \subset \mathrm{Out}(G_v)$ as the subgroup consisting of automorphisms equal to the identity on A_i. It is equal to $\mathrm{Mc}(\mathrm{Inc}_v \cup \mathcal{H}_{i,v})$ and contained in $\rho_v(\mathrm{Mc}^0(\mathcal{H}_i))$. We must show that the index is bounded.

The group A_i is invariant under $\rho(\mathrm{Mc}^0(\mathcal{H}_i))$ and we have to bound the order of the image of $\mathrm{Mc}^0(\mathcal{H}_i)$ in $\mathrm{Out}(A_i)$. Any incident edge group $\bar{G}_e$ of G_v contains an edge stabilizer G_e of T_i with finite index, and the image of the map $\rho_e\colon \mathrm{Mc}^0(\mathcal{H}_i) \to \mathrm{Out}(G_e)$ is finite by Lemma 2.3. Since A_i is generated by incident edge groups and elements which are fixed by $\mathrm{Mc}^0(\mathcal{H}_i)$, this implies that the image of $\mathrm{Mc}^0(\mathcal{H}_i)$ in $\mathrm{Out}(A_i)$ is finite. Its cardinality is uniformly bounded because there is a bound for the order of finite subgroups of $\mathrm{GL}(k,\mathbb{Z})$, so the index of $\mathrm{Out}^1(G_v)$ in $\rho_v(\mathrm{Mc}^0(\mathcal{H}_i))$ is bounded.

- We now consider a non-abelian vertex stabilizer G_v. It follows from the way U_i was constructed that G_v is, for each i, the fundamental group of a graph of groups $\Lambda_{i,v}$. This graph is a tree. It has a central vertex v_i, which may be viewed as a vertex of T_i/G with G_{v_i} non-abelian. All edges e join v_i to a vertex u_e carrying a root-closed abelian group, and the index of G_e in G_{u_e} is finite. The graph of groups $\Lambda_{i,v}$ is invariant under the action of $\mathrm{Mc}^0(\mathcal{H}_i)$ on G_v.

We say that G_v (or v) is *rigid with sockets* or *QH with sockets*, depending on the type of v_i as a vertex of T_i (since the number of vertices of T_i/G is bounded, we may assume that this type is independent of i).

- If G_v is rigid with sockets, we define $\mathrm{Out}^1(G_v)$ as the trivial group and we have to explain why $\rho_v(\mathrm{Mc}^0(\mathcal{H}_i))$ is a finite group of bounded order. Assume first that $U = T_i$ (ie U is also a regular JSJ tree). Lemma 2.3 then implies that $\rho_v(\mathrm{Mc}^0(\mathcal{H}_i))$ is a finite subgroup of G_v, but we need to bound its order only in terms of G (independently of the sequence $\mathcal{H}_i$). To get this uniform bound, we note that there are only finitely many possibilities for G_v up to isomorphism by [24]. Moreover, $\mathrm{Out}(G_v)$ is virtually torsion-free by [25, Corollary 4.5], so there is a bound for the order of its finite subgroups.

In general (ie without assuming $U = T_i$), we study $\rho_v(\mathrm{Mc}^0(\mathcal{H}_i))$ through its action on the graph of groups $\Lambda_{i,v}$ as in Section 2.3 (note that edges are not permuted). The group of twists is trivial because edge groups are maximal abelian in G_{v_i} and terminal vertex groups are abelian (see [27, Proposition 3.1]), so we only have to control the action of $\mathrm{Mc}^0(\mathcal{H}_i)$ on vertex groups of $\Lambda_{i,v}$.

Applying Lemma 2.3 to the JSJ decomposition T_i, we get finiteness of the image of $\mathrm{Mc}^0(\mathcal{H}_i)$ in $\mathrm{Out}(G_{v_i})$ and in $\mathrm{Out}(G_e)$ for every edge e of T_i, and hence of $\Lambda_{i,v}$. The action of an automorphism on the edge groups of $\Lambda_{i,v}$ determines the action on the abelian vertex groups because they contain the incident edge group with finite index. This proves that $\rho_v(\mathrm{Mc}^0(\mathcal{H}_i))$ is finite, and boundedness follows as above.

- There remains the case when G_v is QH with sockets. The group G_{v_i} is then isomorphic to the fundamental group of a compact surface Σ_i and incident edge groups are boundary subgroups. The topology of Σ_i may vary with i, but the number of boundary components of Σ_i is bounded (by a simple accessibility argument, or because the rank of G_{v_i} as a free group is bounded, by [24]).

If J is a subgroup of G, denote by $\mathcal{U}_i(J)$ the set of elements of J that are $\mathcal{H}_i^{+\mathrm{ab}}$–universally elliptic (ie elliptic in every G–tree with abelian edge stabilizers which is relative to $\mathcal{H}_i$ and to non-cyclic abelian subgroups). We view it as a union of J–conjugacy classes. Since $\mathcal{H}_i \subset \mathcal{H}_{i+1}$, we have $\mathcal{U}_i(J) \subset \mathcal{U}_{i+1}(J)$. We shall show that the sequence $\mathcal{U}_i(G_v)$ stabilizes.

We first study $\mathcal{U}_i(G_{v_i})$: we claim that $\mathcal{U}_i(G_{v_i})$ is the union of the conjugacy classes of boundary subgroups of $G_{v_i} = \pi_1(\Sigma_i)$. Indeed, any boundary subgroup is an incident edge group of v_i (up to conjugacy) or has a finite-index subgroup conjugate to a group in $\mathcal{H}_i$ (otherwise, G would be freely decomposable relative to $\mathcal{H}_i$; see [21, Proposition 7.5]). It follows that $\mathcal{U}_i(G_{v_i})$ contains all boundary subgroups (incident edge groups are $\mathcal{H}_i^{+\mathrm{ab}}$–universally elliptic because T_i is a JSJ tree relative to $\mathcal{H}_i^{+\mathrm{ab}}$). Conversely, by [21, Proposition 7.6], any $g \in \mathcal{U}_i(G_{v_i})$ is contained in a boundary subgroup of $\pi_1(\Sigma_i)$. This proves our claim and shows, in particular, that $\mathcal{U}_i(G_{v_i})$ is the union of a bounded number of conjugacy classes of maximal cyclic subgroups $L_j(i)$ of G_{v_i}.

We now consider $\mathcal{U}_i(G_v)$. The $\mathcal{H}_i^{+\mathrm{ab}}$–universally elliptic elements of G_v are contained (up to conjugacy) in G_{v_i} or in one of the terminal vertex groups of $\Lambda_{i,v}$, so $\mathcal{U}_i(G_v)$ is the union of the conjugates of the root-closures (in G_v) of the groups $L_j(i)$. Since $\mathcal{H}_i \subset \mathcal{H}_{i+1}$, we have $\mathcal{U}_i(G_v) \subset \mathcal{U}_{i+1}(G_v)$. As $\mathcal{U}_i(G_v)$ is the union of the conjugates of a bounded number of cyclic subgroups, we may assume that $\mathcal{U}_i(G_v) = \mathcal{U}(G_v)$ does not depend on i.

Elements of $\rho_v(\mathrm{Mc}^0(\mathcal{H}_i))$ send each cyclic group in $\mathcal{U}(G_v)$ to a conjugate (conjugacy classes are not permuted because the action on T_i/G is trivial). They act trivially on groups in $\mathcal{H}_{i,v}$, but they may map an element g belonging to a terminal vertex group of $\Lambda_{v,i}$ to g^{-1} (geometrically, they correspond to homeomorphisms of Σ_i which may reverse orientation on boundary components).

We define $\mathrm{Out}^1(G_v) \subset \mathrm{Out}(G_v)$ as the group of automorphisms acting trivially on each cyclic group in $\mathcal{U}(G_v)$ (geometrically, we restrict to homeomorphisms of Σ_i equal to the identity on the boundary). It is contained in $\mathrm{Mc}(\mathrm{Inc}_v \cup \mathcal{H}_{i,v})$, because $\mathcal{U}_i(G_v)$ contains the incident edge groups of G_v in U, hence contained in $\rho_v(\mathrm{Mc}^0(\mathcal{H}_i))$, and the index is bounded in terms of the number of conjugacy classes of cyclic subgroups in $\mathcal{U}(G_v)$. $\square$

Remark 5.2 Groups of the form $\mathrm{Out}(G;\mathcal{H})$, with $\mathcal{H}$ a finite family of abelian groups, do not satisfy the descending chain condition: consider $G = \mathbb{Z}^2 = \langle x, y\rangle$ and $\mathcal{H}_i = \{\langle x, y^{2^i}\rangle\}$.

6 Proof of the other results

We first note the following consequence of the chain condition:

Proposition 6.1 *If $\mathcal{C}$ is an infinite family of conjugacy classes, there exists a finite subfamily $\mathcal{C}' \subset \mathcal{C}$ such that* $\mathrm{Mc}(\mathcal{C}) = \mathrm{Mc}(\mathcal{C}')$.

Recall that $\mathrm{Mc}(\mathcal{C})$ is the group of outer automorphisms fixing all conjugacy classes belonging to $\mathcal{C}$.

Proof Write $\mathcal{C}$ as an increasing union of finite families $\mathcal{C}_i$ and note that $\mathrm{Mc}(\mathcal{C})$ is the intersection of the descending chain $\mathrm{Mc}(\mathcal{C}_i)$. □

To prove Corollary 1.6, saying in particular that every McCool group is an elementary McCool group, we need the following fact:

Lemma 6.2 *Let G be a toral relatively hyperbolic group. Let H be a subgroup and $\alpha \in \mathrm{Aut}(G)$. If $\alpha(h)$ and h are conjugate in G for every $h \in H$, then α acts on H as conjugation by some $g \in G$.*

Proof We may assume that there is a non-trivial $h \in H$ such that $\alpha(h) = h$. If H is abelian, malnormality of maximal abelian subgroups implies that α is the identity on H. If not, the result follows from [31, Lemma 5.2] (which is valid for any homomorphism $\varphi\colon H \to G$, not just automorphisms of H); see also [2, Corollary 7.4]. □

Corollary 1.6 *Let G be a toral relatively hyperbolic group. If $\mathcal{H}$ is any family of subgroups of G, there exists a finite set of conjugacy classes such that* $\mathrm{Mc}(\mathcal{H}) = \mathrm{Mc}(\mathcal{C})$.

Recall that $\mathrm{Mc}(\mathcal{H})$ is also denoted by $\mathrm{Out}(G;\mathcal{H}^{(t)})$. We favor the notation $\mathrm{Mc}(\mathcal{H})$ in this subsection.

Proof Given an arbitrary family $\mathcal{H}$, let $\mathcal{C}_{\mathcal{H}}$ be the set of all conjugacy classes having a representative belonging to some H_i. By Lemma 6.2, $\mathrm{Mc}(\mathcal{H}) = \mathrm{Mc}(\mathcal{C}_{\mathcal{H}})$. We apply Proposition 6.1 to get $\mathrm{Mc}(\mathcal{H}) = \mathrm{Mc}(\mathcal{C})$ with $\mathcal{C}$ finite. □

Together with Theorem 3.11, this implies our most general finiteness result.

Corollary 6.3 *Let G be a toral relatively hyperbolic group. Let $\mathcal{H}$ be an arbitrary collection of subgroups of G. Let $\mathcal{K}$ be a finite collection of abelian subgroups of G. Let T be a simplicial tree on which G acts with abelian edge stabilizers, with each group in $\mathcal{H} \cup \mathcal{K}$ fixing a point.*

Then the group $\mathrm{Out}(T, \mathcal{H}^{(\mathrm{t})}, \mathcal{K}) = \mathrm{Out}(T) \cap \mathrm{Out}(G; \mathcal{H}^{(\mathrm{t})}, \mathcal{K})$ of automorphisms leaving T invariant, acting trivially on each group of $\mathcal{H}$ and sending each $K \in \mathcal{K}$ to a conjugate (in an arbitrary way) is of type VF.

Proof By Corollary 1.6, we may write $\mathrm{Out}(G; \mathcal{H}^{(\mathrm{t})}) = \mathrm{Mc}(\mathcal{C})$ for some finite family of conjugacy classes $[c_i]$, with each c_i belonging to a group of $\mathcal{H}$ and hence elliptic in T. Defining $\mathcal{L} = \{\langle c_i \rangle\}$, we see that $\mathrm{Mc}(\mathcal{C})$ is a finite-index subgroup of $\mathrm{Out}(G; \mathcal{L})$, so $\mathrm{Out}(T, \mathcal{H}^{(\mathrm{t})}, \mathcal{K})$ is a finite-index subgroup of $\mathrm{Out}(T, \mathcal{K} \cup \mathcal{L})$. By Theorem 3.11, this group has type VF and therefore so does $\mathrm{Out}(T, \mathcal{H}^{(\mathrm{t})}, \mathcal{K})$. $\square$

Proposition 1.7 and Theorem 1.8 will be proved at the end of the section.

Proposition 1.10 *Given a toral relatively hyperbolic group G, there exists a number C such that, if a subgroup $\widehat{M} \subset \mathrm{Out}(G)$ contains a group $\mathrm{Mc}(\mathcal{H})$ with finite index, then the index $[\widehat{M} : \mathrm{Mc}(\mathcal{H})]$ is bounded by C.*

Proof By Corollary 1.6, we may write $\mathrm{Mc}(\mathcal{H}) = \mathrm{Mc}(\mathcal{C}')$ for some finite set $\mathcal{C}'$. Let $\mathcal{C}$ be the orbit of $\mathcal{C}'$ under $\widehat{M}$. Since $\mathrm{Mc}(\mathcal{C}')$ fixes $\mathcal{C}'$, this is a finite $\widehat{M}$–invariant collection of conjugacy classes. We thus have

$$\mathrm{Mc}(\mathcal{C}) \subset \mathrm{Mc}(\mathcal{C}') \subset \widehat{M} \subset \widehat{\mathrm{Mc}}(\mathcal{C})$$

and it suffices to bound the index $[\widehat{\mathrm{Mc}}(\mathcal{C}) : \mathrm{Mc}(\mathcal{C})]$.

As in the beginning of Section 5, let $G = G_1 * \cdots * G_n * F_r$ be a Grushko decomposition of G relative to $\mathcal{C}$ and let $\mathcal{G} = \{G_1, \dots, G_n\}$. The group $\widehat{\mathrm{Mc}}(\mathcal{C})$ permutes the conjugacy classes of the groups in $\mathcal{G}$. Since the cardinality of $\mathcal{G}$ is bounded and G has finitely many free factors up to isomorphism, we may assume that G is one-ended relative to $\mathcal{C}$.

We now consider the JSJ decomposition T_{can} over abelian groups relative to $\mathcal{C}$ and non-cyclic abelian groups. It is invariant under $\widehat{\mathrm{Mc}}(\mathcal{C})$, so we may study $\widehat{\mathrm{Mc}}(\mathcal{C})$ through its action on T_{can} (see Section 2.3).

The number of edges of $\Gamma_{\mathrm{can}} = T_{\mathrm{can}}/G$ being bounded by the first case of Proposition 4.5, we may replace $\widehat{\mathrm{Mc}}(\mathcal{C})$ and $\mathrm{Mc}(\mathcal{C})$ by their subgroups $\widehat{\mathrm{Mc}}^0(\mathcal{C})$ and $\mathrm{Mc}^0(\mathcal{C})$ acting trivially on Γ. The group of twists $\mathcal{T}$ is contained in $\mathrm{Mc}^0(\mathcal{C})$, so as in the proof of

Lemma 5.1 it suffices to construct $\mathrm{Out}^1(G_v) \subset \mathrm{Mc}_{G_v}(\mathrm{Inc}_v \cup \mathcal{C}_{\|G_v})$ with the index of $\mathrm{Out}^1(G_v)$ in $\rho_v(\widehat{\mathrm{Mc}}^0(\mathcal{C}))$ uniformly bounded. We distinguish the same cases as in the proof of Lemma 5.1.

If G_v is abelian, isomorphic to $\mathbb{Z}^k$ with $k \geq 2$, let $H < G_v$ be the set of elements whose orbit under $\rho_v(\widehat{\mathrm{Mc}}^0(\mathcal{C}))$ is finite. This is a subgroup of G_v, isomorphic to some $\mathbb{Z}^p$, which is invariant under $\rho_v(\widehat{\mathrm{Mc}}^0(\mathcal{C}))$ and contains the incident edge groups by Lemma 2.3. We define $\mathrm{Out}^1(G_v) = \mathrm{Mc}_{G_v}(\{H\})$. It is contained in $\mathrm{Mc}_{G_v}(\mathrm{Inc}_v \cup \mathcal{C}_{\|G_v})$. The image of $\rho_v(\widehat{\mathrm{Mc}}^0(\mathcal{C}))$ in $\mathrm{Aut}(H) = \mathrm{GL}(p, \mathbb{Z})$ is finite, and its order bounds the index of $\mathrm{Out}^1(G_v)$ in $\rho_v(\widehat{\mathrm{Mc}}^0(\mathcal{C}))$. This concludes the proof in this case, since there is a bound for the order of finite subgroups of $\mathrm{GL}(p, \mathbb{Z})$.

If G_v is rigid, we let $\mathrm{Out}^1(G_v)$ be trivial. The image of $\widehat{\mathrm{Mc}}^0(\mathcal{C})$ in $\mathrm{Out}(G_v)$ is finite by Lemma 2.3, and bounded by [24] as in the proof of Lemma 5.1.

If $G_v = \pi_1(\Sigma)$ is QH, we define $\mathrm{Out}^1(G_v) = \mathcal{PM}^+(\Sigma) = \mathrm{Mc}_{G_v}(\mathrm{Inc}_v \cup \mathcal{C}_{\|G_v})$. Elements of $\rho_v(\widehat{\mathrm{Mc}}^0(\mathcal{C}))$ may reverse orientation, or permute boundary components of Σ. □

Corollary 6.4 *Extended elementary McCool groups $\widehat{\mathrm{Mc}}(\mathcal{C})$ of G satisfy a uniform chain condition.*

Proof Given a descending chain $\widehat{\mathrm{Mc}}(\mathcal{C}_i)$, define $\mathcal{C}'_i = \mathcal{C}_0 \cup \cdots \cup \mathcal{C}_i$ and note that

$$\mathrm{Mc}(\mathcal{C}'_i) = \bigcap_{j \leq i} \mathrm{Mc}(\mathcal{C}_j) \subset \widehat{\mathrm{Mc}}(\mathcal{C}_i) = \bigcap_{j \leq i} \widehat{\mathrm{Mc}}(\mathcal{C}_j) \subset \widehat{\mathrm{Mc}}(\mathcal{C}'_i).$$

The corollary follows from Theorem 1.5, since by Proposition 1.10 the index of $\mathrm{Mc}(\mathcal{C}'_i)$ in $\widehat{\mathrm{Mc}}(\mathcal{C}'_i)$ is bounded. □

We now prove Corollary 1.11, stating that, for any $A < \mathrm{Out}(G)$, there is a subgroup $A_0 < A$ of bounded finite index such that, for the action of A_0 on the set of conjugacy classes of G, every orbit is a singleton or is infinite.

Proof of Corollary 1.11 Let $\mathcal{C}_A$ be the (possibly infinite) set of conjugacy classes of G whose A–orbit is finite. Partition $\mathcal{C}_A$ into A–orbits and let $\mathcal{C}_p$ be the union of the first p orbits. The image of A in the group of permutations of $\mathcal{C}_p$ is contained in that of $\widehat{\mathrm{Mc}}(\mathcal{C}_p)$, so by Proposition 1.10 its order is bounded by some fixed C. This C also bounds the order of the image of A in the group of permutations of $\mathcal{C}_A$. □

Recall that $\mathrm{Ac}(\mathcal{H}, H_0) \subset \mathrm{Aut}(G)$ is the group of automorphisms acting trivially on $\mathcal{H}$ (in the sense of Definition 1.2, ie by conjugation) and fixing the elements of H_0. Proposition 1.13 states that, if G is non-abelian, then $\mathrm{Ac}(\mathcal{H}, H_0)$ is an extension

$$1 \longrightarrow K \longrightarrow \mathrm{Ac}(\mathcal{H}, H_0) \longrightarrow \mathrm{Mc}(\mathcal{H}') \longrightarrow 1$$

with $\mathrm{Mc}(\mathcal{H}') \subset \mathrm{Out}(G)$ a McCool group and K the centralizer of H_0. Corollary 1.14 states that the groups $\mathrm{Ac}(\mathcal{H}, H_0)$ are of type VF and satisfy a uniform chain condition.

Proof of Proposition 1.13 Let $\mathcal{H}' = \mathcal{H} \cup \{H_0\}$. Map $\mathrm{Ac}(\mathcal{H}, H_0) \subset \mathrm{Aut}(G)$ to $\mathrm{Out}(G)$. The image is $\mathrm{Mc}(\mathcal{H}')$. The kernel K is the set of inner automorphisms equal to the identity on H_0. Since G has trivial center, it is isomorphic to the centralizer of H_0. □

Proof of Corollary 1.14 The group $\mathrm{Mc}(\mathcal{H}')$ has type VF by Theorem 1.3. The group K is abelian or equal to G, so has type F because G does [10]. Proposition 1.13 and Corollary 3.2 imply that $\mathrm{Ac}(\mathcal{H}, H_0)$ has type VF. Moreover, a chain of centralizers has length at most 2 since the centralizer of H_0 is trivial, G or a maximal abelian subgroup. The uniform chain condition for McCool groups (Theorem 1.5) then implies the uniform chain condition for groups of the form $\mathrm{Ac}(\mathcal{H}, H_0)$. □

We now deduce the bounded chain condition for fixed subgroups.

Proof of Theorem 1.8 Let $J_0 \subsetneq J_1 \subsetneq \cdots \subsetneq J_p$ be a strictly ascending chain of fixed subgroups. Let $\mathrm{Ac}(\varnothing, J_i)$ be the subgroup of $\mathrm{Aut}(G)$ consisting of automorphisms equal to the identity on J_i. Since J_i is a fixed subgroup, $\mathrm{Ac}(\varnothing, J_i) \supsetneq \mathrm{Ac}(\varnothing, J_{i+1})$. Corollary 1.14 then gives a bound on the length of the chain. □

Remark One can adapt the arguments of Section 5 to prove Theorem 1.8 directly (without passing through McCool groups).

We now prove Proposition 1.7, saying that $\mathrm{Out}(F_n)$ contains infinitely many non-isomorphic McCool groups for $n \geq 4$ and infinitely many non-conjugate McCool groups for $n \geq 3$.

Proof of Proposition 1.7 Let H be the free group on three generators a, b, c. Given a non-trivial element $w \in \langle a, b \rangle$, let P_w be the cyclic HNN extension $P_w = \langle a, b, c, t \mid tct^{-1} = w \rangle$. It is free of rank 3, with basis a, b, t. Let φ_w be the automorphism of P_w fixing a and b and mapping t to wt (it equals the identity on H since it fixes $c = t^{-1}wt$). The image Φ_w of φ_w in $\mathrm{Out}(P_w)$ preserves the Bass–Serre tree T of the HNN extension (it belongs to its group of twists $\mathcal{T}$).

We apply this construction with $w = a^k b^k$ for k a positive integer. As k varies, the cyclic subgroups $\langle \Phi_w \rangle$ are pairwise non-conjugate in $\mathrm{Out}(P_w) \simeq \mathrm{Out}(F_3)$, as seen by considering the action on the abelianization.

We shall now prove the second assertion of the proposition for $n = 3$, by showing that $\langle \Phi_w \rangle$ is a McCool group of P_w, namely $\langle \Phi_w \rangle = \mathrm{Mc}_{P_w}(\{H\}) \subset \mathrm{Out}(F_3)$. The extension to $n > 3$ is straightforward, by adding generators to H.

Consider splittings of P_w over abelian (ie cyclic) subgroups relative to H. The tree T is a JSJ tree because its vertex stabilizers are universally elliptic [21, Lemma 4.7]; in particular, P_w is freely indecomposable relative to H. Moreover, T equals its tree of cylinders (up to adding redundant vertices) because w is not a proper power, so T is the canonical JSJ tree T_{can}. The McCool group $\mathrm{Mc}_{P_w}(\{H\})$ therefore leaves T invariant and it is easily checked using [27] that $\mathrm{Mc}_{P_w}(\{H\}) = \mathcal{T} = \langle \Phi_w \rangle$.

To prove the first assertion of the proposition, consider $R_w = P_w * \langle d \rangle \simeq F_4$, the family $\mathcal{H} = \{H, \langle d \rangle\}$ and the McCool group $\mathrm{Mc}_{R_w}(\mathcal{H}) \subset \mathrm{Out}(F_4)$. The decomposition $R_w = P_w * \langle d \rangle$ is a Grushko decomposition of R_w relative to $\mathcal{H}$ because P_w is freely indecomposable relative to H. This decomposition is invariant under $\mathrm{Mc}_{R_w}(\mathcal{H})$ because it is a one-edge splitting (see [14, Corollary 1.3]).

The stabilizer $\mathrm{Out}(T)$ of the Bass–Serre tree T in $\mathrm{Out}(R_w)$ is naturally isomorphic to

$$\mathrm{Aut}(P_w) \times \mathrm{Aut}(\langle d \rangle) \simeq \mathrm{Aut}(P_w) \times \mathbb{Z}/2\mathbb{Z}$$

(see [27]); the natural map $\mathrm{Out}(T) \to \mathrm{Out}(P_w)$ kills the factor $\mathbb{Z}/2\mathbb{Z}$ and coincides with the quotient map $\mathrm{Aut}(P_w) \to \mathrm{Out}(P_w)$ on the other factor. The McCool group $\mathrm{Mc}_{R_w}(\mathcal{H})$ is isomorphic to the preimage of $\mathrm{Mc}_{P_w}(\{H\}) = \langle \Phi_w \rangle$ in $\mathrm{Aut}(P_w)$, hence to the mapping torus

$$Q_w = \langle a, b, t, u \mid ua = au,\ ub = bu,\ utu^{-1} = a^k b^k t \rangle.$$

The abelianization of Q_w is $\mathbb{Z}^3 \times \mathbb{Z}/k\mathbb{Z}$, so the isomorphism type of Q_w changes when k varies. This proves the first assertion of the proposition for $n = 4$. The extension to larger n is again straightforward. $\square$

Appendix: Groups with finitely many McCool groups

In this appendix we describe cases when $\mathrm{Out}(G)$ only contains finitely many McCool subgroups. In particular, we show that the values of n given in Proposition 1.7 are optimal.

Proposition A.1 *If G is a torsion-free, one-ended hyperbolic group, then* $\mathrm{Out}(G)$ *only contains finitely many McCool groups up to conjugacy.*

Proposition A.2 $\mathrm{Out}(F_2)$ *only contains finitely many McCool groups up to conjugacy.*

Proposition A.3 $\mathrm{Out}(F_3)$ *only contains finitely many McCool groups up to isomorphism.*

The proof of Proposition A.1 requires the fact that $\mathrm{Out}(G)$, and, more generally, extended McCool groups $\widehat{\mathrm{Mc}}(\mathcal{C})$, only contain finitely many conjugacy classes of finite subgroups. This will appear in [17].

Proof of Proposition A.1 We assume that $\mathrm{Out}(G)$ contains infinitely many non-conjugate elementary McCool groups $\mathrm{Mc}(\mathcal{C}_i)$ and we derive a contradiction (this implies the proposition, by Corollary 1.6).

It is proved in [33, Corollary 4.9] that there are only finitely many minimal actions of G on trees with cyclic edge stabilizers, up to the action of $\mathrm{Out}(G)$, so we may assume that the canonical cyclic JSJ tree relative to $\mathcal{C}_i$ (the tree T_{can} of Section 2.2) is a given tree T. This tree is invariant under all groups $\mathrm{Mc}(\mathcal{C}_i)$, so $\mathrm{Mc}(\mathcal{C}_i) \subset \mathrm{Out}(T)$. In this proof, we cannot restrict to $\mathrm{Out}^0(T)$.

Given a vertex v of T, we define $\mathcal{C}_{i,v}$ as the restriction $\mathcal{C}_{i|G_v}$ if G_v is cyclic and as $\mathcal{C}_{i\|G_v}$ if G_v is not cyclic (recall from Section 2.1 that conjugacy classes represented by elements fixing an edge of T do not belong to $\mathcal{C}_{i\|G_v}$). The tree being bipartite, $\mathcal{C}_i$ is the disjoint union of the $\mathcal{C}_{i,v}$.

We say that v is *used* if $\mathcal{C}_{i,v}$ is non-empty. Since there are finitely many G–orbits of vertices, we may assume that usedness is independent of i; we let V_u be a set of representatives of orbits of used vertices. We may also assume that the type of vertices with non-cyclic stabilizer (rigid or QH) is independent of i (QH vertices with Σ a pair of pants are rigid; we do not consider them as QH).

We claim that QH vertices G_v of T are not used. Indeed, any boundary subgroup of G_v is an incident edge stabilizer of T: otherwise, G_v would split as a free product relative to Inc_v, contradicting one-endedness of G. Elements in $\mathcal{C}_i$ are universally elliptic (relative to $\mathcal{C}_i$) and the only universally elliptic subgroups of G_v are contained in boundary subgroups of G_v because G_v is flexible (see [21, Proposition 7.6]), so $\mathcal{C}_{i\|G_v}$ is empty.

For $v \in V_u$, define $\mathrm{Out}_i(G_v) \subset \mathrm{Out}(G_v)$ as the set of automorphisms which fix each conjugacy class in $\mathcal{C}_{i,v}$ and leave the set of incident edge stabilizers globally invariant. Any automorphism in $\mathrm{Mc}(\mathcal{C}_i)$ is an automorphism of T which leaves G_v invariant (up to conjugacy) and induces an automorphism belonging to $\mathrm{Out}_i(G_v)$. Conversely, any automorphism of T satisfying these properties for every $v \in V_u$ lies in $\mathrm{Mc}(\mathcal{C}_i)$. This means that $\mathrm{Mc}(\mathcal{C}_i)$ is completely determined by the knowledge of the groups $\mathrm{Out}_i(G_v)$ for $v \in V_u$.

We complete the proof by showing that there are only finitely many possibilities for each $\mathrm{Out}_i(G_v)$. This is clear if G_v is cyclic, and QH vertices are not used, so there remains to consider the case where G_v is rigid.

In this case, $\mathrm{Out}_i(G_v)$ is finite by Lemma 2.3 (otherwise G_v would have a cyclic splitting relative to Inc_v and $\mathcal{C}_{i,v}$, contradicting rigidity). Since G_v is hyperbolic, $\mathrm{Out}(G_v)$ has finitely many conjugacy classes of finite subgroups [17]. We deduce that there are finitely many possibilities for $\mathrm{Out}_i(G_v)$, up to conjugacy in $\mathrm{Out}(G_v)$. Unfortunately, this is not enough to get finiteness for $\mathrm{Mc}(\mathcal{C}_i)$ up to conjugacy in $\mathrm{Out}(G)$, because the conjugator may fail to extend to an automorphism of G.

To remedy this, we consider $\mathrm{Mc}(\mathrm{Inc}_v)$ and $\widehat{\mathrm{Mc}}(\mathrm{Inc}_v)$, with Inc_v the family of incident edge groups as in Section 2.1 and $\widehat{\mathrm{Mc}}(\mathrm{Inc}_v) = \widehat{\mathrm{Out}}(G_v; \mathrm{Inc}_v)$ the set of outer automorphisms of G_v preserving Inc_v (see Definition 2.1; edge groups may be permuted and the generator of an edge group may be mapped to its inverse).

The group $\mathrm{Out}_i(G_v) \subset \mathrm{Out}(G_v)$ is finite and contained in $\widehat{\mathrm{Mc}}(\mathrm{Inc}_v)$ (but not necessarily in $\mathrm{Mc}(\mathrm{Inc}_v)$). By [17], $\widehat{\mathrm{Mc}}(\mathrm{Inc}_v)$ has only finitely many conjugacy classes of finite subgroups. It follows that there are only finitely many possibilities for $\mathrm{Out}_i(G_v)$ up to conjugation by an element of $\widehat{\mathrm{Mc}}(\mathrm{Inc}_v)$, hence also up to conjugation by an element of $\mathrm{Mc}(\mathrm{Inc}_v)$ since $\mathrm{Mc}(\mathrm{Inc}_v)$ has finite index in $\widehat{\mathrm{Mc}}(\mathrm{Inc}_v)$.

We may therefore assume that $\mathrm{Out}_i(G_v)$ is independent of i if G_v is cyclic and $v \in V_u$, and that all groups $\mathrm{Out}_i(G_v)$ are conjugate by elements of $\mathrm{Mc}(\mathrm{Inc}_v)$ if $v \in V_u$ is rigid. Any element of $\mathrm{Mc}(\mathrm{Inc}_v)$ extends "by the identity" to an automorphism of G which leaves T invariant and acts trivially (as conjugation by an element of G) on G_w if w is not in the orbit of v. Since $\mathrm{Mc}(\mathcal{C}_i)$ is determined by the groups $\mathrm{Out}_i(G_v)$ for $v \in V_u$, we conclude that all groups $\mathrm{Mc}(\mathcal{C}_i)$ are conjugate in $\mathrm{Out}(G)$. □

Proof of Proposition A.2 We view $\mathrm{Out}(F_2) \simeq \mathrm{GL}(2, \mathbb{Z})$ as the mapping class group of a punctured torus Σ (with orientation-reversing maps allowed). Let c be a peripheral conjugacy class (representing the commutator of basis elements of F_2).

We consider a McCool group $\mathrm{Mc}(\mathcal{H}) \subset \mathrm{Out}(F_2)$. We may assume that $\mathrm{Mc}(\mathcal{H})$ is infinite. By the classification of elements of $\mathrm{GL}(2, \mathbb{Z})$ or by the Bestvina–Paulin method and Rips theory, F_2 then splits over a cyclic group relative to $\mathcal{H}$ and c (see for instance [25, Theorem 3.9]). Such a splitting is dual to a non-peripheral simple closed curve $\gamma \subset \Sigma$.

If there are two different splittings, they are dual to curves γ and γ' whose union fills Σ, so $\mathcal{H}$ only contains peripheral subgroups. It follows that $\mathrm{Mc}(\mathcal{H})$ is either $\mathrm{Out}(F_2) \simeq \mathrm{GL}(2,\mathbb{Z})$ or $\mathrm{SL}(2,\mathbb{Z})$. If the splitting is unique, $\mathrm{Mc}(\mathcal{H})$ fixes γ (viewed as an unoriented curve up to isotopy). Since the splitting dual to γ is relative to $\mathcal{H}$, the Dehn twist T_γ around γ is contained in $\mathrm{Mc}(\mathcal{H})$. The stabilizer $\mathrm{Stab}(\gamma)$ of γ in the mapping class group contains $\langle T_\gamma \rangle$ with finite index (the index is 4 because a homeomorphism may reverse the orientation of Σ and/or of γ). We thus have $\langle T_\gamma \rangle \subset \mathrm{Mc}(\mathcal{H}) \subset \mathrm{Stab}(\gamma)$, with both indices finite. Finiteness of $\mathrm{Mc}(\mathcal{H})$ up to conjugacy follows, since γ is unique up to the action of the mapping class group. □

The remainder of this appendix is devoted to the proof of Proposition A.3. We first record a few useful facts.

Lemma A.4 *Fix n. Up to isomorphism,* $\mathrm{Out}(F_n)$ *only contains finitely many virtually solvable subgroups.*

Proof Virtually solvable subgroups are virtually abelian [1; 5]. More precisely, they contain $\mathbb{Z}^k$ with $k \le 2n-3$ as a subgroup of bounded index (see [5, Proof of Theorem 1.1, page 94]). This implies finiteness, for instance by [32, Theorem 8.6]. □

Lemma A.5 *Let A be virtually cyclic and B be virtually F_n for some n. Up to isomorphism, there are only finitely many groups which are extensions of A by B.*

Proof This follows from standard extension theory [8, Sections III.10 and IV.6], noting that $\mathrm{Out}(A)$ is finite and B has a finite-index subgroup with trivial H^2. □

Proof of Proposition A.3 Now consider a McCool group $\mathrm{Mc}(\mathcal{H}) \subset \mathrm{Out}(F_3)$. The first step is to reduce to the case where F_3 is freely indecomposable relative to $\mathcal{H}$. If this does not hold, let Γ be a Grushko decomposition relative to $\mathcal{H}$ (see Section 2.2). It is not unique; we choose one with as few edges as possible.

If all vertex groups are cyclic, groups in $\mathcal{H}$ are generated (up to conjugacy) by powers of elements belonging to some fixed basis of F_3, and finiteness holds. Otherwise, there is a vertex group $G_v \simeq F_2$. Our choice of Γ implies that Γ has a single edge (it is an HNN extension, or an amalgam $F_2 * \mathbb{Z}$ with a finite-index subgroup of $\mathbb{Z}$ belonging to $\mathcal{H}$). It follows that Γ is $\mathrm{Mc}(\mathcal{H})$–invariant [14; 28] and $\mathrm{Mc}(\mathcal{H})$ is determined by its image in $\mathrm{Out}(F_2)$. This image is the McCool group $\mathrm{Mc}(\mathcal{H}_{|F_2})$, so finiteness follows from Proposition A.2.

We continue the proof under the assumption that F_3 is freely indecomposable relative to $\mathcal{H}$. Let Γ_{can} be the canonical $\mathrm{Mc}(\mathcal{H})$–invariant cyclic JSJ decomposition relative to $\mathcal{H}$ (see Section 2.2). Vertex groups G_v are cyclic, rigid or QH.

One easily checks the formula $\sum_v(\operatorname{rk} G_v - 1) = 2$. In particular, $\operatorname{rk} G_v \leq 3$ for all v and, if some G_v is isomorphic to F_3, then all other vertex groups are cyclic.

If $G_v \simeq \pi_1(\Sigma)$ is a QH vertex group, it is isomorphic to F_2 or F_3, so there are 9 possibilities for the compact surface Σ:

(1) Pair of pants.

(2) Sphere with 4 boundary components.

(3) Projective plane with 2 boundary components.

(4) Projective plane with 3 boundary components.

(5) Torus with 1 boundary component.

(6) Torus with 2 boundary components.

(7) Klein bottle with 1 boundary component.

(8) Klein bottle with 2 boundary components.

(9) Non-orientable surface of genus 3 with 1 boundary component.

Each incident edge group G_e is (up to conjugacy) a boundary subgroup of $\pi_1(\Sigma)$. Conversely, there are two possibilities for a boundary subgroup C. If it is an incident edge group, it equals G_e for a unique incident edge. If not, we say that the corresponding boundary component of Σ is *free*; in this case, some finite-index subgroup of C belongs to $\mathcal{H}$.

As in Section 2.3, the finite-index subgroup $\mathrm{Mc}^0(\mathcal{H})$ of $\mathrm{Mc}(\mathcal{H})$ acting trivially on Γ_{can} maps to $\prod_v \mathrm{Out}(G_v)$ with kernel the group of twists $\mathcal{T}$. The image in $\mathrm{Out}(G_v)$ is finite if G_v is cyclic or rigid, and virtually the mapping class group of Σ if G_v is QH, and $\mathcal{T}$ is isomorphic to some $\mathbb{Z}^k$ (see [25, Section 4.3]).

By mapping class group, we mean the group of isotopy classes of homeomorphisms of a compact surface Σ mapping each boundary component to itself in an orientation-preserving way. We denote it by $\mathcal{PM}^+(\Sigma)$ as in Section 2.2.

By Lemma A.4, we may assume that there is a QH vertex v with $\mathcal{PM}^+(\Sigma)$ non-solvable. As explained above, there are 9 possibilities for Σ. Cases 1, 3 and 7 are ruled out because $\mathcal{PM}^+(\Sigma)$ is virtually cyclic (see [34], or argue as in the proof of Proposition A.2, noting that a finite-index subgroup of $\mathcal{PM}^+(\Sigma)$ fixes a conjugacy class of F_2 which is not a power of the commutator).

If Γ_{can} is trivial (ie if the QH subgroup G_v is the whole group), $\mathrm{Mc}(\mathcal{H})$ is the mapping class group of Σ. We therefore assume that Γ_{can} is non-trivial.

Lemma A.6 *If G_v has rank 3, then Σ has a free boundary component.*

Proof This follows from [4, Lemma 4.1], a generalization of the standard fact that a cyclic amalgam $A *_{\langle c\rangle} B$ of free groups is free only if c belongs to a basis in A or B. □

This lemma rules out case 9.

Now suppose that all vertices of Γ_{can} other than v are terminal vertices carrying $\mathbb{Z}$ (by Lemma A.6, this holds in cases 6 and 8). In this case the group of twists $\mathcal{T}$ is trivial (see [27, Proposition 3.1]). The group $\mathrm{Mc}(\mathcal{H})$ contains $\mathcal{PM}^+(\Sigma)$ with finite index and there are finitely many possibilities: they depend on whether edges of Γ_{can} may be permuted and whether elements in edge groups may be mapped to their inverse.

We must now deal with cases 2, 4 and 5. We start with 4. The only possibility left is that Γ_{can} has two vertices v and w joined by 2 edges, with G_w cyclic. Every automorphism leaving Γ_{can} invariant maps G_v to itself (up to conjugacy), and we consider the natural map from $\mathrm{Mc}(\mathcal{H})$ to $\mathrm{Out}(G_v)$. As above, the image contains $\mathcal{PM}^+(\Sigma)$ with finite index and there are finitely many possibilities. The kernel is the group of twists $\mathcal{T}$, which is isomorphic to $\mathbb{Z}$. Since $\mathcal{PM}^+(\Sigma)$ is isomorphic to F_3 by [34, Theorem 7.5], we conclude by Lemma A.5.

The argument in case 2 is similar. Besides v and w, there may be another vertex w', with $G_{w'}$ cyclic and a single edge between v and w'. The group $\mathcal{PM}^+(\Sigma)$ is again free; it is isomorphic to F_2 (see for instance [13, Section 4.2.4]).

In case 5 (a once-punctured torus), there is a single edge incident to v. Collapsing all other edges yields a $\mathrm{Mc}(\mathcal{H})$–invariant decomposition as an amalgam $F_3 = G_v *_{\langle a\rangle} G_w$ with $G_w \simeq F_2$. By the standard fact recalled above, a belongs to a basis of G_w (and is equal to a commutator in G_v). The group $\mathrm{Mc}(\mathcal{H})$ acts trivially on the graph underlying this amalgam and the map ρ (see Section 2.3) maps $\mathrm{Mc}(\mathcal{H})$ to $\mathrm{Out}(G_v)\times\mathrm{Out}(G_w)$, with kernel the group of twists $\mathcal{T}$, isomorphic to $\mathbb{Z}$. The image in $\mathrm{Out}(G_v)$ is isomorphic to $\mathrm{GL}(2,\mathbb{Z})$ or $\mathrm{SL}(2,\mathbb{Z})$.

We now consider the image L of $\mathrm{Mc}(\mathcal{H})$ in $\mathrm{Out}(G_w)$. It preserves the conjugacy class of $\langle a\rangle$. If L is finite (necessarily of order at most 6), then $\mathrm{Mc}(\mathcal{H})$ maps onto $\mathrm{GL}(2,\mathbb{Z})$ or $\mathrm{SL}(2,\mathbb{Z})$ with virtually cyclic kernel K; there are finitely many possibilities for K up to isomorphism (it maps to L with cyclic kernel), and we conclude by Lemma A.5. As explained in the proof of Proposition A.2, if L is infinite, it is virtually cyclic, contains a "Dehn twist" T_a and has index at most 4 in the stabilizer of the conjugacy class of $\langle a\rangle$ in $\mathrm{Out}(G_w)$. Since $\mathrm{Mc}(\mathcal{H})$ is determined by its image in $\mathrm{Out}(G_v)\times\mathrm{Out}(G_w)$ and this image contains $\mathrm{SL}(2,\mathbb{Z})\times\langle T_a\rangle$, this leaves only finitely many possibilities. $\square$

References

[1] **E Alibegović**, *Translation lengths in* $\mathrm{Out}(F_n)$, Geom. Dedicata 92 (2002) 87–93 MR1934012

[2] **Y Antolin**, **A Minasyan**, **A Sisto**, *Commensurating endomorphisms of acylindrically hyperbolic groups and applications*, preprint (2013) arXiv:1310.8605

[3] **M Bestvina**, **M Feighn**, *Bounding the complexity of simplicial group actions on trees*, Invent. Math. 103 (1991) 449–469 MR1091614

[4] **M Bestvina**, **M Feighn**, *Outer limits*, preprint (1992) Available at http://andromeda.rutgers.edu/~feighn/papers/outer.pdf

[5] **M Bestvina**, **M Feighn**, **M Handel**, *Solvable subgroups of* $\mathrm{Out}(F_n)$ *are virtually abelian*, Geom. Dedicata 104 (2004) 71–96 MR2043955

[6] **A Borel**, *Density and maximality of arithmetic subgroups*, J. Reine Angew. Math. 224 (1966) 78–89 MR0205999

[7] **A Borel**, **J-P Serre**, *Corners and arithmetic groups*, Comment. Math. Helv. 48 (1973) 436–491 MR0387495

[8] **K S Brown**, *Cohomology of groups*, Graduate Texts in Mathematics 87, Springer, New York (1982) MR672956

[9] **M Culler**, **K Vogtmann**, *Moduli of graphs and automorphisms of free groups*, Invent. Math. 84 (1986) 91–119 MR830040

[10] **F Dahmani**, *Classifying spaces and boundaries for relatively hyperbolic groups*, Proc. London Math. Soc. 86 (2003) 666–684 MR1974394

[11] **F Dahmani**, **D Groves**, *The isomorphism problem for toral relatively hyperbolic groups*, Publ. Math. Inst. Hautes Études Sci. (2008) 211–290 MR2434694

[12] **F Dahmani**, **V Guirardel**, *The isomorphism problem for all hyperbolic groups*, Geom. Funct. Anal. 21 (2011) 223–300 MR2795509

[13] **B Farb**, **D Margalit**, *A primer on mapping class groups*, Princeton Mathematical Series 49, Princeton Univ. Press (2012) MR2850125

[14] **M Forester**, *Deformation and rigidity of simplicial group actions on trees*, Geom. Topol. 6 (2002) 219–267 MR1914569

[15] **R Geoghegan**, *Topological methods in group theory*, Graduate Texts in Mathematics 243, Springer, New York (2008) MR2365352

[16] **V Guirardel**, *Actions of finitely generated groups on $\mathbb{R}$–trees*, Ann. Inst. Fourier (Grenoble) 58 (2008) 159–211 MR2401220

[17] **V Guirardel**, **G Levitt**, *Extension finiteness for relatively hyperbolic groups*, in preparation

[18] **V Guirardel**, **G Levitt**, *Strata and canonical subtrees of $\mathbb{R}$–trees*, in preparation

[19] **V Guirardel**, **G Levitt**, *Deformation spaces of trees*, Groups Geom. Dyn. 1 (2007) 135–181 MR2319455

[20] **V Guirardel**, **G Levitt**, *The outer space of a free product*, Proc. Lond. Math. Soc. 94 (2007) 695–714 MR2325317

[21] **V Guirardel**, **G Levitt**, *JSJ decompositions: definitions, existence, uniqueness, I: The JSJ deformation space*, preprint (2009) `arXiv:0911.3173`

[22] **V Guirardel**, **G Levitt**, *JSJ decompositions: definitions, existence, uniqueness, II: Compatibility and acylindricity*, preprint (2010) `arXiv:1002.4564`

[23] **V Guirardel**, **G Levitt**, *Trees of cylinders and canonical splittings*, Geom. Topol. 15 (2011) 977–1012 MR2821568

[24] **V Guirardel**, **G Levitt**, *Vertex finiteness for splittings of relatively hyperbolic groups*, preprint (2013) `arXiv:1311.2835` to appear in Israel J. Math.

[25] **V Guirardel**, **G Levitt**, *Splittings and automorphisms of relatively hyperbolic groups*, Groups Geom. Dyn. 9 (2015) 599–663 MR3356977

[26] **M Handel**, **L Mosher**, *Subgroup decomposition in* $\mathrm{Out}(F_n)$*: introduction and research announcement*, preprint (2013) `arXiv:1302.2681`

[27] **G Levitt**, *Automorphisms of hyperbolic groups and graphs of groups*, Geom. Dedicata 114 (2005) 49–70 MR2174093

[28] **G Levitt**, *Characterizing rigid simplicial actions on trees*, from: "Geometric methods in group theory", (J Burillo, S Cleary, M Elder, J Taback, E Ventura, editors), Contemp. Math. 372, Amer. Math. Soc. (2005) 27–33 MR2139674

[29] **A Martino**, **E Ventura**, *Fixed subgroups are compressed in free groups*, Comm. Algebra 32 (2004) 3921–3935 MR2097438

[30] **J McCool**, *Some finitely presented subgroups of the automorphism group of a free group*, J. Algebra 35 (1975) 205–213 MR0396764

[31] **A Minasyan**, **D Osin**, *Normal automorphisms of relatively hyperbolic groups*, Trans. Amer. Math. Soc. 362 (2010) 6079–6103 MR2661509

[32] **D Segal**, *Polycyclic groups*, Cambridge Tracts in Mathematics 82, Cambridge Univ. Press (1983) MR713786

[33] **Z Sela**, *Acylindrical accessibility for groups*, Invent. Math. 129 (1997) 527–565 MR1465334

[34] **B Szepietowski**, *A presentation for the mapping class group of a non-orientable surface from the action on the complex of curves*, Osaka J. Math. 45 (2008) 283–326 MR2441942

Institut de Recherche Mathématique de Rennes, Université de Rennes 1 et CNRS (UMR 6625)
263 avenue du Général Leclerc, CS 74205, 35042 Rennes Cedex, France

Laboratoire de Mathématiques Nicolas Oresme, Université de Caen et CNRS (UMR 6139)
BP 5186, 14032 Caen Cedex 5, France

vincent.guirardel@univ-rennes1.fr, levitt@unicaen.fr

Received: 15 October 2014

Algebraic & Geometric Topology 15 (2015) 3535–3567

A generating set for the palindromic Torelli group

NEIL J FULLARTON

A *palindrome* in a free group F_n is a word on some fixed free basis of F_n that reads the same backwards as forwards. The *palindromic automorphism group* ΠA_n of the free group F_n consists of automorphisms that take each member of some fixed free basis of F_n to a palindrome; the group ΠA_n has close connections with hyperelliptic mapping class groups, braid groups, congruence subgroups of $GL(n, \mathbb{Z})$, and symmetric automorphisms of free groups. We obtain a generating set for the subgroup of ΠA_n consisting of those elements that act trivially on the abelianisation of F_n, the *palindromic Torelli group* $\mathcal{PI}_n$. The group $\mathcal{PI}_n$ is a free group analogue of the hyperelliptic Torelli subgroup of the mapping class group of an oriented surface. We obtain our generating set by constructing a simplicial complex on which $\mathcal{PI}_n$ acts in a nice manner, adapting a proof of Day and Putman. The generating set leads to a finite presentation of the principal level 2 congruence subgroup of $GL(n, \mathbb{Z})$.

20F65, 57M07, 57MXX

1 Introduction

Let F_n be the free group of rank n on some fixed free basis X. The palindromic automorphism group of F_n, denoted ΠA_n, consists of automorphisms of F_n that take each member of X to some palindrome, that is, a word on X that reads the same backwards as forwards. Collins [8] introduced the group ΠA_n and proved that it is finitely presented, giving an explicit presentation. Glover and Jensen [15] obtained further results about ΠA_n, utilising a contractible subspace of the auter space of F_n on which ΠA_n acts cocompactly, with finite stabilisers. For instance, they calculate that the virtual cohomological dimension of ΠA_n is $n-1$. The group ΠA_n is a free group analogue of the hyperelliptic mapping class group of an oriented surface; we develop this analogy later in this introduction.

In this paper, we are primarily concerned with the intersection of ΠA_n with the Torelli subgroup of F_n, that is, the subgroup of automorphisms of ΠA_n that act trivially on the abelianisation of F_n. We denote this intersection by $\mathcal{PI}_n$, and refer to it as the *palindromic Torelli group* of F_n. Little appears to be known about the group $\mathcal{PI}_n$: Collins [8] first observed that it is non-trivial, and Jensen, McCammond and Meier

Published: 12 January 2016 DOI: 10.2140/agt.2015.15.3535

[17, Corollary 6.3] showed that $\mathcal{PI}_n$ is not of finite homological type for $n \geq 3$. In Section 2, we introduce non-trivial members of $\mathcal{PI}_n$ $(n \geq 3)$ known as *doubled commutator transvections* and *separating π–twists*. The main theorem of this paper establishes that these generate $\mathcal{PI}_n$.

Theorem A *The group $\mathcal{PI}_n$ $(n \geq 3)$ is generated by doubled commutator transvections and separating π–twists.*

In order to prove Theorem A, we establish finite generating sets for the subgroups of $\Pi\mathrm{A}_n$ consisting of automorphisms that fix each member of some specified subset of the free basis X. These generating sets, which are given precisely in the statement of Proposition 2.2, are obtained by utilising Stallings' graph folding algorithm.

Let $\Gamma_n[2]$ denote the *principal level* 2 *congruence subgroup of* $\mathrm{GL}(n,\mathbb{Z})$, that is, the kernel of the surjection $\mathrm{GL}(n,\mathbb{Z}) \to \mathrm{GL}(n,\mathbb{Z}/2)$ that reduces matrix entries mod 2. In Section 2, we discuss a short exact sequence with kernel the palindromic Torelli group and quotient $\Gamma_n[2]$. For $1 \leq i \neq j \leq n$, let $S_{ij} \in \Gamma_n[2]$ be the matrix that has 1s on the diagonal and 2 in the (i,j) position, with 0s elsewhere, and let $O_i \in \Gamma_n[2]$ differ from the identity only in having -1 in the (i,i) position. The following corollary of Theorem A provides a finite presentation of $\Gamma_n[2]$ for $n \geq 4$.

Corollary 1.1 *The principal level* 2 *congruence group $\Gamma_n[2]$ $(n \geq 4)$ is generated by*

$$\{S_{ij}, O_i \mid 1 \leq i \neq j \leq n\},$$

subject to the defining relators

(1) $O_i{}^2$,
(2) $[O_i, O_j]$,
(3) $(O_i S_{ij})^2$,
(4) $(O_j S_{ij})^2$,
(5) $[O_i, S_{jk}]$,
(6) $[S_{ki}, S_{kj}]$,
(7) $[S_{ij}, S_{kl}]$,
(8) $[S_{ji}, S_{ki}]$,
(9) $[S_{kj}, S_{ji}]S_{ki}{}^{-2}$,
(10) $(S_{ij}S_{ik}{}^{-1}S_{ki}S_{ji}S_{jk}S_{kj}{}^{-1})^2$,

where $1 \leq i, j, k, l \leq n$ are pairwise different.

We note that in the proof of Theorem A it becomes apparent that not every relator of type 10 is needed. In fact, for each choice of three indices i, j and k, we need only select one such word (and disregard the others, for which the indices have been permuted).

We also derive the following similar presentation for $\Gamma_n[2]$ when $n = 2$ or 3; however, these are acquired independently of Theorem A. Indeed, the presentation of $\Gamma_3[2]$ is used to obtain a generating set for $\mathcal{PI}_3$, which forms the base case of an inductive proof of Theorem A.

Proposition 1.2 *The principal level* 2 *congruence group* $\Gamma_n[2]$ $(n=2,3)$ *is generated by*

$$\{S_{ij}, O_i \mid 1 \leq i \neq j \leq n\},$$

subject to the defining relators in the statement of Corollary 1.1 of types

- (1)–(4) *for* $n = 2$,
- (1)–(6), (8)–(10) *for* $n = 3$.

A key tool in the proof of Proposition 1.2 is an "even" version of the division algorithm for the integers. This is the observation that under certain circumstances, the quotient $q \in \mathbb{Z}$ given when dividing $a \in \mathbb{Z}$ by $b \in \mathbb{Z}$ may be chosen to be even, if we sacrifice control of the sign of the remainder $r \in \mathbb{Z}$. More details of this procedure are given in the proofs of Lemma 2.4 and Theorem 5.1.

A similar presentation for $\Gamma_n[2]$ was recently found independently by Kobayashi [18], and was also known to Margalit and Putman. As Margalit and Putman pointed out, this is a natural presentation for $\Gamma_n[2]$, as relators of types (6)–(9) bear a strong resemblance to the Steinberg relations that hold between the transvections generating $\mathrm{SL}(n, \mathbb{Z})$; see Milnor [22, Section 5].

A comparison with mapping class groups While $\Pi\mathrm{A}_n$ is defined entirely algebraically, it may viewed as a free group analogue of a subgroup of the mapping class group of an oriented surface. Let S_g and S_g^1 denote the compact, connected, oriented surfaces of genus g with 0 and 1 boundary components, respectively. We shall use S to denote such a surface, with or without boundary. Recall that the *mapping class group* of the surface S, denoted $\mathrm{Mod}(S)$, consists of isotopy classes of orientation-preserving self-homeomorphisms of S, where isotopies are required to fix any boundary component pointwise at all times. For a self-homeomorphism f of S, we denote its isotopy class by $[f]$.

A *hyperelliptic involution* of the surface S is an order-2 homeomorphism of the surface that acts as $-I$ on $H_1(S, \mathbb{Z})$; see Brendle and Margalit [4, Sections 2 & 4]. Let s denote the homeomorphism of S_g^1 seen in Figure 1. By capping the boundary with a disk, the map s induces a homeomorphism of S_g, which we also denote s, by an abuse of notation. The map s is an example of a hyperelliptic involution of S_g^1 (and S_g). We note that the mapping class of any hyperelliptic involution in $\mathrm{Mod}(S_g)$ $(g \geq 1)$ is conjugate to $[s]$; see Farb and Margalit [12, Proposition 7.15].

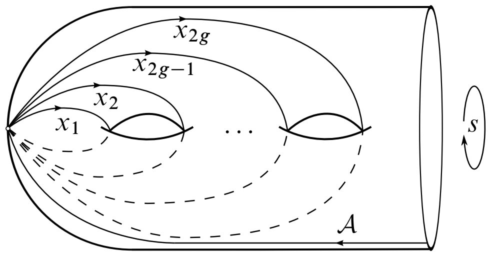


Figure 1: The involution s rotates the surface by π radians. Under the Nielsen embedding, we may view the braid group $B_{2g} \leq \mathrm{SMod}(S_g^1)$ as a subgroup of $\Pi\mathrm{A}_{2g} \leq \mathrm{Aut}(F_{2g})$.

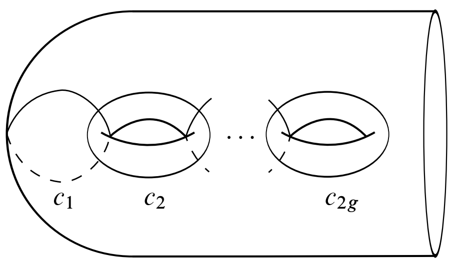


Figure 2: The standard symmetric chain in S_g^1. The Dehn twists about $c_1, \dots, c_{2g}$ generate $\mathrm{SMod}(S_g^1) \cong B_{2g+1}$.

The *hyperelliptic mapping class group of the surface* S_g, denoted $\mathrm{SMod}(S_g)$, is the centraliser of $[s]$ in $\mathrm{Mod}(S_g)$. Although $[s] \notin \mathrm{Mod}(S_g^1)$, as s does not fix the boundary of S_g^1, we define the *hyperelliptic mapping class group of* S_g^1, denoted $\mathrm{SMod}(S_g^1)$, to be the group of isotopy classes of the centraliser of s in $\mathrm{Homeo}^+(S_g^1)$ [12, Chapter 9].

An obvious analogue of a hyperelliptic involution in $\mathrm{Aut}(F_n)$ is an order-2 member of $\mathrm{Aut}(F_n)$ that acts as $-I$ on $H_1(F_n, \mathbb{Z}) = \mathbb{Z}^n$. An example of such an involution in $\mathrm{Aut}(F_n)$ is the automorphism ι that inverts each member of the free basis X. An analogy between s and ι is strengthened by two observations. Firstly, Glover and Jensen [15, Proposition 2.4] showed that any hyperelliptic involution in $\mathrm{Aut}(F_n)$ is conjugate to ι. Secondly, the action of s on $\pi_1(S_g^1) = F_{2g}$, with free basis as seen in Figure 1, is to invert each member of the free basis, as ι does. It is easily verified that $\Pi\mathrm{A}_n$ is the centraliser of ι in $\mathrm{Aut}(F_n)$ [15, Section 2], so we may think of $\Pi\mathrm{A}_n$ as being a free group analogue of the hyperelliptic mapping class groups $\mathrm{SMod}(S_g)$ and $\mathrm{SMod}(S_g^1)$.

The comparison between $\Pi\mathrm{A}_n$ and $\mathrm{SMod}(S_g^1)$ is made more precise using the classical Nielsen embedding $\mathrm{Mod}(S_g^1) \hookrightarrow \mathrm{Aut}(F_{2g})$. Take the $2g$ oriented loops seen in Figure 1 as a free basis for $\pi_1(S_g^1)$. Observe that s acts on these loops by switching

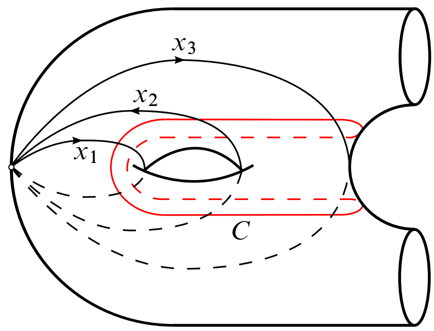


Figure 3: The Dehn twist about the symmetric, separating curve C maps to a separating π–twist in $\mathcal{PI}_{2g}$ under the Nielsen embedding. Note that we only depict a genus-one subsurface of S_g^1, and that x_2 has a different orientation than in Figure 1.

their orientations. In order to use Nielsen's embedding into $\mathrm{Aut}(F_{2g})$, we must take these loops to be based on the boundary; we surger using the arc $\mathcal{A}$ to achieve this. The group $\mathrm{SMod}(S_g^1)$ is isomorphic to the braid group B_{2g+1} by the Birman–Hilden theorem [3], and is generated by Dehn twists about the curves in the standard, symmetric chain on S_g^1, seen in Figure 2. The Dehn twists about the $2g-1$ curves $c_2,\dots,c_{2g}$ generate the braid group B_{2g}. Taking the loops seen in Figure 1 as our free basis X, a straightforward calculation shows that the images of these $2g-1$ twists in $\mathrm{Aut}(F_{2g})$ lie in $\Pi\mathrm{A}_{2g}$. Specifically, the twist about c_{i+1} is taken to the automorphism Q_i of the form

$$x_i \mapsto x_{i+1}, \quad x_{i+1} \mapsto x_{i+1}x_i{}^{-1}x_{i+1}, \quad x_j \mapsto x_j$$

for $1 \le i < 2g$ and $j \ne i, i+1$. This shows that $\Pi\mathrm{A}_n$ contains the braid group B_n as a subgroup, when n is even. This embedding of B_n is a restriction of one studied by Perron and Vannier [24] and Crisp and Paris [9]. When n is odd, we also have $B_n \hookrightarrow \Pi\mathrm{A}_n$, since discarding Q_1 gives a generating set for B_{2g-1} inside $\Pi\mathrm{A}_{2g-1} \le \mathrm{Aut}(F_{2g})$.

Palindromic and hyperelliptic Torelli groups The main focus of our study in this paper is the palindromic Torelli group $\mathcal{PI}_n$. This group arises as a natural analogue of a subgroup of $\mathrm{SMod}(S_g^1)$. The *Torelli subgroup* of $\mathrm{Mod}(S_g^1)$, denoted $\mathcal{I}_g^1$, consists of mapping classes that act trivially on $H_1(S_g^1,\mathbb{Z})$. There is non-trivial intersection between $\mathcal{I}_g^1$ and $\mathrm{SMod}(S_g^1)$; we define $\mathcal{SI}_g^1 := \mathrm{SMod}(S_g^1) \cap \mathcal{I}_g^1$ to be the *hyperelliptic Torelli group*. Brendle, Margalit and Putman [5] recently proved a conjecture of Hain [16], also stated by Morifuji [23], showing that $\mathcal{SI}_g^1$ is generated by Dehn twists about separating simple closed curves of genus one and two that are fixed by s.

Our generating set for $\mathcal{PI}_n$ compares favourably with Brendle, Margalit and Putman's for $\mathcal{SI}_g^1$, in the following way. We shall see in Section 2 that any Dehn twist about a symmetric separating curve of genus one that lies in the pre-image of the Nielsen embedding discussed above, maps to a separating π–twist in $\mathcal{PI}_{2g}$. In fact, up to conjugation by $\Pi\mathrm{A}_{2g}$, this is the definition of a separating π–twist. The Dehn twist about the curve C shown in Figure 3 is an example of such a mapping class. Note that the Dehn twist about C is one of the generators of Brendle, Margalit and Putman. We shall see in Proposition 3.7 that doubled commutator transvections do not suffice to generate $\mathcal{PI}_n$, so we observe that our generating set involves Brendle, Margalit and Putman's generators in a significant way. Thus, the similarity between $\mathcal{SI}_g^1$ and $\mathcal{PI}_n$ is not just a superficial comparison of definitions: the Nielsen embedding gives rise to a deeper connection between these two groups.

One way in which the analogy between $\mathcal{PI}_n$ and $\mathcal{SI}_g^1$ breaks down, however, is their behaviour when $\Pi\mathrm{A}_n$ and $\mathrm{SMod}(S_g^1)$ are abelianised, to $(\mathbb{Z}/2)^3$ and $\mathbb{Z}$, respectively. An immediate corollary of Theorem A is that $\mathcal{PI}_n$ vanishes in the abelianisation of $\Pi\mathrm{A}_n$. In contrast, the image of $\mathcal{SI}_g^1$ in the abelianisation of $\mathrm{SMod}(S_g^1)$ is $4\mathbb{Z}$, which may be shown by calculating the images of Brendle, Margalit and Putman's generators in the abelianisation of $\mathrm{SMod}(S_g^1)$.

Palindromes in right-angled Artin groups In forthcoming work with Anne Thomas [14], we extend Collins' definition of palindromic automorphisms to the right-angled Artin group setting. We obtain generating sets for the analogously defined palindromic automorphism group and palindromic Torelli group of an arbitrary right-angled Artin group.

Approach of the paper To prove Theorem A, we employ a standard, geometric technique: we find a sufficiently connected complex on which $\mathcal{PI}_n$ acts with sufficiently connected quotient, and use a theorem of Armstrong [1] to conclude that $\mathcal{PI}_n$ is generated by the action's vertex stabilisers. This approach is modelled on a proof of Day and Putman [11], which recovers Magnus' finite generating set for the Torelli subgroup of $\mathrm{Aut}(F_n)$.

Conventions We apply functions from right to left. For $g, h \in G$ a group, we let $[g,h] = ghg^{-1}h^{-1}$. In a graph, we denote an edge between vertices x and y by $x - y$. In a group G, we will also conflate a relation $P = Q$ with the relator PQ^{-1} when this is unambiguous.

Outline of the paper In Section 2, the definitions of the palindromic automorphism group and palindromic Torelli group of a free group are given, along with some elementary properties of these groups. In Section 3, we introduce the complex of partial π–bases of F_n, and use it to obtain a generating set for $\mathcal{PI}_n$. In Section 4, we prove key results about the connectivity of the complexes involved in the proof of Theorem A. In Section 5, we obtain a finite presentation for $\Gamma_3[2]$ used in the base case of our inductive proof of Theorem A.

Acknowledgements The results in this paper formed part of the author's PhD thesis [13]. The author thanks his PhD supervisor Tara Brendle for her guidance, and for introducing him to the palindromic Torelli group. The author is grateful for the hospitality of the Institute for Mathematical Research at Eidgenössische Technische Hochschule Zürich, where part of this work was completed. The author also thanks Ruth Charney and Karen Vogtmann for helpful discussions, and Thomas Church, Dan Margalit, Luis Paris, Andrew Putman, Richard Wade and Liam Watson for useful comments on an earlier version of this paper. The author is grateful to the referee for numerous constructive comments.

2 The palindromic automorphism group

Let F_n be the free group of rank n, on some fixed free basis $X := \{x_1, \dots, x_n\}$. For a word $w = l_1 \cdots l_k$ on $X^{\pm 1}$, let w^{rev} denote the *reverse* of w; that is, we have $w^{\mathrm{rev}} = l_k \cdots l_1$. Such a word w is said to be a *palindrome* on X if $w^{\mathrm{rev}} = w$. For example, x_1, ${x_2}^2$ and $x_2 x_1^{-1} x_2$ are all palindromes on X.

An automorphism $\alpha \in \mathrm{Aut}(F_n)$ is said to be *palindromic* (with respect to the fixed free basis X) if for each $x_i \in X$ the word $\alpha(x_i)$ may be written as a palindrome on X. Such automorphisms form a subgroup of $\mathrm{Aut}(F_n)$ which we call the *palindromic automorphism group of* F_n and denote by $\Pi\mathrm{A}_n$. That $\Pi\mathrm{A}_n$ is a group is easily shown by verifying that $\Pi\mathrm{A}_n$ is the centraliser in $\mathrm{Aut}(F_n)$ of the automorphism ι which inverts each member of X. The following proposition gives us information about the form of the palindromes $\alpha(x_i)$.

Proposition 2.1 *Let $\alpha \in \Pi\mathrm{A}_n$ and $x_i \in X$. Then $\alpha(x_i) = w^{\mathrm{rev}} \sigma(x_i)^{\epsilon_i} w$, where w is a word on $X^{\pm 1}$, σ is a permutation of X and $\epsilon_i \in \{\pm 1\}$.*

Proof For a palindrome $p = w^{\mathrm{rev}} x_i^{\epsilon_i} w \in F_n$ of odd length ($w \in F_n$, $x_i \in X$, $\epsilon_i \in \{\pm 1\}$), let $c(p) = x_i$. The following argument is implicit in the work of Collins [8].

Let $\alpha \in \Pi\mathrm{A}_n$. Since $\alpha(X)$ is a free basis, its image under the natural surjection $F_n \to (\mathbb{Z}/2)^n$ must suffice to generate $(\mathbb{Z}/2)^n$. If some $\alpha(x_i)$ is of even length, it will have zero image, and so the image of $\alpha(X)$ could not generate $(\mathbb{Z}/2)^n$. If

$c(\alpha(x_i)) = c(\alpha(x_j))$ for some $i \neq j$, then $\alpha(x_i)$ and $\alpha(x_j)$ will have the same image in $(\mathbb{Z}/2)^n$, and so again $\alpha(X)$ could not generate $\alpha(\mathbb{Z}/2)^n$. $\square$

Finite generation of $\Pi\mathrm{A}_n$ Collins first studied the group $\Pi\mathrm{A}_n$, giving a finite presentation for it. For $i \neq j$, let $P_{ij} \in \Pi\mathrm{A}_n$ map x_i to $x_j x_i x_j$ and fix x_k with $k \neq i$. For each $1 \leq j \leq n$, let $\iota_j \in \Pi\mathrm{A}_n$ map x_j to x_j^{-1} and fix x_k with $k \neq j$. We refer to P_{ij} as an *elementary* palindromic automorphism and to ι_j as an *inversion*. We let $\Omega^{\pm 1}(X)$ denote the group generated by the inversions and the permutations of X. The group generated by all elementary palindromic automorphisms and inversions is called the *pure palindromic automorphism group* of F_n, and is denoted $\mathrm{P}\Pi\mathrm{A}_n$.

Collins showed that $\Pi\mathrm{A}_n \cong \mathrm{E}\Pi\mathrm{A}_n \rtimes \Omega^{\pm 1}(X)$ for $n \geq 2$, where $\mathrm{E}\Pi\mathrm{A}_n = \langle P_{ij} \rangle$. The group $\Omega^{\pm 1}(X)$ acts on $\mathrm{E}\Pi\mathrm{A}_n$ in the natural way, and a defining set of relations for $\mathrm{E}\Pi\mathrm{A}_n$ is given by

(1) $[P_{ik}, P_{jk}] = 1$,

(2) $[P_{ij}, P_{kl}] = 1$,

(3) $P_{ij} P_{jk} P_{ik} = P_{ik}^{-1} P_{jk} P_{ij}$,

where i, j, k, l are pairwise different and the obviously undefined relators are omitted in the $n = 2$ and $n = 3$ cases.

We remark that, as noted by Collins [8], this presentation of $\mathrm{E}\Pi\mathrm{A}_n$ is very similar to one given for the *pure symmetric automorphism group of* F_n, $\mathrm{P}\Sigma A_n$, which consists of automorphisms taking each $x \in X$ to a conjugate of itself. This similarity is not entirely surprising, as we may think of a palindrome yxy as a conjugate yxy^{-1}, working "mod 2" $(x, y \in X)$. The embedding $B_n \hookrightarrow \Pi\mathrm{A}_n$ discussed in Section 1 bears a striking resemblance to Artin's faithful representation of B_n into ΣA_n, the full symmetric automorphism group, whose members take each $x \in X$ to *some* conjugate [2, Corollary 1.8.3]; this similarity arises via the branched double cover map $S_g^1 \to D_{2g+1}$ [12, Figure 9.13].

Using graph folding techniques of Stallings, we obtain a new proof of finite generation of $\Pi\mathrm{A}_n$, as well as finding generating sets for certain fixed-point subgroups of $\Pi\mathrm{A}_n$. We first introduce the notation and terminology of Wade [26] regarding graph folding.

Let R_n denote the wedge of n copies of S^1 at a point o. We canonically identify $\pi_1(R_n, o)$ with F_n by selecting an orientation of each S^1, and labelling the i^{th} copy of S^1 by $x_i \in X$. We shall let $\overline{x}_i$ denote the edge obtained by reversing the orientation of x_i.

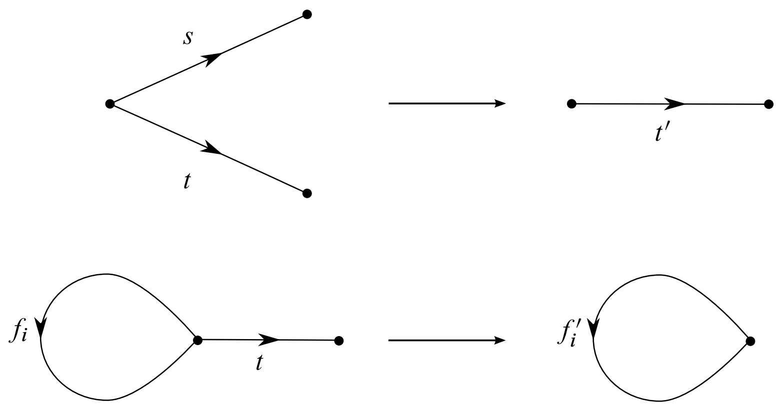


Figure 4: The two types of folding that may occur for our graph morphism ϕ. Wade [26] refers to the top fold as a type 1 fold, and to the bottom as a type 2 fold. The edges are labelled suggestively: we will demand that $s, t \in T$ and $f_i \notin T$.

Now, let Y be a finite graph of rank n with basepoint b. We will view our graphs as combinatorial objects, rather than topological ones. In particular, morphisms between graphs must take edges to edges, rather than edge-paths. A free basis for the (free) fundamental group $\pi_1(Y, b)$ is obtained in the usual way, by selecting a maximal tree T in Y, then choosing an orientation of the edges $f_1, \dots, f_n$ in Y but not T. To be consistent with Wade, we canonically orient an edge e of T by declaring its initial vertex $i(e)$ to be the one closer to the basepoint b under the edge-path metric on T.

Suppose $\theta\colon Y \to R_n$ is a morphism of graphs that induces an isomorphism of fundamental groups. The morphism θ, together with the choice of basepoint b, maximal tree T and an ordering L of the (oriented) edges of $Y \setminus T$ form a *branding* of the graph Y. A graph Y together with a 4-tuple $\mathcal{G} = (b, T, L, \theta)$ form a *branded graph* with *branding* $\mathcal{G}$.

Each branded graph Y with branding $\mathcal{G} = (b, T, L, \theta)$ yields an automorphism $B_{\mathcal{G}} \in \operatorname{Aut}(F_n)$, as follows. For each x_i in the free basis X of F_n, we have

$$B_{\mathcal{G}}(x_i) = \theta_*(y_i),$$

where $\{y_1, \dots, y_n\}$ is the free basis of $\pi_1(Y, b)$ arising from the choices of b, T and L in the branding $\mathcal{G}$, and $\theta_*\colon \pi_1(Y, b) \to \pi_1(R_n, o)$ is the map induced by θ.

If the morphism θ maps a pair of edges e_1 and e_2 with $i(e_1) = i(e_2)$ to the same edge l of R_n, then θ factors through the quotient graph Y' of Y obtained by *folding* e_1 and e_2 together: that is, the graph obtained by identifying e_1 with e_2, and also their terminal vertices, $t(e_1)$ and $t(e_2)$, with each other. In particular, if $q\colon Y \to Y'$ is the quotient map obtained by the folding, then there is a unique graph morphism $\theta'\colon Y' \to R_n$ such that $\theta = \theta' \circ q$. While Stallings considered more general foldings, since we require θ to induce an isomorphism of fundamental groups, only two types of folding may arise for us, which are shown in Figure 4.

If we insist that the edges s and t seen in Figure 4 lie in T, and that the edge f_i does not, carrying out either type of fold induces a branding $\mathcal{G}'$ of the folded graph Y' (it is non-trivial to verify that the image of T in Y' is a maximal tree; we leave this to Wade). It may also be the case that we wish to carry out a fold of type 1 or type 2, but that s or t does not lie in T. Before folding, we must change the maximal tree so that the relevant edges lie in the new tree. This defines a new branding $\mathcal{G}''$ of Y. In either case, it may be shown via a careful consideration of $\pi_1(Y,b)$ (see [26, Propositions 3.2 and 3.3]) that $B_{\mathcal{G}} = B_{\mathcal{G}'} \cdot W'$ and $B_{\mathcal{G}} = B_{\mathcal{G}''} \cdot W''$, where W' and W'' are specified *Whitehead automorphisms* of F_n. These are automorphisms which fix some $x \in X$ and send each $x_i \in X \setminus \{x\}$ to one of x_i, $x_i x^{\epsilon_i}$, $x^{\epsilon_i} x_i$ or $x^{\epsilon_i} x_i x^{-\epsilon_i}$ for some $\epsilon_i \in \{\pm 1\}$.

Stallings' folding algorithm allows us to repeatedly fold the graph Y and its quotients, beginning with the morphism $\theta\colon Y \to R_n$, then continuing to fold via $\theta'\colon Y' \to R_n$, and so on. This procedure eventually terminates when we exhaust the edges we are able to fold; in this case, Stallings showed that the quotient graph is R_n, and so the morphism $\psi\colon R_n \to R_n$ obtained by repeatedly folding via θ simply permutes and perhaps inverts the n loops in R_n. This folding procedure allows us to write the automorphism $B_{\mathcal{G}}$ we began with as a product of Whitehead automorphisms and permutations and inversions of X.

With the details of folding established, we now put the algorithm to use to find generators for $\Pi\mathrm{A}_n$.

Proposition 2.2 *Fix $0 \le k \le n$, and let $\Pi\mathrm{A}_n(k)$ consist of automorphisms which fix $x_1, \dots, x_k$. (Our convention is that $\Pi\mathrm{A}_n(0) = \Pi\mathrm{A}_n$). A finite generating set for $\Pi\mathrm{A}_n(k)$ is*

$$\left[\Omega^{\pm 1}(X) \cap \Pi\mathrm{A}_n(k)\right] \cup \{P_{ij} \mid i > k\}.$$

Proof The idea behind this proof was inspired by a proof of Wade [26, Theorem 4.1].

We begin by introducing some terminology. Let $\phi\colon S \to T$ be an isomorphism of finite trees. For a vertex or edge r of S, denote by r' the image of r under ϕ. Choose a distinguished vertex v of S, of valence 1. An *arch of S at v* (see Figure 5) is the graph formed by gluing S to T along v and v', then, for each vertex $r \in S$, adding some number of edges (possibly zero) between r and r' (we allow $r = v$). We refer

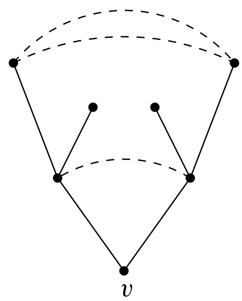


Figure 5: An example of an arch, with base point v. The dashed edges indicate the bridges that have been added to the trees that were glued together at the base point.

to these new edges as *bridges*. The image of v in the arch forms a natural base point, and any edge with v as one of its endpoints is called a *stem*. By a *wedge of arches* we mean a collection of arches glued together at their base points. Note that each of the trees S_i and T_i of each arch sit inside Y as subgraphs, and Y is the union of these subgraphs, together with any bridges inside each arch.

Let $\theta\colon Y \to R_n$ be a graph morphism, with Y a wedge of arches. We call θ *symmetric* if for each edge s_i in each tree S_i in each arch of Y we have $\theta(s_i') = \theta(\overline{s_i})$. We shall define two new types of folding that we may carry out to any symmetric morphism $\theta\colon Y \to R_n$, with the resulting morphism $\theta'\colon Y' \to R_n$ on the folded graph Y' also being symmetric.

Let $\alpha \in \Pi\mathrm{A}_n(k)$. We may realise α as a morphism of graphs $\theta\colon Z \to R_n$, where Z is the result of subdividing each S^1 of R_n into the appropriate number of edges, and "spelling out" the word $\alpha(x_i)$ on the i^{th} copy of S^1. Precisely, the j^{th} edge of the oriented, subdivided S^1 corresponding to $\alpha(x_i)$ is mapped to the loop in R_n corresponding to the j^{th} letter of $\alpha(x_i)$, correctly oriented. Note that Z is a wedge of arches, and θ is symmetric by construction. We thus have $\alpha = B_{\mathcal{G}}$, where $\mathcal{G}$ is the branding of Z arising from the maximal tree that excludes the (appropriately ordered) middle subdivided edge of each copy of S^1. We now use graph folding to write α as a product of permutations, inversions and elementary palindromic automorphisms.

Let $\theta\colon Y \to R_n$ be symmetric, for some wedge of arches Y, built out of trees S_i, T_i $(1 \le i \le k)$. Since θ is symmetric, foldings of Y come together in natural pairs. Consider folds of type 1. For instance, if we are able to fold together two edges $h_i \in S_i$ and $h_j \in S_j$ since $\theta(h_i) = \theta(h_j)$ (allowing $i = j$), then we will also be able to fold together h_i' and h_j', as they will also both have the same image under θ, namely $\theta(\overline{h_i}) = \theta(\overline{h_j})$. We call this pair of folds a *type A 2–fold*.

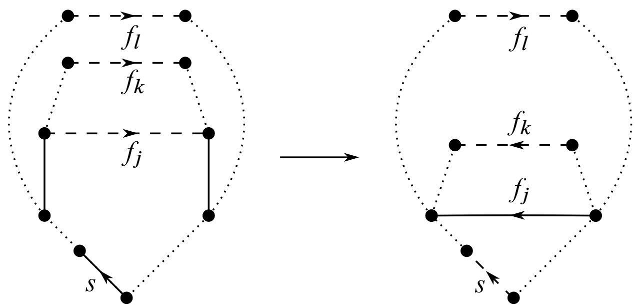


Figure 6: The two adjacent solid edges are folded onto f_j. The dashed edges represent edges excluded from the graph's chosen maximal tree. In order to record what effect this type B 2–fold has on the branded graph's associated automorphism, we must swap f_j into the maximal tree, in place of the stem s.

We may also have a sequence of edges (h_{j-1}, h_j, h_{j+1}) mapped under θ to the sequence $(\overline{x}, x, \overline{x})$ where x is an oriented edge of R_n, $h_{j-1} \in S_i$, $h_{j+1} = h'_{j-1}$ and h_j is a bridge. We fold h_{j-1} and h_{j+1} onto h_j, and call this pair of folds a *type B 2–fold*. Such a fold is seen in Figure 6.

Doing either of these 2–folds to Y yields another, different wedge of arches, Y', say. A type B 2–fold simply removes an edge of valence one from S_i (and its corresponding edge in T_i) by folding it onto a bridge, producing new trees S'_i and T'_i which we use to construct Y' as a wedge of arches. A type A 2–fold similarly alters the trees S_i, S_j, T_i and T_j, producing new trees S'_i and T'_i in a description of Y' as a wedge of arches. The morphism $\theta'\colon Y' \to R_n$ induced by the folding of Y is again symmetric: any edges s_i and s'_i that were not folded still satisfy $\theta'(s'_i) = \theta'(\overline{s_i})$ by construction of θ', but so do the images of any folded edges, given how we decompose Y' as a wedge of arches using the new trees S'_i and T'_i.

In order to see what effect these 2–folds have on $\alpha \in \Pi\mathrm{A}_n$, we must keep track of a preferred maximal tree T we define on each wedge of arches Y. The edges of Y not in T are the bridges coming from each arch. In order to carry out a type B 2–fold we must swap the bridge f_j (seen in Figure 6) into the maximal tree. Let $p_{i(f_j)}$ denote the unique reduced path in T joining the base point to the initial vertex of f_j. Apart from one degenerate case, which we deal with separately, we may always swap f_j into the maximal tree T by excluding the stem appearing in $p_{i(f_j)}$. Using calculations of Wade [26, Propositions 3.2 and 3.3], it is straightforward to verify that the effect of swapping maximal trees in this way, doing a type B 2–fold, then swapping back to the maximal tree where all bridges are excluded is to carry out an elementary palindromic automorphism $P_{ij}^{\epsilon_k}$ to some members of X. Precisely, let $\theta\colon Y_1 \to R_n$ be a symmetric morphism of graphs, where Y_1 has branding $\mathcal{G}_1$ and let $\mathcal{G}_2$ be the induced branding of the graph Y_2 obtained by carrying out the above series of tree swaps and folds. Then

$$\phi_{\mathcal{G}_1} = \phi_{\mathcal{G}_2} \cdot P,$$

where $\phi_{\mathcal{G}_i}$ is the automorphism of F_n associated to $\mathcal{G}_i$ $(i = 1, 2)$ and P is a product of elementary palindromic automorphisms.

The only degenerate case of the above is when one (and hence both) of the edges we want to fold onto a bridge is a stem. In this case, we do one of two things. If the bridge is a loop at the base point v, we carry out two type 2 folds. Otherwise, we change maximal trees as before then fold one of the stems onto the bridge with a type 1 fold. This causes the other stem to become a loop, around which we fold the bridge using a type 2 fold. As before, the automorphism of F_n associated to these sequences of steps is a product of elementary palindromic automorphisms.

Carrying out a sequence of 2–folds of types A and B eventually produces a map $R_n \to R_n$, and so we complete the folding algorithm by applying the appropriate automorphism from $\Omega^{\pm 1}(X)$. Since $\alpha \in \Pi\mathrm{A}_n(k)$, the graph Z we constructed has a single loop at the base point for each x_i $(1 \le i \le k)$, as $\alpha(x_i) = x_i$, so the first k ordered loops of R_n were not subdivided to form Z. Thus, while folding such a graph Y, we only need Collins' generators that fix the first k members of the free basis X. The proposition is thus proved. □

Corollary 2.3 *The group* $\mathrm{P}\Pi\mathrm{A}_n(k)$ *of pure palindromic automorphisms that fix* $x_1, \dots, x_k$ $(0 \le k \le n)$ *is generated by the set* $\{P_{ij}, \iota_i \mid i > k\}$.

The principal level 2 congruence subgroup of GL(n, $\mathbb{Z}$) Recall that $\Gamma_n[2]$ denotes the *principal level* 2 *congruence subgroup* of $\mathrm{GL}(n, \mathbb{Z})$, that is, the kernel of the map $\mathrm{GL}(n, \mathbb{Z}) \to \mathrm{GL}(n, \mathbb{Z}/2)$ given by reducing matrix entries mod 2. Let S_{ij} be the matrix with 1s on the diagonal, 2 in the (i, j) position and 0s elsewhere, and let O_i be the matrix which differs from the identity matrix only in having a -1 in the (i, i) position. The following lemma verifies a well-known generating set for $\Gamma_n[2]$ (see, for example, McCarthy and Pinkall [21, Corollary 2.3]). We include a proof here to introduce the idea of an "even division algorithm", which we utilise in the proof of Theorem 5.1.

Lemma 2.4 *The set* $\{O_i, S_{ij} \mid 1 \le i \ne j \le n\}$ *generates* $\Gamma_n[2]$.

Proof Observe that we may think of the matrices S_{ij} as corresponding to carrying out "even" row operations, that is, adding an even multiple of one matrix row to another. Let u be the first column of some matrix in $\Gamma_n[2]$, and denote by $u^{(i)}$ the i^{th} row of u. Let v_1 be the standard column vector with a 1 in the first entry and 0s elsewhere.

Claim The column u can be reduced to $\pm v_1$ using even row operations.

We use induction on $|u^{(1)}|$. For $|u^{(1)}| = 1$, the claim is obvious. Now suppose $|u^{(1)}| > 1$. As in the proof of Proposition 2.1, we deduce that there must be some $u^{(j)}$ which is not a multiple of $u^{(1)}$. By the division algorithm, there exist $q, r \in \mathbb{Z}$ such that $u^{(j)} = q|u^{(1)}| + r$, with $0 \leq r < |u^{(1)}|$. If q is not even, we instead write $u^{(j)} = (q+1)|u^{(1)}| + (r - |u^{(1)}|)$. Note that if q is odd, then $r \neq 0$, since $u^{(1)}$ is odd and $u^{(j)}$ is even, and so $-|u^{(1)}| < r - |u^{(1)}|$. Depending on the parity of q, we do the appropriate number of even row operations to replace $u^{(j)}$ with r or $r - |u^{(1)}|$. In both cases, we have replaced $u^{(j)}$ with an integer of absolute value smaller than $|u^{(1)}|$. It is clear that now we may reduce the absolute value of $u^{(1)}$ by either adding or subtracting twice the (new) j^{th} row from the first row, and so by induction we have proved the claim.

We now induct on n to prove the lemma. It is clear that $\Gamma_1[2] = \langle O_1 \rangle$. Using the above claim, we may assume that we have reduced $M \in \Gamma_n[2]$ to the form

$$\left[\begin{array}{c|c} \pm 1 & * \\ \hline 0 & N \end{array}\right],$$

where $N \in \Gamma_{n-1}[2]$. Our aim is to further reduce M to the identity matrix using the set of matrices in the statement of the lemma. By induction, we may assume that N can be reduced to the identity matrix using the appropriate members of $\{S_{ij}, O_i \mid i, j > 1\}$. Then we simply use even row operations to fix the top row, and finish by applying O_1 if necessary. $\square$

By Lemma 2.4, the restriction of the canonical map $\mathrm{Aut}(F_n) \to \mathrm{GL}(n, \mathbb{Z})$ gives the short exact sequence

$$1 \longrightarrow \mathcal{PI}_n \longrightarrow \mathrm{P\Pi A}_n \longrightarrow \Gamma_n[2] \longrightarrow 1,$$

since P_{ij} maps to S_{ji} and ι_i maps to O_i.

The rest of the paper is concerned with finding a generating set for the palindromic Torelli group $\mathcal{PI}_n$. In order to describe our generating set, we introduce some terminology.

Let Y be the image of the free basis X under some automorphism $\alpha \in \Pi\mathrm{A}_n$. The set Y is also a free basis for F_n, whose members are palindromes on X; thus, we refer to Y as a *π–basis*. An automorphism $\phi \in \mathcal{PI}_n$ is a *doubled commutator transvection* if, for some y_1, y_2, y_3 in some π–basis Y, ϕ maps y_1 to $[y_2, y_3]^{\mathrm{rev}} y_1 [y_2, y_3]$, and fixes the other members of Y. Observe that $\phi \in \mathcal{PI}_n$ is a doubled commutator transvection if and only if ϕ is conjugate in $\Pi\mathrm{A}_n$ to the commutator $\chi_1 := [P_{12}, P_{13}]$.

An automorphism $\phi \in \mathcal{PI}_n$ is a *separating π–twist* if, for some y_1, y_2, y_3 in some π–basis Y, ϕ is given by

$$\phi(y_i) = \begin{cases} d^{\mathrm{rev}} y_1 d & \text{if } i = 1, \\ d^{-1} y_2 (d^{\mathrm{rev}})^{-1} & \text{if } i = 2, \\ d^{\mathrm{rev}} y_3 d & \text{if } i = 3, \\ y_i & \text{otherwise,} \end{cases}$$

where $d = {y_1}^{-1} {y_2}^{-1} {y_3}^{-1} y_1 y_2 y_3 \in F_n$. It is a straightforward, if lengthy, calculation to verify that $\phi \in \mathcal{PI}_n$ is a separating π–twist if and only if ϕ is conjugate in $\Pi\mathrm{A}_n$ to the automorphism

$$\chi_2 := (P_{23} {P_{13}}^{-1} P_{31} P_{32} P_{12} {P_{21}}^{-1})^2 \in \mathcal{PI}_n.$$

The definition of a separating π–twist may seem unwieldy; however, it belies a hidden geometry. The automorphism χ_2 is the image in $\mathcal{PI}_n$ under the Nielsen embedding of the Dehn twist about the curve C seen in Figure 3. We call such automorphisms separating π–twists to reflect this geometric interpretation.

Theorem A states that doubled commutator transvections and separating π–twists suffice to generate $\mathcal{PI}_n$. To prove this, we construct a new complex on which $\mathcal{PI}_n$ acts in a suitable way. We then apply a theorem of Armstrong [1] to conclude that $\mathcal{PI}_n$ is generated by the action's vertex stabilisers. In the following section, we define the complex and use it to prove Theorem A.

3 The complex of partial π–bases

Day and Putman [11] use the *complex of partial bases* of F_n, denoted $\mathcal{B}_n$, to derive a generating set for IA_n. We build a complex modelled after $\mathcal{B}_n$, and follow their approach to find a generating set for $\mathcal{PI}_n$.

Fix $X := \{x_1, \dots, x_n\}$ as a free basis of F_n. A *π–basis*, as discussed above, is a set of palindromes on X which also forms a free basis of F_n. A *partial π–basis* is a set of palindromes on X which may be extended to a π–basis. The *complex of partial π–bases* of F_n, denoted $\mathfrak{B}_n^\pi$, is defined to be the simplicial complex whose $(k-1)$–simplices correspond to partial π–bases $\{w_1, \dots, w_k\}$. We postpone until Section 4 the proof of the following theorem on the connectedness of $\mathfrak{B}_n^\pi$.

Theorem 3.1 *For $n \geq 3$, the complex $\mathfrak{B}_n^\pi$ is simply connected.*

Our complex $\mathfrak{B}_n^\pi$ is not a subcomplex of $\mathcal{B}_n$, as the vertices of $\mathcal{B}_n$ are taken to be conjugacy classes, rather than genuine members of F_n. We remove this technicality, as it can be shown that two odd-length palindromes are conjugate if and only if they are equal. Given this, it is clear, however, that $\mathfrak{B}_n^{\pi}$ is isomorphic to a subcomplex of $\mathcal{B}_n$.

There is an obvious simplicial action of ΠA_n on $\mathfrak{B}_n^{\pi}$. This action is, by definition, transitive on the set of k–simplices, for each $0 \leq k < n$. Further, $\mathcal{PI}_n$ acts without rotations, that is, if $\phi \in \mathcal{PI}_n$ stabilises a simplex s of $\mathfrak{B}_n^{\pi}$, then it fixes s pointwise. Following work of Charney [7] on related complexes, we obtain that the quotient of $\mathfrak{B}_n^{\pi}$ by $\mathcal{PI}_n$ is highly connected.

Theorem 3.2 *For $n \geq 3$, the quotient $\mathfrak{B}_n^{\pi}/\mathcal{PI}_n$ is $(n-3)$–connected.*

The proof of this theorem is discussed in Section 4.

Theorems 3.1 and 3.2 allow us to apply the following theorem of Armstrong [1] to the action of $\mathcal{PI}_n$ on $\mathfrak{B}_n^{\pi}$, for $n \geq 4$. The statement of the theorem is as given by Day and Putman [11].

Theorem 3.3 *Let G act simplicially on a simply connected simplicial complex X, without rotations. Then G is generated by the vertex stabilisers of the action if and only if X/G is simply connected.*

We analyse the vertex stabilisers of $\mathcal{PI}_n$ using an inductive argument. It is known that $\mathcal{PI}_1 = 1$ and $\mathcal{PI}_2 = 1$; the latter equality follows from the fact that $\mathrm{IA}_2 = \mathrm{Inn}(F_2)$ and $\mathrm{Inn}(F_n) \cap \Pi A_n = 1$ for $n \geq 1$. We treat the $n = 3$ case differently, as the quotient $\mathfrak{B}_3^{\pi}/\mathcal{PI}_3$ is not simply connected, and so does not allow us to apply Armstrong's theorem directly. This treatment is postponed until Section 5.

A Birman exact sequence We require a version of the free group analogue of the Birman exact sequence, as developed by Day and Putman [10]. Recall that $\mathrm{P}\Pi A_n(k)$ consists of the pure palindromic automorphisms fixing $x_1, \dots, x_k$.

Proposition 3.4 *For $0 \leq k \leq n$, there exists the split short exact sequence*

$$1 \longrightarrow \mathcal{J}_n(k) \longrightarrow \mathrm{P}\Pi A_n(k) \longrightarrow \mathrm{P}\Pi A_{n-k} \longrightarrow 1,$$

where $\mathcal{J}_n(k)$ is the normal closure in $\mathrm{P}\Pi A_n(k)$ of the set $\{P_{ij} \mid i > k, j \leq k\}$.

Proof A map $\theta_*\colon \mathrm{P}\Pi A_n(k) \to \mathrm{P}\Pi A_{n-k}$ is induced by the map $\theta\colon F_n \to F_{n-k}$ that trivialises each $x_1, \dots, x_k$. Let $\{y_{k+1}, \dots, y_n\}$ be a free basis for F_{n-k}, where $\theta(x_i) = y_i$ for $k+1 \leq i \leq n$. Denote by Q_{ij} and η_i the elementary palindromic automorphism sending y_i to $y_j y_i y_j$ and the inversion sending y_i to ${y_i}^{-1}$, respectively $(k+1 \leq i \neq j \leq n)$.

By Corollary 2.3, we know that $\mathrm{P\Pi A}_n(k)$ is generated by the set

$$\mathcal{S} := \{P_{ij}, \iota_i \mid i > k, 1 \leq j \leq n\}.$$

If $j \leq k$, then $\theta_*(P_{ij})$ is trivial. If $i, j \geq k+1$, then $\theta_*(P_{ij}) = Q_{ij}$ and $\theta_*(\iota_i) = \eta_i$, so we have that θ_* is surjective, by examining Collins' generators for $\mathrm{P\Pi A}_{n-k}$. Indeed, the map θ_* has a section, taking Q_{ij} to P_{ij} and η_i to ι_i, which we know is well-defined by Collins' finite presentation for $\mathrm{P\Pi A}_{n-k}$. Thus, we obtain a split short exact sequence via the epimorphism θ_*.

All that is left to establish is the kernel of θ_*. Notice that we have a presentation for $\mathrm{P\Pi A}_{n-k}$ in terms of the generating set $\theta_*(\mathcal{S})$: explicitly, we add the relations $\theta_*(P_{ij}) = 1$ for $j \leq k$ to Collins' relations on the set $\{Q_{ij}, \eta_i\}$. It is a standard fact (see, for example, Magnus, Karrass and Solitar [20, proof of Theorem 2.1]) that the kernel of θ_* is the normal closure in $\mathrm{P\Pi A}_n(k)$ of the obvious lifts of the defining relators on $\theta_*(\mathcal{S})$. The only defining relators with non-trivial lifts in $\mathrm{P\Pi A}_n(k)$ are the relators $\theta_*(P_{ij})$ with $j \leq k$, thus the kernel is $\mathcal{J}_n(k)$ as in the statement of the proposition. □

Our "Birman kernel" $\mathcal{J}_n(k)$ is rather worse behaved than the analogous Birman kernel of Day and Putman. Their kernel, denoted $\mathcal{K}_{n,k,l}$, is finitely generated, whereas it may be shown by adapting the proof of their Theorem E that $\mathcal{J}_n(k)$ is not. This difference is due in part to the fact that their version of $\mathrm{P\Pi A}_n(k)$ need only fix each of $x_1, \dots, x_k$ up to conjugacy. The lack of finite generation of $\mathcal{J}_n(k)$ is, however, not an obstacle to the goal of the current paper; we only require that $\mathcal{J}_n(k)$ is *normally* generated by a finite set.

Our Birman exact sequence projects into $\mathrm{GL}(n,\mathbb{Z})$ in an obvious way, made precise in the following lemma. Let v_i denote the image of $x_i \in F_n$ under the abelianisation map. We denote by $\Gamma_n[2](k)$ the members of $\Gamma_n[2]$ which fix $v_1, \dots, v_k \in \mathbb{Z}^n$, and by $\mathcal{H}_n(k)$ the group $\mathrm{Hom}(\mathbb{Z}^{n-k}, (2\mathbb{Z})^k)$.

Lemma 3.5 *Fix $0 \leq k \leq n$. Then there exists the commutative diagram*

$$\begin{tikzcd}
1 \arrow[r] & \mathcal{J}_n(k) \arrow[r] \arrow[d, two heads] & \mathrm{P\Pi A}_n(k) \arrow[r] \arrow[d, two heads] & \mathrm{P\Pi A}_{n-k} \arrow[r] \arrow[l, bend left, "s"] \arrow[d, two heads] & 1 \\
1 \arrow[r] & \mathcal{H}_n(k) \arrow[r] & \Gamma_n[2](k) \arrow[r] & \Gamma_{n-k}[2] \arrow[r] \arrow[l, bend left, "t"] & 1
\end{tikzcd}$$

of split short exact sequences, where s and t are the obvious splitting homomorphisms.

Proof The top row is given by Proposition 3.4. A generating set for $\Gamma_n[2](k)$ follows from the proof of Lemma 2.4; it is precisely the image in $\mathrm{GL}(n,\mathbb{Z})$ of $\{P_{ij}, \iota_i \mid i > k\}$, the generating set of $\mathrm{P\Pi A}_n(k)$ given by Corollary 2.3. The bottom row then follows by an argument similar to the proof of Proposition 3.4, noting that the kernel is generated by the images of P_{ij} $(i > k,\ j \leq k)$. It is straightforward to verify that this kernel is $\mathrm{Hom}(\mathbb{Z}^{n-k}, (2\mathbb{Z})^k)$. Intuitively, $\alpha \in \mathrm{Hom}(\mathbb{Z}^{n-k}, (2\mathbb{Z})^k)$ encodes how many (even) multiples of v_j $(1 \leq i \leq k)$ are added to each v_i $(k < j \leq n)$.

The only vertical map left to consider is the right-most one. Its existence and surjectivity follow from Lemma 2.4. It is clear that all the arrows commute, and that the splitting homomorphisms s and t are compatible with the commutative diagram, so the proof is complete. $\square$

A generating set for $\mathcal{J}_n(1) \cap \mathcal{PI}_n$ By mapping $\mathrm{P\Pi A}_n(k)$ into $\Gamma_n[2](k)$ then conjugating the normal subgroup $\mathcal{H}_n(k)$, we obtain a homomorphism $\alpha_k\colon \mathrm{P\Pi A}_n(k) \to \mathrm{Aut}(\mathcal{H}_n(k))$. Setting $k = 1$, we obtain the following lemma.

Lemma 3.6 *The group $\mathcal{J}_n(1) \cap \mathcal{PI}_n$ is normally generated in $\mathcal{J}_n(1)$ by the set*

$$\big\{[P_{ij}, P_{i1}], [P_{ij}, P_{j1}]P_{i1}^2 \mid 1 < i \neq j \leq n\big\}.$$

Proof By Lemma 3.5, there is a short exact sequence

$$1 \longrightarrow \mathcal{J}_n(1) \cap \mathcal{PI}_n \longrightarrow \mathcal{J}_n(1) \longrightarrow \mathcal{H}_n(1) \longrightarrow 1.$$

The set $Y := \{\phi P_{j1} \phi^{-1} \mid \phi \in \mathrm{P\Pi A}_n(1), 1 < j \leq n\}$ generates $\mathcal{J}_n(1)$ by Proposition 3.4. Let a_j denote the image of P_{j1} in $\mathrm{GL}(n,\mathbb{Z})$. A direct calculation verifies that the set $\{a_j\}$ is a free abelian basis for $\mathcal{H}_n(1)$.

For $\phi \in \mathrm{P\Pi A}_n(1)$, let $\bar{\phi}$ denote the image of ϕ in $\Gamma_n[2](1)$, and let $\bar{Y}$ denote the image of Y. The set of relations

$$\big\{[a_i, a_j] = 1, \bar{\phi} a_i \bar{\phi}^{-1} = \alpha_1(\phi)(a_i) \mid 1 < i \neq j \leq n, \phi \in \mathrm{P\Pi A}_n(1)\big\},$$

together with the generating set $\bar{Y}$, forms a presentation for $\mathcal{H}_n(k)$. It is clear that the image of any member of Y in $\mathcal{H}_n(1)$ is a word on the free abelian basis $\{a_i\}$, and that this word is determined by the homomorphism α_1.

The group $\mathcal{J}_n(1) \cap \mathcal{PI}_n$ is normally generated in $\mathcal{J}_n(1)$ by the obvious lifts of the (infinitely many) relators in the given presentation for $\mathcal{H}_n(1)$. The relators of the form $[a_i, a_j]$ have trivial lift, and so are not required in the generating set. Let C be the finite generating set for $\mathrm{P\Pi A}_n(1)$ given by Corollary 2.3. It can be shown that the obvious lift of the finite set of relators

$$D := \{\bar{c} a_j \bar{c}^{-1} \alpha_1(c)(a_j)^{-1} \mid c \in C^{\pm 1}, 1 < j \leq n\}$$

suffices to normally generate $\mathcal{J}_n(1) \cap \mathcal{PI}_n$. This may be seen using a simple induction argument on the length of a given expression of $\phi \in \mathrm{P\Pi A}_n(1)$ on $C^{\pm 1}$.

All that remains is to verify that the obvious lift of D is the set given in the statement of the lemma; this is a straightforward calculation. □

We now prove Theorem A using the action of $\mathcal{PI}_n$ on $\mathfrak{B}_n^\pi$.

Proof of Theorem A Recall that the set of doubled commutator transvections in $\mathcal{PI}_n$ is precisely the conjugacy class of $[P_{12}, P_{13}]$ in $\Pi\mathrm{A}_n$, and that the set of separating π–twists in $\mathcal{PI}_n$ is precisely the conjugacy class of

$$(P_{23}{P_{13}}^{-1}P_{31}P_{32}P_{12}{P_{21}}^{-1})^2$$

in $\Pi\mathrm{A}_n$.

The group $\mathcal{PI}_n$ acts on $\mathfrak{B}_n^\pi$ simplicially and without rotations. Combining Theorems 3.1, 3.2 and 3.3, we conclude that, for $n \geq 4$, $\mathcal{PI}_n$ is generated by the vertex stabilisers of the action on $\mathfrak{B}_n^\pi$.

Let $\mathcal{PI}_n(1)$ denote the stabiliser of the vertex x_1. Since $\Pi\mathrm{A}_n$ acts transitively on the vertices of $\mathfrak{B}_n^\pi$, the stabiliser in $\mathcal{PI}_n$ of any vertex is conjugate in $\Pi\mathrm{A}_n$ to $\mathcal{PI}_n(1)$. Lemma 3.5 gives us the split short exact sequence

$$1 \longrightarrow \mathcal{J}_n(1) \cap \mathcal{PI}_n \longrightarrow \mathcal{PI}_n(1) \longrightarrow \mathcal{PI}_{n-1} \longrightarrow 1.$$

We induct on n. By the above split short exact sequence, to generate $\mathcal{PI}_n(1)$ it suffices to combine a generating set of $\mathcal{J}_n(1) \cap \mathcal{PI}_n(1)$ with a lift of one of $\mathcal{PI}_{n-1}$.

We begin with the base case, $n = 3$. In Section 5, we verify that the presentation of $\Gamma_3[2]$ given in Corollary 1.1 is correct when $n = 3$. Given the short exact sequence

$$1 \longrightarrow \mathcal{PI}_3 \longrightarrow \mathrm{P\Pi A}_3 \longrightarrow \Gamma_3[2] \longrightarrow 1,$$

we may take the obvious lifts of the relators in this presentation as a normal generating set for $\mathcal{PI}_3$ in $\mathrm{P\Pi A}_3$. Relators 1–7 are trivial when lifted. Relator 8 lifts to $[P_{ij}, P_{ik}]$ and relator 9 lifts to $[P_{jk}, P_{ij}]{P_{ik}}^{-2}$, which equals $P_{ik}[P_{ij}, P_{ik}]{P_{ik}}^{-1}$. Thus the lifts of relators 8 and 9 are conjugate to $[P_{12}, P_{13}]$ in $\Pi\mathrm{A}_3$. Finally, relator 10 lifts to

$$(P_{23}{P_{13}}^{-1}P_{31}P_{32}P_{12}{P_{21}}^{-1})^2,$$

so the base case $n = 3$ is true, as each relator lifts to either a doubled commutator transvection, a separating π–twist or the identity automorphism.

Now suppose $n > 3$. By induction, the group $\mathcal{PI}_{n-1}$ is generated by the purported generating set. We lift this generating set to $\mathcal{PI}_n(1)$ in the obvious way.

By Lemma 3.6, we need only add in $\mathcal{J}_n(1)$–conjugates of the words $[P_{ij}, P_{i1}]$ and $[P_{ij}, P_{j1}]P_{i1}^2$, for $1 < i \neq j \leq n$. The former are clearly conjugate in $\Pi\mathrm{A}_n$ to the doubled commutator transvection $[P_{12}, P_{13}]$. For the latter, observe that

$$[P_{ij}, P_{j1}]P_{i1}^2 = [P_{ij}, P_{i1}^{-1}],$$

which again is conjugate in $\Pi\mathrm{A}_n$ to $[P_{12}, P_{13}]$, so we are done. □

Theorem A allows us to conclude that $\mathcal{PI}_n$ is normally generated in $\Pi\mathrm{A}_n$ by the automorphisms $\chi_1 = [P_{12}, P_{13}]$ and

$$\chi_2 = (P_{23}{P_{13}}^{-1}P_{31}P_{32}P_{12}{P_{21}}^{-1})^2.$$

Let $\Omega_n \leq \Pi\mathrm{A}_n$ denote the symmetric group on X. The presentation for $\Gamma_n[2] \cong \mathrm{P}\Pi\mathrm{A}_n/\mathcal{PI}_n$ given in Corollary 1.1 follows from Theorem A by adding the Ω_n–orbits of χ_1 and χ_2 to Collins' presentation for $\mathrm{P}\Pi\mathrm{A}_n$ as relators, then applying the obvious Tietze transformations.

We now demonstrate that the presence of separating π–twists in our generating set for $\mathcal{PI}_n$ is necessary.

Proposition 3.7 *For $n \geq 3$, the group generated by doubled commutator transvections is a proper subgroup of $\mathcal{PI}_n$.*

Proof Let $\mathcal{D}$ denote the subgroup of $\mathcal{PI}_n$ generated by doubled commutator transvections. In other words, $\mathcal{D}$ is the normal closure of $\chi_1 = [P_{12}, P_{13}]$ in $\Pi\mathrm{A}_n$. Then the Ω_n–orbit of χ_1 is a normal generating set for $\mathcal{D}$ in $\mathrm{P}\Pi\mathrm{A}_n$. Adding the members of this orbit to the presentation of $\mathrm{P}\Pi\mathrm{A}_n$ as relators yields a finite presentation $\mathcal{Q}$ of $\mathrm{P}\Pi\mathrm{A}_n/\mathcal{D}$, which may be altered using Tietze transformations so that it looks like the presentation in Corollary 1.1, with relator 10 (and relator 7, if $n = 3$) removed (where we interpret S_{ij} and O_i as formal symbols, rather than matrices). We shall show that the relations of $\mathcal{Q}$ are not a complete set of relations on the generating set $\{S_{ij}, O_i\}$ for $\Gamma_n[2] \cong \mathrm{P}\Pi\mathrm{A}_n/\mathcal{PI}_n$, and so conclude that $\mathcal{D} \neq \mathcal{PI}_n$.

It is easily shown that for

$$\xi := (S_{32}{S_{31}}^{-1}S_{13}S_{23}S_{21}{S_{12}}^{-1})^2,$$

the image of χ_2 in $\Gamma_n[2]$, is trivial, but we shall show that ξ is non-trivial in the group presented by $\mathcal{Q}$. Observe that by trivialising all the generators of $\Gamma_n[2]$ except for S_{12} and S_{21}, we surject $\Gamma_n[2]$ onto the free Coxeter group generated by the images of S_{12} and S_{21}, say A and B, respectively. This is easily verified by examining the relators of $\mathcal{Q}$. The image of ξ under this map is $ABAB \neq 1$, and so ξ is non-trivial in the group presented by $\mathcal{Q}$. Therefore $\mathcal{D}$ is a proper subgroup of $\mathcal{PI}_n$. □

Note that in the proof of Proposition 3.7 we also showed that relators 1–9 of Corollary 1.1 are not a sufficient set of relators that hold between the O_i and S_{jk}, as relator 10 is not a consequence of the others. This allows us to conclude that the quotient space $\mathfrak{B}_3^\pi/\mathcal{PI}_3$ is not simply connected.

Corollary 3.8 *The complex $\mathfrak{B}_3^\pi/\mathcal{PI}_3$ is not simply connected.*

Proof By Theorem 3.3, the complex $\mathfrak{B}_3^\pi/\mathcal{PI}_3$ is simply connected if and only if $\mathcal{PI}_3$ is generated by the vertex stabilisers of the action of $\mathcal{PI}_3$ on $\mathfrak{B}_3^\pi$. As in the proof of Theorem A, the group generated by the vertex stabilisers of this action may be normally generated in ΠA_3 by the group $\mathcal{PI}_3(1)$. The same calculations as in the proof of Theorem A show that $\mathcal{PI}_3(1)$ is the normal closure of the doubled commutator transvection $[P_{12}, P_{13}]$. However, Proposition 3.7 showed that this normal closure is a proper subgroup of $\mathcal{PI}_3$, so the quotient $\mathfrak{B}_3^\pi/\mathcal{PI}_3$ is not simply connected. □

4 The connectivity of $\mathfrak{B}_n^\pi$ and its quotient

In this section, we determine the levels of connectivity of $\mathfrak{B}_n^\pi$ and $\mathfrak{B}_n^\pi/\mathcal{PI}_n$. The former is found to be simply connected, following the same approach as Day and Putman [11], while the latter is shown to be closely related to a complex already studied by Charney [7], which is $(n-3)$–connected.

The connectivity of $\mathfrak{B}_n^\pi$ First, we recall the definition of the Cayley graph of a group. Let G be a group with finite generating set S. The *Cayley graph of G with respect to S*, denoted $\mathrm{Cay}(G,S)$, is the graph with vertex set G and edge set $\{(g,gs) \mid g \in G, s \in S^{\pm 1}\}$, where an ordered pair (x,y) indicates that vertices x and y are joined by an edge. If $s \in S$ has order 2, we identify each pair of edges (g,gs) and (g,gs^{-1}) for each $g \in G$, to ensure that the Cayley graph is simplicial. Similarly, we also insist that the identity element of G is excluded from S.

We establish Theorem 3.1 by constructing a map Ψ from the Cayley graph of ΠA_n to $\mathfrak{B}_n^\pi$ and demonstrating that the induced map of fundamental groups is both surjective and trivial. We require the Cayley graph of ΠA_n with respect to a particular generating set, which we now describe. Assume that $n \geq 3$. For $1 \leq i \neq j < n$, let t_{ij} permute x_i and x_j, fixing x_k with $k \neq i,j$. Using the symmetric group action on X, we deduce from Proposition 2.2 that we may generate ΠA_n using the set

$$Z := \{t_{ij}, \iota_2, \iota_3, P_{21}, P_{23}, P_{31}, P_{34} \mid 1 \leq i \neq j \leq n\}.$$

We may use the symmetric group action on X to streamline the presentation of ΠA_n given in Section 2, to obtain the following list of defining relators for ΠA_n on the generating set Z:

(1) $t_{ij} = t_{ji}$,

(2) ${t_{ij}}^2 = 1$,

(3) $u t_{ij} u^{-1} = t_{u(i)u(j)}$,

(4) ${\iota_2}^2 = 1$,

(5) $(\iota_2\iota_3)^2 = 1$,

(6) $[\iota_2, P_{31}] = 1$,

(7) $(\iota_2 P_{21})^2 = 1$,

(8) $(\iota_3 P_{23})^2 = 1$,

(9) $P_{23}P_{31}P_{21} = {P_{21}}^{-1}P_{31}P_{23}$,

(10) $[P_{21}, P_{31}] = 1$,

(11) $[P_{21}, P_{34}] = 1$,

(12) $\iota_3 = t_{23}\iota_2 t_{23}$,

(13) $P_{31} = t_{23}P_{21}t_{23}$,

(14) $P_{23} = t_{13}P_{21}t_{13}$,

(15) $P_{34} = t_{14}t_{23}P_{21}t_{23}t_{14}$,

(16) $P_{21} = wP_{21}w^{-1}$ for $w \in \mathcal{W}$,

(17) $\iota_2 = v\iota_2 v^{-1}$ for $v \in \mathcal{V}$,

where $1 \le i \ne j \le n$, $u \in \{t_{ij}\}$, and $\mathcal{W}$ and $\mathcal{V}$ are the sets of words on $\{t_{ij}\}$ that fix both x_1 and x_2, and only x_2, respectively. The relations of type 16 and 17 arise due to the streamlining of the presentation of $\Pi\mathrm{A}_n = \mathrm{E}\Pi\mathrm{A}_n \rtimes \Omega^{\pm 1}(X)$ given in Section 2. Note that relations 1–3 are a complete set of relations for the symmetric group, when generated by the transpositions $\{t_{ij}\}$ [25].

We now consider the Cayley graph $\mathrm{Cay}(\Pi\mathrm{A}_n, Z)$. Observe that for each $z \in Z$ either $z(x_1) = x_1$ or $\{x_1, z(x_1)\}$ forms a partial π–basis for F_n. This allows us to construct a map of complexes from the star of the vertex 1 in $\mathrm{Cay}(\Pi\mathrm{A}_n, Z)$ to $\mathfrak{B}_n^{\pi}$, by mapping an edge $z \in Z^{\pm 1}$ to the edge $v_1 - z(v_1)$ (which may be degenerate). Using the actions of $\Pi\mathrm{A}_n$ on $\mathrm{Cay}(\Pi\mathrm{A}_n, Z)$ and $\mathfrak{B}_n^{\pi}$, we can extend this map to a map of complexes $\Psi\colon \mathrm{Cay}(\Pi\mathrm{A}_n, Z) \to \mathfrak{B}_n^{\pi}$. Explicitly, Ψ takes a vertex $z_1 \cdots z_r$ of $\mathrm{Cay}(\Pi\mathrm{A}_n, Z)$ ($z_i \in Z^{\pm 1}$) to the vertex $z_1 \cdots z_r(x_1)$.

Proof of Theorem 3.1 This proof is modelled on Day and Putman's proof of [11, Theorem A]. Let

$$\Psi_*\colon \pi_1(\mathrm{Cay}(\Pi\mathrm{A}_n, Z), 1) \to \pi_1(\mathfrak{B}_n^{\pi}, x_1)$$

be the map of fundamental groups induced by Ψ. Explicitly, the image of a loop $z_1 \cdots z_k$ ($z_i \in Z^{\pm 1}$) in $\pi_1(\mathrm{Cay}(\Pi\mathrm{A}_n, Z), 1)$ under Ψ_* is

$$x_1 - z_1(x_1) - z_1z_2(x_1) - \cdots z_1z_2 \cdots - z_k(x_1) = x_1.$$

We first show that Ψ_* is the trivial map, then show that it is also surjective.

Recall that the Cayley graph C of a group G with presentation $\langle X \mid R\rangle$ forms the 1–skeleton of its *Cayley complex*, which we obtain by attaching disks along the loops in C corresponding to all conjugates in G of the words in R. It is well-known that the Cayley complex of a group G is simply connected [19, Proposition 4.2]. We now verify that the loops in $\pi_1(\mathrm{Cay}(\Pi\mathrm{A}_n, Z), 1)$ corresponding to the relators in the above streamlined presentation for $\Pi\mathrm{A}_n$ have trivial image under Ψ_*. This allows us to extend Ψ to a map from the (simply connected) Cayley complex of $\Pi\mathrm{A}_n$ (rel Z), and so conclude that Ψ_* is trivial.

Note that in the following we confuse a relator with the loop in $\pi_1(\mathrm{Cay}(\Pi\mathrm{A}_n, Z), 1)$ to which it corresponds. Many of the relators 1–17 map to x_1 in $\mathfrak{B}_n^\pi$, as they are words on members of Z that fix x_1. The only ones we need to check are 1–3 and 14–17. Relators 1–3 map into the contractible simplex spanned by $x_1, \dots, x_n$, so are trivial. Relators 14 and 15 are mapped into the simplices $x_1 - x_3$ and $x_1 - x_4$, respectively. We rewrite relators 16 and 17 as $P_{21}w = wP_{21}$ and $\iota_2 v = v\iota_2$. It is clear, then, that relators of type 16 map into the contractible subcomplex of $\mathfrak{B}_n^\pi$ spanned by $x_1, \dots, x_n$ and $x_1x_2x_1$, and relators of type 17 map into the contractible subcomplex spanned by $x_1, {x_2}^{\pm 1}, \dots, x_n$. All relators have now been dealt with, so we conclude that Ψ_* is the trivial map.

We argue as in Day and Putman's proof [11] for the surjectivity of Ψ_*. We represent a loop $\omega \in \pi_1(\mathfrak{B}_n^\pi, x_1)$ as

$$x_1 = w_0 - w_1 - \cdots - w_k = x_1,$$

for some $k \geq 0$. We will demonstrate that for any such path (not necessarily with $w_k = x_1$), there exist $\phi_1, \dots, \phi_k \in \Pi\mathrm{A}_n(1)$ such that

$$w_i = \phi_1 t_{12} \phi_2 t_{12} \cdots \phi_i t_{12}(x_1)$$

for $0 \leq i \leq k$. We use induction. In the case $k = 0$, there is nothing to prove. Now suppose $k > 0$. Consider the subpath

$$w_0 - w_1 - \cdots - w_{k-1}.$$

By induction, to prove the claim all we need find is $\phi_k \in \Pi\mathrm{A}_n(1)$ such that

$$w_k = \phi_1 t_{12} \cdots \phi_k t_{12}(x_1).$$

We know that $w_{k-1} = \phi_1 t_{12} \cdots \phi_{k-1} t_{12}(x_1)$ and w_k form a partial π–basis, therefore so do x_1 and $(\phi_1 t_{12} \cdots \phi_{k-1} t_{12})^{-1}(w_k)$. By construction, the action of $\Pi\mathrm{A}_n$ is transitive on the set of two-element partial π–bases, so there exists $\phi_k \in \Pi\mathrm{A}_n(1)$ mapping x_2 to $(\phi_1 t_{12} \cdots \phi_{k-1} t_{12})^{-1}(w_k)$. Therefore

$$w_k = \phi_1 t_{12} \cdots \phi_k t_{12}(x_1),$$

as required.

Now we define

$$\phi_{k+1} = (\phi_1 t_{12} \cdots \phi_k t_{12})^{-1},$$

so that

$$R := \phi_1 t_{12} \cdots \phi_k t_{12} \phi_{k+1} = 1$$

is a relation in ΠA_n. Observe that since $w_k = x_1$, we have $\phi_{k+1} \in \Pi A_n(1)$. Also, the generating set Z contains a subset that generates $\Pi A_n(1)$, by Proposition 2.2. We are thus able to write

$$\phi_i = z_1^i \cdots z_{p_i}^i,$$

for some $z_j^i \in Z^{\pm 1}$ $(1 \leq i \leq k+1,\ 1 \leq j \leq p_i)$, each of which fixes x_1. We see that $R \in \pi_1(\mathrm{Cay}(\Pi A_n, Z), 1)$ maps to $\omega \in \pi_1(\mathfrak{B}_n^\pi, x_1)$. Removing repeated vertices, R maps to

$$x_1 - \phi_1 t_{12}(x_1) - \cdots - \phi_1 t_{12} \cdots \phi_k t_{12}(x_1) = x_1,$$

which equals ω by construction. Hence Ψ_* is surjective as well as trivial, and hence $\pi_1(\mathfrak{B}_n^\pi, x_1) = 1$. □

The connectivity of $\mathfrak{B}_n^\pi / \mathcal{PI}_n$ A complex analogous to $\mathfrak{B}_n^\pi$ may be defined when working over $\mathbb{Z}^n$ rather than F_n. We write $\mathcal{B}_n(\mathbb{Z})$ for the *complex of partial bases of* $\mathbb{Z}^n$, whose $(k-1)$–simplices correspond to subsets $\{u_1, \ldots, u_k\}$ of free abelian bases of $\mathbb{Z}^n$. Writing members of $\mathbb{Z}^n$ multiplicatively, there is an analogous notion of an odd palindrome on some fixed free abelian basis V, and so also of a partial π–basis. The *complex of partial π–bases of* $\mathbb{Z}^n$ is defined in the obvious way, and denoted $\mathfrak{B}_n^\pi(\mathbb{Z})$. Just as ΠA_n acts transitively on the set of π–bases of F_n, so does $\Gamma_n[2]$ act transitively on the set of π–bases of $\mathbb{Z}^n$, as we now verify.

Lemma 4.1 *The group $\Gamma_n[2]$ acts transitively on the set of π–bases of $\mathbb{Z}^n$.*

Proof By definition, any π–basis is of the form $\{Mv_1, \ldots, Mv_n\}$, for $M \in \Gamma_n[2]$ and $\{v_1, \ldots, v_n\}$ the standard basis of $\mathbb{Z}^n$, where v_i has 1 in the i^{th} position and 0s elsewhere. Thus, we have a well-defined action of $\Gamma_n[2]$ on the set of π–bases of $\mathbb{Z}^n$ by left-multiplication of basis elements, which is transitive, as every π–basis lies in the same orbit as $\{v_1, \ldots, v_n\}$. □

We first show that $\mathfrak{B}_n^\pi / \mathcal{PI}_n \cong \mathfrak{B}_n^\pi(\mathbb{Z})$, then show that $\mathfrak{B}_n^\pi(\mathbb{Z})$ is $(n-3)$–connected using a related complex studied by Charney. To prove the former, the following lemma is required.

Lemma 4.2 *Fix $\{u_1, \ldots, u_n\}$ as a π–basis for $\mathbb{Z}^n$, and let $\rho\colon F_n \to \mathbb{Z}^n$ be the abelianisation map. Let $\widetilde{U} = \{\tilde{u}_1, \ldots, \tilde{u}_k\}$ be a partial π–basis of F_n such that $\rho(\tilde{u}_i) = u_i$ for each $1 \leq i \leq k$. Then we can extend $\widetilde{U}$ to a π–basis of F_n, $\{\tilde{u}_1, \ldots, \tilde{u}_n\}$, such that $\rho(\tilde{u}_i) = u_i$ for $1 \leq i \leq n$.*

Proof Extend $\{\tilde{u}_1, \dots, \tilde{u}_k\}$ to a full π–basis of F_n, $\{\tilde{u}_1, \dots, \tilde{u}_k, \tilde{u}'_{k+1}, \dots, \tilde{u}'_n\}$, and define $u'_j = \rho(\tilde{u}'_j)$ for $k+1 \le j \le n$. Then $\{u_1, \dots, u_k, u'_{k+1}, \dots, u'_n\}$ is a π–basis for $\mathbb{Z}^n$. By Lemma 4.1, the group $\Gamma_n[2]$ acts transitively on the set of π–bases of $\mathbb{Z}^n$, so there exists $\phi \in \Gamma_n[2](k)$ such that $\phi(u'_j) = u_j$ for $k+1 \le j \le n$. By Lemma 3.5, ϕ lifts to some $\tilde{\phi} \in \mathrm{P\Pi A}_n(k)$, and the π–basis $\{\tilde{u}_1, \dots, \tilde{u}_k, \tilde{\phi}(\tilde{u}'_{k+1}), \dots, \tilde{\phi}(\tilde{u}'_n)\}$ projects onto $\{u_1, \dots, u_n\}$ as desired. □

Now we establish an isomorphism of simplicial complexes $\mathfrak{B}_n^\pi / \mathcal{PI}_n \cong \mathfrak{B}_n^\pi(\mathbb{Z})$.

Theorem 4.3 *The spaces $\mathfrak{B}_n^\pi / \mathcal{PI}_n$ and $\mathfrak{B}_n^\pi(\mathbb{Z})$ are isomorphic as simplicial complexes.*

Proof Let $\rho\colon F_n \to \mathbb{Z}^n$ be the abelianisation map, and define a map of simplicial complexes $\Phi\colon \mathfrak{B}_n^\pi \to \mathfrak{B}_n^\pi(\mathbb{Z})$ on simplices by $\{w_1, \dots, w_k\} \mapsto \{\rho(w_1), \dots, \rho(w_k)\}$ for $1 \le k \le n$. The map Φ is surjective: by Lemma 4.2, each π–basis of $\mathbb{Z}^n$ is the image of some π–basis of F_n, and π–bases of $\mathbb{Z}^n$ correspond to maximal simplices of $\mathfrak{B}_n^\pi(\mathbb{Z})$.

It is clear that the map Φ is invariant under the action of $\mathcal{PI}_n$ on $\mathfrak{B}_n^\pi$, and so Φ factors through $\mathfrak{B}_n^\pi / \mathcal{PI}_n$. To establish the theorem, all we need do is show that the induced map from $\mathfrak{B}_n^\pi / \mathcal{PI}_n \to \mathfrak{B}_n^\pi(\mathbb{Z})$ is injective. In other words, we must show that if two simplices s, s' of $\mathfrak{B}_n^\pi$ have the same image under Φ, then s and s' differ by the action of some member of $\mathcal{PI}_n$.

Suppose that $s = \{w_1, \dots, w_k\}$ and $s' = \{w'_1, \dots, w'_k\}$ have the same image under Φ. We may assume that $\rho(w_i) = \rho(w'_i)$ for $1 \le i \le k$. Let $\Phi(s) = \{\bar{w}_1, \dots, \bar{w}_k\}$, and extend this partial π–basis of $\mathbb{Z}^n$ to a full π–basis $W = \{\bar{w}_1, \dots, \bar{w}_n\}$. By Lemma 4.2, we may extend $\{w_1, \dots, w_k\}$ to $\{w_1, \dots, w_n\}$ and $\{w'_1, \dots, w'_k\}$ to $\{w'_1, \dots, w'_n\}$ such that both of these full π–bases map onto W. Define $\theta \in \mathrm{\Pi A}_n$ by $\theta(w_i) = w'_i$ for $1 \le i \le n$. By construction, $\theta(s) = s'$ and $\theta \in \mathcal{PI}_n$, so the theorem is proved. □

This more explicit description of $\mathfrak{B}_n^\pi / \mathcal{PI}_n$ as $\mathfrak{B}_n^\pi(\mathbb{Z})$ enables us to investigate the quotient's connectivity.

Proof of Theorem 3.2 By a *unimodular sequence* in $\mathbb{Z}^n$, we mean an (ordered) sequence $(u_1, \dots, u_k) \subset (\mathbb{Z}^n)^k$ whose entries form a basis of a direct summand of $\mathbb{Z}^n$. Observe that this is just an ordered version of the notion of a partial basis of $\mathbb{Z}^n$. The set of all such sequences of length at least one form a poset under subsequence inclusion. Charney considers (among others) the subposet of sequences $(u_1, \dots, u_k)$ such that each u_i is congruent to a standard basis vector v_j under mod 2 reduction of the entries of u_i. We denote by $\mathcal{X}_n$ the poset complex given by the subposet of such sequences. Theorem 2.5 of Charney [7] says that $\mathcal{X}_n$ is $(n-3)$–connected.

Let $\mathfrak{B}_n^\pi(\mathbb{Z})^*$ denote the barycentric subdivision of $\mathfrak{B}_n^\pi(\mathbb{Z})$. Label each vertex of $\mathfrak{B}_n^\pi(\mathbb{Z})^*$ by the partial π–basis associated to the simplex of $\mathfrak{B}_n^\pi(\mathbb{Z})$ to which the vertex corresponds. Define a simplicial map $h\colon \mathcal{X}_n \to \mathfrak{B}_n^\pi(\mathbb{Z})^*$ by $(u_1,\dots,u_k)\mapsto \{u_1,\dots,u_k\}$. We may think of h as "forgetting the order" of each unimodular sequence. Comparing the definitions of $\mathcal{X}_n$ and $\mathfrak{B}_n^\pi(\mathbb{Z})$, it is not immediately clear that h is well-defined, as there might be some vertex $(u_1,\dots,u_k)$ of $\mathcal{X}_n$ such that $\{u_1,\dots,u_k\}$ extends to a full basis of $\mathbb{Z}^n$, but not a full π–basis. However, viewing the full basis of $\mathbb{Z}^n$ as a matrix in $\Gamma_n[2]$, a straightforward column operations argument shows that this cannot be the case, so h is well-defined.

We see that h induces a map $\pi_i(\mathcal{X}_n)\to\pi_i(\mathfrak{B}_n^\pi(\mathbb{Z})^*)$ for $i\geq 0$, and show that the induced map is surjective. Set a consistent lexicographical order on the vertices of $\mathfrak{B}_n^\pi(\mathbb{Z})^*$, and view $\omega\in\pi_i(\mathfrak{B}_n^\pi(\mathbb{Z})^*)$ as a simplicial i–sphere. The chosen lexicographical ordering allows us to lift ω to $\pi_i(\mathcal{X}_n)$, so the induced maps are surjective. The statement of the theorem follows immediately, since $\pi_i(\mathcal{X}_n)=1$ for $0\leq i\leq n-3$. □

5 A presentation for $\Gamma_3[2]$

In order to apply Armstrong's theorem [1], it must be the case that $\mathfrak{B}_n^\pi/\mathcal{PI}_n\cong\mathfrak{B}_n^\pi(\mathbb{Z})$ is simply connected. However, as we have seen from Corollary 3.8, the space $\mathfrak{B}_3^\pi(\mathbb{Z})$ has non-trivial fundamental group. The case $n=3$ forms the base case of our inductive proof of Theorem A, so we require an alternative approach to find a generating set for $\mathcal{PI}_3$. Our approach is to find a specific finite presentation of $\Gamma_3[2]$, and use the short exact sequence

$$1\longrightarrow\mathcal{PI}_3\longrightarrow \mathrm{P\Pi A}_3\longrightarrow\Gamma_3[2]\longrightarrow 1$$

to lift the relators in the presentation of $\Gamma_3[2]$ to a normal generating set for $\mathcal{PI}_3$.

The augmented partial π–basis complex for $\mathbb{Z}^3$ By adding simplices to the complex $\mathfrak{B}_3^\pi(\mathbb{Z})$, we obtain a simply connected complex that $\Gamma_3[2]$ acts on. This action allows us to present $\Gamma_3[2]$.

Recall that $\mathcal{B}_n(\mathbb{Z})$ is the partial basis complex of $\mathbb{Z}^n$. We represent its vertices by column vectors $u=(u^{(1)},\dots,u^{(n)})^T$. For use in the proof of Theorem 5.1, we follow Day and Putman [11] and define the *rank of* u to be $|u^{(n)}|$, and denote it by $R(u)$. Let $\mathcal{Y}$ denote the full subcomplex of $\mathcal{B}_3(\mathbb{Z})$ spanned by $\mathfrak{B}_3^\pi(\mathbb{Z})$ and vertices u for which $u^{(1)}$ and $u^{(2)}$ are odd and $u^{(3)}$ is even. We call $\mathcal{Y}$ the *augmented partial π–basis complex for* $\mathbb{Z}^3$. We now demonstrate that $\mathcal{Y}$ is simply connected.

Theorem 5.1 *The complex $\mathcal{Y}$ is simply connected.*

Proof By Theorem 2.5 of Charney [7], we know that $\mathfrak{B}_3^{\pi}(\mathbb{Z})$ is 0–connected, and hence so is $\mathcal{Y}$. To show that $\mathcal{Y}$ is simply connected, we adapt the proof of Theorem B of Day and Putman [11].

Let u be a vertex of a simplicial complex C. The *link of u in C*, denoted $\mathrm{lk}_C(u)$, is the full subcomplex of C spanned by vertices joined by an edge to u. Let $v_3 \in \mathbb{Z}^3$ be the standard basis vector with third entry 1 and 0s elsewhere. Observe that for any vertex $u \in \mathcal{Y}$ we have $\mathrm{lk}_{\mathcal{Y}}(u) \cong \mathrm{lk}_{\mathcal{Y}}(v_3)$. This is because the group generated by $\Gamma_3[2]$ and the matrix

$$E = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

acts simplicially on $\mathcal{Y}$ and transitively on the 0–skeleton of $\mathcal{Y}$. This action is transitive on vertices because $\Gamma_3[2]$ acts transitively on the vertices of $\mathfrak{B}_3^{\pi}(\mathbb{Z})$, and any vertex of $\mathcal{Y} \setminus \mathfrak{B}_3^{\pi}(\mathbb{Z})$ may be taken to a vertex of $\mathfrak{B}_3^{\pi}(\mathbb{Z})$ by acting on it with E.

We begin by establishing that $\mathrm{lk}_{\mathcal{Y}}(v_3)$ is connected (and hence, by the above, so is the link of any vertex of $\mathcal{Y}$). By considering what the columns of $M \in \mathrm{GL}(3,\mathbb{Z})$ whose final column is v_3 must look like, we see that a necessary and sufficient condition for $(u^{(1)}, u^{(2)}, u^{(3)})^T$ to be a member of $\mathrm{lk}_{\mathcal{Y}}(v_3)$ is that $(u^{(1)}, u^{(2)})^T$ is a vertex of $\mathcal{B}_2(\mathbb{Z})$. The link $\mathrm{lk}_{\mathcal{Y}}(v_3)$ may thus be described as follows: it has one vertex for each pair (a,b), where a is a vertex of $\mathcal{B}_2(\mathbb{Z})$ and $b \in 2\mathbb{Z}$, with vertices (a,b) and (c,d) joined by an edge if and only if a and c are joined by an edge in $\mathcal{B}_2(\mathbb{Z})$. Hence $\mathrm{lk}_{\mathcal{Y}}(v_3)$ is connected, though note that its fundamental group is an infinite-rank free group.

Now, let $\omega \in \pi_1(\mathcal{Y}, v_3)$. We represent ω by the sequence of vertices

$$w_0 - w_1 - \cdots - w_r,$$

where w_i $(1 \le i \le r)$ are vertices of $\mathcal{Y}$, and $w_0 = w_r = v_3$. Our goal is to systematically homotope this loop so that the rank of each vertex in the sequence is 0. Such a loop may be contracted to the vertex v_3, and so is trivial in $\pi_1(\mathcal{Y})$.

Consider a vertex w_i for some $1 < i < r$, with $R(w_i) \neq 0$. Since $\mathrm{lk}_{\mathcal{Y}}(w_i)$ is connected, there is some path

$$w_{i-1} - q_1 - q_2 - \cdots - q_s - w_{i+1}$$

in $\mathrm{lk}_{\mathcal{Y}}(w_i)$, as seen in Figure 7. Fix attention on some q_j $(1 \le j \le s)$. By the division algorithm, there exist $a_j, b_j \in \mathbb{Z}$ such that $R(q_j) = a_j \cdot R(w_i) + b_j$, with $0 \le b_j < R(w_i)$. As in the proof of Lemma 2.4, we wish to ensure that a_j is even, if possible. In all but

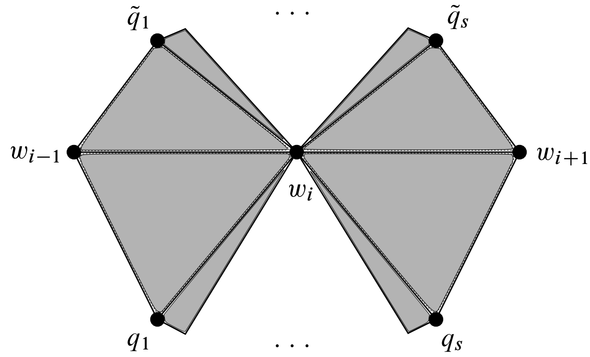


Figure 7: We find two homotopic paths that bound a disk inside $\mathrm{lk}_{\mathcal{Y}}(w_i)$, where the "upper" path seen here is constructed so that $R(\tilde{q}_j) < R(q_j)$ for $1 \leq j \leq s$.

one case, we will be able to rewrite the division algorithm as $R(q_j) = A_j \cdot R(w_i) + B_j$, for some $A_j, B_j \in \mathbb{Z}$ such that A_j is even and $0 \leq |B_j| < R(w_i)$. We do a case-by-case parity analysis. Note that since q_j and w_i are joined by an edge, $R(q_j)$ and $R(w_i)$ cannot both be odd, otherwise q_j and w_i would both map to the same member of $(\mathbb{Z}/2)^3$ when we reduce their entries mod 2. This would prohibit $\{q_j, w_i\}$ from extending to a basis J of $\mathbb{Z}^3$, otherwise the image of J in $(\mathbb{Z}/2)^3$ would generate despite only having two members. If $R(q_j)$ and $R(w_i)$ have different parities and a_j is odd, we may take $A_j = a_j + 1$ and $B_j = b_j - R(w_i)$. In that case, $|B_j| < R(w_i)$, since b_j must be odd and hence non-zero. If both $R(q_j)$ and $R(w_i)$ are even, we may still do this, unless $b_j = 0$.

We now associate to each q_j a new vertex, $\tilde{q}_j$, defined by

$$\tilde{q}_j = \begin{cases} q_j - a_j \cdot w_i & \text{if } a_j \text{ even,} \\ q_j - A_j \cdot w_i & \text{if } a_j \text{ odd, } b_j \neq 0, \\ q_j - a_j \cdot w_i & \text{if } a_j \text{ odd, } b_j = 0. \end{cases}$$

Note that $R(\tilde{q}_j) = 0$ when $b_j = 0$, and under the conditions given, $\tilde{q}_j$ is always well-defined as a vertex of $\mathcal{Y}$. The path

$$w_{i-1} - q_1 - \cdots - q_s - w_{i+1}$$

is homotopic inside $\mathrm{lk}_{\mathcal{Y}}(w_i)$ to the path

$$w_{i-1} - \tilde{q}_1 - \cdots - \tilde{q}_s - w_{i+1},$$

as seen in Figure 7. By construction, $R(\tilde{q}_j) < R(w_i)$. Iterating this procedure continually homotopes ω until it is inside the contractible (full) subcomplex spanned by v_3 and $\mathrm{lk}_{\mathcal{Y}}(v_3)$, and hence is trivial. Therefore $\pi_1(\mathcal{Y}) = 1$. $\square$

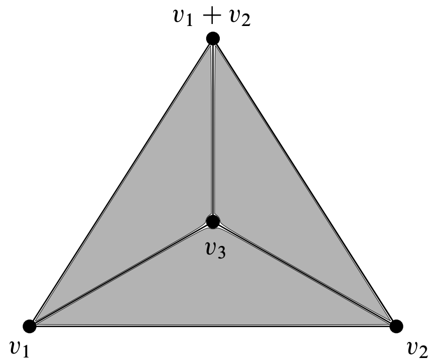


Figure 8: The quotient complex of $\mathcal{Y}$ under the action of $\Gamma_3[2]$. We have labelled its vertices using representatives from the vertex set of $\mathcal{Y}$.

The complex $\mathfrak{B}_3^{\pi}(\mathbb{Z})$ is not simply connected It may be tempting to try to use the method in the above proof to show that $\mathfrak{B}_3^{\pi}(\mathbb{Z})$ is simply connected; however, we know by Corollary 3.8 that $\mathfrak{B}_3^{\pi}(\mathbb{Z})$ has non-trivial fundamental group. The obstruction to the above proof going through occurs when defining $\tilde{q}_j$ in the case that a_j is odd and $b_j = 0$, as $\tilde{q}_j \notin \mathfrak{B}_3^{\pi}(\mathbb{Z})$. When a_j is odd and $b_j = 0$, there is no even multiple of w_i that can be added to q_j to decrease its rank, so this method of homotoping loops to a point will not work.

Presenting $\Gamma_3[2]$ Let $\Gamma_3[2](w_1, \dots, w_k)$ denote the stabiliser of the ordered tuple $(w_1, \dots, w_k)$ of vertices of $\mathcal{Y}$. Having demonstrated that $\mathcal{Y}$ is simply connected, we now turn our attention to the obvious action of $\Gamma_3[2]$ on $\mathcal{Y}$. This action is simplicial, does not invert edges, and the quotient complex under the action is contractible, as seen in Figure 8. The quotient lifts to a subcomplex W of $\mathcal{Y}$ via the vertex labels seen in Figure 8. This subcomplex is what Brown [6] refers to as a *fundamental domain* for the action, and so a theorem of Brown [6, Theorem 3] allows us to conclude that $\Gamma_3[2]$ is the free product of the stabilisers of the vertices of W modulo *edge relations*, which identify the copies of the edge stabiliser $\Gamma_3[2](a, b)$ inside the vertex stabilisers $\Gamma_3[2](a)$ and $\Gamma_3[2](b)$, where $a, b \in \{v_1, v_2, v_3, v_1 + v_2\}$ are distinct.

We obtain a finite presentation for $\Gamma_3[2](v_1)$ using the semi-direct production decomposition of $\Gamma_3[2](v_1)$ given by Lemma 3.5 (noting that $\Gamma_2[2] \cong \mathrm{P\Pi A}_2$). The group $\Gamma_3[2](v_1)$ is generated by the set $\{O_2, O_3, S_{23}, S_{32}, S_{12}, S_{13}\}$, with a complete list of relators given by all relators of the form 1–9 (excluding 7, as it is not defined when $n = 3$) seen in Corollary 1.1. By permuting the indices accordingly, we also obtain finite presentations for the stabiliser groups $\Gamma_3[2](v_2)$ and $\Gamma_3[2](v_3)$. Identifying the edge stabiliser subgroups of these three groups appropriately, we obtain the presentation seen in Corollary 1.1 without relators 7 and 10; we denote this presentation by $\mathcal{P}$.

We now see that the effect of identifying the edge stabiliser subgroups of $\Gamma_n[2](v_1+v_2)$ with the corresponding copies inside the other three vertex stabiliser groups is to include one additional relator: relator 10. Since $\Gamma_3[2](v_1+v_2)$ and $\Gamma_3[2](v_1)$ are conjugate inside $\mathrm{GL}(3,\mathbb{Z})$, we take a formal presentation for $\Gamma_3[2](v_1+v_2)$ by adding a "hat" to each of the symbols in the presentation of $\Gamma_3[2](v_1)$.

The members of $\Gamma_3[2](v_1+v_2)$ are not, however, strings of formal symbols, but are members of $\Gamma_3[2]$. To express them as such, we observe that

$$\Gamma_3[2](v_1+v_2) = E_{21}\cdot\Gamma_3[2](v_1)\cdot {E_{21}}^{-1},$$

where E_{21} is the elementary matrix with 1 in the $(2,1)$ position. In Table 1 we see the conjugates of the generators of $\Gamma_3[2](v_1)$ by E_{21}. These give expressions for the formal symbols generating $\Gamma_3[2](v_1+v_2)$. For example,

$$\widehat{S}_{12} = E_{21}S_{12}{E_{21}}^{-1} = O_1O_2S_{21}{S_{12}}^{-1}.$$

Generator M of $\Gamma_3[2](v_1)$	The conjugate $\widehat{M} = E_{21}\cdot M\cdot {E_{21}}^{-1}$
O_2	$S_{21}O_2$
O_3	O_3
S_{12}	$O_1O_2S_{21}{S_{12}}^{-1}$
S_{13}	$S_{13}S_{23}$
S_{23}	S_{23}
S_{32}	$S_{32}{S_{31}}^{-1}$

Table 1: The conjugates of the generating set of $\Gamma_3[2](v_1)$ by E_{21}

Let f_i be the edge joining v_1+v_2 to v_i ($1\le i\le 3$), and let J_i be the stabiliser of f_i. We consider these each in turn. Observe that

$$J_2 = E_{21}\cdot\Gamma_3[2](v_1,v_2)\cdot {E_{21}}^{-1},$$

so J_2 is generated by $\{O_3, S_{13}S_{23}, S_{23}\}$. We have expressed those three generators in terms of the generators of $\Gamma_3[2](v_1)$. To obtain the relations corresponding to this edge stabiliser, we must express them using the generators of $\Gamma_3[2](v_1+v_2)$, and set them to be equal accordingly. Consulting Table 1, we get the edge relations

$$\widehat{O}_3 = O_3,\quad \widehat{S}_{13} = S_{13}S_{23}\quad\text{and}\quad \widehat{S}_{23} = S_{23}.$$

Note that these relations simply reiterate the expressions we had already determined for $\widehat{O}_3$, $\widehat{S}_{13}$ and $\widehat{S}_{23}$. Similarly, as we obtain J_3 by conjugating $\Gamma_3[2](v_1,v_3)$ by E_{21}, the edge relations arising from the edge f_3 are

$$\widehat{O}_2 = S_{21}O_2,\quad \widehat{S}_{12} = O_1O_2S_{21}{S_{12}}^{-1}\quad\text{and}\quad \widehat{S}_{32} = S_{32}{S_{31}}^{-1}.$$

Finally, to obtain J_1, we conjugate $\Gamma_3[2](v_1, v_2)$ by the elementary matrix E_{12}. We obtain that J_1 is generated by $\{O_3, S_{13}, S_{13}S_{23}\}$, which gives edge relations $\widehat{O}_3 = O_3$, $S_{13} = \widehat{S}_{13}\widehat{S}_{23}^{-1}$ and $\widehat{S}_{13} = S_{13}S_{23}$. Note that these relations all arise as consequences of the edge relations coming from the edges f_2 and f_3, so are not required.

We now use these edge relations to replace the formal relators defining $\Gamma_3[2](v_1 + v_2)$ with words on the generating set $\{S_{ij}, O_k\}$. Using Tietze transformations and Brown's Theorem 3 [6], we may then conclude that a complete presentation for $\Gamma_3[2]$ is obtained by adding these relators to the presentation $\mathcal{P}$. For example, the relator $\widehat{O}_2^2$ becomes $(S_{21}O_2)^2$. All but one of these additional relators are consequences of ones already in $\mathcal{P}$. The one relator that is not is $[\widehat{S}_{13}, \widehat{S}_{32}]\widehat{S}_{12}^{-2}$, which becomes

$$[S_{13}S_{23}, S_{32}{S_{31}}^{-1}](O_1O_2S_{21}{S_{12}}^{-1})^{-2}.$$

Using the other relations in $\Gamma_3[2]$, this word may be rewritten in the form of relator 10 in Corollary 1.1; we have thus verified that the presentation given in Corollary 1.1 is correct when $n = 3$. This proves Proposition 1.2.

References

[1] **M A Armstrong**, *On the fundamental group of an orbit space*, Proc. Cambridge Philos. Soc. 61 (1965) 639–646 MR0187244

[2] **J S Birman**, *Braids, links, and mapping class groups*, Annals of Mathematics Studies 82, Princeton Univ. Press (1974) MR0375281

[3] **J S Birman**, **H M Hilden**, *On isotopies of homeomorphisms of Riemann surfaces*, Ann. of Math. 97 (1973) 424–439 MR0325959

[4] **T E Brendle**, **D Margalit**, *Point pushing, homology, and the hyperelliptic involution*, Michigan Math. J. 62 (2013) 451–473 MR3102525

[5] **T Brendle**, **D Margalit**, **A Putman**, *Generators for the hyperelliptic Torelli group and the kernel of the Burau representation at $t = -1$*, preprint (2014) `arXiv:1211.4018`

[6] **K S Brown**, *Presentations for groups acting on simply-connected complexes*, J. Pure Appl. Algebra 32 (1984) 1–10 MR739633

[7] **R Charney**, *On the problem of homology stability for congruence subgroups*, Comm. Algebra 12 (1984) 2081–2123 MR747219

[8] **D J Collins**, *Palindromic automorphisms of free groups*, from: "Combinatorial and geometric group theory", (A J Duncan, N D Gilbert, J Howie, editors), London Math. Soc. Lecture Note Ser. 204, Cambridge Univ. Press (1995) 63–72 MR1320275

[9] **J Crisp**, **L Paris**, *Artin groups of type B and D*, Adv. Geom. 5 (2005) 607–636 MR2174484

[10] **M Day**, **A Putman**, *A Birman exact sequence for* $\mathrm{Aut}(F_n)$, Adv. Math. 231 (2012) 243–275 MR2935388

[11] **M Day**, **A Putman**, *The complex of partial bases for* F_n *and finite generation of the Torelli subgroup of* $\mathrm{Aut}(F_n)$, Geom. Dedicata 164 (2013) 139–153 MR3054621

[12] **B Farb**, **D Margalit**, *A primer on mapping class groups*, Princeton Mathematical Series 49, Princeton Univ. Press (2012) MR2850125

[13] **N J Fullarton**, *Palindromic automorphisms of free groups and rigidity of automorphism groups of right-angled Artin groups*, PhD thesis, University of Glasgow (2014) MR3389544 Available at `http://search.proquest.com/docview/1689623221`

[14] **N J Fullarton**, **A Thomas**, *Palindromic automorphisms of right-angled Artin groups*, preprint (2015) `arXiv:1510.03939`

[15] **H H Glover**, **C A Jensen**, *Geometry for palindromic automorphism groups of free groups*, Comment. Math. Helv. 75 (2000) 644–667 MR1789180

[16] **R Hain**, *Finiteness and Torelli spaces*, from: "Problems on mapping class groups and related topics", (B Farb, editor), Proc. Sympos. Pure Math. 74, Amer. Math. Soc. (2006) 57–70 MR2264131

[17] **C Jensen**, **J McCammond**, **J Meier**, *The Euler characteristic of the Whitehead automorphism group of a free product*, Trans. Amer. Math. Soc. 359 (2007) 2577–2595 MR2286046

[18] **R Kobayashi**, *A finite presentation of the level* 2 *principal congruence subgroup of* $\mathrm{GL}(n;\mathbb{Z})$, preprint (2015) `arXiv:1407.1612`

[19] **R C Lyndon**, **P E Schupp**, *Combinatorial group theory*, Ergeb. Math. Grenzgeb. 89, Springer, Berlin (1977) MR0577064

[20] **W Magnus**, **A Karrass**, **D Solitar**, *Combinatorial group theory*, 2nd edition, Dover, New York (1976) MR0422434

[21] **J D McCarthy**, **U Pinkall**, *Representing homology automorphisms of nonorientable surfaces* (2004) Revised version of Max Planck Inst. preprint MPI/SFB 85-11 Available at `http://users.math.msu.edu/users/mccarthy/publications/repr.hom.auts.pdf`

[22] **J Milnor**, *Introduction to algebraic K–theory*, Annals of Mathematics Studies 72, Princeton Univ. Press (1971) MR0349811

[23] **T Morifuji**, *On Meyer's function of hyperelliptic mapping class groups*, J. Math. Soc. Japan 55 (2003) 117–129 MR1939188

[24] **B Perron**, **J P Vannier**, *Groupe de monodromie géométrique des singularités simples*, Math. Ann. 306 (1996) 231–245 MR1411346

[25] **A Putman**, *Obtaining presentations from group actions without making choices*, Algebr. Geom. Topol. 11 (2011) 1737–1766 MR2821439

[26] **R D Wade**, *Folding free-group automorphisms*, Q. J. Math. 65 (2014) 291–304 MR3179663

Department of Mathematics, Rice University
MS 136, 6100 Main Street, Houston, TX 77005, USA

neil.fullarton@rice.edu

Received: 4 November 2014 Revised: 14 April 2015

Algebraic & Geometric Topology 15 (2015) 3569–3598

Character varieties of double twist links

KATHLEEN L PETERSEN
ANH T TRAN

We compute both natural and smooth models for the $\mathrm{SL}_2(\mathbb{C})$ character varieties of the two-component double twist links, an infinite family of two-bridge links indexed as $J(k,l)$. For each $J(k,l)$, the component(s) of the character variety containing characters of irreducible representations are birational to a surface of the form $C \times \mathbb{C}$, where C is a curve. The same is true of the canonical component. We compute the genus of this curve, and the degree of irrationality of the canonical component. We realize the natural model of the canonical component of the $\mathrm{SL}_2(\mathbb{C})$ character variety of the $J(3, 2m+1)$ link as the surface obtained from $\mathbb{P}^1 \times \mathbb{P}^1$ as a series of blow-ups.

57M25; 57N10, 14J26

1 Introduction

Given a complete orientable finite-volume hyperbolic 3–manifold with cusps, the $\mathrm{SL}_2(\mathbb{C})$ character variety of M, $X(M)$, is an affine complex algebraic set associated to representations $\pi_1(M) \to \mathrm{SL}_2(\mathbb{C})$. Thurston [14] showed that any irreducible component of such a variety containing the character of a discrete faithful representation has complex dimension equal to the number of cusps of M. Such components are called *canonical components* and are denoted $X_0(M)$. Character varieties have been fundamental tools in studying the topology of M (we refer the reader to Shalen [13] for more), and canonical components encode a wealth of topological information about M, including containing subvarieties associated to Dehn fillings of M and identifying boundary slopes of essential surfaces; see Culler and Shalen [3].

We consider the two-component double twist links $J(k,l)$ and compute the character varieties of their complements in S^3. As pictured in Figure 1, the integers k and l determine the number of half-twists in the boxes; positive numbers correspond to right-handed twists and negative numbers correspond to left-handed twists. The link $J(k,l)$ is a two-component link when kl is odd and a knot when kl is even. Macasieb, Petersen and van Luijk [8] determined and analyzed character varieties of the $J(k,l)$ knots. In this paper, we extend this work to the two-component $J(k,l)$ links. These

Published: 12 January 2016 DOI: 10.2140/agt.2015.15.3569

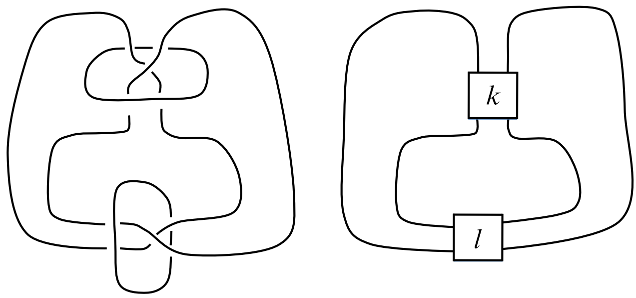


Figure 1: The link $J(k,l)$ is the result of $-1/k$ and $-1/l$ surgery on the four-component link pictured on the left.

are hyperbolic exactly when $|k|$ and $|l|$ are greater than one; the $J(\pm 1, l) = J(l, \pm 1)$ links are torus links. We will now exclusively consider the hyperbolic $J(k,l)$ links.

In Definition 3.5 we define the Chebyshev polynomials S_j which are used throughout the paper. Our first theorem establishes natural models for the $\mathrm{SL}_2(\mathbb{C})$ character varieties of the double twist links. With $\pi_1(k,l) = \pi_1(S^3 - J(k,l))$, let $X_{\mathrm{irr}}(k,l)$ denote the closure of the set of all irreducible characters χ_ρ of representations $\rho\colon \pi_1(k,l) \to \mathrm{SL}_2(\mathbb{C})$. Let $X_0(k,l)$ denote a canonical component. In fact, a consequence of this work is that for a given double twist link, there is only one canonical component. For this natural model, we use the presentation for $\pi_1(k,l)$ in Section 3 with $x = \chi_\rho(a)$, $y = \chi_\rho(b)$ and $z = \chi_\rho(ab^{-1})$. The vanishing set of the characters of reducible representations $\pi_1(k,l) \to \mathrm{SL}_2(\mathbb{C})$ is well-known and is given by

$$xyz + 4 - x^2 - y^2 - z^2$$

in $\mathbb{C}^3(x,y,z)$. These are all characters of abelian representations.

Theorem 1.1 *Let $k = 2m+1$ and $l = 2n+1$. A natural model for the algebraic set $X_{\mathrm{irr}}(k,l)$ is the vanishing set of*

$$S_n(t)S_{m-1}(z) - S_{n-1}(t)S_m(z)$$

in $\mathbb{C}^3(x,y,z)$, where

$$t = (xS_m(z) - yS_{m-1}(z))(yS_m(z) - xS_{m-1}(z)) - z(S_m^2(z) + S_{m-1}^2(z)) + 4S_m(z)S_{m-1}(z).$$

The expression t is the trace of $\rho(\beta)$, with the loop β as pictured in Figure 2. In terms of the presentation for the fundamental group in Section 3, the loop β corresponds to the word w_k.

Our next theorem establishes smooth models for these algebraic sets.

Theorem 1.2 *Let $k=2m+1$ and $l=2n+1$. The algebraic set $X_{\mathrm{irr}}(k,l)$ is birational to $C(k,l)\times\mathbb{C}$, where the curve $C(k,l)\subset\mathbb{C}^2(t,z)$ is given by*

$$C(k,l)=\{S_n(t)S_{m-1}(z)-S_{n-1}(t)S_m(z)=0\}.$$

If $k\neq l$ then $C(k,l)$ is smooth and irreducible as considered in $\mathbb{P}^1(t)\times\mathbb{P}^1(z)$, and $X_0(k,l)=X_{\mathrm{irr}}(k,l)$ is birational to $C(k,l)\times\mathbb{C}$.

The curve $C(3,3)=C(-3,-3)$ is given by $t=z$. If $k=l$ and $|l|>3$ then $C(l,l)$ is the union of exactly two components: $C_0(l,l)$, given by $t=z$, and $C_1(l,l)$, the scheme-theoretic complement of $C_0(l,l)$ in $C(l,l)$. Both are smooth and irreducible as considered in $\mathbb{P}^1(t)\times\mathbb{P}^1(z)$. The algebraic set $X_{\mathrm{irr}}(k,l)$ is given by the union $X_0(l,l)\cup X_1(l,l)$, where $X_0(l,l)$ is birational to $C_0(l,l)\times\mathbb{C}$ and $X_1(l,l)$ is birational to $C_1(l,l)\times\mathbb{C}$.

We next compute some invariants of these algebraic sets. Since $X_{\mathrm{irr}}(k,l)$ is birational to the product of a curve $C(k,l)$ and $\mathbb{C}$, we compute the genus of this curve.

Theorem 1.3 *Let $k=2m+1$ and $l=2n+1$ with $|k|,|l|>1$. When $k\neq l$ the genus of $C(k,l)$ is*

$$\left(\left\lfloor\tfrac{|k|}{2}\right\rfloor-1\right)\left(\left\lfloor\tfrac{|l|}{2}\right\rfloor-1\right).$$

The genus of $C_0(l,l)$ is zero, and when $|l|>3$ the genus of $C_1(l,l)$ is $\left(\left\lfloor\tfrac{|l|}{2}\right\rfloor-2\right)^2$.

The degree of irrationality of an irreducible n–dimensional complex algebraic set X is defined to be the minimal degree of any rational map from X to a dense subset of $\mathbb{C}^n$. This is denoted $\gamma(X)$ and is a birational invariant. When X is a curve this is called the gonality of X. See Petersen and Reid [11] for a discussion on how gonality and genus behave in families of Dehn fillings. In light of this, since $J(k,l)$ is $-1/k$ and $-1/l$ filling of the four-component link in Figure 1, we compute the degree of irrationality of the surfaces $X_0(k,l)$ and $X_1(l,l)$.

Theorem 1.4 *Let $k=2m+1$ and $l=2n+1$. The degree of irrationality of $X_0(k,l)$ is $\min\left\{\left\lfloor\tfrac{|k|}{2}\right\rfloor,\left\lfloor\tfrac{|l|}{2}\right\rfloor\right\}$ when $k\neq l$. The degree of irrationality of $X_0(l,l)$ is 1, and when $|l|>3$ the degree of irrationality of $X_1(l,l)$ is $\left\lfloor\tfrac{|l|}{2}\right\rfloor-1$.*

Finally, we study the $J(3,2m+1)$ links realizing $X_0(3,2m+1)$ as a series of blow-ups of $\mathbb{P}^1\times\mathbb{P}^1$ and show the following.

Theorem 1.5 *The desingularization of the natural model for the canonical component of the* $\mathrm{SL}_2(\mathbb{C})$ *character variety of the double twist link* $J(3, 2m+1)$ *is the conic bundle over the projective line* $\mathbb{P}^1$ *which is isomorphic to the surface obtained from* $\mathbb{P}^1 \times \mathbb{P}^1$ *by repeating a one-point blow-up* $9m$ *times if* $m \geq 1$*, and* $-(6+9m)$ *times if* $m \leq -2$*. Equivalently, it is isomorphic to the surface obtained from* $\mathbb{P}^2$ *by repeating a one-point blow-up* $1+9m$ *times if* $m \geq 1$*, and* $-(5+9m)$ *times if* $m \leq -2$*.*

Remark 1.6 For $m \geq 1$, the link $J(3, 2m+1)$ is obtained by $1/m$ Dehn surgery on the magic manifold. Hence Theorem 1.5 confirms Conjecture 3.1.3 in Landes' thesis [7].

Acknowledgement This work was partially supported by a grant from the Simons Foundation (number 209226 to Kathleen Petersen).

2 Character varieties

We will define our notation, but refer the reader to [8] for a detailed discussion of character varieties. Let M be a complete finite-volume hyperbolic 3–manifold. The $\mathrm{SL}_2(\mathbb{C})$ character variety of M is the set of all characters of representations $\rho\colon \pi_1(M) \to \mathrm{SL}_2(\mathbb{C})$. The character associated to ρ is $\chi_\rho\colon \pi_1(M) \to \mathbb{C}$, defined by $\chi_\rho(\gamma) = \operatorname{tr} \rho(\gamma)$.

Let $X(M)$ denote the $\mathrm{SL}_2(\mathbb{C})$ character variety, that is

$$X(M) = \{\chi_\rho \mid \rho\colon \pi_1(M) \to \mathrm{SL}_2(\mathbb{C})\}.$$

The characters of reducible representations themselves form an algebraic set, which is a subset of $X(M)$. We will call this set $X_{\mathrm{red}}(M)$. The closure of the set of characters of irreducible representations will be denoted by $X_{\mathrm{irr}}(M)$. Any irreducible component of $X(M)$ which contains the character of a discrete faithful representation is contained in $X_{\mathrm{irr}}(M)$ and is called a canonical component and denoted $X_0(M)$.

Thurston [14] showed that the complex dimension of any canonical component equals the number of cusps of M. Canonical components encode much of the topology of M, often seen through the trace functions. Canonical components containing subvarieties corresponding to Dehn fillings of M and their ideal points can be used to determine essential surfaces in M (see [3]).

When M has only one cusp $X_0(M)$ is a curve. Several infinite families of these have been studied. (See [1; 8; 16], for explicit computations. See [10] and [2] for examples of families of manifolds with many components in their character varieties.) When M has at least two cusps the algebraic geometry becomes more demanding, and only a few solitary examples have been computed. Landes [6; 7] computed a smooth model for the canonical component of the $\mathrm{SL}_2(\mathbb{C})$ character variety of the complement of the Whitehead link, a two-component link. (She explicitly showed that it is a rational surface homeomorphic to the projective plane blown up at 10 points.) Harada [5] computed the character varieties of the four arithmetic two-bridge link complements (including the Whitehead link and the figure-8 knot). Our computation of the character varieties of the double twist links is the first result to compute character varieties for infinitely many 3–manifolds with two cusps.

3 Double twist links

Let $J(k,l)$ be the double twist link indicated in the right-hand side of Figure 1. This link is $-1/k$ and $-1/l$ filling on two components of the four-component link shown in the left-hand side of Figure 1. This is a knot when kl is even and a two-component link when kl is odd. The link $J(k,l)$ corresponds to the continued fraction $[k,-l]$. It is hyperbolic, unless $|k|$ or $|l|$ is 1. Let $X(k,l)$ denote the $\mathrm{SL}_2(\mathbb{C})$ character variety of $S^3 - J(k,l)$.

In [8] the character varieties of the $J(k,l)$ knots were computed. We now consider the $J(k,l)$ links with two components, so both k and l are odd. Suppose $k=2m+1$ and $l=2n+1$. The link group of $J(k,l)$ is $\pi_1(k,l)=\pi_1(S^3-J(k,l))$ and has presentation

$$\pi_1(k,l)=\langle a,b \mid a w_k^n b = w_k^{n+1}\rangle,$$

where $w_k=(ab^{-1})^m ab(a^{-1}b)^m$ [8].

Definition 3.1 Let $F_{a,b}=\langle a,b\rangle$ be the free group on two letters a and b. For a word u in $F_{a,b}$ let $\overleftarrow{u}$ denote the word obtained from u by writing the letters in u in reversed order.

We begin by simplifying the presentation of the link group.

Lemma 3.2 *With $w_k=(ab^{-1})^m ab(a^{-1}b)^m$ and $r=w_k^n(ab^{-1})^m$, we have*

$$\pi_1(k,l)=\langle a,b \mid r=\overleftarrow{r}\rangle.$$

Proof We can rewrite the presentation of $\pi_1(k,l)$ as

$$\begin{aligned}\pi_1(k,l)&=\langle a,b \mid a w_k^n b=(ab^{-1})^m ab(a^{-1}b)^m w_k^{n-1}(ab^{-1})^m ab(a^{-1}b)^m\rangle\\ &=\langle a,b \mid w_k^n=(b^{-1}a)^m b(a^{-1}b)^m w_k^{n-1}(ab^{-1})^m a(ba^{-1})^m\rangle\\ &=\langle a,b \mid w_k^n(ab^{-1})^m=(b^{-1}a)^m b(a^{-1}b)^m w_k^{n-1}(ab^{-1})^m a\rangle.\end{aligned}$$

Let $c=(ab^{-1})^m a$ and $d=b(a^{-1}b)^m$. Then $w_k=cd$. It follows that

$$b(a^{-1}b)^m w_k^{n-1}(ab^{-1})^m a = d(cd)^{n-1}c=(dc)^n=\overleftarrow{(cd)^n}=\overleftarrow{w_k^n}.$$

Hence

$$\begin{aligned}\pi_1(k,l)&=\langle a,b \mid w_k^n(ab^{-1})^m=\overleftarrow{(ab^{-1})^m}\,\overleftarrow{w_k^n}\rangle\\ &=\langle a,b \mid w_k^n(ab^{-1})^m=\overleftarrow{w_k^n(ab^{-1})^m}\rangle.\end{aligned}$$

Since $r=w_k^n(ab^{-1})^m$, the lemma follows. □

With coordinates $x=\operatorname{tr}\rho(a)$, $y=\operatorname{tr}\rho(b)$ and $z=\operatorname{tr}\rho(ab^{-1})$, the character variety of the free group $F_{a,b}$ is isomorphic to $\mathbb{C}^3[x,y,z]$ by the Fricke–Klein–Vogt theorem [4; 17]. Consider a word u in $F_{a,b}$. Define the polynomial $P_u\in\mathbb{C}[x,y,z]$ to be $P_u(x,y,z)=\operatorname{tr}\rho(u)$. It follows that for every word u in $F_{a,b}$ the polynomial P_u is the unique polynomial such that for any representation $\rho\colon F_{a,b}\to \mathrm{SL}_2(\mathbb{C})$ we have $\operatorname{tr}\rho(u)=P_u(x,y,z)$.

We now consider representations $\rho\colon \pi_1(k,l)\to\mathrm{SL}_2(\mathbb{C})$. By Lemma 3.2 the group $\pi_1(k,l)$ has a presentation with two generators and one relation and therefore is a quotient of $F_{a,b}$. First, we establish some notation which we will use throughout the manuscript.

Definition 3.3 Let $k=2m+1$ and $l=2n+1$. For $\rho\colon \pi_1(k,l)\to\mathrm{SL}_2(\mathbb{C})$ define

$$x=\operatorname{tr}\rho(a),\quad y=\operatorname{tr}\rho(b)\quad\text{and}\quad z=\operatorname{tr}\rho(ab^{-1}),$$

and for a word u in $F_{a,b}$ define the polynomial $P_u(x,y,z)=\operatorname{tr}\rho(u)\in\mathbb{C}[x,y,z]$. Further, let $t=P_{w_k}$ and

$$\varphi(x,y,z)=P_{rab}-P_{\overleftarrow{r}ab}.$$

For every representation $\rho\colon \pi_1(k,l)\to\mathrm{SL}_2(\mathbb{C})$, we consider x, y and z as functions of ρ. Using the presentation above for $\pi_1(k,l)$ with two generators and one relation, we conclude that $P_{rab}=P_{\overleftarrow{r}ab}$, which is simply $\varphi(x,y,z)=0$, in $X(k,l)$. In fact, by [16, Theorem 1] $X(k,l)$ is exactly the zero set of $\varphi(x,y,z)$. (See also [12, Theorem 2.1]). Moreover, because of the format of the defining word, $P_{\overleftarrow{r}ab}=P_{bar}$ [16, Theorem 1]. (That is, these polynomials in $\mathbb{C}^3[x,y,z]$ are identical.) Therefore, $\varphi(x,y,z)=P_{rab}-P_{bar}$. We summarize this discussion in the following proposition.

Proposition 3.4 *The polynomial $\varphi(x,y,z)$ is given by $P_{rab}-P_{bar}$. The character variety $X(k,l)$ is the zero set of $\varphi(x,y,z)$ in $\mathbb{C}^3(x,y,z)$.*

We wish to obtain a nice format for φ. We introduce a family of Chebyshev polynomials, often called the Fibonacci polynomials, that will be essential to our computation of φ. (These are slightly different polynomials than were used in [8]; the indices are shifted by one.)

Definition 3.5 Let $S_j(\omega)$ be the Chebyshev polynomials defined by

$$S_0(\omega)=1, \quad S_1(\omega)=\omega \quad \text{and} \quad S_{j+1}(\omega)=\omega S_j(\omega)-S_{j-1}(\omega)$$

for all integers j.

It is elementary to verify the following lemmas.

Lemma 3.6 *With $\omega=\sigma+\sigma^{-1}$ we have*

$$S_j(\omega)=\frac{\sigma^{j+1}-\sigma^{-j-1}}{\sigma-\sigma^{-1}}.$$

The degree of S_j is j if $j>-1$ and $-j-2$ if $j<-1$.

Lemma 3.7 *Suppose the sequence $\{f_j\}_{j\in\mathbb{Z}}$ satisfies the recurrence relation $f_{j+1}=\omega f_j-f_{j-1}$ for all integers j. Then $f_j=S_j(\omega)f_0-S_{j-1}(\omega)f_{-1}$.*

The following lemma can be verified by using Lemma 3.6.

Lemma 3.8 *We have*

(a) $S_j^2(\omega)+S_{j-1}^2(\omega)-\omega S_j(\omega)S_{j-1}(\omega)=1$,

(b) $S_j^2(\omega)-S_{j-1}^2(\omega)=S_{2j}(\omega)$,

(c) $S_{m-1}(\omega)\big(\omega+(\omega^2-4)S_{m-1}(\omega)S_m(\omega)\big)+S_m(\omega)=S_{3m}(\omega)$.

We now simplify the polynomial φ by writing the trace polynomials in terms of these Chebyshev polynomials.

Proposition 3.9 *We have*

$$t=(xS_m(z)-yS_{m-1}(z))(yS_m(z)-xS_{m-1}(z)) \\ -z(S_m^2(z)+S_{m-1}^2(z))+4S_m(z)S_{m-1}(z).$$

Proof By definition, $t = P_{w_k}$. By applying Lemma 3.7 twice, we have

$$\begin{aligned} P_{w_k} &= P_{(ab^{-1})^m ab (a^{-1}b)^m} \\ &= S_m^2(z) P_{ab} + S_{m-1}^2(z) P_{(ab^{-1})^{-1}ab(a^{-1}b)^{-1}} \\ &\qquad\qquad - S_m(z) S_{m-1}(z) (P_{(ab^{-1})^{-1}ab} + P_{ab(a^{-1}b)^{-1}}) \\ &= S_m^2(z) P_{ab} + S_{m-1}^2(z) P_{ba} - S_m(z) S_{m-1}(z) (P_{b^2} + P_{a^2}) \\ &= (S_m^2(z) + S_{m-1}^2(z))(xy - z) - S_m(z) S_{m-1}(z)(x^2 + y^2 - 4). \end{aligned}$$

The proposition follows. $\square$

Proposition 3.10 *The polynomial $\varphi(x,y,z) \in \mathbb{C}^3[x,y,z]$ is*

$$\varphi(x,y,z) = (xyz + 4 - x^2 - y^2 - z^2)\big(S_n(t) S_{m-1}(z) - S_{n-1}(t) S_m(z)\big),$$

where t is as in Proposition 3.9.

Proof As mentioned above, by [16, Theorem 1] $X(k,l)$ is the zero set of $\varphi(x,y,z)$ and $P_{\overleftarrow{r}ab} = P_{bar}$. By applying Lemma 3.7 we have

$$\begin{aligned} P_{rab} - P_{bar} &= P_{w_k^n (ab^{-1})^m ab} - P_{baw_k^n (ab^{-1})^m} \\ &= S_n(t)(P_{(ab^{-1})^m ab} - P_{ba(ab^{-1})^m}) \\ &\qquad\qquad - S_{n-1}(t)(P_{w_k^{-1}(ab^{-1})^m ab} - P_{baw_k^{-1}(ab^{-1})^m}) \\ &= S_n(t)(P_{(ab^{-1})^m ab} - P_{ba(ab^{-1})^m}) \\ &\qquad\qquad - S_{n-1}(t)(P_{(a^{-1}b)^m} - P_{ab(a^{-1}b)^m(ba)^{-1}}), \end{aligned}$$

where

$$\begin{aligned} P_{(ab^{-1})^m ab} - P_{ba(ab^{-1})^m} &= S_m(z)(P_{ab} - P_{ba}) \\ &\qquad - S_{m-1}(z)(P_{(ab^{-1})^{-1}ab} - P_{ba(ab^{-1})^{-1}}) \\ &= -S_{m-1}(z)(P_{b^2} - P_{baba^{-1}}) \\ &= S_{m-1}(z)(xyz + 4 - x^2 - y^2 - z^2), \\ P_{(a^{-1}b)^m} - P_{ab(a^{-1}b)^m(ba)^{-1}} &= S_m(z)(P_1 - P_{ab(ba)^{-1}}) \\ &\qquad - S_{m-1}(z)(P_{(a^{-1}b)^{-1}} - P_{ab(a^{-1}b)^{-1}(ba)^{-1}}) \\ &= S_m(z)(xyz + 4 - x^2 - y^2 - z^2). \end{aligned}$$

Hence

$$P_{rab} - P_{bar} = (xyz + 4 - x^2 - y^2 - z^2)(S_n(t) S_{m-1}(z) - S_{n-1}(t) S_m(z)). \quad \square$$

The character variety $X(k,l)$ is clearly reducible. The set of reducible characters, $X_{\mathrm{red}}(k,l)$, can easily be determined, as in [1], for example. We have the following, from which Theorem 1.1 follows immediately.

Proposition 3.11 *The vanishing set of*

$$xyz + 4 - x^2 - y^2 - z^2$$

in $\mathbb{C}^3(x,y,z)$ *is the set of characters of reducible representations* $\pi_1(k,l) \to \mathrm{SL}_2(\mathbb{C})$.

A natural model for the algebraic set $X_{\mathrm{irr}}(k,l)$ *is the vanishing set of*

$$S_n(t)S_{m-1}(z) - S_{n-1}(t)S_m(z)$$

in $\mathbb{C}^3(x,y,z)$, *where* t *is as in Proposition 3.9.*

In light of this, we wish to understand the vanishing set of the difference $S_n(t)S_{m-1}(z) - S_{n-1}(t)S_m(z)$. The equation $S_n(t)S_{m-1}(z) = S_{n-1}(t)S_m(z)$ can be written as

$$\frac{S_n(t)}{S_{n-1}(t)} = \frac{S_m(z)}{S_{m-1}(z)}$$

when $S_{n-1}(t)S_{m-1}(z) \neq 0$, so we can think of it as lying in a product of projective lines. We will make use of this approach when proving smoothness and irreducibility.

Definition 3.12 Let $V(k,l)$ be the vanishing set of $S_n(t)S_{m-1}(z) - S_{n-1}(t)S_m(z)$ in $\mathbb{C}^3(x,y,z)$.

By Proposition 3.11 the components of $X(k,l)$ containing characters of irreducible representations, those included in $X_{\mathrm{irr}}(k,l)$, are contained in $V(k,l)$ and $V(k,l)$ is a natural model for this set.

4 The structure of $V(k,l)$

The set $V(k,l)$ is the closure of the set of characters of irreducible representations. The equation $S_n(t)S_{m-1}(z) - S_{n-1}(t)S_m(z)$ is relatively simple, except that t itself is a function of the natural variables x, y, and z. Explicitly, by Proposition 3.9,

$$t = (xS_m(z) - yS_{m-1}(z))(yS_m(z) - xS_{m-1}(z)) - z(S_m^2(z) + S_{m-1}^2(z)) + 4S_m(z)S_{m-1}(z).$$

We will show that there is a relatively simple model for $X_{\mathrm{irr}}(k,l)$ up to birational equivalence.

Definition 4.1 Let $u = xS_m(z) - yS_{m-1}(z)$ and $v = yS_m(z) - xS_{m-1}(z)$.

It follows that

$$t = uv - z(S_m^2(z) + S_{m-1}^2(z)) + 4S_m(z)S_{m-1}(z).$$

By the definitions of u and v,

$$x = \frac{uS_m(z) + vS_{m-1}(z)}{S_m^2(z) - S_{m-1}^2(z)} \quad \text{and} \quad y = \frac{vS_m(z) + uS_{m-1}(z)}{S_m^2(z) - S_{m-1}^2(z)}.$$

We will show that this substitution of u and v for x and y corresponds to a birational map, simplifying the definition of t. Then we will show that substituting t for u is another birational map, thus eliminating the problem of having nested variables. This has the fortunate consequence that the equation $S_n(t)S_{m-1}(z) - S_{n-1}(t)S_m(z)$ contains no u, so we can conclude that the algebraic set $V(k,l)$ is birational to the product of a curve and $\mathbb{C}$.

Definition 4.2 Let $U(k,l)$ be the vanishing set of

$$S_n(t)S_{m-1}(z) - S_{n-1}(t)S_m(z)$$

in $\mathbb{C}^3(u,v,z)$, where

$$t = uv - z(S_m^2(z) + S_{m-1}^2(z)) + 4S_m(z)S_{m-1}(z).$$

Before showing that $V(k,l)$ is birational to $U(k,l)$ we prove a lemma.

Lemma 4.3 *On $V(k,l)$, $S_m^2(z) - S_{m-1}^2(z) = 0$ only for a set of codimension one.*

Proof By definition $S_j(z)$ is a Chebyshev polynomial, and by Lemma 3.8 we have that $S_m^2(z) - S_{m-1}^2(z) = S_{2m}(z)$. Moreover, letting $z = \sigma + \sigma^{-1}$, we can write

$$S_{2m}(\sigma + \sigma^{-1}) = \frac{\sigma^{2m+1} - \sigma^{-2m-1}}{\sigma - \sigma^{-1}}.$$

Therefore, if $S_m^2(z) - S_{m-1}^2(z) = 0$ then $\sigma^{2m+1} - \sigma^{-2m-1} = 0$ and so $\sigma^{4m+2} = 1$. It follows that

$$\sigma = e^{2\pi i s/(4m+2)} = e^{\pi i s/(2m+1)}$$

for some $0 \le s \le 4m+2$. When $s = 2r$ is even ($1 \le r \le m$),

$$z = \sigma + \sigma^{-1} = 2\mathrm{Re}(\sigma) = 2\cos\Big(\frac{2\pi r}{2m+1}\Big)$$

and z is a root of $S_m(z)+S_{m-1}(z)$. When $s=2r+1$ is odd ($0\le r\le m-1$),

$$z=2\cos\Big(\frac{(2r+1)\pi}{2m+1}\Big)$$

and z is a root of $S_m(z)-S_{m-1}(z)$.

First, we will show that $S_m(z)-S_{m-1}(z)=0$ only for a set of dimension one on $V(k,l)$. By Lemma 3.8,

$$S_m^2(z)+S_{m-1}^2(z)-zS_m(z)S_{m-1}(z)=1.$$

Since $S_m(z)=S_{m-1}(z)$, we obtain $S_m^2(z)=1/(2-z)$ and

$$t=-S_m^2(z)((x-y)^2+2z-4)=\frac{1}{z-2}((x-y)^2+2(z-2))=\frac{(x-y)^2}{z-2}+2.$$

We conclude that

$$(x-y)^2=(z-2)(t-2).$$

On $V(k,l)$, $S_n(t)S_{m-1}(z)-S_{n-1}(t)S_m(z)=0$. Since $S_m(z)=S_{m-1}(z)$ we get

$$S_m(z)(S_n(t)-S_{n-1}(t))=0.$$

Since z is as above, we see that $S_m(z)\neq 0$ since $S_m^2(z)=1/(2-z)$. Hence $S_n(t)-S_{n-1}(t)=0$. It follows that

$$t=2\cos\Big(\frac{(2s+1)\pi}{2n+1}\Big),$$

where $0\le s\le n-1$. We conclude that

$$(x-y)^2=4\Big(\cos\Big(\frac{(2r+1)\pi}{2m+1}\Big)-1\Big)\Big(\cos\Big(\frac{(2s+1)\pi}{2n+1}\Big)-1\Big).$$

This defines $x-y$ explicitly, and therefore determines a set of dimension one in $V(k,l)$. Since the dimension of $V(k,l)$ is two, this is a codimension-one set.

We complete the proof by showing that $S_m(z)+S_{m-1}(z)=0$ only for a set of dimension one on $V(k,l)$. Note that $z=2\cos\big(\frac{2\pi r}{2m+1}\big)$, where $1\le r\le m$. We have

$$S_m^2(z)+S_{m-1}^2(z)-zS_m(z)S_{m-1}(z)=1.$$

Since $S_m(z)=-S_{m-1}(z)$, we obtain $S_m^2(z)=1/(2+z)$ and

$$t=S_m^2(z)((x+y)^2-2z-4)=\frac{1}{2+z}((x+y)^2-2(z+2))=\frac{(x+y)^2}{2+z}-2.$$

We conclude that

$$(x+y)^2=(t+2)(z+2).$$

On $V(k,l)$ we have $S_n(t)S_{m-1}(z) - S_{n-1}(t)S_m(z) = 0$, and hence

$$S_m(z)(S_n(t) + S_{n-1}(t)) = 0.$$

Since z is as above, we conclude that $S_n(t)+S_{n-1}(t)=0$. This means $t=2\cos\bigl(\frac{2\pi s}{2n+1}\bigr)$ (where $1 \le s \le n$). Hence

$$(x+y)^2 = 4\Bigl(\cos\Bigl(\frac{2\pi r}{2m+1}\Bigr)+1\Bigr)\Bigl(\cos\Bigl(\frac{2\pi s}{2n+1}\Bigr)+1\Bigr).$$

This defines $x+y$ explicitly, and therefore determines a set of dimension one in $V(k,l)$. Since the dimension of $V(k,l)$ is two, this is a codimension-one set. $\square$

The next result now easily follows.

Proposition 4.4 *The set $V(k,l) \subset \mathbb{C}^3(x,y,z)$ is birational to $U(k,l) \subset \mathbb{C}^3(u,v,z)$.*

Proof As discussed above, the substitution defines a rational map between $V(k,l)$ and $U(k,l)$, namely

$$(x,y,z) \mapsto \biggl(\frac{xS_m(z)+yS_{m-1}(z)}{S_m^2(z)-S_{m-1}^2(z)}, \frac{yS_m(z)+xS_{m-1}(z)}{S_m^2(z)-S_{m-1}^2(z)}, z\biggr),$$

with inverse

$$(u,v,z) \mapsto \bigl(uS_m(z)-vS_{m-1}(z), vS_m(z)-uS_{m-1}(z), z\bigr).$$

It suffices to see that $S_m^2(z)-S_{m-1}^2(z) = 0$ only for a set of codimension one on $V(k,l)$, which follows from Lemma 4.3. $\square$

We now wish to perform one more birational transformation.

Definition 4.5 Let $W(k,l)$ be the vanishing set of

$$S_n(t)S_{m-1}(z) - S_{n-1}(t)S_m(z)$$

in $\mathbb{C}^3(t,v,z)$.

For each odd integer l, let $W_0(l,l)$ denote the component of $W(l,l)$ given by $t=z$ and if $|l|>3$ let $W_1(l,l)$ denote the projective closure of the scheme-theoretic complement of $W_0(l,l)$ in $W(l,l)$.

First, we prove a lemma.

Lemma 4.6 *On $U(k,l)$, $v=0$ only for a set of dimension zero.*

Proof If $v = 0$ then since

$$t = uv - z(S_m^2(z) + S_{m-1}^2(z)) + 4S_m(z)S_{m-1}(z)$$

we conclude that

$$t = -z(S_m^2(z) + S_{m-1}^2(z)) + 4S_m(z)S_{m-1}(z).$$

The defining polynomial for $U(k,l)$ is $S_n(t)S_{m-1}(z) - S_{n-1}(t)S_m(z)$. Upon substituting the above polynomial in $\mathbb{Z}[z]$ for t we see that this defining polynomial can be expressed as a polynomial in $\mathbb{Z}[z]$. As a result, this has a finite number of roots. For each of these z values, there is one associated t, and hence we have a finite number of points on $U(k,l)$ where $v = 0$. $\square$

Now we are prepared to show the following.

Proposition 4.7 *The set $U(k,l) \subset \mathbb{C}^3(u,v,z)$ is birational to $W(k,l) \subset \mathbb{C}^3(t,v,z)$.*

Proof Since t is linear in u, we define the rational map from $\mathbb{C}^3(u,v,z)$ to $\mathbb{C}^3(t,v,z)$ by this replacement. That is, define the rational map

$$(u,v,z) \mapsto \left(\frac{(u + z(S_m^2(z) + S_{m-1}^2(z)) - 4S_m(z)S_{m-1}(z))}{v}, v, z \right)$$

which has rational inverse

$$(t,v,z) \mapsto \big(tv - z(S_m^2(z) + S_{m-1}^2(z)) + 4S_m(z)S_{m-1}(z), v, z\big).$$

The result now follows from Lemma 4.6. $\square$

Definition 4.8 Let $C(k,l)$ be the curve given by the vanishing set of

$$S_n(t)S_{m-1}(z) - S_{n-1}(t)S_m(z)$$

in $\mathbb{C}^2(t,z)$. For each odd integer l, let $C_0(l,l)$ denote the component of $C(l,l)$ given by $t = z$ and if $|l| > 3$ let $C_1(l,l)$ denote the projective closure of the scheme-theoretic complement of $C_0(l,l)$ in $C(l,l)$.

With this definition, the surface $W(k,l)$ is a product of the curve $C(k,l)$ and $\mathbb{C}$. We have shown that $V(k,l)$ is birational to $W(k,l)$, which is equivalent to the following, proving the first portion of Theorem 1.2.

Theorem 4.9 *The algebraic set $X_{\mathrm{irr}}(k,l)$ is birational to $W(k,l)$, which is, in turn, isomorphic to $C(k,l) \times \mathbb{C}$.*

5 Smoothness and irreducibility of $W(k,l)$

We will show that if $k \neq l$ then $W(k,l)$ is smooth and irreducible, and if $k = l$ then $W(l,l)$ has two irreducible components. Since $W(k,l)$ is the product of $C(k,l)$ and $\mathbb{C}$, we will focus on the curve $C(k,l)$. Our proof is similar to [8], but with small modifications. Recall that $k = 2m+1$ and $l = 2n+1$. The equation $S_n(t)S_{m-1}(z) = S_{n-1}(t)S_m(z)$ can be written as

$$\frac{S_n(t)}{S_{n-1}(t)} = \frac{S_m(z)}{S_{m-1}(z)}$$

when $S_{n-1}(t)S_{m-1}(z) \neq 0$.

Definition 5.1 Let

$$h_j = S_j/S_{j-1}, \quad \Delta_j = S_j' S_{j-1} - S_j S_{j-1}' \quad \text{and} \quad H_n = S_j'' S_{j-1} - S_{j-1}'' S_j.$$

We can rewrite the defining equation for $W(k,l)$ as $h_n(t) = h_m(z)$, and with this notation the derivative is $h_j' = \Delta_j / S_{j-1}^2$.

The following lemma can be verified by using Lemma 3.6.

Lemma 5.2 *We have*

(a) $(\omega^2-4)\Delta_j(\omega) = S_{2j}(\omega) - (2j+1)$,

(b) $(\omega^2-4)^2 H_j(\omega) = (2j-2)\omega S_{2j}(\omega) - (4j+2)S_{2j-1}(\omega) + (4j+2)\omega$.

We will need the following lemma (see [8, Lemma 2.6]) to connect smoothness and irreducibility.

Lemma 5.3 *Let $C \subset \mathbb{P}^1 \times \mathbb{P}^1$ be a smooth projective curve of bidegree (a,b) with $a,b > 0$. Then C is irreducible and its genus is $(a-1)(b-1)$.*

The proof of smoothness will follow from comparing valuations at potential critical points. We begin with a few lemmas. In the case that $mn < 0$ we use the following lemma.

Lemma 5.4 *Let $\omega \in \mathbb{C}$ be a root of Δ_n. If $n > 0$ then $|h_n(\omega)| > 1$, and if $n < 0$ then $|h_n(\omega)| < 1$.*

Proof Suppose that $\Delta_n(\omega) = 0$. By Lemma 5.2, $S_{2n}(\omega) = 2n+1$. We have $S_{n-1}(\omega) \neq 0$ (otherwise

$$S_n(\omega)S'_{n-1}(\omega) = S'_n(\omega)S_{n-1}(\omega) - \Delta_n(\omega) = 0,$$

which cannot occur, since S_{n-1} is separable and relatively prime to S_n in $\mathbb{C}[\omega]$). Hence $h_n(\omega) = S_n(\omega)/S_{n-1}(\omega)$ is well-defined. Write $\omega = \sigma + \sigma^{-1}$. We have $S_{2n}(\omega) = 2n+1$, ie

$$\sigma^{2n+1} - \sigma^{-(2n+1)} = (2n+1)(\sigma - \sigma^{-1}).$$

Assume $n > 0$. Then $\sigma^{2n+1} - \sigma^{-(2n+1)}$ and $\sigma - \sigma^{-1}$ are in the same half-plane. It follows that $\sigma^{2n+1} - \bar{\sigma}^{2n+1}$ and $\sigma - \bar{\sigma}$ are in the same half-plane. Since both these values are purely imaginary, we conclude $(\sigma^{2n+1} - \bar{\sigma}^{2n+1})(\sigma - \bar{\sigma}) \leq 0$, with equality if and only if σ^{2n+1} is real.

Let $\alpha = \sigma\bar{\sigma} = |\sigma|^2 > 0$. We have

$$\begin{aligned}
&|\sigma^{n+1} - \sigma^{-(n+1)}|^2 - |\sigma^n - \sigma^{-n}|^2 \\
&\quad = (\sigma^{n+1} - \sigma^{-(n+1)})(\bar{\sigma}^{n+1} - \bar{\sigma}^{-(n+1)}) - (\sigma^n - \sigma^{-n})(\bar{\sigma}^n - \bar{\sigma}^{-n}) \\
&\quad = (\alpha^{n+1} + \alpha^{-(n+1)} - (\alpha^n + \alpha^{-n})) - (\sigma^{2n+1} - \bar{\sigma}^{2n+1})(\sigma - \bar{\sigma})/\sigma^{n+1}\bar{\sigma}^{n+1} \\
&\quad = (\alpha - 1)(\alpha^{2n+1} - 1)/\alpha^{n+1} - (\sigma^{2n+1} - \bar{\sigma}^{2n+1})(\sigma - \bar{\sigma})/\alpha^{n+1} \geq 0.
\end{aligned}$$

Equality holds if and only if $|\sigma|^2 = \alpha = 1$ and σ^{2n+1} is real, so if and only if $\sigma^{2n+1} = \pm 1$. If $\sigma^{2n+1} = \pm 1$, the equation $\sigma^{2n+1} - \sigma^{-(2n+1)} = (2n+1)(\sigma - \sigma^{-1})$ implies that $\sigma = \sigma^{-1}$, so $\sigma = \pm 1$ and $\omega = \pm 2$. If $\omega = \pm 2$ then

$$|h_n(\omega)| = |S_n(\omega)/S_{n-1}(\omega)| = (n+1)/n > 1.$$

The proof for $n < 0$ is similar. In that case $\sigma^{2n+1} - \bar{\sigma}^{2n+1}$ and $\sigma - \bar{\sigma}$ are in opposite half-planes and $(\alpha - 1)(\alpha^{2n+1} - 1) \leq 0$. □

In the remaining case ($mn > 0$) we can use non-archimedean places instead of complex absolute values. For any root ω of Δ_n, we have $S_{2n}(\omega) = 2n+1$. It follows that

$$h_n^2(\omega) - 1 = \left(\frac{S_n(\omega)}{S_{n-1}(\omega)}\right)^2 - 1 = \frac{S_{2n}(\omega)}{S_{n-1}^2(\omega)} = \frac{2n+1}{S_{n-1}^2(\omega)}.$$

Lemma 5.5 *For any field $\mathbb{F}$ with characteristic not dividing $2n$, the polynomial S_{n-1} is separable over $\mathbb{F}$ and we have $(\Delta_n, S_{n-1}) = (1)$ in $\mathbb{F}[\omega]$.*

Proof We have $(\sigma^{n+1} - \sigma^{n-1})S_{n-1} = \sigma^{2n} - 1$, and the reduction of this polynomial to $\mathbb{F}$ is separable. It follows that S_{n-1} is separable over $\mathbb{F}$, ie $(S_{n-1}, S'_{n-1}) = (1)$. Since $\Delta_n = S'_n S_{n-1} - S_n S'_{n-1}$, we have $(\Delta_n, S_{n-1}) = (S_n S'_{n-1}, S_{n-1}) = (1)$. □

Lemma 5.6 *Let p be a prime dividing $2n+1$. Let K be a number field containing a root ω of Δ_n. Let v be a valuation on K with $v(p)=1$. Then $v(S_{n-1}(\omega))=0$.*

Proof The polynomial Δ_n is monic, so ω is an algebraic integer. Let $\mathfrak{p}$ be the prime associated with v, and $\mathbb{F}_{\mathfrak{p}}$ be its residue field. Then the characteristic p of $\mathbb{F}_{\mathfrak{p}}$ does not divide $2n$, so by Lemma 5.5 the reduction of $S_{n-1}(\omega)$ to $\mathbb{F}_{\mathfrak{p}}$ is not 0. This implies $v(S_{n-1}(\omega))=0$. □

We now address smoothness.

Proposition 5.7 *Let k and l be any odd integers with $k \neq l$. Then $C(k,l)$ is smooth over $\mathbb{Q}$.*

Proof Suppose $P=(t_0,z_0)$ is a singular point on the affine part of $C(k,l)$. Then $S_{n-1}(t_0)\neq 0$ and $S_{m-1}(z_0)\neq 0$. (If $S_{n-1}(t_0)=0$ then $S_{m-1}(z_0)=0$. Since P is a singular point, we also have $S'_{n-1}(t_0)=0$ and $S'_{m-1}(z_0)=0$. This is impossible since S_j is separable.) Then $C(k,l)$ can be given around P by $h_n(t)=h_m(z)$. The fact that P is a singular point is then equivalent to the fact that t_0 and z_0 are critical points for h_n and h_m, respectively. (We have $\Delta_n(t_0)=\Delta_m(z_0)=0$, ie $h'_n(t_0)=h'_m(z_0)=0$.)

First, consider the case when $kl<0$. The points at infinity are smooth by [8, Lemma 5.6]. The proposition follows from Lemma 5.4. That is, the values of h_k at its critical points are all different from each other, and they are also different from the values of h_l at all its critical points when $k\neq l$.

Now, assume that $kl>0$ but $k\neq l$. Assume $P(t_0,z_0)$ is a singular point over $\overline{\mathbb{Q}}$ of the standard affine part of $C(k,l)$. Let K be the number field $\mathbb{Q}(t_0,z_0)$. We have $\Delta_n(t_0)=\Delta_m(z_0)=0$ and $C(k,l)$ is given around P by $h_n(t_0)=h_m(z_0)$. It follows that $h_n^2(t_0)-1=h_m^2(z_0)-1$, ie

$$(*) \qquad \frac{2n+1}{S^2_{n-1}(t_0)}=\frac{2m+1}{S^2_{m-1}(z_0)}.$$

Let p be any prime such that $v_p(2n+1)\neq v_p(2m+1)$. By symmetry we may assume $v_p(2n+1)>v_p(2m+1)$. Let $\mathfrak{p}$ be any prime of K above p, and let v be the valuation on K associated to p, normalized so that v restricts to v_p on $\mathbb{Q}$. By Lemma 5.6, we have

$$v\left(\frac{2n+1}{S^2_{n-1}(t_0)}\right)=v(2n+1)>v(2m+1)\geq v\left(\frac{2m+1}{S^2_{m-1}(z_0)}\right).$$

This contradicts the equality $(*)$, and we conclude that no singular point P exists on the affine part. By [8, Lemma 5.6] there are no singular points at infinity. □

Proposition 5.8 *Let l be any odd integer. Then the curve $C_1(l,l)$ is smooth over $\mathbb{Q}$.*

Proof Let $F = S_n(t)S_{n-1}(z) - S_{n-1}(t)S_n(z)$ and $G = F/(z-t)$. Then $C_1(l,l)$ is defined by $G(t,z)=0$. Any singular point of $C_1(l,l)$ is also a singular point of $C(l,l)$. By [8, Lemma 5.6] we find that $C(l,l)$ is smooth at all points at infinity, so $C_1(l,l)$ is as well. Assume that $P=(t_0,z_0)$ is a singular point of the standard affine part of $C_1(l,l)$. Then P is also a singular point of $C(l,l)$. Note that $\Delta_n(t_0)=0$ and $\Delta_n(z_0)=0$, and we may rewrite $F(P)=0$ as $h_n(t_0)=h_n(z_0)$. Recall $(\omega^2-4)\Delta_n(\omega)=S_{2n}(\omega)-(2n+1)$.

Since $S_n^2(\omega) - \omega S_n(\omega)S_{n-1}(\omega) + S_{n-1}^2(\omega) = 1$ and $S_n^2(\omega) - S_{n-1}^2(\omega) = S_{2n}(\omega)$, we have

$$h_n(\omega) + h_n^{-1}(\omega) = \omega + \frac{1}{S_n(\omega)S_{n-1}(\omega)}$$

and

$$h_n(\omega) - h_n^{-1}(\omega) = \frac{S_{2n}(\omega)}{S_n(\omega)S_{n-1}(\omega)}.$$

Since $h_n(t_0)=h_n(z_0)$ and $S_{2n}(t_0)=S_{2n}(z_0)=2n+1$, we conclude that $t_0=z_0$.

Recall that $H_n = S_n'' S_{n-1} - S_n S_{n-1}''$. By l'Hôpital's rule, we have

$$\begin{aligned} -H_n(t_0) &= F_{zz}(t_0,t_0) \\ &= \lim_{z\to t_0} \frac{F_z(t_0,z)}{z-t_0} = 2\lim_{z\to t_0}\frac{F(t_0,z)}{(z-t_0)^2} = 2\lim_{z\to t_0}\frac{G(t_0,z)}{z-t_0} \\ &= 2G_z(t_0,t_0). \end{aligned}$$

The fact that $C_1(l,l)$ is singular at $P=(t_0,t_0)$ implies that $0=G_z(P)=-\frac{1}{2}H_n(t_0)$. Hence, by Lemma 5.2 we have

$$(2n-2)t_0S_{2n}(t_0) - (4n+2)S_{2n-1}(t_0) + (4n+2)t_0 = (t_0^2-4)^2H_n(t_0) = 0.$$

Since $S_{2n}(t_0)=2n+1$, we obtain $S_{2n-1}(t_0)=nt_0$. Since

$$S_{2n}^2(t_0) - t_0S_{2n}(t_0)S_{2n-1}(t_0) + S_{2n-1}^2(t_0) = 1,$$

we conclude that $t_0=\pm 2$. This is a contradiction, since $\Delta_n(\pm 2)\neq 0$ by direct calculation. We are done. □

Proposition 5.9 *The algebraic set $C(k,l)$ is smooth and has one irreducible component if $k\neq l$. The curve $C(3,3)=C(-3,-3)$ is given by $t=z$. If $k=l$ and $|l|>3$ then $C(k,l)$ has two irreducible components, $C_0(l,l)$ and $C_1(l,l)$. Both $C_0(l,l)$ and $C_1(l,l)$ are smooth.*

Proof By Lemma 5.3 it suffices to show that $C(k,l)$ is smooth. If $k \neq l$, then $C(k,l)$ is smooth by Proposition 5.7. If $k = l$ then $C_1(l,l)$ is smooth by Proposition 5.8. The proposition follows since $C_0(l,l)$ is given by $t = z$ and is smooth. □

We have shown that if $k \neq l$ then $X_{\mathrm{irr}}(k,l)$ is a single irreducible component. When $k = l$ and $|l| > 3$, we have shown that $X_{\mathrm{irr}}(k,l)$ comprises two irreducible components, and we now identify the canonical component.

Lemma 5.10 *If $k \neq l$ then $X_0(k,l)$ is birational to $C_0(k,l) \times \mathbb{C}$. The curve $C(3,3) = C(-3,-3)$ is given by $t = z$ and $X_0(3,3)$ is birational to $C(3,3) \times \mathbb{C}$. If $k = l$ and $|l| > 3$ then $X_0(l,l)$ is birational to $C_0(l,l) \times \mathbb{C}$ and there is one more irreducible component of $X_{\mathrm{irr}}(l,l)$, birational to $C_1(l,l) \times \mathbb{C}$.*

Proof By Theorem 4.9, $X_{\mathrm{irr}}(k,l)$ is birational to $C(k,l) \times \mathbb{C}$. Proposition 5.9 shows that $X_0(k,l) = X_{\mathrm{irr}}(k,l)$ when $k \neq l$.

By the definition of $C_0(l,l)$ it suffices to show that $t = z$ corresponds to the canonical component. By construction, $z = \chi_\rho(ab^{-1})$ corresponds to the loop α pictured in Figure 2. Moreover, $t = \chi_\rho(w_k)$ corresponds to the loop β pictured in the figure. When $k = l$ the symmetry induced by flipping the four-plat upside down swaps these loops. On the level of the character variety this symmetry induces the identity $t = z$. (The symmetry acts trivially on $x = \chi_\rho(a)$ and $y = \chi_\rho(b)$.) For any discrete faithful representation, $t = z$ must hold on the level of characters since the loops corresponding to z and t must have the same length (since they are swapped by the symmetry). The symmetry sends each meridian to a loop freely homotopic to itself, with the reverse orientation, and does the same for each longitude. Therefore, the symmetry induces a symmetry on any Dehn filling of the link. We conclude that $t = z$ must be satisfied by all Dehn fillings as well. By work of Thurston [14], all but finitely many Dehn fillings of one cusp of the link are on canonical components, and so are dense in $X_0(k,l)$. (See [9] and also [8, Section 2.3]) The fact that there are exactly two irreducible components in this case follows from Proposition 5.9. □

We summarize this section in the following theorem.

Theorem 1.2 *Let $k = 2m+1$ and $l = 2n+1$. The algebraic set $X_{\mathrm{irr}}(k,l)$ is birational to $C(k,l) \times \mathbb{C}$, where the curve $C(k,l) \subset \mathbb{C}^2(t,z)$ is given by*

$$C(k,l) = \{S_n(t)S_{m-1}(z) - S_{n-1}(t)S_m(z) = 0\}.$$

If $k \neq l$ then $C(k,l)$ is smooth and irreducible as considered in $\mathbb{P}^1(t) \times \mathbb{P}^1(z)$, and $X_0(k,l) = X_{\mathrm{irr}}(k,l)$ is birational to $C(k,l) \times \mathbb{C}$.

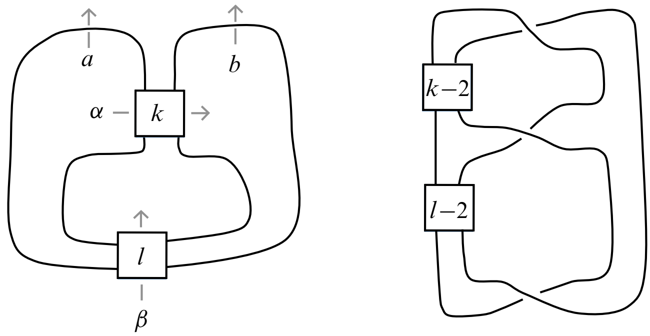


Figure 2: Meridian loops on double twist links and the four-plat presentation

The curve $C(3,3)=C(-3,-3)$ is given by $t=z$. If $k=l$ and $|l|>3$ then $C(l,l)$ is the union of exactly two components: $C_0(l,l)$, given by $t=z$, and $C_1(l,l)$, the scheme-theoretic complement of $C_0(l,l)$ in $C(l,l)$. Both are smooth and irreducible as considered in $\mathbb{P}^1(t)\times\mathbb{P}^1(z)$. The algebraic set $X_{\mathrm{irr}}(k,l)$ is given by the union $X_0(l,l)\cup X_1(l,l)$, where $X_0(l,l)$ is birational to $C_0(l,l)\times\mathbb{C}$ and $X_1(l,l)$ is birational to $C_1(l,l)\times\mathbb{C}$.

We conclude this section with a few remarks about symmetries. The proof of Lemma 5.10 relied on analysis of the symmetry which flips the four-plat upside down. For all k and l, the link complement $S^3-J(k,l)$ has a non-trivial symmetry group. In the case when $k\neq l$ this is generated by two involutions. The first is the flip about a vertical axis through the k half-twists. (In the left projection in Figure 2 this axis is the vertical axis through the middle of the diagram.) The second symmetry is the analogous symmetry through an axis through the l half-twists. (In the left projection in Figure 2 this axis is a circle through the middle of the l half-twists which goes horizontally through the k box.)

These symmetries both take the loop corresponding to a to a loop freely homotopic to the loop corresponding to b^{-1}, and fix the free homotopy class of the un-oriented loop corresponding to ab^{-1}. Since they are involutions, the effect on the character variety is that $x=\chi_\rho(a)$ is sent to $y=\chi_\rho(b)=\chi_\rho(b^{-1})$ and $z=\chi_\rho(ab^{-1})$ is fixed. By definition, $t=\chi_\rho(w_k)$ is given by

$$\begin{aligned}(xS_m(z)-yS_{m-1}(z))(yS_m(z)-xS_{m-1}(z))\\ -z(S_m^2(z)+S_{m-1}^2(z))+4S_m(z)S_{m-1}(z).\end{aligned}$$

We conclude that these symmetries fix t. Therefore, the induced action of the symmetry group on $\mathbb{C}^3[x,y,z]$ when $k\neq l$ is given by $(x,y,z)\mapsto(y,x,z)$. (This is the action of an index-two subgroup when $k=l$.)

Recall that

$$\varphi(x, y, z) = (xyz + 4 - x^2 - y^2 - z^2)\big(S_n(t)S_{m-1}(z) - S_{n-1}(t)S_m(z)\big).$$

The abelian component of the character variety is given by $(xyz + 4 - x^2 - y^2 - z^2)$ and is preserved by this action. As there are points on this component where $x \neq y$, we conclude that the action preserves this component set-wise but not point-wise. (For example, the point $(2, -2, 2)$ is sent to $(-2, 2, 2)$.) The set of irreducible representations is given by $(S_n(t)S_{m-1}(z) - S_{n-1}(t)S_m(z))$. Since t and z are fixed, this component (or in the case when $k = l$, these two components) is fixed by these symmetries. When $k \neq l$, since $x \neq y$ for infinitely many representations on this component, we conclude that the action preserves this component set-wise but not point-wise. Similarly, when $k = l$ these symmetries preserve both $z = t$ and the other component set-wise but not point-wise.

We conclude that even the non-geometric representations algebraically preserve this symmetry. However, when $k = l$ the additional symmetry fixes the un-oriented free homotopy class of loops corresponding to a and similarly for b, but takes the un-oriented loop corresponding to ab^{-1} to a loop freely homotopic to one corresponding to w_k. This is not freely homotopic to $(ab^{-1})^{\pm 1}$. It is this that induces the factoring of the defining equation, φ. In this case, when $|l| > 3$ there is a component which corresponds to necessarily non-geometric representations which do not algebraically preserve this symmetry.

6 Further invariants

We have shown in Theorem 1.2 that, when $k \neq l$, $X_0(k, l)$ is birational to $C_0(k, l) \times \mathbb{C}$, and that $C_0(k, l)$ is smooth and irreducible in $\mathbb{P}^1 \times \mathbb{P}^1$. We have also shown that $X_{\mathrm{irr}}(l, l)$ is birational to the union of $C_0(l, l) \times \mathbb{C}$ and $C_1(l, l) \times \mathbb{C}$. We now compute the genus of these curves, and the degree of irrationality of $X_0(k, l)$ and $X_1(l, l)$.

Lemma 6.1 *When $k \neq l$ the bidegree of $C(k, l)$ is $\big(\lfloor \frac{|k|}{2} \rfloor, \lfloor \frac{|l|}{2} \rfloor\big)$. The bidegree of $C_1(l, l)$ is $\big(\lfloor \frac{|l|}{2} \rfloor - 1, \lfloor \frac{|l|}{2} \rfloor - 1\big)$.*

Proof By Lemma 3.6, $S_{-1} = 0$ and the degree of S_j is j when $j > 0$ and $-j - 2$ when $j < -1$. Therefore, the bidegree of $C(k, l)$ is (a, b), where $a = n$ if $n > 0$ and $a = -n - 1$ if $n < -1$, and $b = m$ if $m > 0$ and $b = -m - 1$ if $m < -1$. This is equivalent to $a = \lfloor \frac{|k|}{2} \rfloor$ and $b = \lfloor \frac{|l|}{2} \rfloor$. The computation for $C_1(l, l)$ follows from this using the definition of $C_1(l, l)$. □

Theorem 1.3 *Let $|k|, |l| > 1$. When $k \neq l$ the genus of $C(k,l)$ is*

$$\left(\left\lfloor \tfrac{|k|}{2} \right\rfloor - 1\right)\left(\left\lfloor \tfrac{|l|}{2} \right\rfloor - 1\right).$$

The genus of $C_0(l,l)$ is zero, and for $|l| > 3$ the genus of $C_1(l,l)$ is $\left(\left\lfloor \frac{|l|}{2} \right\rfloor - 2\right)^2$.

Proof The result follows from the following, by Lemma 6.1. If C is a smooth projective curve in $\mathbb{P}^1 \times \mathbb{P}^1$ of bidegree (a,b) then the genus is $(a-1, b-1)$ (see [8]). □

Definition 6.2 Let X be an irreducible (affine or projective) complex variety of dimension n. The degree of irrationality of X, $\gamma(X)$, is the minimal degree of any rational map from X to a dense subset of $\mathbb{C}^n$. When X is a curve, this is also called the gonality of X.

The gonality, in its relation to character varieties and Dehn filling, is discussed at length in [11]. Moreover, the gonality of the components of the $\mathrm{SL}_2(\mathbb{C})$ and $\mathrm{PSL}_2(\mathbb{C})$ character varieties are computed (Theorem 9.2, Theorem 9.4). We now compute the degree of irrationality of our sets.

Theorem 1.4 *The degree of irrationality of $X_0(k,l)$ is $\min\left\{\left\lfloor \frac{|k|}{2} \right\rfloor, \left\lfloor \frac{|l|}{2} \right\rfloor\right\}$ when $k \neq l$. The degree of irrationality of $X_0(l,l)$ is 1, and the degree of irrationality of $X_1(l,l)$ is $\left\lfloor \frac{|l|}{2} \right\rfloor - 1$.*

Proof The degree of irrationality of a surface of the form $C \times \mathbb{C}$ is equal to the gonality of C; see [18, Proposition 1] and [15]. (If C is a non-singular projective curve then $C \times \mathbb{C}$ is a non-singular projective surface since the fibers have genus zero.) Following [11, Lemma 9.1] if C is a smooth irreducible curve in $\mathbb{P}^1 \times \mathbb{P}^1$ of bidegree (a,b) with $ab \neq 0$ then the gonality of C is $\min\{a,b\}$. The result now follows from Lemma 6.1. □

7 Desingularization of $X_0(3, 2m+1)$

The simplest subfamily of the hyperbolic two-component double twist links is when $k = 3$ (so $n = 1$). This family includes the Whitehead link $5^2_1 = (8/3)$ which is $J(3,3)$, and $6^2_2 = (10/3)$ which is $J(3,-3)$. In this section we first determine the singular points of the natural model of $X_0(3,l)$, where $l = 2m+1$ in Proposition 7.5. In Proposition 7.7 we determine the degenerate fibers of the map $\phi\colon S \to \mathbb{P}^1$, $(x:y:u, z:w) \mapsto (z:w)$. We then show in Theorem 1.5 that the desingularization of the natural model for $X_0(3, 2m+1)$ is a series of blowups of $\mathbb{P}^1 \times \mathbb{P}^1$.

By Theorem 1.2, $X_0(3,l) = X_{\mathrm{irr}}(3,l)$ is birational to $C(3,l) \times \mathbb{C}$, where $C(3,l)$ is given by $tS_{m-1}(z) = S_m(z)$ in $\mathbb{C}^2(t,z)$. Since this defining polynomial is linear in t we conclude that $C(3,l)$ is itself birational to $\mathbb{C}$ and $X_0(3,l)$ is indeed birational to $\mathbb{C}^2$. The Whitehead link, $J(-3,-3) = J(3,3)$ is a degenerate case of the $J(3,l)$ links, where $X_0(3,3) = X_{\mathrm{irr}}(3,3)$ is given by $t = z$ up to birational equivalence.

We begin by homogenizing the defining polynomial for $X_0(3,l)$, where $l = 2m+1$. Recall that

$$t = (xS_m(z) - yS_{m-1}(z))(yS_m(z) - xS_{m-1}(z)) \\ - z(S_m^2(z) + S_{m-1}^2(z)) + 4S_m(z)S_{m-1}(z).$$

Since $S_m^2(z) + S_{m-1}^2(z) - zS_m(z)S_{m-1}(z) = 1$, this simplifies to

$$t = xy - z + (xyz + 4 - x^2 - y^2 - z^2)S_m(z)S_{m-1}(z).$$

The defining polynomial for the natural model of $X_0(3,l)$ is $tS_{m-1}(z) - S_m(z)$ in $\mathbb{C}[x,y,z]$. We now homogenize it.

Definition 7.1 Let $T_j = T_j(z,w) = w^j S_j(z/w)$.

The following is a direct consequence of the Chebyshev identity

$$S_j^2(\omega) + S_{j-1}^2(\omega) - \omega S_j(\omega)S_{j-1}(\omega) = 1.$$

Lemma 7.2 *We have*

$$T_j^2 + w^2 T_{j-1}^2 - z\, T_j T_{j-1} = w^{2j}.$$

It is now elementary to determine the homogenous defining polynomial.

Lemma 7.3 *The homogenization of the defining polynomial $tS_{m-1}(z) - S_m(z)$ in $\mathbb{P}^2 \times \mathbb{P}^1 = \{(x:y:u, z:w)\}$ is*

$$F = \big[(xyw - u^2z)w^{2m} + (xyzw + 4u^2w^2 - x^2w^2 - y^2w^2 - u^2z^2)T_mT_{m-1}\big]T_{m-1} \\ - u^2w^{2m}T_m.$$

We now determine the singular points in the projective closure of our natural model in $\mathbb{P}^2 \times \mathbb{P}^1$. To find singular points, we consider solutions $(x:y:u, z:w)$ of $F = F_x = F_y = F_u = F_z = F_w = 0$.

First, we compute these partial derivatives; the results are elementary to verify by direct calculations.

Lemma 7.4 *The first-order partials of F as in Lemma 7.3 are given by the following:*

$$\begin{aligned}
F_x &= (yw^{2m} + (yz - 2xw)T_m T_{m-1})w\, T_{m-1},\\
F_y &= (xw^{2m} + (xz - 2yw)T_m T_{m-1})w\, T_{m-1},\\
F_u &= -2u\big[(zw^{2m} + (z^2 - 4w^2)T_m T_{m-1})T_{m-1} + w^{2m} T_m\big],\\
F_z &= \big[-u^2w^{2m} + (xyw - 2u^2z)T_m T_{m-1}\\
&\qquad + (xyzw + 4u^2w^2 - x^2w^2 - y^2w^2 - u^2z^2)(T_m T_{m-1})_z\big]T_{m-1}\\
&\qquad + \big[(xyw - u^2z)w^{2m} + (xyzw + 4u^2w^2 - x^2w^2 - y^2w^2 - u^2z^2)T_m T_{m-1}\big]\\
&\qquad\quad \times (T_{m-1})_z - u^2w^{2m}(T_m)_z,\\
F_w &= \big[(2m+1)xyw^{2m} - 2mu^2zw^{2m-1} + (xyz + 8u^2w - 2x^2w - 2y^2w)T_m T_{m-1}\\
&\qquad + (xyzw + 4u^2w^2 - x^2w^2 - y^2w^2 - u^2z^2)(T_m T_{m-1})_w\big]T_{m-1}\\
&\qquad + \big[(xyw - u^2z)w^{2m} + (xyzw + 4u^2w^2 - x^2w^2 - y^2w^2 - u^2z^2)T_m T_{m-1}\big]\\
&\qquad\quad \times (T_{m-1})_w - u^2(2mw^{2m-1}T_m + w^{2m}(T_m)_w).
\end{aligned}$$

We can now determine the singular points.

Proposition 7.5 *The singular points $(x : y : u, z : w) \in \mathbb{P}^2 \times \mathbb{P}^1$ of F are*

- $(1:0:0, 1:0)$, $(0:1:0, 1:0)$,
- $(1:0:0, z:1)$, $(0:1:0, z:1)$, *where z is a root of $S_{m-1}(z)$,*
- $(1:1:0, z:1)$, *where z is a root of $S_m(z) - S_{m-1}(z)$,*
- $(1:-1:0, z:1)$, *where z is a root of $S_m(z) + S_{m-1}(z)$.*

The number of singularities is $4m$ if $m \geq 1$, and is $-(2+4m)$ if $m \leq -2$.

Proof We break the analysis down into cases.

First, we consider the case when $(w : z) = (0 : 1)$. We have $F_x = F_y = 0$, $F = -u^2$ and $F_u = -2u$. Hence $u = 0$. Now we have $F_z = 0$ and $F_w = xy$. Thus $xy = 0$. In this case, there are two singular points, $(1:0:0, 1:0)$ and $(0:1:0, 1:0)$.

Next, we consider the case when $w = 1$. First we assume that $S_{m-1}(z) = 0$. Then

$$F_x = F_y = 0, \quad F = -u^2 S_m(z), \quad F_u = -2u S_m(z).$$

Since $S_m(z) \neq 0$, we have $u = 0$. Then $F_z = xyS'_{m-1}(z)$ and $F_w = xy(T_{m-1})_w$. Since $S'_{m-1}(z) \neq 0$, we must have $xy = 0$. In this case, the singular points are $(1:0:0, z:1)$, $(0:1:0, z:1)$, where z is a root of $S_{m-1}(z)$.

Finally, we assume that $w = 1$ and $S_{m-1}(z) \neq 0$. We have

$$F_x = y + (yz - 2x)S_m(z)S_{m-1}(z) = y(S_m^2(z) + S_{m-1}^2(z)) - 2xS_m(z)S_{m-1}(z),$$
$$F_y = x + (xz - 2y)S_m(z)S_{m-1}(z) = x(S_m^2(z) + S_{m-1}^2(z)) - 2yS_m(z)S_{m-1}(z).$$

If x and y are not simultaneously equal to 0, we must have $S_m^2(z) - S_{m-1}^2(z) = 0$.

We first consider the subcase when $x = y = 0$, so $(x : y : u) = (0 : 0 : 1)$. Then, by Lemma 3.8,

$$F = S_{m-1}(z)\bigl(-z + (4 - z^2)S_{m-1}(z)S_m(z)\bigr) - S_m(z) = -S_{3m}(z).$$

Since $S_{3m}(z)$ is separable in $\mathbb{C}[z]$, there are no singular points in this case.

Therefore, we may assume that $xy \neq 0$ and $S_m^2(z) - S_{m-1}^2(z) = 0$. We consider the cases that $S_m(z) - S_{m-1}(z) = 0$ and $S_m(z) + S_{m-1}(z) = 0$ separately.

First assume that $S_m(z) - S_{m-1}(z) = 0$. Then $F_x = F_y = 0$ is equivalent to $x = y$. Since $S_m^2(z) = 1/(2 - z)$, we have $F = u^2 S_m(z)$ and $F_u = 2uS_m(z)$. Hence $u = 0$. Now we have

$$F_z = \bigl[S_m(z)S_{m-1}(z) + (z - 2)(S_m(z)S_{m-1}(z))'\bigr]x^2 S_{m-1}(z).$$

From $S_m^2(z) + S_{m-1}^2(z) - zS_m(z)S_{m-1}(z) = 1$ and $S_m(z) = S_{m-1}(z)$ we get

$$(z - 2)(S_m'(z) + S_{m-1}'(z)) = -S_m(z).$$

It follows that $F_z = 0$. We have

$$F_w = \bigl[(2m + 1) + (z - 4)S_m(z)S_{m-1}(z) + (z - 2)(T_m T_{m-1})_w\bigr]x^2 S_{m-1}(z).$$

From $T_m^2 + w^2 T_{m-1}^2 - z\, T_m T_{m-1} = w^{2m}$ (by Lemma 7.2) and $S_m(z) = S_{m-1}(z)$ we get

$$(2 - z)\bigl((T_m)_w + (T_{m-1})_w\bigr)S_m(z) + 2S_m^2(z) = 2m.$$

It follows that

$$(2m + 1) + (z - 4)S_m(z)S_{m-1}(z) + (z - 2)(T_m T_{m-1})_w = 1 + (z - 2)S_m^2(z) = 0.$$

Hence $F_w = 0$. The corresponding singular points are $(1 : 1 : 0, z : 1)$, where z is a root of $S_m(z) - S_{m-1}(z)$.

Finally, assume that $xy = 0$ and $S_m(z) + S_{m-1}(z) = 0$. Similar to the above, the singular points are $(1 : 1 : 0, z : 1)$, where z is a root of $S_m(z) + S_{m-1}(z)$. □

Definition 7.6 Let $S = \mathcal{Z}(F) \subset \mathbb{P}^2 \times \mathbb{P}^1$ be the vanishing set of F and $\widetilde{S}$ be the desingularization of S.

Now we determine the degenerate fibers; we determine all $(z:w) \in \mathbb{P}^1$ such that $F = F_x = F_y = F_u = 0$ has at least one solution $(x:y:u) \in \mathbb{P}^2$.

Proposition 7.7 *The degenerate fibers of $\phi\colon S \to \mathbb{P}^1$, $(x:y:u, z:w) \mapsto (z:w)$, are*

- $\phi^{-1}(1:0) = \{(x:y:u) \in \mathbb{P}^2 \mid u^2 = 0\}$,
- $\phi^{-1}(z:1) = \{(x:y:u) \in \mathbb{P}^2 \mid u^2 = 0\}$, *where z is a root of $S_{m-1}(z)$,*
- $\phi^{-1}(z:1) = \{(x:y:u) \in \mathbb{P}^2 \mid (xS_m(z) - yS_{m-1}(z))(yS_m(z) - xS_{m-1}(z)) = 0\}$, *where z is a root of $S_{3m}(z)$,*
- $\phi^{-1}(z:1) = \{(x:y:u) \in \mathbb{P}^2 \mid (x-y)^2 - (2-z)u^2 = 0\}$, *where z is a root of $S_m(z) - S_{m-1}(z)$,*
- $\phi^{-1}(z:1) = \{(x:y:u) \in \mathbb{P}^2 \mid (x+y)^2 - (2+z)u^2 = 0\}$, *where z is a root of $S_m(z) + S_{m-1}(z)$.*

Proof We break the analysis down into cases.

First, we consider the case when $(z:w) = (0:1)$. We have $F_x = F_y = 0$, $F = -u^2$ and $F_u = -2u$. Hence $u = 0$. Note that $\phi^{-1}(1:0) = \{(x:y:u) \in \mathbb{P}^2 \mid u^2 = 0\}$.

Next, we consider the case when $w = 1$. First we assume that $S_{m-1}(z) = 0$. Then

$$F_x = F_y = 0, \quad F = -u^2 S_m(z), \quad F_u = -2uS_m(z).$$

Hence $u = 0$. In this case $\phi^{-1}(z:1) = \{(x:y:u) \in \mathbb{P}^2 \mid u^2 = 0\}$.

Finally, we assume that $w = 1$ and $S_{m-1}(z) \neq 0$. Note that if x and y are not simultaneously equal to 0, we must have $S_m^2(z) - S_{m-1}^2(z) = 0$.

We first consider the subcase when $x = y = 0$, so $(x:y:u) = (0:0:1)$. Then

$$F_x = F_y = 0, \quad F = -S_{3m}(z), \quad F_u = -2S_{3m}(z).$$

Hence $S_{3m}(z) = 0$. In this case

$$\phi^{-1}(z:1) = \big\{(x:y:u) \in \mathbb{P}^2 \mid (xS_m(z) - yS_{m-1}(z))(yS_m(z) - xS_{m-1}(z)) = 0\big\}.$$

As a result we may assume that $xy \neq 0$. Therefore

$$S_m(z) - S_{m-1}(z) = 0 \quad \text{or} \quad S_m(z) + S_{m-1}(z) = 0.$$

If $S_m(z) - S_{m-1}(z) = 0$ then $F = F_x = F_y = F_u = 0$ is equivalent to $x = y$ and $u = 0$. In this case

$$\phi^{-1}(z:1) = \{(x:y:u) \in \mathbb{P}^2 \mid (x-y)^2 - (2-z)u^2 = 0\}.$$

If $S_m(z)+S_{m-1}(z)=0$ then $F=F_x=F_y=F_u=0$ is equivalent to $x=-y$ and $u=0$. In this case

$$\phi^{-1}(z:1)=\{(x:y:u)\in\mathbb{P}^2 \mid (x+y)^2-(2+z)u^2=0\}. \qquad \square$$

Next, we consider desingularization. Since S is birational to $\mathbb{P}^1\times\mathbb{P}^1$, we can blow down $\widetilde{S}$ over $\mathbb{P}^1$ some number of times so that it becomes a fiber bundle $\mathbb{P}^1\times\mathbb{P}^1$ over $\mathbb{P}^1$.

Definition 7.8 In the following, let χ denote the Euler characteristic of a surface. Let S_{sing} be the set of singular points of S and $N_{\mathrm{sing}}=|S_{\mathrm{sing}}|$. Furthermore, let N be such that $\widetilde{S}$ is obtained from $\mathbb{P}^1\times\mathbb{P}^1$ by N one-point blow-ups.

We have

$$\chi(\widetilde{S})=\chi(S-S_{\mathrm{sing}})+N_{\mathrm{sing}}\,\chi(\mathbb{P}^1)=\chi(S)+N_{\mathrm{sing}}$$

(see [5, Lemma 2.2]).

By definition, $\widetilde{S}$ is obtained from $\mathbb{P}^1\times\mathbb{P}^1$ by N one-point blow-ups. Then since $\chi(\mathbb{P}^1\times\mathbb{P}^1)=4$, using $\mathbb{P}^1\times\mathbb{P}^1$ in place of S in the above, we have

$$\chi(\widetilde{S})=\chi(\mathbb{P}^1\times\mathbb{P}^1)+N=4+N.$$

It follows that $N=\chi(S)+N_{\mathrm{sing}}-4$. We summarize this as a lemma.

Lemma 7.9 *We have $N=\chi(S)+N_{\mathrm{sing}}-4$.*

Proposition 7.10 *The Euler characteristic of S is* $\chi(S)=\begin{cases}4+5m & \text{if } m\geq 1,\\ -5m & \text{if } m\leq -2.\end{cases}$

Proof Let $\varphi\colon S\hookrightarrow\mathbb{P}^2\times\mathbb{P}^1 \dashrightarrow \mathbb{P}^1\times\mathbb{P}^1$ be the rational map defined by

$$(x:y:u,z:w)\mapsto(x:y,z:w).$$

Let P be the set of points $(0:0:1,z:1)$ where z is a root of $S_{3m}(z)$. The map φ is not defined at points in P. Let $U:=S-P$. We now determine $\varphi(U)$.

Write $F=G+u^2H$, where

$$\begin{aligned}G&=\big(xyw^{2m+1}+(xyzw-x^2w^2-y^2w^2)T_mT_{m-1}\big)T_{m-1},\\ H&=\big(-zw^{2m}+(4w^2-z^2)T_mT_{m-1}\big)T_{m-1}-w^{2m}T_m.\end{aligned}$$

Note that $\varphi(U)$ is the collection of all points $(x:y, z:w) \in \mathbb{P}^1 \times \mathbb{P}^1$ except those for which $F(x:y,z:w) \in \mathbb{C}[u]$ is a nonzero constant. The polynomial $F(x:y,z:w) \in \mathbb{C}[u]$ is a nonzero constant whenever $H = 0$ and $G \neq 0$, which is equivalent to

$$w = 1, \quad S_{3m}(z) = 0 \quad \text{and} \quad (xS_m(z) - yS_{m-1}(z))(yS_m(z) - xS_{m-1}(z)) \neq 0.$$

Hence $\varphi(U) = \mathbb{P}^1 \times \mathbb{P}^1 - Q$, where Q is the set of points $(x:y, z:1) \in \mathbb{P}^1 \times \mathbb{P}^1$ satisfying $S_{3m}(z) = 0$ and

$$(xS_m(z) - yS_{m-1}(z))(yS_m(z) - xS_{m-1}(z)) \neq 0.$$

Note that $\chi(Q) = 0$.

Let L be the set of points $(x:y, z:1) \in \mathbb{P}^1 \times \mathbb{P}^1$ satisfying $S_{3m}(z) = 0$ and

$$(xS_m(z) - yS_{m-1}(z))(yS_m(z) - xS_{m-1}(z)) = 0.$$

Note that $\{G = H = 0\} \subset \mathbb{P}^1 \times \mathbb{P}^1$ is equal to L. Hence

$$\chi(L) = \chi(\varphi^{-1}(L)) = \begin{cases} 6m & \text{if } m \geq 1, \\ -(6m+4) & \text{if } m \leq -2. \end{cases}$$

Recall that

$$G = \big(xyw^{2m} + (xyz - x^2w - y^2w)T_m T_{m-1}\big)wT_{m-1}.$$

Since $T_m^2 + w^2T_{m-1}^2 - zT_mT_{m-1} = w^{2m}$, we have

$$G = (xT_m - ywT_{m-1})(yT_m - xwT_{m-1})wT_{m-1}.$$

Let $B := \mathcal{Z}(G)$ be the zero set of G in $\mathbb{P}^1 \times \mathbb{P}^1$. Then $B = B_1 \cup B_2 \cup B_3$, where

$$\begin{aligned} B_1 &= \mathcal{Z}(w) = \mathbb{P}^1 \times \{(1:0)\}, \\ B_2 &= \mathcal{Z}(T_{m-1}) = \mathbb{P}^1 \times \{(z:1) \mid S_{m-1}(z) = 0\}, \\ B_3 &= \mathcal{Z}(xT_m - ywT_{m-1}) \cup \mathcal{Z}(yT_m - xwT_{m-1}) \end{aligned}$$

are subsets in $\mathbb{P}^1 \times \mathbb{P}^1$.

We have $B_3 = B_{31} \cup B_{32}$, where

$$B_{31} = \mathcal{Z}(xT_m - ywT_{m-1}) \quad \text{and} \quad B_{32} = \mathcal{Z}(yT_m - xwT_{m-1}).$$

Note that $(x:y, z:w) \in B_{31} \cap B_{32}$ if and only if $x = y$ and $T_m = wT_{m-1}$, or $x = -y$ and $T_m = -wT_{m-1}$. Hence

$$B_{31} \cap B_{32} = \{(1:1, z:1) \mid S_m(z) - S_{m-1}(z) = 0\} \\ \cup \{(1:-1, z:1) \mid S_m(z) + S_{m-1}(z) = 0\}.$$

It follows that

$$\chi(B_{31} \cap B_{32}) = \begin{cases} 2m & \text{if } m \geq 1, \\ -(2m+2) & \text{if } m \leq -2. \end{cases}$$

Then

$$\chi(B_3) = \chi(B_{31}) + \chi(B_{32}) - \chi(B_{31} \cap B_{32}) = \begin{cases} 4-2m & \text{if } m \geq 1, \\ 6+2m & \text{if } m \leq -2. \end{cases}$$

We have $B_1 \cap B_2 = \varnothing$, $B_1 \cap B_3 = \{(1:0,1:0),(0:1,1:0)\}$, and

$$B_2 \cap B_3 = \{(1:0,z:1),(0:1,z:1) \mid S_{m-1}(z) = 0\}.$$

Hence

$$\begin{aligned} \chi(B) &= \chi(B_1) + \chi(B_2) + \chi(B_3) - \chi(B_1 \cap B_2) - \chi(B_1 \cap B_3) \\ &\qquad\qquad - \chi(B_2 \cap B_3) + \chi(B_1 \cap B_2 \cap B_3) \\ &= \begin{cases} 2 + (2m-2) + (4-2m) - 0 - 2 - (2m-2) + 0 = 4-2m & \text{if } m \geq 1, \\ 2 - (2m+2) + (6+2m) - 0 - 2 + (2m+2) + 0 = 6+2m & \text{if } m \leq -2. \end{cases} \end{aligned}$$

It follows that

$$\begin{aligned} \chi(U) &= 2\chi(\mathbb{P}^1 \times \mathbb{P}^1 - (B \sqcup Q)) + \chi(B - L) + \chi(\varphi^{-1}(L)) \\ &= 2\chi(\mathbb{P}^1 \times \mathbb{P}^1) - \chi(B) - 2\chi(Q) - \chi(L) + \chi(\varphi^{-1}(L)) \\ &= \begin{cases} 4+2m & \text{if } m \geq 1, \\ 2-2m & \text{if } m \leq -2. \end{cases} \end{aligned}$$

Then

$$\chi(S) = \chi(U) + \chi(P) = \begin{cases} (4+2m) + 3m = 4+5m & \text{if } m \geq 1, \\ (2-2m) - (3m+2) = -5m & \text{if } m \leq -2. \end{cases} \qquad \square$$

Proposition 7.10 and Proposition 7.5 along with the fact that

$$N = \chi(S) + N_{\text{sing}} - 4$$

give

$$N = \chi(S) + N_{\text{sing}} - 4 = \begin{cases} (4+5m) + 4m - 4 = 9m & \text{if } m \geq 1, \\ (-5m) + (-(2+4m)) - 4 = -(6+9m) & \text{if } m \leq -2. \end{cases}$$

This calculation completes the proof of Theorem 1.5.

References

[1] **K L Baker**, **K L Petersen**, *Character varieties of once-punctured torus bundles with tunnel number one*, Internat. J. Math. 24 (2013) MR3078072

[2] **S Boyer**, **E Luft**, **X Zhang**, *On the algebraic components of the* $\mathrm{SL}(2,\mathbb{C})$ *character varieties of knot exteriors*, Topology 41 (2002) 667–694 MR1905834

[3] **M Culler**, **P B Shalen**, *Varieties of group representations and splittings of* 3*–manifolds*, Ann. of Math. 117 (1983) 109–146 MR683804

[4] **R Fricke**, *Ueber die Theorie der automorphen Modulgruppen*, Nachr. Ges. Wiss. Göttingen, Math.-Phys. Kl. 1896 (1896) 91–101

[5] **S Harada**, *Canonical components of character varieties of arithmetic two bridge link complements*, preprint (2012) `arXiv:1112.3441`

[6] **E Landes**, *Identifying the canonical component for the Whitehead link*, Math. Res. Lett. 18 (2011) 715–731 MR2831837

[7] **E Landes**, *On the canonical components of character varieties of hyperbolic* 2*–bridge link complements*, PhD thesis, University of Texas at Austin (2011) Available at `http://hdl.handle.net/2152/ETD-UT-2011-08-2877`

[8] **M L Macasieb**, **K L Petersen**, **R M van Luijk**, *On character varieties of two-bridge knot groups*, Proc. Lond. Math. Soc. 103 (2011) 473–507 MR2827003

[9] **T Ohtsuki**, *Ideal points and incompressible surfaces in two-bridge knot complements*, J. Math. Soc. Japan 46 (1994) 51–87 MR1248091

[10] **T Ohtsuki**, **R Riley**, **M Sakuma**, *Epimorphisms between* 2*–bridge link groups*, from: “The Zieschang Gedenkschrift”, (M Boileau, M Scharlemann, R Weidmann, editors), Geom. Topol. Monogr. 14 (2008) 417–450 MR2484712

[11] **K L Petersen**, **A W Reid**, *Gonality and genus of canonical components of character varieties*, preprint (2014) `arXiv:1408.3665`

[12] **K Qazaqzeh**, *The character variety of a family of one-relator groups*, Internat. J. Math. 23 (2012) MR2888942

[13] **P B Shalen**, *Representations of* 3*–manifold groups*, from: “Handbook of geometric topology”, (R J Daverman, R B Sher, editors), North-Holland, Amsterdam (2002) 955–1044 MR1886685

[14] **W P Thurston**, *The geometry and topology of three-manifolds*, lecture notes, Princeton University (1979) Available at `http://msri.org/publications/books/gt3m`

[15] **H-o Tokunaga**, **H Yoshihara**, *Degree of irrationality of abelian surfaces*, J. Algebra 174 (1995) 1111–1121 MR1337188

[16] **A T Tran**, *The universal character ring of the* $(-2, 2m+1, 2n)$*–pretzel link*, Internat. J. Math. 24 (2013) MR3103879

[17] **H Vogt**, *Sur les invariants fondamentaux des équations différentielles linéaires du second ordre*, Ann. Sci. École Norm. Sup. 6 (1889) 3–71 MR1508833

[18] **H Yoshihara**, *Degree of irrationality of a product of two elliptic curves*, Proc. Amer. Math. Soc. 124 (1996) 1371–1375 MR1327053

Department of Mathematics, Florida State University
208 Love Building, 1017 Academic Way, Tallahassee, FL 32306-4510, USA

Department of Mathematical Sciences, The University of Texas at Dallas
800 W Campbell Rd FO 35, Richardson, TX 75080, USA

petersen@math.fsu.edu, att140830@utdallas.edu

Received: 13 November 2014 Revised: 16 April 2015

Algebraic & Geometric Topology 15 (2015) 3599–3612

The L^2–Alexander torsion is symmetric

JÉRÔME DUBOIS
STEFAN FRIEDL
WOLFGANG LÜCK

We show that the L^2–Alexander torsion of a 3–manifold is a symmetric function. This can be viewed as a generalization of the symmetry of the Alexander polynomial of a knot.

57M27; 57Q10

1 Introduction

An *admissible triple* (N, ϕ, γ) consists of an irreducible, orientable, compact 3–manifold $N \neq S^1 \times D^2$ with empty or toroidal boundary, a class $\phi \neq 0 \in H^1(N;\mathbb{Z}) = \mathrm{Hom}(\pi_1(N), \mathbb{Z})$ and a homomorphism $\gamma\colon \pi_1(N) \to G$ such that ϕ factors through γ. In [4; 5] we used the L^2–torsion (see for example Lück [14]) to associate to an admissible triple (N, ϕ, γ) the L^2–Alexander torsion $\tau^{(2)}(N, \phi, \gamma)$ which is a function

$$\tau^{(2)}(N, \phi, \gamma)\colon \mathbb{R}_{>0} \to \mathbb{R}_{\geq 0}$$

that is well defined up to multiplication by a function of the type $t \mapsto t^m$ for some $m \in \mathbb{Z}$. We recall the definition in Section 6.

The goal of this paper is to show that the L^2–Alexander torsion is symmetric. In order to state the precise symmetry result we need to recall that given a 3–manifold N the *Thurston norm* [16] of some $\phi \in H^1(N;\mathbb{Z}) = \mathrm{Hom}(\pi_1(N), \mathbb{Z})$ is defined as

$$x_N(\phi) := \min\{\chi_-(S) \mid S \subset N \text{ properly embedded surface dual to } \phi\}.$$

Here, given a surface S we define its complexity as $\chi_-(S) := -\chi(S')$, where S' is the result of deleting all components from S that are disks or spheres. Thurston [16] showed that x_N is a (possibly degenerate) norm on $H^1(N;\mathbb{Z})$. Now we can formulate the main result of this paper.

Published: 12 January 2016 DOI: 10.2140/agt.2015.15.3599

Theorem 1.1 *Let (N,ϕ,γ) be an admissible triple. Then for any representative τ of $\tau^{(2)}(N,\phi,\gamma)$ there exists an $n\in\mathbb{Z}$ with $n\equiv x_N(\phi)$ mod 2 such that*

$$\tau(t^{-1})=t^n\cdot\tau(t)\ \text{for any}\ t\in\mathbb{R}_{>0}.$$

It is worth looking at the case that $N=S^3\setminus\nu K$ is the complement of a tubular neighborhood νK of an oriented knot $K\subset S^3$. We denote by $\phi_K\colon\pi_1(N)\to\mathbb{Z}$ the epimorphism sending the oriented meridian to 1. Let $\gamma\colon\pi_1(N)\to G$ be a homomorphism such that ϕ_K factors through γ. We define

$$\tau^{(2)}(K,\gamma):=\tau^{(2)}(S^3\setminus\nu K,\phi_K,\gamma).$$

If we take $\gamma=\mathrm{id}$ to be the identity, then we showed in [4] that

$$\tau^{(2)}(K,\mathrm{id})=\Delta_K^{(2)}(t)\cdot\max\{1,t\},$$

where $\Delta_K^{(2)}(t)\colon\mathbb{R}_{>0}\to\mathbb{R}_{\geq 0}$ denotes the L^2–Alexander invariant of Li and Zhang [12; 13], which was also studied by Dubois and Wegner [6; 7] and Aribi [1; 2].

If we take $\gamma=\phi_K$, then we showed in [4] that the L^2–Alexander torsion $\tau^{(2)}(K,\phi_K)$ is fully determined by the Alexander polynomial $\Delta_K(t)$ of K and that in turn $\tau^{(2)}(K,\phi_K)$ almost determines the Alexander polynomial $\Delta_K(t)$. In this sense the L^2–Alexander torsion can be viewed as a "twisted" version of the Alexander polynomial, and at least morally it is related to the twisted Alexander polynomial of Wada [20] and to the higher-order Alexander polynomials of Cochran [3] and Harvey [10]. We refer to [5] for more on the relationship and similarities between the various twisted invariants.

If K is a knot, then any Seifert surface is dual to ϕ_K and it immediately follows that $x(\phi_K)\leq\max\{2\cdot\mathrm{genus(K)}-1,0\}$. In fact an elementary argument shows that for any *non-trivial* knot we have the equality $x(\phi_K)=2\cdot\mathrm{genus(K)}-1$. In particular the Thurston norm of ϕ_K is odd. We thus obtain the following corollary to Theorem 1.1.

Theorem 1.2 *Let $K\subset S^3$ be an oriented non-trivial knot and let $\gamma\colon\pi_1(N)\to G$ be a homomorphism such that ϕ_K factors through γ. Then there exists an odd n with*

$$\tau^{(2)}(K,\gamma)(t^{-1})=t^n\cdot\tau^{(2)}(K,\gamma)(t)\quad\text{for any}\ t\in\mathbb{R}_{>0}.$$

The proof of Theorem 1.1 has many similarities with the proof of the main theorem in Friedl, Kim and Kitayama [9], which in turn builds on the ideas of Turaev [17; 18; 19]. In an attempt to keep the proof as short as possible we will on several occasions refer to [9] and [17] for definitions and results.

Conventions All manifolds are assumed to be connected, orientable and compact. All CW–complexes are assumed to be finite and connected. If G is a group then we equip $\mathbb{C}[G]$ with the involution given by complex conjugation and by $\bar{g} := g^{-1}$ for $g \in G$. We extend this involution to matrices over $\mathbb{C}[G]$ by applying the involution to each entry. Given a ring R we will view all modules as left R–modules, unless we say explicitly otherwise. Furthermore, given a matrix $A \in M_{m,n}(R)$ we denote by $A\colon R^m \to R^n$ the R–homomorphism of left R–modules obtained by right multiplication with A and thinking of elements in R^m as the only row in a $(1,m)$–matrix.

Acknowledgments The second author gratefully acknowledges the support provided by the SFB 1085 Higher Invariants at the University of Regensburg, funded by the Deutsche Forschungsgemeinschaft (DFG). The paper is also financially supported by the Leibniz Prize of the third author granted by the DFG. We are also grateful to the referees for carefully reading an earlier version of this paper.

2 Euler structures

In this section we recall the notion of an Euler structure of a pair of CW–complexes and manifolds which is due to Turaev. We refer to [17; 18; 9] for full details. Throughout this paper, given a space X, we denote by $\mathcal{H}_1(X)$ the first integral homology group viewed as a multiplicative group.

2A Euler structures on CW–complexes

Let X be a CW–complex of dimension m and let Y be a proper subcomplex. We denote by $p\colon \widetilde{X} \to X$ the universal covering of X and we write $\widetilde{Y} := p^{-1}(Y)$. An *Euler lift* is a set of cells in $\widetilde{X}$ such that each cell of $X \setminus Y$ is covered by precisely one of the cells in the Euler lift.

Using the canonical left action of $\pi = \pi_1(X)$ on $\widetilde{X}$ we obtain a free and transitive action of π on the set of cells of $\widetilde{X} \setminus \widetilde{Y}$ lying over a fixed cell in $X \setminus Y$. If c and c' are two Euler lifts, then we can order the cells such that $c = \{c_{ij}\}$ and $c' = \{c'_{ij}\}$ and such that for each i and j the cells c_{ij} and c'_{ij} lie over the same i–cell in $X \setminus Y$. In particular there exist unique $g_{ij} \in \pi$ such that $c'_{ij} = g_{ij} \cdot c_{ij}$. We denote the projection map $\pi \to \mathcal{H}_1(X)$ by Ψ. We define

$$c'/c := \prod_{i=0}^{m} \prod_{j} \Psi(g_{ij})^{(-1)^i} \in \mathcal{H}_1(X).$$

We say that c and c' are *equivalent* if $c'/c \in \mathcal{H}_1(X)$ is trivial. An equivalence class of Euler lifts will be referred to as an *Euler structure*. We denote by $\mathrm{Eul}(X,Y)$ the set of Euler structures. If $Y = \varnothing$ then we will also write $\mathrm{Eul}(X) = \mathrm{Eul}(X,Y)$.

Given $g \in \mathcal{H}_1(X)$ and $e \in \mathrm{Eul}(X,Y)$ we define $g \cdot e \in \mathrm{Eul}(X,Y)$ as follows: pick representatives c for e and $\widetilde{g} \in \pi_1(X)$ for g, then act on one i–cell of c by $g^{(-1)^i}$. The resulting Euler lift represents an element in $\mathrm{Eul}(X,Y)$ which is independent of the choice of the cell. We denote by $g \cdot e$ the Euler structure represented by this new Euler lift. This defines a free and transitive $\mathcal{H}_1(X)$–action on $\mathrm{Eul}(X,Y)$, with $(g \cdot e)/e = g$.

If (X',Y') is a cellular subdivision of (X,Y), then there exists a canonical $\mathcal{H}_1(X)$–equivariant bijection $\sigma\colon \mathrm{Eul}(X,Y) \to \mathrm{Eul}(X',Y')$ which is defined as follows. Let $e \in \mathrm{Eul}(X,Y)$ and pick an Euler lift for (X,Y) which represents e. There exists a unique Euler lift for (X',Y') such that the cells in the Euler lift of (X',Y') are contained in the cells of the Euler lift of (X,Y). We denote by $\sigma(e)$ the Euler structure represented by this Euler lift. This map equals the map of Turaev [17, Section 1.2].

2B Euler structures of smooth manifolds

Let N be a manifold and let $\partial_0 N \subset \partial N$ be a union of components of ∂N such that $\chi(N, \partial_0 N) = 0$. A *triangulation* of N is a pair (X,t) where X is a simplicial complex and $t\colon |X| \to N$ is a homeomorphism. Throughout this section we write $Y := t^{-1}(\partial_0 N)$. For the most part we will suppress t from the notation. Following [17, Section I.4.1] we consider the projective system of sets $\{\mathrm{Eul}(X,Y)\}_{(X,t)}$, where (X,t) runs over all C^1–triangulations of N and where the maps are the $\mathcal{H}_1(N)$–equivariant bijections between these sets induced either by C^1–subdivisions or by smooth isotopies in N. We define $\mathrm{Eul}(N, \partial_0 N)$ by identifying the sets $\{\mathrm{Eul}(X,Y)\}_{(X,t)}$ via these bijections. We refer to $\mathrm{Eul}(N, \partial_0 N)$ as the set of *Euler structures on* $(N, \partial_0 N)$. For a C^1–triangulation X of N we get a canonical $\mathcal{H}_1(N)$–equivariant bijection $\mathrm{Eul}(X,Y) \to \mathrm{Eul}(N, \partial_0 N)$.

3 The L^2–torsion of a manifold

3A The L^2–torsion of a chain complex

First we recall some key properties of the Fuglede–Kadison determinant and the definition of the L^2–torsion of a chain complex of free based left $\mathbb{C}[G]$–modules. Throughout the section we refer to [14] and to [4] for details and proofs.

We fix a group G. Let A be a $k \times l$–matrix over $\mathbb{C}[G]$. There exists the notion of A being of *determinant class*. (To be more precise, we view the $k \times l$–matrix A as a map

$\mathcal{N}(G)^l \to \mathcal{N}(G)^k$, where $\mathcal{N}(G)$ is the von Neumann algebra of G, and then there is the notion of being of *determinant class*.) We treat this entirely as a black box, but we note that if G is residually amenable, eg a 3–manifold group [11] or solvable, then by [8] any matrix over $\mathbb{Q}[G]$ is of determinant class. If the matrix A is not of determinant class then for the purpose of this paper we define $\det_{\mathcal{N}(G)}(A) = 0$. On the other hand, if A is of determinant class, then we define

$$\det_{\mathcal{N}(G)}(A) := \text{Fuglede–Kadison determinant of } A \in \mathbb{R}_{>0}.$$

Here we do not assume that A is a square matrix. In an attempt to keep the paper as short as possible we will not provide the (somewhat lengthy) definition of the Fuglede–Kadison determinant. Instead we summarize a few key properties in the following theorem which is a consequence of [14, Example 3.12] and [14, Theorem 3.14].

Theorem 3.1 (1) *If A is a square matrix with complex entries such that the usual determinant $\det(A) \in \mathbb{C}$ is non-zero, then $\det_{\mathcal{N}(G)}(A) = |\det(A)|$.*

(2) *The determinant does not change if we swap two rows or two columns.*

(3) *Right multiplication of a column by $\pm g$, $g \in G$ does not change the determinant.*

(4) *For any matrix A over $\mathbb{C}[G]$ we have $\det_{\mathcal{N}(G)}(A) = \det_{\mathcal{N}(G)}(\bar{A}^t)$.*

Note that (2) implies that when we study Fuglede–Kadison determinants of homomorphisms we can work with unordered bases. Now let

$$C_* = (0 \to C_l \xrightarrow{\partial_l} C_{l-1} \xrightarrow{\partial_{l-1}} \cdots \xrightarrow{\partial_2} C_1 \xrightarrow{\partial_1} C_0 \to 0)$$

be a chain complex of free left $\mathbb{C}[G]$–modules. We refer to [14] for the definition of the L^2–Betti numbers $b_i^{(2)}(C_*) \in \mathbb{R}_{\geq 0}$. Now suppose that the chain complex is equipped with bases $B_i \subset C_i$, $i = 0, \dots, l$. If one of the L^2–Betti numbers $b_i^{(2)}(C_*)$ is non-zero or if one the boundary maps is not of determinant class, then we define the L^2–torsion $\tau^{(2)}(C_*, B_*) := 0$. Otherwise we define the *L^2–torsion* to be

$$\tau^{(2)}(C_*, B_*) := \prod_{i=1}^{l} \det_{\mathcal{N}(G)}(A_i)^{(-1)^i} \in \mathbb{R}_{>0},$$

where the A_i denote the boundary matrices corresponding to the given bases. This definition is the multiplicative inverse of the exponential of the L^2–torsion as defined in [14, Definition 3.29].

3B The twisted L^2–torsion of CW–complexes and manifolds

Let (X,Y) be a pair of CW–complexes and let $e \in \mathrm{Eul}(X,Y)$. We denote by $p\colon \widetilde{X} \to X$ the universal covering of X and we write $\widetilde{Y} := p^{-1}(Y)$. The deck transformation turns $C_*(\widetilde{X},\widetilde{Y})$ naturally into a chain complex of left $\mathbb{Z}[\pi_1(X)]$–modules.

Now let G be a group and let $\varphi\colon \pi(X) \to \mathrm{GL}(d,\mathbb{C}[G])$ be a representation. We view elements of $\mathbb{C}[G]^d$ as row vectors. Right multiplication via $\varphi(g)$ thus turns $\mathbb{C}[G]^d$ into a right $\mathbb{Z}[\pi_1(X)]$–module. We consider the chain complex

$$C_*^{\varphi}(X,Y;\mathbb{C}[G]^d) := \mathbb{C}[G]^d \otimes_{\mathbb{Z}[\pi_1(X)]} C_*(\widetilde{X},\widetilde{Y})$$

of left $\mathbb{C}[G]$–modules. Let $e \in \mathrm{Eul}(X,Y)$. We pick an Euler lift $\{c_{ij}\}$ that represents e. Throughout this paper we denote by $v_1,\dots,v_d$ the standard basis for $\mathbb{C}[G]^d$. We equip the chain complex $C_*^{\varphi}(X,Y;\mathbb{C}[G]^d)$ with the basis provided by the $v_k \otimes c_{ij}$. Therefore we can define

$$\tau^{(2)}(X,Y,\varphi,e) := \tau^{(2)}\big(C_*^{\varphi}(X,Y;\mathbb{C}[G]^d),\{v_k \otimes c_{ij}\}\big) \in \mathbb{R}_{\geq 0}.$$

Lemma 3.2 (1) *The number $\tau^{(2)}(X,Y,\varphi,e)$ is well defined.*

(2) *If $g \in \mathcal{H}_1(X)$, then*

$$\tau^{(2)}(X,Y,\varphi,ge) = \det_{\mathcal{N}(G)}(\varphi(g^{-1}))\cdot\tau^{(2)}(X,Y,\varphi,e).$$

(3) *If (X',Y') is a cellular subdivision of (X,Y) and if $e' \in \mathrm{Eul}(X',Y')$ is the Euler structure corresponding to e, then*

$$\tau^{(2)}(X',Y',\varphi,e') = \tau^{(2)}(X,Y,\varphi,e).$$

The proofs are completely analogous to the proofs for ordinary Reidemeister torsion as given in [18; 9]. In the interest of space we will not provide the proofs.

Finally let N be a manifold and let $\partial_0 N \subset \partial N$ be a union of components of ∂N with $\chi(N,\partial_0 N) = 0$. Let G be a group and let $\varphi\colon \pi(N) \to \mathrm{GL}(d,\mathbb{C}[G])$ be a representation. Let $e \in \mathrm{Eul}(N,\partial_0 N)$. Recall that for any C^1–triangulation $f\colon X \to N$ we get a bijection $\mathrm{Eul}(X,Y) \xrightarrow{f_*} \mathrm{Eul}(N,\partial_0 N)$. We define

$$\tau^{(2)}(N,\partial_0 N,\varphi,e) := \tau^{(2)}(X,Y,\varphi\circ f_*, f_*^{-1}(e)).$$

By Lemma 3.2(3) and the discussion in [17] the invariant $\tau^{(2)}(N,\partial_0 N,\varphi,e) \in \mathbb{R}_{\geq 0}$ is well defined, ie independent of the choice of the triangulation.

4 Duality for torsion of manifolds equipped with Euler structures

4A The algebraic duality theorem for L^2–torsion

Let G be a group and let V be a right $\mathbb{C}[G]$–module. We denote by $\overline{V}$ the left $\mathbb{C}[G]$–module with the same underlying abelian group but with the module structure given by $p \cdot_{\overline{V}} v := v \cdot_V \bar{p}$ for $p \in \mathbb{C}[G]$ and $v \in V$. If V is a left $\mathbb{C}[G]$–module then we can consider $\operatorname{Hom}_{\mathbb{C}[G]}(V, \mathbb{C}[G])$, the set of all left $\mathbb{C}[G]$–module homomorphisms. Since the range $\mathbb{C}[G]$ is a $\mathbb{C}[G]$–bimodule we can naturally view $\operatorname{Hom}_{\mathbb{C}[G]}(V, \mathbb{C}[G])$ as a right $\mathbb{C}[G]$–module.

In the following let C_* be a chain complex of length m of left $\mathbb{C}[G]$–modules with boundary operators ∂_i. Suppose that C_* is equipped with a basis B_i for each C_i. We denote by $C^\#$ the *dual chain complex* whose chain groups are the $\mathbb{C}[G]$–left modules $C_i^\# := \overline{\operatorname{Hom}_{\mathbb{C}[G]}(C_{m-i}, \mathbb{C}[G])}$ and where the boundary map $\partial_i^\#\colon C_{i+1}^\# \to C_i^\#$ is given by $(-1)^{m-i}\partial_{m-i-1}^*$. This means that for any $c \in C_{m-i}$ and $d \in C_{i+1}^\#$ we have $\partial_i^\#(d)(c) = (-1)^{m-i} d(\partial_{m-i}(c))$. We denote by $B_*^\#$ the bases of $C^\#$ dual to the bases B_*.

Lemma 4.1 *If $\tau^{(2)}(C_*, B_*) = 0$, then $\tau^{(2)}(C_*^\#, B_*^\#) = 0$, otherwise we have*

$$\tau^{(2)}(C_*, B_*) = \tau^{(2)}(C_*^\#, B_*^\#)^{(-1)^{m+1}}.$$

Proof By the proof of [14, Theorem 1.35(3)] the L^2–Betti numbers of C_* vanish if and only if the L^2–Betti numbers of $C_*^\#$ vanish. In particular, if either L^2–Betti number does not vanish, then both torsions are zero.

Now we suppose that the L^2–Betti numbers of C_* vanish. We denote by A_i the corresponding matrices of the boundary maps of C_*. The boundary matrices of the chain complex $C_*^\#$ with respect to the basis $B_*^\#$ are given by $(-1)^{m-i}\overline{A}_i^t$. Now the lemma is an immediate consequence of the definitions and of Theorem 3.1(4). □

4B The duality theorem for manifolds

Before we state our main technical duality theorem we need to introduce two more definitions.

(1) Let G be a group and let $\varphi\colon \pi \to \mathrm{GL}(d, \mathbb{C}[G])$ be a representation. We denote by $\varphi^\dagger$ the representation which is given by $g \mapsto \overline{\varphi(g^{-1})}^t$.

(2) Let N be an m–manifold and let $e \in \operatorname{Eul}(N, \partial N)$. Pick a triangulation X for N. We denote by Y the subcomplex corresponding to ∂N. Let $X^\dagger$ be the CW–complex that is given by the cellular decomposition of N dual to X. Pick

an Euler lift $\{c_{ij}\}$ that represents $e \in \operatorname{Eul}(X,Y) = \operatorname{Eul}(N,\partial N)$. For any i–cell c in $\widetilde{X} \setminus \widetilde{Y}$ we denote by $c^\dagger$ the unique oriented $(m-i)$–cell in $\widetilde{X}^\dagger$ which has intersection number $+1$ with c. The Euler lift $\{c_{ij}^\dagger\}$ defines an element in $\operatorname{Eul}(X^\dagger) = \operatorname{Eul}(N)$ that we denote by $e^\dagger$. This map is an $\mathcal{H}_1(N)$–equivariant bijection and we denote the inverse map $\operatorname{Eul}(N,\partial N) \to \operatorname{Eul}(N)$ again by $e \mapsto e^\dagger$. We refer to [15, Chapter 70], [18, Section 14] and [9, Section 4] for details.

Theorem 4.2 *Let N be an m–manifold. Let G be a group and let $\varphi\colon \pi(N) \to \mathrm{GL}(d,\mathbb{C}[G])$ be a representation. Let $e \in \operatorname{Eul}(N,\partial N)$. Then either $\tau^{(2)}(N,\partial N,\varphi,e)$ and $\tau^{(2)}(N,\varphi^\dagger,e^\dagger)$ are both zero, or the following equality holds:*

$$\tau^{(2)}(N,\partial N,\varphi,e) = \tau^{(2)}(N,\varphi^\dagger,e^\dagger)^{(-1)^{m+1}}.$$

Proof Pick a triangulation X for N and denote by Y the subcomplex corresponding to ∂N. Let $X^\dagger$ be the CW–complex which is given by the cellular decomposition of N dual to X. We identify $\pi = \pi_1(X) = \pi_1(N) = \pi_1(X^\dagger)$. We pick an Euler lift $\{c_{ij}\}$ which represents $e \in \operatorname{Eul}(N,\partial N) = \operatorname{Eul}(X,Y)$. We denote by $c_{ij}^\dagger$ the corresponding dual cells. The theorem follows from the definitions and the following claim.

Claim *Either both $\tau^{(2)}(C_*^\varphi(X,Y;\mathbb{C}[G]^d),\{v_k \otimes c_{ij}\})$ and $\tau^{(2)}(C_*^{\varphi^\dagger}(X^\dagger;\mathbb{C}[G]^d),\{v_k \otimes c_{ij}^\dagger\})$ are zero, or the following equality holds:*

$$\tau^{(2)}\big(C_*^\varphi(X,Y;\mathbb{C}[G]^d),\{v_k \otimes c_{ij}\}\big) = \tau^{(2)}\big(C_*^{\varphi^\dagger}(X^\dagger;\mathbb{C}[G]^d),\{v_k \otimes c_{ij}^\dagger\}\big)^{(-1)^{m+1}}.$$

In order to prove the claim we first note that there is a unique, sesquilinear paring

$$\begin{aligned} C_{m-i}(\widetilde{X},\widetilde{Y}) \times C_i(\widetilde{X}^\dagger) &\to \mathbb{Z}[\pi], \\ (a,b) &\mapsto \langle a,b\rangle := \sum_{g\in\pi} (a\cdot gb)g^{-1} \end{aligned}$$

such that $a \cdot b^\dagger = \delta_{ab}$ for any two cells a and b of $\widetilde{X} \setminus \widetilde{Y}$. Here *sesquilinear* means that for any $a \in C_{m-i}(\widetilde{X},\widetilde{Y})$, $b \in C_i(\widetilde{X}^\dagger)$ and $p,q \in \mathbb{Z}[\pi]$ we have $\langle pa,qb\rangle = q\langle a,b\rangle\,\bar{p}$. It is straightforward to see that the pairing is non-singular. It follows immediately from [18, Claim 14.4]) that these maps give rise to well-defined maps

$$\begin{aligned} C_i(\widetilde{X},\widetilde{Y}) &\to \overline{\operatorname{Hom}_{\mathbb{Z}[\pi]}(C_{m-i}(\widetilde{X}^\dagger),\mathbb{Z}[\pi])}, \\ a &\mapsto (b \mapsto \langle a,b\rangle) \end{aligned}$$

that define an isomorphism of based chain complexes of right $\mathbb{Z}[\pi]$–modules. In fact it follows easily from the definitions that the maps define an isomorphism

$$(C_*(\widetilde{X},\widetilde{Y}),\{c_{ij}\}) \to \big(\overline{\operatorname{Hom}_{\mathbb{Z}[\pi]}(C_{m-*}(\widetilde{X}^\dagger),\mathbb{Z}[\pi])},\{(c_{ij}^\dagger)^*\}\big)$$

of based chain complexes of left $\mathbb{Z}[\pi]$–modules. Tensoring these chain complexes with $\mathbb{C}[G]^d$ we obtain an isomorphism

$$\begin{aligned}\big(\mathbb{C}[G]^d \otimes_{\mathbb{Z}[\pi]} C_*(\widetilde{X},\widetilde{Y}), \{v_k \otimes c_{ij}\}\big) \\ \to \big(\mathbb{C}[G]^d \otimes_{\mathbb{Z}[\pi]} \overline{\operatorname{Hom}_{\mathbb{Z}[\pi]}(C_{m-*}(\widetilde{X}^\dagger), \mathbb{Z}[\pi])}, \{v_k \otimes (c_{ij}^\dagger)^*\}\big)\end{aligned}$$

of based chain complexes of $\mathbb{C}[G]$–modules. Furthermore the maps

$$\begin{aligned}\mathbb{C}[G]^d \otimes_{\mathbb{Z}[\pi]} \overline{\operatorname{Hom}_{\mathbb{Z}[\pi]}(C_i(\widetilde{X}^\dagger), \mathbb{Z}[\pi])} &\to \overline{\operatorname{Hom}_{\mathbb{C}[G]}\big(C_i^{\varphi^\dagger}(X^\dagger;\mathbb{C}[G]^d), \mathbb{C}[G]\big)}, \\ v \otimes f &\mapsto \begin{pmatrix} C_i^{\varphi^\dagger}(X^\dagger;\mathbb{C}[G]^d) \to \mathbb{C}[G], & \\ & w \otimes \sigma \mapsto v\varphi(\overline{f(\sigma)})\overline{w}^t \end{pmatrix}\end{aligned}$$

induce an isomorphism

$$\big(C_*^{\varphi}(X,Y;\mathbb{C}[G]^d), \{v_k \otimes c_{ij}\}\big) \to \big(C_*^{\varphi^\dagger}(X^\dagger;\mathbb{C}[G]^d)^\#, \{(v_k \otimes c_{ij}^\dagger)^\#\}\big)$$

of based chain complexes of $\mathbb{C}[G]$–modules. The claim follows from Lemma 4.1. □

5 Twisted L^2–torsion of 3–manifolds

5A Canonical structures on tori

Let T be a torus. We equip T with a CW–structure with one 0–cell p, two 1–cells x and y and one 2–cell s. We write $\pi = \pi_1(T,p)$ and by a slight abuse of notation we denote by x and y the elements in π represented by x and y. We denote by $\widetilde{T}$ the universal cover of T. There exist lifts of the cells such that the chain complex $C_*(\widetilde{T})$ of left $\mathbb{Z}[\pi]$–modules with respect to the bases given by these lifts is of the form

$$0 \to \mathbb{Z}[\pi] \xrightarrow{(y-1\ \ 1-x)} \mathbb{Z}[\pi]^2 \xrightarrow{\begin{pmatrix}1-x\\1-y\end{pmatrix}} \mathbb{Z}[\pi] \to 0. \tag{1}$$

We refer to the corresponding Euler structure of T as the *canonical Euler structure on* T. Given a group G we say that a representation $\varphi\colon \pi \to \operatorname{GL}(1,\mathbb{C}[G])$ is *monomial* if for any $x \in \pi$ we have $\varphi(x) = zg$ for some $z \in \mathbb{C}$ and $g \in G$. The following is [4, Lemma 5.6].

Lemma 5.1 *Let $\varphi\colon \pi_1(T) \to \operatorname{GL}(1,\mathbb{C}[G])$ be a monomial representation such that $b_*^{(2)}(T;\mathbb{C}[G]) = 0$ and e be the canonical Euler structure on T. Then $\tau^{(2)}(T,\varphi,e) = 1$.*

5B Chern classes on 3–manifolds with toroidal boundary

Let N be a 3–manifold with toroidal incompressible boundary and let $e \in \mathrm{Eul}(N, \partial N)$. Let X be a triangulation for N. We denote the subcomplexes corresponding to the boundary components of N by $S_1 \cup \dots \cup S_b$. We denote by $p\colon \widetilde{X} \to X$ and $p_i\colon \widetilde{S}_i \to S_i, i = 1, \dots, b$ the universal covering maps of X and $S_i, i = 1, \dots, b$. For each i we identify a component of $p^{-1}(S_i)$ with $\widetilde{S}_i$

Pick an Euler lift c that represents e. For each boundary torus S_i pick an Euler lift $\widetilde{s_i}$ to $\widetilde{S}_i \subset p^{-1}(S_i) \subset \widetilde{X}$ that represents the canonical Euler structure. The set of cells $\{\tilde{s}_1, \dots, \tilde{s}_b, c\}$ defines an Euler structure $K(e)$ for N, which only depends on e. Put differently, we defined a map $K\colon \mathrm{Eul}(N, \partial N) \to \mathrm{Eul}(N)$ which is easily seen to be $\mathcal{H}_1(N)$–equivariant. Given $e \in \mathrm{Eul}(N)$ there exists a unique element $g \in \mathcal{H}_1(N)$ such that $e = g \cdot K(e^\dagger)$. Following Turaev [19, page 11] and [9, Section 6.3] we define $c_1(e) := g \in H_1(N; \mathbb{Z})$ and we refer to $c_1(e)$ as the *Chern class of* e.

5C Torsions of 3–manifolds

Let π and G be groups and let $\varphi\colon \pi \to \mathrm{GL}(1, \mathbb{C}[G])$ be a monomial representation. By the multiplicativity of the Fuglede–Kadison determinant, see [14, Theorem 3.14], given $g \in \pi$ the invariant $\det_{\mathcal{N}(G)}(\varphi(g))$ only depends on the homology class of g. Put differently, $\det_{\mathcal{N}(G)} \circ \varphi\colon \pi \to \mathbb{R}_{\geq 0}$ descends to a map $\det_{\mathcal{N}(G)} \circ \varphi\colon H_1(\pi; \mathbb{Z}) \to \mathbb{R}_{\geq 0}$.

Theorem 5.2 *Let N be a 3–manifold which is either closed or which has toroidal, incompressible boundary. Let G be a group and let $\varphi\colon \pi(N) \to \mathrm{GL}(1, \mathbb{C}[G])$ be a monomial representation such that $b_*^{(2)}(\partial N; \mathbb{C}[G]) = 0$. For any $e \in \mathrm{Eul}(N)$ we have*

$$\tau^{(2)}(N, \partial N, \varphi, e^\dagger) = \det_{\mathcal{N}(G)}(\varphi(c_1(e))) \cdot \tau^{(2)}(N, \varphi, e).$$

Proof The assumption that $b_*^{(2)}(\partial N; \mathbb{C}[G]) = 0$ together with the proof of [14, Theorem 1.35(2)] implies that $b_*^{(2)}(N; \mathbb{C}[G]) = 0$ if and only if $b_*^{(2)}(N, \partial N; \mathbb{C}[G]) = 0$. If both are non-zero, then both torsions $\tau^{(2)}(N, \partial N, \varphi, e^\dagger)$ and $\tau^{(2)}(N, \varphi, e)$ are zero. For the remainder of this proof we assume that $b_*^{(2)}(N; \mathbb{C}[G]) = 0$.

Pick a triangulation X for N. As usual denote by Y the subcomplex corresponding to ∂N. Let $e \in \mathrm{Eul}(N)$. Pick an Euler lift c_* which represents $e^\dagger \in \mathrm{Eul}(N, \partial N) = \mathrm{Eul}(X, Y)$. Denote the components of Y by $Y_1 \cup \dots \cup Y_b$ and pick $\tilde{s}_*^1, \dots, \tilde{s}_*^b$ as in the previous section. We write $\tilde{s}_* = \tilde{s}_*^1 \cup \dots \cup \tilde{s}_*^b$. Denote by $\{\tilde{s}_* \cup c_*\}$ the resulting Euler lift for X. Recall that this Euler lift represents $K(e^\dagger)$.

Claim $\tau^{(2)}\big(C_*^\varphi(X, Y; \mathbb{C}[G]), \{c_*\}\big) = \tau^{(2)}\big(C_*^\varphi(X; \mathbb{C}[G]), \{\tilde{s}_* \cup c_*\}\big).$

We consider the following short exact sequence of chain complexes

$$0 \to \bigoplus_{i=1}^{b} C_*^{\varphi}(Y_i; \mathbb{C}[G]) \to C_*^{\varphi}(X; \mathbb{C}[G]) \to C_*^{\varphi}(X, Y; \mathbb{C}[G]) \to 0,$$

with the bases $\{s_*^i\}_{i=1,\dots,b}$, $\{\tilde{s}_* \cup c_*\}$ and $\{c_*\}$. These bases are in fact compatible, in the sense that the middle basis is the image of the left basis together with a lift of the right basis. By Lemma 5.1 we have $\tau^{(2)}(C_*^{\varphi}(Y_i; \mathbb{C}[G]), \{\widetilde{s}_*^i\}) = 1$ for $i = 1, \dots, b$. Now it follows from the multiplicativity of torsion, see [14, Theorem 3.35], that

$$\tau^{(2)}\big(C_*^{\varphi}(X, Y; \mathbb{C}[G]), \{c_*\}\big) = \tau^{(2)}\big(C_*^{\varphi}(X; \mathbb{C}[G]), \{c_* \cup \tilde{s}_*\}\big).$$

Here we used that the complexes are acyclic. This concludes the proof of the claim.

Finally it follows from this claim, the definitions and Lemma 3.2 that

$$\begin{aligned}\tau^{(2)}(N, \partial N, \varphi, e^{\dagger}) &= \tau^{(2)}\big(C_*^{\varphi}(X, Y; \mathbb{C}[G]), \{c_*\}\big) = \tau^{(2)}\big(C_*^{\varphi}(X; \mathbb{C}[G]), \{\tilde{s}_* \cup c_*\}\big) \\ &= \tau^{(2)}(N, \varphi, K(e^{\dagger})) = \tau^{(2)}(N, \varphi, c_1(e)^{-1} e) \\ &= \det\nolimits_{\mathcal{N}(G)}(\varphi(c_1(e))) \cdot \tau^{(2)}(N, \varphi, e). \end{aligned}$$ □

6 The symmetry of the L^2–Alexander torsion

Let $(N, \phi, \gamma\colon \pi_1(N) \to G)$ be an admissible triple and let $e \in \mathrm{Eul}(N)$. Given $t \in \mathbb{R}_{>0}$ we consider the representation $\gamma_t\colon \pi_1(N) \to \mathrm{GL}(1, \mathbb{C}[G])$ that is given by $\gamma_t(g) := (t^{\phi(g)} \gamma(g))$. We denote by $\tau^{(2)}(N, \phi, \gamma, e)$ the function

$$\begin{aligned}\tau^{(2)}(N, \phi, \gamma, e)\colon \mathbb{R}_{>0} &\to \mathbb{R}_{\geq 0}, \\ t &\mapsto \tau^{(2)}(N, \gamma_t, e).\end{aligned}$$

For another $e' \in \mathrm{Eul}(N)$ we have $e' = ge$ for some $g \in \mathcal{H}_1(N)$. By Lemma 3.2

$$\tau^{(2)}(N, \phi, \gamma, ge)(t) = t^{-\phi(g)} \tau^{(2)}(N, \phi, g, e)(t) \text{ for all } t \in \mathbb{R}_{>0}.$$

Put differently, the functions $\tau^{(2)}(N, \phi, \gamma, e)$ and $\tau^{(2)}(N, \phi, \gamma, ge)$ are equivalent. We denote by $\tau^{(2)}(N, \phi, \gamma)$ the equivalence class of the functions $\tau^{(2)}(N, \phi, \gamma, e)$ and we refer to $\tau^{(2)}(N, \phi, \gamma)$ as the *L^2–Alexander torsion of* (N, ϕ, γ).

Proof of Theorem 1.1 Let $e \in \mathrm{Eul}(N)$ and $t \in \mathbb{R}_{>0}$. We write $\tau = \tau^{(2)}(N, \gamma, \phi, e)$. Note that $(\gamma_t)^\dagger = \gamma_{t^{-1}}$. It follows from Theorems 4.2 and 5.2 that

$$\begin{aligned}\tau(t) &= \tau^{(2)}(N, \gamma, \phi, e) = \tau^{(2)}(N, \gamma_t, e)\\ &= \tau^{(2)}(N, \partial N, (\gamma_t)^\dagger, e^\dagger) = \tau^{(2)}(N, \partial N, \gamma_{t^{-1}}, e^\dagger)\\ &= \det_{\mathcal{N}(G)}(\gamma_{t^{-1}}(c_1(e))) \cdot \tau^{(2)}(N, \gamma_{t^{-1}}, e)\\ &= \det_{\mathcal{N}(G)}(t^{-\phi(c_1(e))} c_1(e)) \cdot \tau^{(2)}(N, \gamma_{t^{-1}}, e)\\ &= t^{-\phi(c_1(e))} \cdot \tau^{(2)}(N, \gamma_{t^{-1}}, e) = t^{-\phi(c_1(e))} \cdot \tau(t^{-1}).\end{aligned}$$

Now it suffices to show that for any $\phi \in H^1(N;\mathbb{Z})$ we have $\phi(c_1(e)) = x_N(\phi)$ mod 2.

So let S be a Thurston norm minimizing surface which is dual to some $\phi \in H^1(N;\mathbb{Z})$. Since N is irreducible and since $N \neq S^1 \times D^2$ we can arrange that S has no disk components. Therefore we have

$$x_N(\phi) \equiv \chi_-(S) \equiv b_0(\partial S) \mod 2\mathbb{Z}.$$

On the other hand, by [19, Lemma VI.1.2] and [19, Section XI.1] we have that $b_0(\partial S) \equiv c_1(e) \cdot S \mod 2\mathbb{Z}$ where $c_1(e) \cdot S$ is the intersection number of $c_1(e) \in H_1(N) = \mathcal{H}_1(N)$ with S. Since S is dual to ϕ, we obtain the desired equality

$$\phi(c_1(e)) \equiv c_1(e) \cdot S \equiv b_0(\partial S) \equiv \chi_-(S) \equiv x_N(\phi) \mod 2\mathbb{Z}. \qquad \square$$

Finally, a *real admissible triple* (N, ϕ, γ) is defined like an admissible triple, except that now we also allow ϕ to lie in $H^1(N;\mathbb{R}) = \mathrm{Hom}(\pi_1(N), \mathbb{R})$. The same definition as in Section 6 associates to (N, ϕ, e) a function $\tau^{(2)}(N, \phi, e)\colon \mathbb{R}_{>0} \to \mathbb{R}_{\geq 0}$ that is well defined up to multiplication by a function of the form $t \mapsto t^r$ for some $r \in \mathbb{R}$. The same argument as in the proof of Theorem 1.1 gives us the following result.

Theorem 6.1 *Let (N, ϕ, γ) be a real admissible triple. Then for any representative τ of $\tau^{(2)}(N, \phi, \gamma)$ there exists an $r \in \mathbb{R}$ such that $\tau(t^{-1}) = t^r \cdot \tau(t)$ for any $t \in \mathbb{R}_{>0}$.*

The only difference to Theorem 1.1 is that for real cohomology classes $\phi \in H^1(N;\mathbb{R})$ we cannot relate the exponent r to the Thurston norm of ϕ.

References

[1] **F Ben Aribi**, *The L^2–Alexander invariant detects the unknot*, C. R. Math. Acad. Sci. Paris 351 (2013) 215–219 MR3089681

[2] **F Ben Aribi**, *The L^2–Alexander invariant detects the unknot*, C. R. Math. Acad. Sci. Paris 351 (2013) 215–219 MR3089681

[3] **T D Cochran**, *Noncommutative knot theory*, Algebr. Geom. Topol. 4 (2004) 347–398 MR2077670

[4] **J Dubois**, **S Friedl**, **W Lück**, *The L^2–Alexander torsion of 3–manifolds*, preprint (2014) `arXiv:1410.6918`

[5] **J Dubois**, **S Friedl**, **W Lück**, *Three flavors of twisted knot invariants*, preprint (2014) `arXiv:1410.6924`

[6] **J Dubois**, **C Wegner**, *L^2–Alexander invariant for torus knots*, C. R. Math. Acad. Sci. Paris 348 (2010) 1185–1189 MR2738924

[7] **J Dubois**, **C Wegner**, *Weighted L^2–invariants and applications to knot theory*, Commun. Contemp. Math. 17 (2015) 1450010, 29 MR3291974

[8] **G Elek**, **E Szabó**, *Hyperlinearity, essentially free actions and L^2–invariants. The sofic property*, Math. Ann. 332 (2005) 421–441 MR2178069

[9] **S Friedl**, **T Kim**, **T Kitayama**, *Poincaré duality and degrees of twisted Alexander polynomials*, Indiana Univ. Math. J. 61 (2012) 147–192 MR3029395

[10] **S L Harvey**, *Higher-order polynomial invariants of 3–manifolds giving lower bounds for the Thurston norm*, Topology 44 (2005) 895–945 MR2153977

[11] **J Hempel**, *Residual finiteness for 3–manifolds*, from: "Combinatorial group theory and topology", (S M Gersten, J R Stallings, editors), Ann. of Math. Stud. 111, Princeton Univ. Press, Princeton, NJ (1987) 379–396 MR895623

[12] **W Li**, **W Zhang**, *An L^2–Alexander–Conway invariant for knots and the volume conjecture*, from: "Differential geometry and physics", (M-L Ge, W Zhang, editors), Nankai Tracts Math. 10, World Sci. Publ., Hackensack, NJ (2006) 303–312 MR2327174

[13] **W Li**, **W Zhang**, *An L^2–Alexander invariant for knots*, Commun. Contemp. Math. 8 (2006) 167–187 MR2219611

[14] **W Lück**, *L^2–invariants: Theory and applications to geometry and K–theory*, Ergeb. Math. Grenzgeb. 44, Springer, Berlin (2002) MR1926649

[15] **J R Munkres**, *Elements of algebraic topology*, Addison-Wesley Publishing Company, Menlo Park, CA (1984) MR755006

[16] **W P Thurston**, *A norm for the homology of 3–manifolds*, Mem. Amer. Math. Soc. 339, Amer. Math. Soc. (1986) MR823443

[17] **V G Turaev**, *Euler structures, nonsingular vector fields, and Reidemeister-type torsions*, Izv. Akad. Nauk SSSR Ser. Mat. 53 (1989) 607–643, 672 MR1013714 In Russian; translated in Math. USSR-Izv. 34 (1990) 627–662

[18] **V G Turaev**, *Introduction to combinatorial torsions*, Birkhäuser, Basel (2001) MR1809561

[19] **V G Turaev**, *Torsions of 3–dimensional manifolds*, Progress in Mathematics 208, Birkhäuser, Basel (2002) MR1958479

[20] **M Wada**, *Twisted Alexander polynomial for finitely presentable groups*, Topology 33 (1994) 241–256 MR1273784

Laboratoire de Mathématiques UMR 6620 – CNRS, Université Blaise Pascal
Campus des Cézeaux, BP 80026, 63171 Aubière, France

Fakultät für Mathematik, Universität Regensburg
D-93053 Regensburg, Germany

Mathematisches Institut, Rheinische Wilhelms-Universität Bonn
Endenicher Allee 60, D-53115 Bonn, Germany

jerome.dubois@math.univ-bpclermont.fr, sfriedl@gmail.com,
wolfgang.lueck@him.uni-bonn.de

Received: 20 November 2014 Revised: 7 April 2015

Algebraic & Geometric Topology 15 (2015) 3613–3652

Cup products, the Johnson homomorphism and surface bundles over surfaces with multiple fiberings

NICK SALTER

Let $\Sigma_g \to E \to \Sigma_h$ be a surface bundle over a surface with monodromy representation $\rho\colon \pi_1\Sigma_h \to \mathrm{Mod}(\Sigma_g)$ contained in the Torelli group $\mathcal{I}_g$. We express the cup product structure in $H^*(E,\mathbb{Z})$ in terms of the Johnson homomorphism $\tau\colon \mathcal{I}_g \to \bigwedge^3(H_1(\Sigma_g,\mathbb{Z}))/H_1(\Sigma_g,\mathbb{Z})$. This is applied to the question of obtaining an upper bound on the maximal n such that $p_1\colon E \to \Sigma_{h_1}, \dots, p_n\colon E \to \Sigma_{h_n}$ are fibering maps realizing E as the total space of a surface bundle over a surface in n distinct ways. We prove that any nontrivial surface bundle over a surface with monodromy contained in the Johnson kernel $\mathcal{K}_g$ fibers in a unique way.

57R22; 57R95

1 Introduction

The theory of the Thurston norm gives a detailed picture of the set of possible ways that a compact, oriented 3–manifold M can fiber as a surface bundle. If $b_1(M) > 1$, then M admits infinitely many such fibrations $\Sigma_g \to M \to S^1$, finitely many for each $g \geq 2$. The purpose of the present paper is to take up a similar sort of inquiry for 4–manifolds $\Sigma_g \to E \to \Sigma_h$ fibering as a surface bundle over a surface of genus $g \geq 2$.

When $h = 1$ (ie the base surface is a torus), a similar story as in the 3–manifold setting unfolds; if M^3 is a 3–manifold admitting infinitely many fiberings $p\colon M \to S^1$, then $p \times \mathrm{id}\colon M^3 \times S^1 \to S^1 \times S^1$ admits infinitely many fiberings as well. However, in stark contrast with the 3–dimensional setting and with the case of surface bundles over the torus, F E A Johnson [8] showed that if $\Sigma_g \to E \to \Sigma_h$ is a surface bundle over a surface with g, $h \geq 2$, then there are only finitely many distinct fibrations $p_i\colon E \to \Sigma_{h_i}$ realizing E as the total space of a surface bundle over a surface (see Proposition 2.1 for a precise definition of what is meant by "distinct"). Hillman [7] contains a treatment of results of this type, as does Rivin [12], in which the case of surface bundles over surfaces is situated in the larger context of "fibering rigidity" for a wide class of manifolds.

A particularly simple example of a surface bundle over a surface admitting two fiberings is that of a trivial bundle, ie a product of surfaces $\Sigma_g \times \Sigma_h$. At the time of

Published: 12 January 2016 DOI: 10.2140/agt.2015.15.3613

Johnson's result, there was essentially one known method for producing nontrivial surface bundles over surfaces with multiple fiberings, due independently to Atiyah [1] and Kodaira [9] (see also the summary in [11]). Their construction is built by taking a certain cyclic branched covering $p\colon E \to \Sigma_g \times \Sigma_h$ of a product of surfaces. The two fibering maps are inherited from the projections of $\Sigma_g \times \Sigma_h$ onto either factor. While Johnson's argument produces a bound on the number of possible fiberings of a surface bundle E that is super-exponential in the Euler characteristic $\chi(E)$, until recently all known examples of surface bundles over surfaces had at most two fiberings, leaving a large gap between the upper and lower bounds on the number of possible fiberings.

The author [14] gave a new method for constructing surface bundles over surfaces with multiple fiberings, including the first examples of bundles admitting an arbitrarily large number of fiberings. In fact, the methods of [14] are capable of producing families E_n of surface bundles admitting exponentially many fiberings as a function of $\chi(E_n)$. The results of this paper can be seen as a complement to that work, in that our concern here is in addressing the question of when surface bundles over surfaces admit *unique* fiberings.

A central theme in the study of surface bundles is the "monodromy–topology dictionary". For any reasonable base space M, there is a well-known correspondence (see eg Farb and Margalit [3])

$$\left\{\begin{matrix}\text{bundle isomorphism classes of}\\ \text{oriented } \Sigma_g\text{–bundles over } M\end{matrix}\right\} \longleftrightarrow \left\{\begin{matrix}\text{conjugacy classes of represen-}\\ \text{tations } \pi_1(M) \to \operatorname{Mod}(\Sigma_g)\end{matrix}\right\}. \tag{1}$$

This raises the question of translating between topological and geometric properties of surface bundles on the one hand and, on the other, algebraic or geometric properties of the monodromy representation. Certain entries in this dictionary are well established, for instance Thurston's landmark result that a fibered 3–manifold $\Sigma_g \to M_\phi \to S^1$ admits a complete hyperbolic metric if and only if the monodromy is a so-called "pseudo-Anosov" element of $\operatorname{Mod}(\Sigma_g)$. In this paper we add to the dictionary by relating the cohomology ring of a surface bundle over a surface to its monodromy representation, then apply these results to give various obstructions for the surface bundle to admit more than one fibering.

From the perspective of the monodromy representation, the phenomenon of multiple fibering remains mysterious. The central result of this paper shows that there is a strong interaction between the existence of multiple fiberings and the theory of the Torelli group $\mathcal{I}_g$. Recall that the *Torelli group* is the kernel of the symplectic representation

$\Psi\colon \mathrm{Mod}(\Sigma_g) \to \mathrm{Sp}_{2g}(\mathbb{Z})$ and that the *Johnson kernel* $\mathcal{K}_g$ is defined as the group generated by Dehn twists T_γ with γ a *separating* curve.[1]

Theorem 1.1 *Let $\pi\colon E \to B$ be a surface bundle over a surface with monodromy in the Johnson kernel $\mathcal{K}_g$. If E admits two distinct fiberings then E is diffeomorphic to $B \times B'$, the product of the base spaces. In other words, any nontrivial surface bundle over a surface with monodromy in $\mathcal{K}_g$ admits a unique fibering.*

The surface bundles over surfaces of [14] can be constructed so as to have monodromy contained in $\mathcal{I}_g$. It follows that the hypothesis in Theorem 1.1 that the monodromy be contained in $\mathcal{K}_g$ is effectively *sharp* with respect to the *Johnson filtration* (see [3, Chapter 6] for the definition of the Johnson filtration).

Theorem 1.1 is proved by first relating the monodromy representation of a surface bundle over a surface $E^4 \to B^2$ to the cohomology ring $H^*(E)$. This analysis will show that the integral cohomology of a surface bundle over a surface with monodromy in $\mathcal{K}_g$ is as simple as possible. It is then shown that, in these circumstances, obstructions to possessing alternative fiberings can be extracted from $H^*(E)$.

In a similar spirit we also have the following general criterion, which we believe to be of independent interest, for a surface bundle over a surface to possess a unique fibering. It can be viewed as the 4–manifold analogue of a well-known fact about fibered 3–manifolds (see Remark 3.6).

Theorem 3.5 *Let $p\colon E \to B$ be a surface bundle over a surface B of genus $g \geq 2$ with monodromy representation $\rho\colon \pi_1 B \to \mathrm{Mod}(\Sigma_g)$. Suppose that the space of invariant cohomology $(H^1(F,\mathbb{Q}))^\rho$ (equivalently, the coinvariant homology of the fiber $(H_1(F,\mathbb{Q}))_\rho$) vanishes. Then E admits a unique fibering.*

The paper is organized as follows. In Section 2, we give various characterizations of the notion of equivalence under consideration. In Section 3, we prove Theorem 3.5. Sections 4–7 are devoted to the proof of Theorem 1.1. Section 4 is devoted to a lemma in differential topology that features in later stages of the proof of Theorem 1.1. The technical heart of the paper is Section 5. In it, we first give an overview of the classical description of the Johnson homomorphism τ in terms of the intersection theory of surfaces in 3–manifolds that fiber over S^1. Using this description of τ, we then carry out a construction of 3–manifolds embedded in surface bundles over surfaces that

[1] As discussed further in Section 5.1, there is an alternative characterization of $\mathcal{K}_g$ as the kernel of the Johnson homomorphism (to be defined there). We will pass between these two perspectives as the situation dictates.

realizes the relationship between the Johnson homomorphism and the intersection product in the homology of the surface bundle. We give a complete description of the product structure in (co)homology for a surface bundle over a surface with monodromy in $\mathcal{I}_g$. These methods of Section 5 extend to an arbitrary surface bundle over a surface, but we do not state them in this level of generality since we have no need for them here.

Section 6 is devoted to some technical results concerning multisections of surface bundles, and their connection to splittings on rational cohomology. These results are used in the course of proving Theorem 1.1.

In Section 7 we turn finally to the proof of Theorem 1.1. The result follows from an analysis of the intersection product structure in $H_*(E)$ for a surface bundle over a surface $\Sigma_g \to E \to \Sigma_h$ with monodromy in $\mathcal{K}_g$. The results of Section 5 are applied to show that if the monodromy of $\Sigma_g \to E \to \Sigma_h$ is contained in $\mathcal{K}_g$, then E, which necessarily has $H^* E \approx H^*\Sigma_g \otimes H^*\Sigma_h$ as an additive group, in fact has $H^* E \approx H^*\Sigma_g \otimes H^*\Sigma_h$ (with $\mathbb{Z}$ coefficients) as a graded ring. This condition is then exploited to prove Theorem 1.1.

Acknowledgements The author would like to express his gratitude to Tom Church, Sebastian Hensel, Jonathan Hillman, Andy Putman, and Alden Walker for illuminating discussions at various stages of this work. He is grateful to the anonymous referees for many helpful suggestions. He would also like to extend his warmest thanks to Benson Farb for his extensive comments as well as his invaluable support from start to finish.

2 Equivalence

If E is a smooth n–manifold and $p_i\colon E \to B_i$, $i = 1, \dots, k$, are projection maps for various fiber bundle structures on E, we can consider the product of all the projection maps

$$p_1 \times \cdots \times p_k\colon E \to B_1 \times \cdots \times B_k.$$

In particular, if E^4 is the total space of a surface bundle over a surface with two fiberings, the *bi-projection* $p_1 \times p_2\colon E \to B_2 \times B_2$ is defined. As remarked in the introduction, ultimately we are concerned with fiberwise diffeomorphism classes of surface bundles. However, it is convenient to consider a more restrictive notion of equivalence, which will turn out to have the advantage of being describable purely on the level of the fundamental group.

We say that two fiberings $p_1\colon E \to B_1$ and $p_2\colon E \to B_2$ are *π_1–fiberwise diffeomorphic* if (1) they are fiberwise diffeomorphic, ie there exists a commutative diagram

$$\begin{array}{ccc} E & \xrightarrow{\phi} & E \\ {\scriptstyle p_1}\downarrow & & \downarrow{\scriptstyle p_2} \\ B_1 & \xrightarrow[\alpha]{} & B_2 \end{array}$$

with ϕ, α diffeomorphisms, and (2) $\phi_*(\pi_1 F_1) = \pi_1 F_1$ (here, as always, F_i denotes a fiber of p_i). Certainly if p_1 and p_2 are π_1–fiberwise diffeomorphic bundle structures, then they are fiberwise diffeomorphic in the usual sense. We are interested in this notion because we want to always regard the trivial bundle $\Sigma_g \times \Sigma_h$ as having two distinct fiberings. In the setting of fiberwise diffeomorphism, the projections onto either factor of $\Sigma_g \times \Sigma_g$ yield equivalent fiberings via the factor-swapping map $\phi(x,y) = (y,x)$, which covers the identity on Σ_g, but $\phi_*(\pi_1(\Sigma_g \times \{p\})) \neq \pi_1(\Sigma_g \times \{p\})$. The following proposition asserts that π_1–fiberwise diffeomorphism classes are in correspondence with the fiber subgroups $\pi_1 F \lhd \pi_1 E$. Recall that this is the setting in which F E A Johnson proved his finiteness result (see [8]).

Proposition 2.1 *Suppose E is the total space of a surface bundle over a surface in two ways, $p_1\colon E \to B_1$ and $p_2\colon E \to B_2$. Let F_1 and F_2 denote fibers of p_1 and p_2, respectively. Then the following are equivalent:*

1. *The fiberings p_1 and p_2 are π_1–fiberwise diffeomorphic.*
2. *The fiber subgroups $\pi_1 F_1$, $\pi_1 F_2 \le \pi_1 E$ are equal.*

If $\deg(p_1 \times p_2) \neq 0$ then the bundle structures p_1 and p_2 are distinct.

Proof First suppose that p_1 and p_2 are equivalent. Appealing to the long exact sequence in homotopy, we see that:

$$\begin{array}{ccccccccc} 1 & \longrightarrow & \pi_1 F_1 & \longrightarrow & \pi_1 E & \longrightarrow & \pi_1 B_1 & \longrightarrow & 1 \\ & & {\scriptstyle \phi_*}\downarrow & & {\scriptstyle \phi_*}\downarrow & & {\scriptstyle \alpha_*}\downarrow & & \\ 1 & \longrightarrow & \pi_1 F_2 & \longrightarrow & \pi_1 E & \longrightarrow & \pi_1 B_2 & \longrightarrow & 1 \end{array}$$

By assumption, $\phi_*(\pi_1 F_1) = \pi_1 F_1$, so that (1) implies (2).

Conversely, suppose that $\pi_1 F_1 = \pi_1 F_2$. Then the bundle structures p_1 and p_2 give rise to the same splitting

$$1 \to \pi_1 F \to \pi_1 E \to \pi_1 B \to 1$$

on fundamental groups. The monodromy for each bundle can be obtained from this sequence via the map $\pi_1 B \to \mathrm{Out}(\pi_1 F) \approx \mathrm{Mod}(\Sigma_g)$. This shows that the monodromies for the two bundle structures are conjugate and so, via the correspondence (1), there is a bundle isomorphism $\phi\colon E \to E$ covering the identity on B. To see that $\phi_*(\pi_1 F_1) = \pi_1 F_1$, consider the induced map on the long exact sequence in homotopy coming from ϕ:

$$\begin{tikzcd}
1 \arrow[r] & \pi_1 F_1 \arrow[r] \arrow[d, "\phi_*"'] & \pi_1 E \arrow[r] \arrow[d, "\phi_*"'] & \pi_1 B \arrow[r] \arrow[d, equal] & 1 \\
1 \arrow[r] & \pi_1 F_2 \arrow[r] & \pi_1 E \arrow[r] & \pi_1 B \arrow[r] & 1
\end{tikzcd}$$

This shows $\phi_*(\pi_1 F_1) = \pi_1 F_2$, and $\pi_1 F_1 = \pi_1 F_2$ by assumption, so (2) implies (1).

It remains to show that if $\deg(p_1 \times p_2) \neq 0$ then p_1 and p_2 are distinct. We establish the contrapositive. Suppose that $\pi_1 F_1 = \pi_1 F_2$. For $i = 1, 2$, we view $\pi_1 B_i$ as the quotient $\pi_1 B_i \approx \pi_1 E/\pi_1 F_i$. If $p_1 \times p_2$ is the bi-projection then, in this notation,

$$(p_1 \times p_2)_*\colon \pi_1 E \to \pi_1 B_1 \times \pi_1 B_2$$

is given by

$$(p_1 \times p_2)_*(x) = (x\,\pi_1 F_1, x\,\pi_1 F_2) = ([x], [x]),$$

where $[x] = x \pmod{\pi_1 F_1} = x \pmod{\pi_1 F_2}$. As $\pi_1 F_1 = \pi_1 F_2$, the quotients $\pi_1 B_1$ and $\pi_1 B_2$ are isomorphic, and as they are $K(G,1)$ spaces, there is a homotopy equivalence

$$f\colon B_1 \to B_2.$$

Let g be the map

$$g = (f \times \mathrm{id}) \circ (p_1 \times p_2)\colon E \to B_2 \times B_2.$$

By the above,

$$\mathrm{Im}(g) = \Delta = \{(x,x) \mid x \in B_2\}.$$

Being nonsurjective, g has degree 0. As $p_1 \times p_2$ is the composition of g with a homotopy equivalence, we conclude that also $\deg(p_1 \times p_2) = 0$. □

In general the condition $\deg(p_1 \times p_2) = 0$ on a bi-projection does not imply that the associated fiberings are equivalent. However, in the setting of the Johnson kernel, this is indeed the case.

Proposition 2.2 *Suppose E is the total space of a surface bundle over a surface in two ways, $p_1\colon E \to B_1$ and $p_2\colon E \to B_2$. Let F_1 and F_2 denote fibers of p_1 and p_2, respectively. Suppose that $\rho_1\colon \pi_1 B_1 \to \mathrm{Mod}(F_1)$ is contained in the Johnson kernel $\mathcal{K}_g$. Then the following are equivalent:*

(1) *The fiberings* p_1 *and* p_2 *are not* π_1*–fiberwise diffeomorphic.*

(2) *The fiber subgroups* $\pi_1 F_1$, $\pi_1 F_2 \leq \pi_1 E$ *are distinct.*

(3) $\deg(p_1 \times p_2) \neq 0$.

(4) E *is diffeomorphic to* $B_1 \times B_2$.

The additional assertions in Proposition 2.2 will be proved in the course of establishing Theorem 1.1 (see Remark 7.6).

3 Surface bundles over surfaces with unique fiberings

In this section, we prove Theorem 3.5. The additive structure of H^*E is central to everything that follows in the paper, so we begin with a review of the relevant results. The following theorem was formulated and proved by Morita [10] for the case of field coefficients of characteristic not dividing $\chi(F)$; subsequently this was improved to integral coefficients in the cohomological setting by Cavicchioli, Hegenbarth and Repovš [2].

Proposition 3.1 (Morita, Cavicchioli–Hegenbarth–Repovš) *Let* F *be a closed surface of genus* $g \geq 2$. *The Serre spectral sequence (with twisted coefficients) of any surface bundle* $F \to E \to B$ *collapses at the* E_2 *page. Consequently, there are noncanonical isomorphisms for all* k,

$$H_k(E,\mathbb{Q}) = H_k(B,\mathbb{Q}) \oplus H_{k-1}(B, H_1(F,\mathbb{Q})) \oplus H_{k-2}(B,\mathbb{Q}),$$
$$H^k(E,\mathbb{Z}) = H^k(B,\mathbb{Z}) \oplus H^{k-1}(B, H^1(F,\mathbb{Z})) \oplus H^{k-2}(B,\mathbb{Z}).$$

The $H_{k-2}B$ *summand of* $H_k E$ *is canonical and is realized by the Gysin map* $p^!$, *which associates to a homology class* $x \in B$ *the induced sub-bundle* E_x *sitting over* x. *Similarly, the* $H^k B$ *summand is canonical via the pullback map* $p^*\colon H^k B \to H^k E$.

If $F \to E \to B$ *has monodromy in* $\mathcal{I}_g$, *then the coefficient system is untwisted and* $H^*(E,\mathbb{Z}) \approx H^*(B,\mathbb{Z}) \otimes H^*(F,\mathbb{Z})$ *additively. In particular,* $H^*(E,\mathbb{Z})$ *is torsion-free and so, by the universal coefficients theorem, there is also an isomorphism* $H_*(E,\mathbb{Z}) \approx H_*(B,\mathbb{Z}) \otimes H_*(F,\mathbb{Z})$.

Because the surface bundles we will be considering in this paper have monodromy lying in $\mathcal{I}_g$, we will subsequently take all coefficients to be $\mathbb{Z}$ without further mention. A remark, which is obvious from Proposition 3.1, is that if $*$ generates $H_0(B)$ then $p^!(*)$ is a primitive class; we will use this fact later on. Here and throughout, we will use the notation

$$[F] = p^!(*) \in H_2(E)$$

to denote the (pushforward of the) fundamental class of the fiber.

The following result is a well-known application of the theory of the Gysin homomorphism and we state it without proof.

Proposition 3.2 *Let $p: E \to B$ be a surface bundle with fiber F. If $\chi(F) \neq 0$, then there are injections*

$$p^*: H^*(B, \mathbb{Q}) \to H^*(E, \mathbb{Q}),$$
$$p^!: H_k(B, \mathbb{Q}) \to H_{k+2}(E, \mathbb{Q}).$$

In the case where $H_*(E, \mathbb{Z})$ is torsion-free, the same statements hold with $\mathbb{Z}$ coefficients. In particular, this is true whenever E has monodromy lying in $\mathcal{I}_g$, since in this case $H^*(E, \mathbb{Z})$ is isomorphic to $H^*(F, \mathbb{Z}) \otimes H^*(B, \mathbb{Z})$ as an abelian group (see Proposition 3.1).

For surface bundles over surfaces with multiple fiberings, there is an extension of the previous result.

Lemma 3.3 *Let E be a 4–manifold with two distinct surface bundle structures $p_1: E \to B_1$ and $p_2: E \to B_2$. Then*

$$p_1^*(H^1(B_1, \mathbb{Q})) \cap p_2^*(H^1(B_2, \mathbb{Q})) = \{0\}$$

and so, by Proposition 3.2, there is a canonical injection

$$p_1^* \times p_2^*: H^1(B_1, \mathbb{Q}) \oplus H^1(B_2, \mathbb{Q}) \hookrightarrow H^1(E, \mathbb{Q}).$$

Proof By the universal coefficients theorem, for any space X there is an identification

$$H^1(X, \mathbb{Q}) \approx \mathrm{Hom}(\pi_1 X, \mathbb{Q}).$$

Under this identification, a character $\alpha \in \mathrm{Hom}(\pi_1 B_i, \mathbb{Q})$ is pulled back to $p_i^*(\alpha)$ in $\mathrm{Hom}(\pi_1 E, \mathbb{Q})$ by precomposition with $(p_i)_*$. In particular, $p_i^*(\alpha)$ vanishes on $\pi_1 F_i = \ker(p_i)_*$. Therefore, any character $\alpha \in p_1^*(H^1(B_1, \mathbb{Q})) \cap p_2^*(H^1(B_2, \mathbb{Q}))$ must vanish on the subgroup generated by $(\pi_1 F_1)(\pi_1 F_2)$.

By Lemma 3.4 below, $(\pi_1 F_1)(\pi_1 F_2)$ has finite index in $\pi_1 E$. For any group Γ, any character $\alpha: \Gamma \to \mathbb{Q}$ vanishing on a finite-index subgroup must vanish identically, proving the claim. □

Lemma 3.4 *Let E be a surface bundle over a surface with two distinct fiberings $p_i: E \to B_i$, $i = 1, 2$; let the fibers be F_1 and F_2, respectively. Then $(\pi_1 F_1)(\pi_1 F_2)$ has finite index in $\pi_1 E$.*

Proof Consider the cross-projection $\pi_1 F_1 \to \pi_1 B_2$. Let the image of $\pi_1 F_1$ in $\pi_1 B_2$ be denoted by Γ. This is a finitely generated normal subgroup of $\pi_1 B_2$. For any surface group of genus $g \geq 2$, any nontrivial finitely generated normal subgroup has finite index (see [8, Property $(\mathcal{D}6)$]). If Γ is the trivial group, then $\pi_1 F_1 \leq \pi_1 F_2$, necessarily again of finite index. In this case, the image of $\pi_1 F_2$ in $\pi_1 B_1$ is therefore finite, but $\pi_1 B_1$ is torsion-free. We conclude that $\Gamma \leq \pi_1 B_2$ has finite index. The kernel of the map $\pi_1 E \to (\pi_1 B_2/\Gamma)$ is exactly $(\pi_1 F_1)(\pi_1 F_2)$. □

Recall that if $\rho\colon G \to \mathrm{GL}(V)$ is a representation then the *invariant* space V^ρ is defined by

$$V^\rho = \{v \in V \mid \rho(g)(v) = v \text{ for all } g \in G\}.$$

The space of *coinvariants* V_ρ of the representation is defined as

$$V_\rho = V/W, \quad \text{where } W = \{v - \rho(g)(v) \mid v \in V,\ g \in G\}.$$

Theorem 3.5 *Let $p\colon E \to B$ be a surface bundle over a surface B of genus $g \geq 2$ with monodromy representation $\rho\colon \pi_1 B \to \mathrm{Mod}(\Sigma_g)$. Suppose that the space of invariant cohomology $(H^1(F,\mathbb{Q}))^\rho$ (equivalently, the coinvariant homology of the fiber $(H_1(F,\mathbb{Q}))_\rho$) vanishes. Then E admits a unique fibering.*

Proof For any surface bundle $p\colon E \to B$ with monodromy ρ and any choice of coefficients, there is a (noncanonical) splitting

$$H^1(E) = p^*(H^1(B)) \oplus (H^1(F))^\rho$$

(see Proposition 3.1). If $(H^1(F,\mathbb{Q}))^\rho = 0$, then this reduces to

$$H^1(E,\mathbb{Q}) = p^* H^1(B,\mathbb{Q}).$$

If $p_2\colon E \to B_2$ is a second, distinct fibering, the above shows that

$$p_2^*(H^1(B_2,\mathbb{Q})) \leq p^* H^1(B,\mathbb{Q}).$$

However, this contradicts Lemma 3.3. □

Remark 3.6 Recall that a surface bundle over S^1, viewed as the mapping torus M of some diffeomorphism ϕ of a surface F, admits a unique fibering if and only if $b_1(M) = 1$. This is the case exactly when $(H_1(F,\mathbb{Q}))_\phi = 0$, so Theorem 3.5 is the counterpart to this fact in dimension 4. Moreover, a random element $\phi \in \mathrm{Mod}(\Sigma_g)$ satisfies $(H_1(F,\mathbb{Q}))_\phi = 0$ (see [13]). It easily follows that a generic monodromy representation will also have $(H_1(F,\mathbb{Q}))_\rho = 0$: "most" surface bundles over surfaces have a single fibering. The proof of Theorem 3.5 is special to the case of surface bundles over surfaces and it is not clear if Theorem 3.5 is true in greater generality.

4 Bi-projections

In this section we state and prove the key lemma from differential topology needed for the proof of Theorem 1.1.

Proposition 4.1 *Let E be a 4–manifold with surface bundle structures $p_1\colon E \to B_1$ and $p_2\colon E \to B_2$. Let F_1 and F_2 denote fibers of p_1 and p_2 lying over a regular value of $p_1 \times p_2$. If $\deg(p_1 \times p_2\colon E \to B_1 \times B_2) \neq 0$, then the following five quantities are equal:*

(1) $\deg(p_1 \times p_2\colon E \to B_1 \times B_2)$.

(2) $\deg(p_1|_{F_2}\colon F_2 \to B_1)$.

(3) $\deg(p_2|_{F_1}\colon F_1 \to B_2)$.

(4) *The algebraic intersection number $I_E(F_1, F_2)$.*

(5) *The cardinality of the intersection $|F \cap F_2|$.*

As (5) indicates, this quantity is always positive.

Proof As p_1 and p_2 are projection maps for fiber bundle structures on E, they are everywhere regular, and $\ker(dp_1)_x$ is identified with the tangent space to the fiber of p_1 through x. Let $z = (b_1, b_2) \in B_1 \times B_2$ be a regular value for $p_1 \times p_2$. It follows from the assumption that $\deg(p_1 \times p_2\colon E \to B_1 \times B_2) \neq 0$ that $d(p_1 \times p_2)_x$ is an isomorphism for all $x \in (p_1 \times p_2)^{-1}(z)$ (and that this preimage is nonempty). The kernel of $d(p_1 \times p_2)_x$ is just the intersection of the kernels of $d(p_1)_x$ and $d(p_2)_x$. It follows that, for all $x \in (p_1 \times p_2)^{-1}(z)$,

$$T_x E \approx T_x F_1 \oplus T_x F_2. \tag{2}$$

Note that this shows that the fibers F_1 and F_2 over b_1 and b_2, respectively, are transverse.

If orientations on E, B_1 and B_2 are chosen properly, then this specifies an orientation on each fiber of p_1 and p_2 via the following decomposition, where H_x is any complement to $T_x F_1 = \ker d(p_1)_x$:

$$T_x F_1 \oplus H_x \approx T_x E.$$

The orientation on H_x is specified by the isomorphism $H_x \approx T_{p_1(x)} B_1$. Of course an analogous convention orients each fiber of p_2. In particular, it follows from Equation (2) that at any regular point for $p_1 \times p_2$ we can take $H_x = T_x F_2$ and that the restriction of $d(p_1)_x$ to $T_x F_2$ is an isomorphism.

Recall that if $f\colon X^n \to Y^n$ is a smooth map of oriented closed n–manifolds then

$$\deg(f) = \sum_{x \in f^{-1}(y)} \varepsilon(x),$$

where y is any regular value of f, and $\varepsilon(x) = 1$ if the orientation on $T_y Y$ induced by df_x agrees with the pre-chosen orientation on Y and $\varepsilon(x) = -1$ otherwise. If Y and Z are smoothly embedded and transversely intersecting oriented submanifolds of the oriented manifold X such that $\dim(X) = \dim(Y) + \dim(Z)$, then the algebraic intersection number of Y and Z is computed as

$$I_X(Y, Z) = \sum_{w \in Y \cap Z} \varepsilon(w),$$

where $\varepsilon(w) = 1$ if the orientation on $T_w X$ given by $T_w Y \oplus T_w Z$ agrees with the pre-chosen orientation on X and $\varepsilon(w) = -1$ otherwise.

It follows from the definitions that

$$(p_1 \times p_2)^{-1}(b_1, b_2) = p_1|_{F_2}^{-1}(b_1) = p_2|_{F_1}^{-1}(b_2) = F_1 \cap F_2.$$

Therefore, each of the sums computing (1)–(5) take place over the same set of points. So it remains only to show that, in each of the contexts (1)–(4), the relevant orientation convention assigns a positive value.

The orientation number assigned to $x \in (p_1 \times p_2)^{-1}(b_1, b_2)$ is given by the sign of the determinant of the map

$$d(p_1 \times p_2)_x\colon T_x E \to T_{b_1} B_1 \oplus T_{b_2} B_2.$$

By the above discussion, our orientation convention stipulates that

$$d(p_1|_{F_2})_x\colon T_x F_2 \to T_{b_1} B_1$$

is an orientation-preserving isomorphism and similarly for $d(p_2|_{F_1})$. This proves the equality of (2) and (3) with (5).

As

$$T_x F_1 = \ker d(p_1)_x \quad \text{and} \quad T_x F_2 = \ker d(p_2)_x,$$

it follows that $d(p_1 \times p_2)_x$ has a block-diagonal decomposition

$$d(p_1 \times p_2)_x = d(p_1)_x \oplus d(p_2)_x\colon T_x F_1 \oplus T_x F_2 \to T_{b_2} B_2 \oplus T_{b_1} B_1,$$

from which it follows that x also carries a positive orientation number in setting (1). Finally, the orientation number for x as a point of intersection between F_1 and F_2

records whether the orientations of $T_x E$ and $T_x F_1 \oplus T_x F_2$ agree, but we have already seen that they necessarily do. □

5 Cup products and the Johnson homomorphism

The goal of this section is to give a construction of embedded submanifolds in a surface bundle over a surface E that will be explicit enough to compute the intersection form on homology or, dually, the cup product structure in cohomology. One of the original definitions of the Johnson homomorphism was via the cup product structure in surface bundles over S^1. In this section we turn this perspective on its head and explain how the Johnson homomorphism computes the cup product structure in a surface bundle over a surface (in fact, these methods extend to surface bundles over arbitrary manifolds). The submanifolds we construct will be codimension-1 (ie 3–manifolds) and built so that their intersection theory is explicitly connected to the Johnson homomorphism.

To this end, in Section 5.1 we give a discussion of the definition of the Johnson homomorphism in the setting of the cup product in surface bundles over S^1. The centerpiece of this is the construction of geometric representatives for classes in H^1, via embedded surfaces which we call "tube-and-cap surfaces". Then, in Section 5.2, we return to the original problem of constructing representatives for classes in H^1 of a surface bundle over a surface as embedded 3–manifolds. The construction is carried out so that the intersection of particular pairs of these 3–manifolds is a tube-and-cap surface, thereby realizing the link between cup products in surface bundles over surfaces and the Johnson homomorphism.

5.1 From the intersection form to the Johnson homomorphism, and back again

In this subsection we will begin to dive into the theory of the Torelli group in earnest, so we begin with a brief review of the relevant definitions. The *Torelli group* $\mathcal{I}_g$ is the kernel of the symplectic representation $\Psi\colon \mathrm{Mod}(\Sigma_g) \to \mathrm{Sp}_{2g}(\mathbb{Z})$. The *Johnson kernel* $\mathcal{K}_g$ is the subgroup of $\mathcal{I}_g$ generated by all Dehn twists T_γ about *separating* curves γ. It is a deep theorem of D Johnson that $\mathcal{K}_g$ can alternately be characterized as the kernel of the Johnson homomorphism τ to be defined below.

Let $\phi \in \mathcal{I}_g$ be a Torelli mapping class and build the mapping torus

$$M_\phi = \Sigma_g \times I/\{(x,1) \sim (\phi(x),0)\}.$$

As $\phi \in \mathcal{I}_g$ for any curve $\gamma \subset \Sigma_g$, the homology class $[\gamma] - \phi_*[\gamma]$ is zero. Thus there exists a map of a surface $i\colon S \to \Sigma_g$ which cobounds $\gamma \cup \phi(\gamma)$. Indeed, there exists an *embedded* surface $S \leq \Sigma_g \times I$ whose boundary is given by

$$\partial S = \gamma \times \{1\} \cup \phi(\gamma) \times \{0\}.$$

To see this, recall that since S^1 is a $K(\mathbb{Z}, 1)$ there is a correspondence

$$H^1(\Sigma_g, \mathbb{Z}) \approx [\Sigma_g, S^1].$$

Via Poincaré duality,

$$H^1(\Sigma_g, \mathbb{Z}) \approx H_1(\Sigma_g, \mathbb{Z}).$$

The induced correspondence

$$H_1(\Sigma_g, \mathbb{Z}) \approx [\Sigma_g, S^1]$$

is realized by taking the preimage of a regular value, which will be an embedded submanifold. Under this correspondence, homotopic maps f, g: $\Sigma_g \to S^1$ yield homologous submanifolds, and conversely. Therefore, the maps f, g: $\Sigma_g \to S^1$ which determine γ and $\phi(\gamma)$ are homotopic. This gives the desired map F: $\Sigma_g \times I \to S^1$ such that the preimage of a regular value is an embedded surface S cobounding γ and $\phi(\gamma)$.

In fact, the choice of S is not unique. Let i': $S' \to M_\phi$ be any map of a *closed* surface to M_ϕ. Then the chain $S + S'$ satisfies $\partial(S + S') = \partial S = \gamma - \phi(\gamma)$. Nonetheless, given any S satisfying $\partial(S) = \gamma - \phi(\gamma)$, we can form a closed submanifold of M_ϕ in the following way. We begin with a tube, diffeomorphic to $S^1 \times I$, embedded into M_ϕ as $\phi(\gamma) \times [0, \frac{1}{3}] \cup \gamma \times [\frac{2}{3}, 1]$. We may then glue in S to $\Sigma_g \times [\frac{1}{3}, \frac{2}{3}]$. The result is a smoothly embedded oriented submanifold $\Sigma_\gamma \subset M_\phi$, which will descend to a homology class Σ_z (here $z = [\gamma]$). See Figure 1.

For convenience, we introduce the following terminology for these surfaces, which we will refer to as *tube surfaces*. The *tube* of a tube surface is the cylinder $S^1 \times I = \phi(\gamma) \times [0, \frac{1}{3}] \cup \gamma \times [\frac{2}{3}, 1]$ and the *cap* is the subsurface S.

We assign an orientation to Σ_γ as follows. The tangent space to a point x contained in the tube has a direct sum decomposition

$$T_x \Sigma_\gamma = V \oplus T_x \gamma, \tag{3}$$

where V is any preimage of $T_{\pi(x)} S^1$ and $T_x \gamma$ is interpreted as the tangent space to the copy of γ sitting in the fiber containing x. Both of the summands in (3) have orientations induced from those on S^1 and γ, respectively, and this endows $T_x \Sigma$ with an orientation. This can then be extended over the cap surface in a coherent way, since S was chosen to be a boundary for $[\gamma] - [\phi(\gamma)]$ with $\mathbb{Z}$ coefficients.

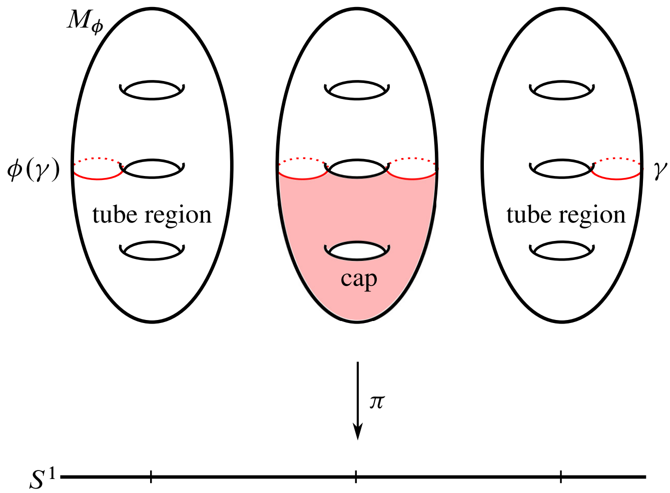


Figure 1: A tube surface

Recall however that the choice of S was not unique. Any closed surface mapping into Σ_g is homologous to some multiple of the fundamental class, so the above procedure really defines a homomorphism $H_1(\Sigma_g) \to H_2(M_\phi)/[F]$, where $[F]$ is the fundamental class of the fiber. If the bundle has a section $\sigma\colon S^1 \to M_\phi$, then we can choose S so that $\operatorname{Im}\sigma$ and Σ_z have zero algebraic intersection, which gives a canonical lift $H_1(\Sigma_g) \to H_2(M_\phi)$. In the absence of such auxiliary data, we instead just choose an *arbitrary* lift and we will account for the consequences later.

Having chosen an embedding $i\colon H_1(\Sigma_g) \hookrightarrow H_2(M_\phi)$ such that $z \mapsto \Sigma_z$, there is an associated direct sum decomposition of $H_2(M_\phi)$, namely

$$H_2(M_\phi) = \langle [F] \rangle \oplus \operatorname{Im} i.$$

Relative to such an embedding, we form the map $\tau(\phi) \in \operatorname{Hom}(\bigwedge^3 H_1(\Sigma_g), \mathbb{Z})$ by

$$\tau(\phi)(x \wedge y \wedge z) = \Sigma_x \cdot \Sigma_y \cdot \Sigma_z,$$

the term on the right being interpreted as the triple algebraic intersection of the given homology classes. Suppose a section exists and that the Σ_x have been constructed accordingly. In this case, D Johnson showed that the map

$$\begin{aligned} \tau\colon \mathcal{I}_{g,*} &\to \operatorname{Hom}\Big(\bigwedge^3 H_1(\Sigma_g), \mathbb{Z}\Big), \\ \phi &\mapsto \tau(\phi), \end{aligned}$$

is a surjective homomorphism. See [3, Chapter 6] for a summary of the Johnson homomorphism, including two alternative definitions. The (pointed) *Johnson kernel* $\mathcal{K}_{g,*}$

is defined, analogously to the case of closed surfaces, as the subgroup of $\mathrm{Mod}(\Sigma_{g,*})$ generated by Dehn twists about separating simple closed curves (scc). As in the closed case, D Johnson established that $\mathcal{K}_{g,*}$ coincides with the kernel of τ. In our context this precisely means that all triple intersections between the various Σ_x vanish.

Having fixed a family of Σ_x, it is then easy to compute the entire intersection form on $\bigwedge^3 H_2(M_\phi)$. Certainly $[F]^2 = 0$. It is also fairly easy to see that

$$[F] \cdot \Sigma_x \cdot \Sigma_y = i(x, y),$$

where $i(x, y)$ denotes the algebraic intersection pairing in $H_1(\Sigma_g)$. Indeed, by picking the choice of fiber to intersect Σ_x on the tube, it is clear that the result is simply the curve x, so that $[F] \cdot \Sigma_x \cdot \Sigma_y$ computes the intersection of x and y on F, at least up to a sign that may be introduced by the (non)compatibilities of the various orientation conventions in play. A quick check reveals this sign to be positive.

We will now be able to account for the ambiguity introduced by our choice of embedding $i\colon H_1(\Sigma_g) \hookrightarrow H_2(M_\phi)$, which will in turn lead to the definition of the Johnson homomorphism on the closed Torelli group $\mathcal{I}_g$. Suppose that $\Sigma'_w = \Sigma_w + k_w[F]$ is some other set of choices that is *coherent* in the sense that $\Sigma'_w + \Sigma'_z = \Sigma'_{w+z}$ (ie $x \mapsto k_x \in H^1(\Sigma_g)$). By linearity,

$$\begin{aligned}\Sigma'_x \cdot \Sigma'_y \cdot \Sigma'_z &= \Sigma_x \cdot \Sigma_y \cdot \Sigma_z + k_x i(y,z) + k_y i(z,x) + k_z i(x,y)\\ &= \tau(\phi)(x \wedge y \wedge z) + k_x i(y,z) + k_y i(z,x) + k_z i(x,y)\\ &= \tau(\phi)(x \wedge y \wedge z) + C^*(k);\end{aligned}$$

here $C\colon \bigwedge^3 H_1(\Sigma_g) \to H_1(\Sigma_g)$ is the contraction with the symplectic form $i(\cdot,\cdot)$ and $k \in \mathrm{Hom}(H_1(\Sigma_g), \mathbb{Z})$ is the form such that $k(w) = k_w$. The upshot of this calculation is that $\tau(\phi)$ is well defined as an element of $\mathrm{Hom}(\bigwedge^3 H_1(\Sigma_g), \mathbb{Z}) / \mathrm{Im}\, C^*$, which can be identified with the more familiar space $\bigwedge^3 H/H$ (here we adopt the usual convention that $H = H_1(\Sigma_g)$). The *Johnson homomorphism* on the closed Torelli group is then given by

$$\begin{aligned}\tau\colon \mathcal{I}_g &\to \mathrm{Hom}\big(\textstyle\bigwedge^3 H_1(\Sigma_g), \mathbb{Z}\big)/\,\mathrm{Im}\, C^* \approx \textstyle\bigwedge^3 H/H,\\ \phi &\mapsto \tau(\phi).\end{aligned}$$

As mentioned above, work of D Johnson shows that the kernel of τ coincides with the previously defined subgroup

$$\mathcal{K}_g = \langle T_\gamma \mid \gamma \text{ separating scc}\rangle.$$

Remark 5.1 The construction given above with the tube-and-cap surfaces is a concrete realization of the isomorphism $H_1(\Sigma_g) \approx H_2(M_\phi)/[F]$ coming from the Serre spectral sequence for $p\colon M_\phi \to S^1$. In fact, this same construction will work for an arbitrary $\phi \in \mathrm{Mod}(\Sigma_g)$, yielding an isomorphism $(H_1(\Sigma_g))^\phi \approx H_2(M_\phi)/[F]$, but we do not pursue this here.

The above discussion shows how to construct the Johnson homomorphism in terms of the intersection form on M_ϕ. Conversely, we will show next how to reconstruct the intersection form on M_ϕ from the data of the Johnson homomorphism $\tau(\phi) \in \bigwedge^3 H/H \approx \mathrm{Hom}(\bigwedge^3 H\Sigma_g, \mathbb{Z})/\operatorname{Im} C^*$. Begin by selecting an *arbitrary* lift $\tilde{\tau}(\phi)$ of $\tau(\phi)$ (of course, the presence of a section gives a canonical such choice). Next, construct a coherent family of homology classes Σ'_x by making choices arbitrarily. Define $\tau'(\phi) \in \mathrm{Hom}(\bigwedge^3 H, \mathbb{Z})$ by

$$\tau'(\phi)(x \wedge y \wedge z) = \Sigma'_x \cdot \Sigma'_y \cdot \Sigma'_z.$$

There is no reason to suspect that $\tau'(\phi) = \tilde{\tau}(\phi)$. However, as we saw above, we do know that $\tau'(\phi) - \tilde{\tau}(\phi) \in \operatorname{Im} C^*$, so there is some functional $\alpha \in H^1(\Sigma_g)$ such that $\tau'(\phi) - \tilde{\tau}(\phi) = C^*(\alpha)$. This functional α will allow us to choose the correct set of Σ_x so that the triple intersections are computed by our choice of $\tilde{\tau}(\phi)$.

Lemma 5.2 *We assume the notation of the above setting. By taking*

$$\Sigma_x = \Sigma'_x - \alpha(x)[F],$$

there is an equality for all x, y and z,

$$\Sigma_x \cdot \Sigma_y \cdot \Sigma_z = \tilde{\tau}(\phi)(x \wedge y \wedge z).$$

Proof We compute:

$$\begin{aligned}\Sigma_x \cdot \Sigma_y \cdot \Sigma_z &= \Sigma'_x \cdot \Sigma'_y \cdot \Sigma'_z - \alpha(x)i(y,z) - \alpha(y)i(z,x) - \alpha(z)i(x,y)\\ &= \tau'(\phi)(x \wedge y \wedge z) - C^*(\alpha)(x \wedge y \wedge z)\\ &= \tilde{\tau}(\phi).\end{aligned}$$

□

5.2 Intersections in surface bundles over surfaces, and beyond

The methods of the previous subsection can be adapted to give a description of certain cup products in $H^1(E)$, where $p\colon E^{n+2} \to B^n$ has monodromy lying in $\mathcal{I}_g$. The idea will be to define an embedding, as before,

$$i\colon H_1(\Sigma_g) \hookrightarrow H_{n+1}(E),$$

by constructing submanifolds M_γ for curves $\gamma \subset \Sigma_g$ by means of a higher-dimensional "tubing construction". Then the triple intersections of collections of $\mathcal{M}_\gamma$ will be partially computable via the Johnson homomorphism in a certain sense, to be described below. In this subsection we will first briefly sketch the properties we require of the submanifolds M_γ, then we will give the construction. Then, in Section 5.3, we will determine much of the intersection pairing in $H_*(E,\mathbb{Z})$.

Our construction will provide, for each simple closed curve $\gamma \subset F$, a submanifold M_γ such that if $[\gamma]=[\gamma']$ then also $[M_\gamma]=[M_{\gamma'}]$. If $[\gamma]=x$, we write M_x in place of $[M_\gamma]$. Let $p\colon E \to B$ be a surface bundle with monodromy in $\mathcal{I}_g$ and let $\rho\colon \pi_1 B \to \mathcal{I}_g$ be the monodromy. By post-composing with $\tau\colon \mathcal{I}_g \to \bigwedge^3 H/H$, we obtain a map from $\pi_1 B$ to an abelian group, so $\tau \circ \rho$ factors through $H_1(B)$. By an abuse of notation we will write $\tau(b)$ for $b \in H_1(B)$.

This map computes (most of) the intersection form in $H_*(E)$. Recall the notation from Proposition 3.1: given a curve $\alpha \subset B$, there is an induced bundle E_α over α, which determines a homology class E_a. A given M_γ can be intersected with E_α to yield a surface $\Sigma_{\alpha,\gamma}$ inside E_α. Our construction will be set up so that

$$M_x \cdot M_y \cdot M_z \cdot E_b = \tau(b)(x \wedge y \wedge z),$$

possibly up to a sign. This is the sense in which $M_x \cdot M_y \cdot M_z$ is partially computable. As an aside, the intersections $M_x \cdot M_y \cdot M_z \cdot X$ for arbitrary $X \in H_3 E$ will all involve intersections with further M_w and are describable (at least in the case of bundles with section) in terms of the higher Johnson invariants

$$\tau\colon H_i(\mathcal{I}_{g,*}) \to \bigwedge\nolimits^{i+2} H,$$

but we will not pursue this point of view further in this paper.

The construction As usual, let $\pi\colon E \to B$ be a surface bundle over a surface with monodromy $\rho\colon \pi_1 B \to \mathcal{I}_g$. We turn now to the question of constructing suitable homology classes $M_x \in H_3(E)$ for $x \in H_1(\Sigma_g)$. The construction will be a higher-dimensional analogue of the construction of tube-and-cap surfaces given in the previous subsection. The reader may find it helpful to consult Figure 2 as they read this subsection.

When the base space B has dimension 2, a new layer of complexity is introduced by the potential absence of sections $\sigma\colon B \to E$, which will require some additional preparatory work in order to construct geometric representatives for homology classes. Our construction method proceeds by exploiting the fact that it is always possible to find sections defined on $B' := \overline{B \setminus D^2}$. We define $E'\colon = \pi^{-1}(B')$ and refer to a section $\sigma\colon B' \to E'$ as a *partial section* of the bundle E. We say that two sections σ_0 and σ_1 of a fiber bundle are *homotopic through sections* if there exists a homotopy σ_t between σ_0 and σ_1 such that σ_t is a section for each fixed t.

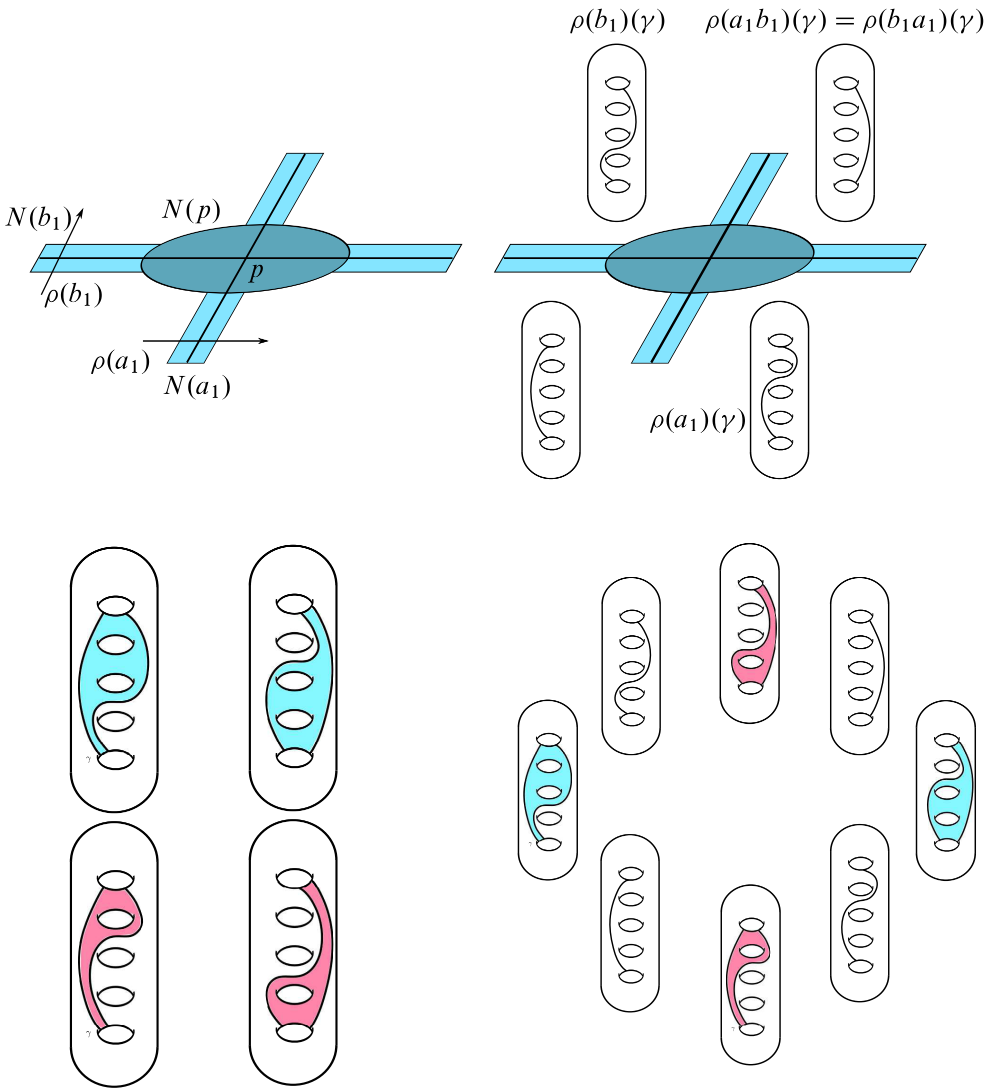


Figure 2: Upper left: the neighborhoods $N(e)$ and $N(p)$. Upper right: M_γ^1 intersected with four different fibers. Lower left: cap surfaces, lying over different portions of N. Lower right: a depiction of $M_\gamma^2 \cap \pi^{-1}(\partial N)$.

Lemma 5.3 *Let $\pi\colon E \to \Sigma_h$ be a surface bundle over a surface with monodromy $\rho\colon \pi_1\Sigma_h \to \mathrm{Mod}(\Sigma_g)$. Let $E' = \pi^{-1}(\overline{\Sigma_h \setminus D^2})$ and note that π restricts to give E' the structure of a Σ_g–bundle over $\overline{\Sigma_h \setminus D^2}$. Then there is a one-to-one correspondence between the set of classes of partial sections $\sigma\colon \overline{\Sigma_h \setminus D^2} \to E'$, up to homotopy through sections, and homomorphisms $\tilde{\rho}\colon F_{2h} \to \mathrm{Mod}(\Sigma_{g,*})$ making the diagram* *below commute:*

$$\begin{array}{ccccccccc} 1 & \longrightarrow & K & \longrightarrow & F_{2h} & \longrightarrow & \pi_1\Sigma_h & \longrightarrow & 1 \\ & & \Big\downarrow{\scriptstyle\tilde\rho} & & \Big\downarrow{\scriptstyle\tilde\rho} & & \Big\downarrow{\scriptstyle\rho} & & \\ 1 & \longrightarrow & \pi_1\Sigma_g & \longrightarrow & \mathrm{Mod}(\Sigma_{g,*}) & \longrightarrow & \mathrm{Mod}(\Sigma_g) & \longrightarrow & 1 \end{array}$$

Proof This follows immediately from the well-known fact that there is a homotopy equivalence

$$K(\mathrm{Mod}(\Sigma_{g,*}),1) \simeq B(\mathrm{Diff}(\Sigma_g,*)),$$

the latter space being the classifying space of Σ_g–bundles with section. □

The kernel $K \lhd F_{2h}$ is normally generated by a single element ω, represented geometrically by the boundary of $\overline{\Sigma_h \setminus D^2}$. The element $\tilde\rho(\omega) \in \pi_1\Sigma_g$ associated to a section σ will be denoted by ω_σ. It is called the *index curve*. The following lemma is immediate from the definitions.

Lemma 5.4 *Assume the notation of Lemma 5.3. Let σ be a partial section of E and let $\omega_\sigma \in \pi_1\Sigma_g$ be the corresponding index curve. Then there exists a local trivialization of E,*

$$t\colon \pi^{-1}(D^2) \to D^2 \times \Sigma_g,$$

relative to which $\sigma(\partial D^2)$ is in the free homotopy class of ω_σ.

The next lemma will be used in the course of the construction in Proposition 5.6.

Lemma 5.5 *Let $S \subset \Sigma_g \times S^1$ be an embedded, closed, oriented subsurface. Suppose $\gamma\colon S^1 \to \Sigma_g \times S^1$ is a section of the projection $\Sigma_g \times S^1 \to S^1$ and that $p_*[\gamma] = 0 \in H_1(\Sigma_g,\mathbb{Z})$ (where $p\colon \Sigma_g \times S^1 \to \Sigma_g$ is the obvious projection). Let $i\colon \Sigma_g \times S^1 \to \Sigma_g \times D^2$ be the natural inclusion. If the algebraic intersection number is $[\gamma]\cdot[S] = 0$ (computed in $\Sigma_g \times S^1$), then there exists an oriented, properly embedded 3–manifold $M \subset \Sigma_g \times D^2$ such that $\partial M = S$.*

Proof The first step is to establish that $i_*[S] = 0$ in $H_2(\Sigma_g \times D^2)$. The Künneth formula establishes natural splittings

$$H_1(\Sigma_g \times S^1) \approx H_1(\Sigma_g) \oplus H_1(S^1),$$
$$H_2(\Sigma_g \times S^1) \approx H_2(\Sigma_g) \oplus (H_1(\Sigma_g) \otimes H_1(S^1)).$$

In these coordinates, the map $i_*\colon H_2(\Sigma_g \times S^1) \to H_2(\Sigma_g \times D^2) \approx H^2(\Sigma_g)$ is given simply by projection onto the $H_2(\Sigma_g)$ factor. The assumptions on γ imply that $[\gamma]$

generates $H_1(S^1) \leq H_1(\Sigma_g \times S^1)$. Under the intersection pairing, $H_1(S^1)$ is orthogonal to $H_1(\Sigma_g) \otimes H_1(S^1)$. From the assumption $[\gamma] \cdot [S] = 0$, it then follows easily that $i_*[S] = 0$. Consequently, there exists a 3–chain C_p in $\Sigma_g \times D^2$ with $\partial C_p = S$.

It remains to explain why C_p can be replaced with a smooth, oriented, properly embedded 3–manifold. This will follow from general results on representing (relative) codimension-1 homology classes by smooth submanifolds (with boundary). The argument proceeds along very similar lines to the construction of embedded cap surfaces in fibered 3–manifolds described above. For an oriented manifold X with boundary, Lefschetz duality gives an isomorphism

$$H_{n-1}(X, \partial X, \mathbb{Z}) \approx H^1(X, \mathbb{Z}) \approx [X, S^1].$$

In our setting, the surface $S \subset \Sigma_g \times S^1$ is represented by a map

$$f\colon \Sigma_g \times S^1 \to S^1$$

such that $S = f^{-1}(*)$ for some regular value $* \in S^1$. Similarly, the (relative) homology class of C_p in $H_3(\Sigma_g \times D^2, \Sigma_g \times S^1, \mathbb{Z})$ corresponds to a map

$$F\colon \Sigma_g \times D^2 \to S^1.$$

Moreover, as $\partial C_p = S$, they represent the same homology class in $H_2(\Sigma_g \times S^1, \mathbb{Z})$. This means that the maps f and $F|_{\Sigma_g \times S^1}$ are homotopic. We can therefore concatenate this homotopy with F to obtain a map

$$\widetilde{F}\colon \Sigma_g \times D^2 \to S^1.$$

On the boundary, $\widetilde{F}$ equals f and is therefore transverse to $* \subset S^1$. In order to replace C_p by a smooth submanifold such that $\partial C_p = C$, we must therefore perturb $\widetilde{F}$ away from a neighborhood of $\partial(\Sigma_g \times D^2)$ and make the result everywhere transverse to $* \subset S^1$. The extension theorem (see [4, page 72]) asserts that we can do precisely this. □

The theory of index curves established above will allow us to construct embedded representatives of homology classes in surface bundles over surfaces when suitable conditions on the monodromy are satisfied.

Proposition 5.6 *Let $\pi\colon E \to B$ be a surface bundle over a surface with monodromy $\rho\colon \pi_1 B \to \mathcal{I}_g$ contained in the Torelli group. Suppose there is a partial section $\sigma\colon B' \to E'$ for which the associated index curve ω_σ lies in the commutator subgroup $[\pi_1\Sigma_g, \pi_1\Sigma_g]$. Then there is an embedding*

$$\iota\colon H_1(F, \mathbb{Z}) \to H_3(E, \mathbb{Z})$$

constructed so that, if $c \in H_1(F, \mathbb{Z})$ is a primitive class, then $\iota(c)$ can be represented by some embedded, oriented, piecewise-smooth 3–submanifold M_c of E.

Proof Let $c \in H_1(F, \mathbb{Z})$ be given. By assumption, c is primitive, so that there exists a simple closed curve $\gamma \subset \Sigma_g$ with $[\gamma] = c$. We will use this to construct a 3–manifold M_γ.

Consider a cell decomposition

$$B = B^0 \subset B^1 \subset B^2$$

of B, where B^0 consists of the single point p, there are $2g$ one-cells $\{a_1, b_1, \dots, a_h, b_h\}$ and a single two-cell D. For each one-cell e, there is an associated element $\rho(e)$ of the monodromy such that the effect of transporting a curve γ across e (from the negative to the positive side, relative to orientations of B and e) sends the isotopy class of γ to $\rho(e)\gamma$. For a one-cell e, let $N(e) \approx e \times I$ be a (closed) regular neighborhood in B. We also let $N(p)$ be a small closed neighborhood of p. If necessary, shrink $N(e)$ so that

$$N := N(a_1) \cup \cdots \cup N(b_h) \setminus N(p)$$

is a union of $2h$ disjoint rectangles.

Let $\gamma \subset F$ be a simple closed curve on a fiber F over a point in

$$D' := \overline{D \setminus (N(p) \cup N(a_1) \cup \cdots \cup N(b_h))}.$$

By construction, D' is nothing more than a closed disk (in the upper-left portion of Figure 2, D' is the closure of the complement of the shaded regions). The submanifold M_γ will be constructed in three stages, denoted by M_γ^i for $i = 1, 2, 3$: first over D', then over N and finally over $N(p)$. Choose a trivialization $\pi^{-1}(D') \approx D' \times F$ and define $M_\gamma^1 = \gamma \times D'$ relative to this trivialization. Then $\partial(M_\gamma^1) \subset \pi^{-1}(\partial D')$. We specify an orientation on M_γ^1 as follows: a point $x \in M_\gamma^1$ has a decomposition of the tangent space

$$T_x M_\gamma^1 \approx T_{\pi(x)} B \oplus T_x \gamma. \tag{4}$$

Both of these two summands carry pre-existing orientations and M_γ^1 is then oriented by specifying the above isomorphism to be orientation-preserving. By analogy with the construction of tube surfaces, we refer to M_γ^1 as the *tube region* of M_γ.

Next we construct M_γ^2. Let e be a one-cell and consider $M_\gamma^1 \cap \pi^{-1}(N(e) \cap N)$. The base space $N(e) \cap N$ is just a rectangle, so the bundle $\pi^{-1}(N(e) \cap N)$ is trivializable. We can therefore find a diffeomorphism

$$\psi : \pi^{-1}(N(e) \cap N) \approx I \times I \times \Sigma_g$$

under which $M^1_\gamma \cap \pi^{-1}(N(e)\cap N)$ is identified with

$$(I\times\{0\}\times\gamma)\cup(I\times\{1\}\times\gamma'),$$

where γ' is some curve in the isotopy class of $\rho(e)(\gamma)$. As we saw in the previous subsection, for each e there exists a family of properly embedded surfaces S_e in $I\times\Sigma_g$ such that $\partial S_e=\{0\}\times\gamma\cup\{1\}\times\gamma'$.

Our choice of S_e will be dictated by the section σ. Applying ψ, the image of σ in $\{t\}\times I\times\Sigma_g$ is a properly embedded arc α_σ. This determines a preferred homology class in $H_2(I\times\Sigma_g,\partial(I\times\Sigma_g),\mathbb{Z})$ among the set of possible S_e, by the relation $[\alpha_\sigma]\cdot[S_e]=0$.

Let S_e be any properly embedded subsurface of $I\times\Sigma_g$ satisfying the conditions $\partial S_e=\{0\}\times\gamma\cup\{1\}\times\gamma'$ and $[\alpha_\sigma]\cdot[S_e]=0$. We can then fill in $\pi^{-1}(N(e)\cap N)$ with $I\times S_e$ for each e, creating M^2_γ. As in the case of a tube surface, the orientation for M^1_γ can be extended over each of these pieces coherently. We refer to $M^2_\gamma\setminus M^1_\gamma$ as the *cap region* of M_γ.

It therefore remains to construct $M^3_\gamma=M_\gamma$. By construction, $\partial M^2_\gamma\subset\pi^{-1}(\partial N(p))$. We would like to be able to fill this boundary in by inserting a "plug" contained in $\pi^{-1}(N(p))$. A priori, there is a homological obstruction to this: if $[\partial M^2_\gamma]\neq 0$ in $H_2(\pi^{-1}(N(p)))$ then this problem is not solvable even on the chain level.

However, the assumption that the index curve ω_σ is in $[\pi_1\Sigma_g,\pi_1\Sigma_g]$ will imply that this obstruction vanishes. Let $t\colon \pi^{-1}(N(p))\to D^2\times\Sigma_g$ be the trivialization of Lemma 5.4 and define $\gamma=t(\sigma(\partial(N(p))))$. Set $S=t(\partial(M^2_\gamma))$. By Lemma 5.4, $[\gamma]=0\in H_1(\pi^{-1}(N(p)))\approx H_1(\Sigma_g)$. We wish to show that $[\gamma]\cdot[S]=0$. By construction, $\partial(M^2_\gamma)$ consists of $4g$ subsurfaces, corresponding to the $2g$ surfaces $S_{a_1},\dots,S_{b_g}$, each appearing twice (once for each component of $N(e)\cap N(p)$). Similarly, γ is comprised of $4g$ segments, again indexed by the components of $N(e)\cap N(p)$. On each one of these components, the relevant S_e was selected to have zero algebraic intersection with the relevant portion of γ, so the same holds true globally: $[\gamma]\cdot[S]=0$.

Applying Lemma 5.5, we obtain a 3–manifold $M_p\subset N(p)\times\Sigma_g$ with $\partial M_p=t(\partial(M^2_\gamma))$. Extending the orientation of M^2_γ over M_p, the result is an oriented, piecewise-smooth submanifold $M_\gamma\subset E$. □

Remark 5.7 It is apparent in the above construction that if γ and γ' are homologous curves, the associated 3–manifolds M_γ and $M_{\gamma'}$ are homologous. Accordingly, if $[\gamma]=[\gamma']=x$, we adopt the notation $M_x=[M_\gamma]=[M_{\gamma'}]$.

While, in general, not every surface bundle over a surface satisfies the hypotheses of Proposition 5.6 (specifically the requirement that there exist a partial section with

$[\omega_\sigma] = 0 \in H_1(\Sigma_g, \mathbb{Z})$), it turns out that this is always the case for surface bundles over surfaces with monodromy in $\mathcal{K}_g$.

Lemma 5.8 *Let $\rho\colon \pi_1\Sigma_h \to \mathcal{K}_g$ be given. Then, for any lift $\tilde{\rho}\colon F_{2h} \to \mathcal{K}_{g,*}$ of ρ, the index curve satisfies $\omega_\sigma \in [\pi_1\Sigma_g, \pi_1\Sigma_g]$.*

Proof When restricted to $\mathcal{K}_g$, the Birman exact sequence takes the form

$$1 \to [\pi_1\Sigma_g, \pi_1\Sigma_g] \to \mathcal{K}_{g,*} \to \mathcal{K}_g \to 1.$$

The result follows. □

An essential feature of the above construction is the relationship between an M_γ and a sub-bundle E_α lying over a curve $\alpha \subset B$. Suppose α is chosen so that, relative to the cell decomposition of B used in constructing M_γ, α is transverse to all the one-cells e and does not pass through $N(p)$. Then a little visual imagination reveals that the intersection of M_γ and E_α is given by a tube surface for γ sitting inside E_α. We call the resulting surface $\Sigma_{\alpha,\gamma}$ and then $[\Sigma_{\alpha,\gamma}]$ is denoted by $\Sigma_{a,x}$, where $[\alpha] = a$ and $[\gamma] = x$.

We define a *family of* M_x to be a choice of M_x for each $x \in H_1(F)$ such that, for all $c \in \mathbb{Z}$ and x, $y \in H_1(F)$,

$$M_{cx+y} = cM_x + M_y.$$

Different choices of M_x lead to different spaces of $\Sigma_{b,x}$ but, conversely, a choice of a family of M_x leads to a corresponding distinguished summand of $H_2(E)$.

5.3 Determination of the intersection form

From this point onwards, we assume without further comment that our surface bundle over a surface, $\pi\colon E \to B$, satisfies the hypotheses of Proposition 5.6 (as a special case, these results apply to all surface bundles over surfaces with monodromy in $\mathcal{K}_g$, by Lemma 5.8). The purpose of this subsection is to give a description of the cup product structure on $H^*(E, \mathbb{Z})$; equivalently, we will describe the intersection form. By Poincaré duality, it suffices to determine, for each X, the set of pairings $X \cdot Y$.

Proposition 5.9 *Let i_B and i_F denote the algebraic intersection pairing on the homology of the base and on the fiber, respectively.*

(1) *There exists a unique class $C \in H_2(E)$ such that $C \cdot [F] = 1$ and $C \cdot \Sigma_{b,z} = 0$ for all $b \in H_1(B)$ and $z \in H_1(\Sigma_g)$. The intersection pairing $H_2(E) \otimes H_2(E) \to \mathbb{Z}$*

is given as follows, where $e = C^2$ by definition:

	C	$[F]$	$\Sigma_{a,z}$
C	e	1	0
$[F]$	1	0	0
$\Sigma_{b,w}$	0	0	$-i_B(a,b)i_F(z,w)$

In the case where the monodromy is contained in the Johnson kernel, we have $e = 0$.

(2) *For any family of M_x, we have*

$$E_a \cdot E_b = i_B(a,b)[F],$$
$$M_x \cdot E_b = \Sigma_{b,x},$$
$$M_z \cdot M_w \cdot [F] = i_F(z,w).$$

(3) *Let $\sigma\colon B' \to E'$ be a partial section for which $[\omega_\sigma] = 0 \in H_1(F)$. Associated to such a section is a lift of $\tau\colon H_1(B) \to \bigwedge^3 H/H$ to $\tilde{\tau}\colon H_1(B) \to \bigwedge^3 H$. The choice of σ gives rise to a splitting*

$$H_3(E) = \pi^!(H_1(B)) \oplus H_1(M) = \{E_b, b \in H_1(B)\} \oplus \{M_z, z \in H_1(F)\}$$

relative to which

$$M_x \cdot M_y \cdot M_z \cdot E_b = M_x \cdot M_y \cdot \Sigma_{b,z} = \tilde{\tau}(b)(x \wedge y \wedge z).$$

In the case where the monodromy is contained in the Johnson kernel, we can take the canonical lift $\tilde{\tau} \equiv 0$ and, for this family of M_x, we have

$$C \cdot M_x = 0 \quad \text{and} \quad C^2 = 0$$

for all $x \in H_1(\Sigma_g)$.

Remark The intersection pairing $H_{n-k}E \otimes H_k E \to \mathbb{Z}$ identifies $H_{n-k}E$ with $\mathrm{Hom}(H_k E, \mathbb{Z})$ and hence with $H^k E$ by the universal coefficients theorem, since the homology of a surface bundle over a surface with monodromy in $\mathcal{I}_g$ is torsion-free (see Proposition 3.1). Therefore, Proposition 5.9 can also be viewed as a description of the cup product in $H^*(E)$.

Proof Before beginning the proof of the statements, a comment on orientations is in order. Recall that if X and Y are embedded surfaces intersecting transversely, then $X \cap Y$ is oriented via the convention that

$$N(X) \oplus N(Y) \oplus T(X \cap Y)$$

should be positively oriented, where, for $W = X$ or $W = Y$, $N(W)$ is oriented by the convention that $N(W) \oplus T(W)$ be positively oriented with respect to the orientation fixed on W. Note that relative to this convention, if X is of odd codimension, then $X \cdot X = 0$; we will often employ this fact without comment in the sequel.

Recall that the submanifolds $\Sigma_x \subset M_\phi$ and $M_z \subset E$ have been oriented using a "base first" convention; see (3) and (4). As remarked already in the proof of Proposition 4.1, E itself is oriented by selecting orientations for B and F. It is a somewhat tedious process to go through and verify the signs on all of the intersections being asserted in this theorem, so we omit the full verification of these results. At the same time, the reader who is interested in verifying the calculations should have no trouble doing so by carefully tracking the orientation conventions we have laid out.

It will turn out to be most natural to construct C after verifying the statements not involving C. We begin with computing $\Sigma_{a,z} \cdot \Sigma_{b,w}$. These are represented by surfaces contained in some E_α and E_β, respectively, where they are tube surfaces constructed from curves γ and δ. We can arrange it so that α and β intersect transversely and such that, over these points, the surfaces intersect in their tube regions. Following the orientation conventions as above, one verifies that the local intersection at such a point (p,q), written $I_{(p,q)}$, is equal to $-I_p I_q$, where I_p denotes the local intersection of α and β relative to the orientation on B and I_q is the local intersection of γ and δ relative to the orientation on F. Summing over all local intersections gives the result in the lower right-hand corner of the table in Proposition 5.9(1).

The relation $[F] \cdot \Sigma_{a,z} = 0$ is easy to verify, by taking $[F]$ to be represented by a fiber not contained in the E_α containing $\Sigma_{a,z}$. This same idea also shows $[F]^2 = 0$, by picking representative fibers over distinct points.

Let us turn now to Proposition 5.9(2). If E_α and E_β intersect transversely at a point, then $E_\alpha \cap E_\beta = F$, the fiber over the point of intersection; a check of the orientation conventions shows that the orientation on F given by the intersection convention agrees with the predetermined orientation, so that

$$E_a \cdot E_b = i_B(a,b)[F],$$

as asserted.

The manifolds M_γ were constructed so as to intersect each E_b in a tube surface, so the relation

$$M_z \cdot E_b = \Sigma_{b,z}$$

can be taken as a definition of the orientation on $\Sigma_{b,z}$. We choose this over the alternative because it can be verified that, under this convention, the orientation on $\Sigma_{b,z}$ agrees with the "base first" convention discussed above.

Now let M_x and M_y be given and consider $M_x \cdot M_y \cdot [F]$. By perturbing the one-skeleton of B, it can be arranged so that the plugs for M_x and M_y are disjoint, the cap regions intersect transversely and the representative fiber intersects M_x and M_y in their tube regions. The local picture therefore becomes the intersection of x and y on F. A check of the orientation convention then shows

$$M_x \cdot M_y \cdot [F] = i_F(x, y).$$

Turning to Proposition 5.9(3), consider now a four-fold intersection

$$M_x \cdot M_y \cdot M_z \cdot [E_\beta].$$

We will assume without further comment that the intersection of representative submanifolds has been made suitably transverse by choosing one-skeleta wisely. The M_w were constructed so that the problem of computing $M_x \cdot M_y \cdot M_z \cdot [E_\beta]$ is exactly the same as the problem of computing the corresponding $\Sigma_x \cdot \Sigma_y \cdot \Sigma_z$ inside the 3–manifold E_β, up to a sign which records whether the orientation on $M_x \cdot [E_\beta]$ agrees with the orientation on the corresponding $\Sigma_x \subset E_\beta$; the convention $M_x \cdot E_b = \Sigma_{x,b}$ makes this sign positive. Lemma 5.2 shows that, within E_b, there exist choices of homology classes Σ_x such that

$$\Sigma_x \cdot \Sigma_y \cdot \Sigma_z = \tilde{\tau}(b)(x \wedge y \wedge z).$$

Recall from Lemma 5.2 that the Σ_x are obtained by starting with an arbitrary family Σ'_x and adding appropriate multiples of $[F]$. By the preceding, if $a \in B$ satisfies $i_B(a, b) = 1$, then

$$(M_z + E_a) \cdot E_b = M_z \cdot E_b + [F].$$

This shows that, by adding appropriate multiples of E_a to M_z (as specified by the formulas in Lemma 5.2), for a given b the formula

$$M_x \cdot M_y \cdot M_z \cdot [E_\beta] = \tilde{\tau}(b)(x \wedge y \wedge z) \tag{5}$$

can be made to hold. By choosing a symplectic basis for $H_1(B)$, this can be made to hold for all $b \in H_1(B)$ simultaneously.

It therefore remains to construct the class C. If x, $y \in H_1(\Sigma_g)$ satisfy $i_F(x, y) = 1$, then $[F] \cdot M_x \cdot M_y = 1$. Similarly, if α and β are loops in B intersecting transversely exactly once and M_x and M_y are as above, then

$$\Sigma_{\alpha,x} \cdot \Sigma_{\beta,y} = \Sigma_{\alpha,x} \cdot M_x \cdot E_\beta = \pm 1. \tag{6}$$

As the space spanned by $[F]$ and the $\Sigma_{b,x}$ classes has codimension one in $H_2(E)$, (5) and (6) together show that the space of classes in $H_2(E)$ pairing trivially with the space of M_x has dimension at most one. We claim that

$$C = M_{x_1} \cdot M_{y_1} + \sum_{(b,z)\in\mathcal{B}\times\mathcal{F}} \tilde{\tau}(b)(x_1 \wedge y_1 \wedge z)\Sigma_{\hat{b}\hat{z}}$$

has all the required properties; here, $\mathcal{B}$ and $\mathcal{F}$ are symplectic bases for $H_1(B)$ and $H_1(F)$, respectively, the map $x \mapsto \hat{x}$ satisfies $i(x,\hat{x}) = 1$, $x_1 \in \mathcal{B}$ and $\hat{x}_1 = y_1$. Recall that C is asserted to have the following properties: $C \cdot [F] = 1$ and $C \cdot \Sigma_{b,z} = 0$ for all $b \in H_1(B)$ and $z \in H_1(\Sigma_g)$. Additionally, when the monodromy of E is contained in the Johnson kernel, we require $C^2 = 0$ and $C \cdot M_x = 0$ for M_x in the family associated to the lift of τ to the zero homomorphism. The proof is a direct calculation. For $C \cdot [F]$, one has, by Proposition 5.9(1) and then Proposition 5.9(2),

$$C \cdot [F] = \left(M_{x_1} \cdot M_{y_1} + \sum_{(b,z)\in\mathcal{B}\times\mathcal{F}} \tilde{\tau}(b)(x_1 \wedge y_1 z)\Sigma_{\hat{b}\hat{z}} \right) \cdot [F] = M_{x_1} \cdot M_{y_1} \cdot [F] = 1.$$

Computation of $C \cdot \Sigma_{b,z}$ proceeds by Proposition 5.9(3) and Proposition 5.9(1), respectively:

$$\begin{aligned} C \cdot \Sigma_{b,z} &= M_{x_1} \cdot M_{y_1} \cdot \Sigma_{b,z} + \tilde{\tau}(b)(x_1 \wedge y_1 \wedge z)(\Sigma_{\hat{b}\hat{z}}) \cdot \Sigma_{b,z} \\ &= \tilde{\tau}(b)(x_1 \wedge y_1 \wedge z) - \tilde{\tau}(b)(x_1 \wedge y_1 \wedge z) \\ &= 0. \end{aligned}$$

When the monodromy of E is contained in $\mathcal{K}_g$, the above formula for C simplifies to $C = M_{x_1} \cdot M_{y_1}$, from which it is apparent that $C^2 = 0$. To see that $C \cdot M_x = 0$ for all x, we will apply Poincaré duality to see that it suffices to show that

$$C \cdot M_x \cdot Y = 0$$

for all classes $Y \in H_3 E$. Since $M_x \cdot E_b = \Sigma_{bx}$ and we have shown $C \cdot \Sigma_{bx} = 0$, it remains only to consider $C \cdot M_z \cdot M_w$. Expanding $M_z \cdot M_w$ in the additive basis for $H_2(E)$,

$$M_z \cdot M_w = \alpha[F] + \beta C + \sum_{(b,z)\in\mathcal{B}\times\mathcal{F}} \gamma_{b,z} \Sigma_{\hat{b},\hat{z}}.$$

As the monodromy of E is contained in $\mathcal{K}_g$, we have $M_z \cdot M_w \cdot \Sigma_{b,x} = 0$; applying this in coordinates for some $(b,x) \in \mathcal{B} \times \mathcal{F}$ gives, by applying the prior formulas,

$$0 = \left(\alpha[F] + \beta C + \sum_{(b,z)\in\mathcal{B}\times\mathcal{F}} \gamma_{b,z} \Sigma_{\hat{b},\hat{z}} \right) \cdot \Sigma_{b,x} = -\gamma_{b,x},$$

so that all $\gamma_{b,z}$ are 0. Consequently, $M_z \cdot M_w = \alpha[F] + \beta C$. Recalling that $[F]^2 = C^2 = 0$ and that $(M_z \cdot M_w)^2 = 0$, this implies $\alpha\beta = 0$.

Also,

$$i_F(z,w) = M_z \cdot M_w \cdot [F] = \beta.$$

Therefore, we conclude that, in the case $i_F(z,w) \neq 0$,

$$M_z \cdot M_w = i_F(z,w)C.$$

As $C^2 = 0$, this shows the result in this case. Now suppose that $i_F(z,w) = 0$. Then we can find z' such that $M_z \cdot M_{z'} = cC$ by the above, with $c \neq 0$, then

$$0 = M_z \cdot M_w \cdot M_z \cdot M_{z'} = cM_z \cdot M_w \cdot C.$$

This shows that $M_z \cdot M_w \cdot C = 0$ for all z and w, finishing the proof of Proposition 5.9. □

6 Multisections and splittings on rational cohomology

Let $p\colon E \to B$ be a surface bundle over an arbitrary base space B equipped with a section $\sigma\colon B \to E$. Then there is an associated splitting of $H^1(E,\mathbb{Z})$ as a direct sum,

$$H^1(E,\mathbb{Z}) = \operatorname{Im} p^* \oplus \ker \sigma^*. \tag{7}$$

The condition that $p\colon E \to B$ admit a section is restrictive. However, recent work of Hamenstädt shows that all surface bundles over surfaces with zero signature admit *multisections* (see Theorem 6.2). In this section, we develop some necessary machinery showing how a multisection of a surface bundle gives rise to a splitting of $H^1(E,\mathbb{Q})$, similarly to (7). The results of this section will be required in the proof of Theorem 1.1.

Remark 6.1 Theorem 6.2 is the only result in this section that requires the base space B to be a surface of genus $g \geq 2$. Lemma 6.3 and Proposition 6.4 are valid for any base space B.

Let $\mathrm{Conf}_n(E)$ denote the configuration space of n unordered distinct points in E and let $\mathrm{PConf}_n(E)$ denote the space of n ordered distinct points in E. The symmetric group S_n on n letters acts freely on $\mathrm{PConf}_n(E)$ by permuting the order of the points, and $\mathrm{PConf}_n(E)/S_n = \mathrm{Conf}_n(E)$.

By a *multisection* of $p\colon E \to B$, we mean a map

$$\sigma\colon B \to \mathrm{Conf}_n(E)$$

for some $n \geq 1$ such that the composition

$$B \to \mathrm{Conf}_n(E) \to B^n/S_n$$

is given by $x \mapsto [x, \dots, x]$. In other words, a multisection selects n distinct unordered points in each fiber. A *pure multisection* is a map

$$\sigma\colon B \to \mathrm{PConf}_n(E)$$

such that the composition

$$B \to \mathrm{PConf}_n(E) \to B^n$$

is given by $x \mapsto (x, \dots, x)$. Our interest in multisections is due to the following result of Hamenstädt (see [5]; also personal communication, 2015):

Theorem 6.2 (Hamenstädt) *Let $p\colon E \to B$ be a surface bundle over a surface such that the signature of E is zero (eg a bundle with at least one fibering with monodromy lying in $\mathcal{I}_g$). Then $p\colon E \to B$ has a multisection σ of cardinality $2g-2$.*

We will use this result to obtain a splitting on $H^*(E, \mathbb{Q})$. As (7) indicates, this is straightforward when the multisection is pure; the work will be to obtain the required maps for general multisections. First note that, by taking a finite cover $\widetilde{B} \to B$, we can pull the bundle back to $\tilde{p}\colon \widetilde{E} \to \widetilde{B}$ so that the multisection pulls back to a pure multisection

$$\psi\colon \widetilde{B} \to \mathrm{PConf}_n(\widetilde{E}).$$

Moreover, we can assume that the covering $\widetilde{B} \to B$ is normal with deck group Γ. By pulling back the Γ action on $\widetilde{B}$, we see that Γ also acts on $\widetilde{E}$, by sending the fiber over b to the fiber over $\gamma(b)$. Then the multisection ψ is in fact Γ–equivariant. This suggests the following lemma:

Lemma 6.3 *Let $\tilde{\sigma}\colon \widetilde{B} \to \widetilde{E}$ be a Γ–equivariant section. Then there is an induced map on Γ–invariant cohomology:*

$$\tilde{\sigma}^*\colon H^*(\widetilde{E}, \mathbb{Q})^{\Gamma} \to H^*(\widetilde{B}, \mathbb{Q})^{\Gamma}.$$

As a result, the transfer map

$$\tau^*\colon H^*(\widetilde{B}, \mathbb{Q}) \to H^*(B, \mathbb{Q})$$

is injective when restricted to $\tilde{\sigma}^*(H^*(\widetilde{E}, \mathbb{Q})^{\Gamma})$.

Proof If $f\colon X \to Y$ is any Γ–equivariant map of topological spaces, then the map $f^*\colon H^*(Y) \to H^*(X)$ will be equivariant, so will restrict to a map on the Γ–invariant subspaces. Transfer (see [6]) gives an identification $H^*(\widetilde{B}, \mathbb{Q})^{\Gamma} \approx H^*(B, \mathbb{Q})$ and the remaining statement follows. □

We now come to the main result of the section. This asserts that, when $p\colon E \to B$ is a surface bundle with a multisection $\sigma : B \to \mathrm{Conf}_n(E)$, there exists a map $\widehat{\sigma}^*\colon H^*(B,\mathbb{Q}) \to H^*(E,\mathbb{Q})$ with many of the same properties as (the pullback of) an actual section map.

Proposition 6.4 *Suppose $\sigma\colon B \to \mathrm{Conf}_n(E)$ is a multisection. Then there exist maps*

$$\widehat{\sigma}^*\colon H^*(E,\mathbb{Q}) \to H^*(B,\mathbb{Q}),$$
$$\widehat{\sigma}_*\colon H_*(B,\mathbb{Q}) \to H_*(E,\mathbb{Q}),$$

with the following properties:

(1)
$$\widehat{\sigma}^* \circ p^*\colon H^*(B) \to H^*(B) = \mathrm{id},$$
$$p_* \circ \widehat{\sigma}_*\colon H_*(B) \to H_*(B) = \mathrm{id}\,.$$

(2) *The maps $\widehat{\sigma}^*$ and $\widehat{\sigma}_*$ are adjoint under the evaluation pairing. That is, for all $\alpha \in H^*(E)$ and $x \in H_*(B)$,*

$$\langle \alpha, \widehat{\sigma}_* x\rangle = \langle \widehat{\sigma}^*\alpha, x\rangle.$$

(3) *If $\alpha \in \ker\widehat{\sigma}^*$ then, for any $\beta \in H^*(E,\mathbb{Q})$ and any $x \in H_*(B,\mathbb{Q})$,*

$$\langle \alpha \smile \beta, \widehat{\sigma}_*(x)\rangle = 0.$$

Consequently, $\widehat{\sigma}^*$ induces a splitting

$$H^1(E,\mathbb{Q}) = \mathrm{Im}\, p^* \oplus \ker\widehat{\sigma}^*. \tag{8}$$

Proof Begin by assuming that the multisection is pure. Let $p_i\colon \mathrm{PConf}_n(E) \to E$ be the projection onto the i^{th} coordinate for $i = 1, \dots, n$. We define

$$\widehat{\sigma}^*(\alpha) = \frac{1}{n}\sum_{i=1}^{n} \sigma^*(p_i^*(\alpha)),$$
$$\widehat{\sigma}_*(x) = \frac{1}{n}\sum_{i=1}^{n} (p_i)_*(\sigma_*(x)).$$

Then properties (1)–(3) follow by direct verification.

In the general case, let $c\colon \widetilde{B} \to B$ be a normal covering such that σ pulls back to a pure multisection ψ. We will use $\bar{c}$ to denote the covering $\widetilde{E} \to E$. Let $\tau^*\colon H^*(\widetilde{B},\mathbb{Q}) \to H^*(B,\mathbb{Q})$ be the transfer map, normalized so that $c^* \circ \tau^* = \mathrm{id}$. Then define $\widehat{\sigma}^*\colon H^*(E,\mathbb{Q}) \to H^*(B,\mathbb{Q})$ by

$$\widehat{\sigma}^* = \tau^* \circ \widehat{\psi}^* \circ \bar{c}^*.$$

Similarly, define $\widehat{\sigma}_*\colon H_*(B,\mathbb{Q}) \to H_*(E,\mathbb{Q})$ by

$$\widehat{\sigma}_* = \bar{c}_* \circ \widehat{\psi}_* \circ \tau_*.$$

For what follows, it will be useful to refer to the following diagram:

$$\begin{tikzcd}
H^*(\widetilde{E}) \arrow[r, shift left, "\tau^*"] \arrow[d, shift left, "\widehat{\psi}^*"] & H^*(E) \arrow[l, shift left, "\bar{c}^*"] \arrow[d, dotted, shift left, "\widehat{\sigma}"] \\
H^*(\widetilde{B}) \arrow[r, shift left, "\tau^*"] \arrow[u, shift left, "\tilde{p}^*"] & H^*(B) \arrow[l, shift left, "c^*"] \arrow[u, shift left, "p^*"]
\end{tikzcd}$$

By definition,

$$\widehat{\sigma}^* \circ p^* = \tau^* \circ \widehat{\psi}^* \circ \bar{c}^* \circ p^*.$$

By commutativity, $\bar{c}^* \circ p^* = \tilde{p}^* \circ c^*$. Then

$$\tau^* \circ \widehat{\psi}^* \circ \bar{c}^* \circ p^* = \tau^* \circ \widehat{\psi}^* \circ \tilde{p}^* \circ c^* = \tau^* \circ c^* = \mathrm{id}.$$

Here we have used the property $\widehat{\psi}^* \circ \tilde{p}^* = \mathrm{id}$ for the pure multisection ψ as well as our normalization convention $\tau^* \circ c^* = \mathrm{id}$ for the transfer map. A similar calculation proves the corresponding result for $\widehat{\psi}_*$ and (1) follows.

Statement (2) follows from the observation that the cohomology and homology transfer maps are adjoint under the evaluation pairing. That is, if $\widetilde{X} \to X$ is a normal covering space with deck group Γ then, for $x \in H_*(X)$ and $\alpha \in H^*(\widetilde{X})$,

$$\langle \alpha, \tau_*(x) \rangle = \langle \tau^*(\alpha), x \rangle.$$

As $\widehat{\psi}^*$ and $\bar{c}^*$ certainly also enjoy this adjointness property, so does $\widehat{\sigma}^*$, and (2) follows.

To establish (3), suppose $\alpha \in \ker \widehat{\sigma}^*$ and take $\beta \in H^*(E,\mathbb{Q})$ and $x \in H_*(B,\mathbb{Q})$. As the transfer map is not a ring homomorphism, (3) does not follow immediately from (2). However, we see that

$$\langle \alpha \smile \beta, \widehat{\sigma}_*(x) \rangle = \langle \widehat{\sigma}^*(\alpha \smile \beta), x \rangle = \big\langle \tau^*((\widehat{\psi}^* \circ \bar{c}^*)(\alpha) \smile (\widehat{\psi}^* \circ \bar{c}^*)(\beta)), x \big\rangle.$$

It therefore suffices to show that $\widehat{\psi}^* \circ \bar{c}^*(\alpha) = 0$. This follows from Lemma 6.3. Indeed, $\bar{c}^*(\alpha) \in H^*(\widetilde{E},\mathbb{Q})^\Gamma$ and $\widehat{\psi}^*$, being a sum of Γ–equivariant maps, is itself Γ–equivariant, so $\widehat{\psi}^* \circ \bar{c}^*$ takes image in $H^*(\widetilde{B},\mathbb{Q})^\Gamma$. On the one hand, we have

$$0 = \widehat{\sigma}^* \alpha = \tau^* \circ \widehat{\psi}^* \circ \bar{c}^*(\alpha)$$

by assumption. Also, by Lemma 6.3, τ^* is injective on the image of $\widehat{\psi}^* \circ \bar{c}^*$, so that $\widehat{\psi}^* \circ \bar{c}^*(\alpha) = 0$ as desired. $\square$

7 Unique fibering in the Johnson kernel

This section is devoted to the proof of Theorem 1.1. The outline is as follows. Let $p_1\colon E \to B_1$ be a surface bundle with monodromy in the Torelli group $\mathcal{I}_g$ and suppose there is a second distinct fibering $p_2\colon E \to B_2$ with fiber F_2. The proof proceeds by analyzing $[F_2]$ in the coordinates on $H_*(E)$ coming from the Torelli fibering p_1. On the one hand, the intersection form in these coordinates is completely understood by virtue of Proposition 5.9. On the other, $[F_2]$ is realizable as an intersection of classes induced from $H_1(B_2)$. Under the assumption that the monodromy of p_1 is contained in $\mathcal{K}_g$ and not merely $\mathcal{I}_g$, it will follow that there is a unique possibility for $[F_2]$. The final step will be to extract the condition that the genera of F_2 and B_1 must be equal from the cohomology ring $H^*(E)$ and to argue that this enforces the triviality of either bundle structure.

The fundamental class of a second fiber In this subsection we will compute $[F_2]$ in the coordinates on H_2 coming from the fibering p_1. The results are formulated under the more general assumption that the monodromy of p_1 lie in $\mathcal{I}_g$ rather than $\mathcal{K}_g$, because we feel that the arguments are clearer in this larger context. The main objective is Lemma 7.3.

Suppose that $p_1\colon E \to B_1$ is a bundle with monodromy lying in $\mathcal{I}_g$. Suppose there is a partial section $\sigma\colon B' \to E'$ such that $[\omega_\sigma] = 0 \in H_1(F)$, giving rise to a lift $\tilde{\tau}$ of the Johnson homomorphism to $\bigwedge^3 H$; then, by Proposition 5.9(3), there is a natural splitting

$$H_3(E) \approx p_1^! H_1(B_1) \oplus H_1(F_1).$$

We use this direct sum decomposition to define the projections

$$P\colon H_3(E) \to p_1^! H_1(B_1) \quad \text{and} \quad Q\colon H_3(E) \to H_1(F)$$

and we consider the restrictions of P and Q to $p_2^!(H_1(B_2))$ for a second fibering $p_2\colon E \to B_2$. Where convenient, we will also define P and Q on $H_1(B_2)$ directly, by precomposing with the injection $p^!$.

Lemma 7.1 *For any second fibering $p_2\colon E \to B_2$, the restriction of Q to $H_1(B_2)$ is a symplectic mapping with respect to $d i_{F_1}$ on $H_1(F_1)$ and i_{B_2} on $H_1(B_2)$, where $d = [F_1]\cdot[F_2]$ is the algebraic intersection number of the two fibers.*

Proof There exist classes $x,\ y \in H_1(B_2)$ such that $x \cdot y = 1 \in H_0(B_2)$, so that $[F_2] = p_2^! x \cdot p_2^! y$ and there are expressions

$$p_2^! x = Px + Qx, \quad p_2^! y = Py + Qy.$$

Consequently,

$$[F_2] = Px \cdot Py + Px \cdot Qy - Py \cdot Qx + Qx \cdot Qy.$$

By Proposition 5.9, $[F_1] \cdot Pz = 0$ for all $z \in H_1(B_2)$, so that

$$d = [F_1] \cdot [F_2] = [F_1] \cdot Qx \cdot Qy,$$

with the first equality holding by assumption. The condition $[F_2] = p_2^! x \cdot p_2^! y$ is equivalent to $i_{B_2}(x, y) = 1$. By Proposition 5.9,

$$d = [F_1] \cdot Qx \cdot Qy = i_{F_1}(Qx, Qy),$$

proving the claim. □

As in the above proof, let $x,\ y \in H_1(B_2)$ satisfy $x \cdot y = 1$. By Poincaré duality, in order to determine $[F_2]$ it suffices to determine the collection of cup products $[F_2] \cdot Z$ for $Z \in H_2(E)$. Relative to the splitting of $H_2(E)$ coming from p_1 (where the monodromy lies in $\mathcal{I}_g$), in particular we must determine $[F_2] \cdot \Sigma_{b,z}$, where $b \in H_1(B_1)$ and $z \in H_1(F_1)$.

Lemma 7.2 *Take $x,\ y \in H_1(B_2)$ satisfying $x \cdot y = 1$. For $b \in H_1(B_1)$ and $z \in H_1(F_1)$, let $\Sigma_{b,z}$ be the associated element of $H_2(E)$. Then*

$$[F_2] \cdot \Sigma_{b,z} = i_{B_1}(Px, b) i_{F_1}(Qy, z) - i_{B_1}(Py, b) i_{F_1}(Qx, z) + \tau(b)(Qx \wedge Qy \wedge z). \tag{9}$$

In particular, if $z \in \langle Qx, Qy \rangle^{\perp}$ then (9) *simplifies to*

$$[F_2] \cdot \Sigma_{b,z} = \tau(b)(Qx \wedge Qy \wedge z). \tag{10}$$

In fact, for all $z \in H_1(F_1)$ there exist pairs $x_z,\ y_z \in H_1(B_2)$ such that $z \in \langle Qx_z, Qy_z \rangle^{\perp}$ holds, so that, for all b and z, (10) *is satisfied for this choice of x_z and y_z.*

Proof The formulas in (9) and (10) follow directly from the description of the intersection form given in Proposition 5.9. The existence of a suitable x and y for a given z is nothing but a matter of symplectic linear algebra. Since we will use some features of the construction later on, we give a detailed explanation. Lemma 7.1 shows that $W = \operatorname{Im} Q$ is a symplectic subspace of $H_1(F_1)$, so we can take a symplectic complement $W^{\perp}$. Any z can therefore be written as $w + w'$ with $w \in W$ and $w' \in W^{\perp}$. If $w = 0$ there is nothing to show. Otherwise, extend w to a symplectic basis for W such that $w = x_1$. As B_2 has genus at least 2, this basis includes x_2 and y_2 and, as $W = \operatorname{Im} Q$, we can select x_z and y_z in $H_1(B_2)$ with $Qx_z = x_2$ and $Qy_z = y_2$. □

We conclude this subsection by amalgamating the work we have done in the previous two propositions in order to give a description of $[F_2]$.

Lemma 7.3 *Let $p_2\colon E \to B_2$ be a second fibering. The choice of partial section $\sigma\colon B' \to E'$ furnishes $H_2(E)$ with the splitting*

$$H_2(E) = \langle [F_1] \rangle \oplus (H_1(B_1) \otimes H_1(F_1)) \oplus H_2(B_1),$$

with $H_1(B_1) \otimes H_1(F_1)$ spanned by the set of $\Sigma_{b,z}$ where b and z range in symplectic bases $\mathcal{B}$ and $\mathcal{F}$ for $H_1(B_1)$ and $H_1(F_1)$, respectively, and $H_2(B_1)$ is spanned by C, as in Proposition 5.9. Relative to this splitting of $H_2(E)$ there is the following expression for $[F_2]$:

$$[F_2] = (\delta - 2de)[F_1] + dC + \sum_{b \in \mathcal{B}, z \in \mathcal{F}} \tilde{\tau}(b)(Qx_z \wedge Qy_z \wedge z)\Sigma_{\hat{b}\hat{z}}. \tag{11}$$

Here, $\delta = i_{B_1}(Px, Py) + Qx \cdot Qy \cdot C$ for any choice of x, $y \in H_1(B_2)$ satisfying $x \cdot y = 1$, $e = C^2$ and $d = [F_1] \cdot [F_2]$ (the algebraic intersection of the two fibers). Also, $\hat{x}$ denotes the symplectic dual of x relative to the chosen symplectic basis.

Proof Suppose V is a free $\mathbb{Z}$–module equipped with a nondegenerate symmetric bilinear pairing $\langle \cdot, \cdot \rangle$. Suppose, moreover, that there exists a generating set $\mathcal{A} = \{a_1, \dots, a_k, b_1, \dots, b_k\}$ with the property that $\langle a_i, a_j \rangle = \langle b_i, b_j \rangle = 0$ for all i and j, $\langle a_i, b_j \rangle = 0$ for $i \neq j$, and $\langle a_i, b_i \rangle = 1$. Then any element $x \in V$ is expressible in the form

$$x = \sum_{i=1}^{k} \langle x, a_i \rangle b_i + \sum_{i=1}^{k} \langle x, b_i \rangle a_i. \tag{12}$$

We will apply this to $V = H_2(E)$ with the intersection pairing; in order to do this we must find a suitable generating set $\mathcal{A}$. Via Proposition 5.9, the space $H_1(B_1) \otimes H_1(F_1)$ is orthogonal under $\cdot$ to $H_2(B_2)$ and to $H_2(F_1)$ and, moreover, the collection of $\Sigma_{b,z}$ for $(b,z) \in \mathcal{B} \times \mathcal{F}$ is such a generating set on this subspace. We also have $[F_1] \cdot C = 1$ as well as $([F_1])^2 = 0$ and $C^2 = e$. Therefore, we can take

$$\mathcal{A} = \{[F_1], C - e[F_1]\} \cup \{\Sigma_{b,z} \mid (b,z) \in \mathcal{B} \times \mathcal{F}\}.$$

The only intersection that remains to be computed is $[F_2] \cdot C$. As $Px \cdot Py = i_{B_1}(Px, Py)[F_1]$, a direct computation gives

$$\begin{aligned} [F_2] \cdot C &= (Px \cdot Py + Px \cdot Qy - Py \cdot Qx + Qx \cdot Qy) \cdot C \\ &= Px \cdot Py \cdot C + Qx \cdot Qy \cdot C \\ &= i_{B_1}(Px, Py) + Qx \cdot Qy \cdot C = \delta. \end{aligned}$$

By assumption, $[F_1] \cdot [F_2] = d$, and (10) computes $[F_2] \cdot \Sigma_{b,z}$. Therefore we may insert these computations into (12) to obtain (11). $\square$

Rigidity in the Johnson kernel We now assume, as is required for Theorem 1.1, that the monodromy of p_1 is contained in $\mathcal{K}_g$. As noted in the previous section, the closed Johnson kernel $\mathcal{K}_g$ coincides with the kernel of $\tau\colon \mathcal{I}_g \to \bigwedge^3 H/H$; similarly, the pointed Johnson kernel $\mathcal{K}_{g,*}$ is the kernel of $\tau\colon \mathcal{I}_{g,*} \to \bigwedge^3 H$. We also noted above that if $\tau \circ \rho\colon H_1(B) \to \bigwedge^3 H/H$ is identically zero then there is a *canonical* lift $\tilde{\tau}\colon H_1(B) \to \bigwedge^3 H$, namely zero. This furnishes the (co)homology of E with a canonical splitting in which all cup products in (10) vanish.

In order to prove the main result of this section, we will compute $[F_2]$ and see that it is "as simple as possible" in the coordinates coming from p_1, the fibering with monodromy in $\mathcal{K}_g$. This will be accomplished via Lemma 7.3. Per our choice of lift $\tilde{\tau}$, the terms expressed via the Johnson homomorphism all vanish, so that

$$[F_2] = a[F_1] + dC$$

for some $a \in \mathbb{Z}$. The coefficient a is determined by $[F_2] \cdot C$ or, equivalently, by $\delta = i_{B_2}(Px, Py)$ (by Proposition 5.9(3), $Qx \cdot Qy \cdot C = 0$). This can be determined from Lemma 7.2.

Lemma 7.4 *Let E be a 4–manifold with two fiberings as a surface bundle over a surface, $p_1\colon E \to B_1$ and $p_2\colon E \to B_2$. Define the projection $P\colon H_1(B_2) \to H_1(B_1)$. Suppose that the monodromy for the bundle structure associated to p_1 lies in $\mathcal{K}_g$. Then $P \equiv 0$ and, consequently $\delta = 0$.*

Proof Returning to (9), in the Johnson kernel setting $[F_2]\cdot\Sigma_{b,z}$ and $\tilde{\tau}(b)(Qx \wedge Qy \wedge z)$ are both zero for all x, y and z. Taking z to be any element satisfying $i_{F_1}(Qy, z) \neq 0$ and $i_{F_1}(Qx, z) = 0$, (9) simplifies to $i_{B_1}(Px, b) = 0$. Since this is true for all b, we conclude that $Px = 0$ and, since any $x \in H_1(B_2)$ has a suitable y such that (9) holds, we conclude that $P \equiv 0$ and $\delta = 0$, as claimed. □

With this in hand, we can apply Lemma 7.3 (recalling from Proposition 5.9(3) that $e = 0$) to see that $[F_2]$ is as simple as possible:

$$[F_2] = dC. \tag{13}$$

As was noted following the statement of Proposition 3.1, $[F_2]$ must be a primitive class, so $d = \pm 1$. We conclude that $d = 1$ (as $d \geq 0$ by Proposition 4.1). We record this fact for later reference:

Lemma 7.5 *Let $p_1\colon E \to B_1$ be a surface bundle over a surface with monodromy in $\mathcal{K}_g$. Suppose there is a second fibering $p_2\colon E \to B_2$. Then*

$$\deg(p_1 \times p_2) = 1.$$

Proposition 4.1 asserts the equality of $\deg(p_1 \times p_2)$ *with* $\deg(p_2|_{F_1}\colon F_1 \to B_2)$ *and* $\deg(p_1|_{F_2}\colon F_2 \to B_1)$. *Consequently,*

$$\deg(p_2|_{F_1}\colon F_1 \to B_2) = \deg(p_1|_{F_2}\colon F_2 \to B_1) = 1.$$

Remark 7.6 Observe that Lemma 7.5 supplies a proof of the missing assertion $(1) \Longrightarrow (3)$ in Proposition 2.2, namely that, if E is a surface bundle over a surface with monodromy in the Johnson kernel, then any second fibering necessarily yields a bi-projection with nonzero degree. Of course, the assertion that any of the conditions (1)–(3) of Proposition 2.2 are equivalent to the bundle E being a product is the content of Theorem 1.1.

Cohomology: Splittings coming from multisections In order to complete the proof of Theorem 1.1, we will combine the work we have done above with an analysis of what the (co)homology of E looks like with respect to the coordinates coming from the second fibering (where the monodromy need not be contained in $\mathcal{I}_g$). The most convenient setting for this portion of the argument is in the *cohomology* ring, so we pause briefly to establish some preliminaries.

Most of what we have established vis-à-vis the intersection pairing on $H_*(E)$ is directly portable to the setting of the cup product in cohomology. In particular, the maps

$$p_i^*\colon H^*(B_i) \to H^*(E)$$

for $i = 1, 2$, are injections. We let $\eta_i \in H^2(B_i)$ be an integral generator compatible with the chosen orientations; it is easy to see that $p_i^*(\eta_i)$ is Poincaré dual to $[F_i]$. Relative to a choice of splitting

$$H^1(E) = p_1^* H^1(B_1) \oplus H^1(F_1),$$

there are the projection maps $P\colon H^1(B_2) \to H^1(B_1)$ and $Q\colon H^1(B_2) \to H^1(F_1)$, and Lemma 7.4 carries over to show that $P \equiv 0$. We can also transport our analysis of the intersection form on $H_*(E)$. In the cohomological setting, we have proved:

Proposition 7.7 *Let* $F_1 \to E \to B_1$ *be a surface bundle over a surface with monodromy in the Johnson kernel* $\mathcal{K}_g$. *Then* E *is an integral cohomology* $B_1 \times F_1$, *ie there exists a canonical isomorphism*

$$H^*(E) \approx H^*(B_1) \otimes H^*(F_1)$$

as graded rings.

We now continue with the proof of Theorem 1.1.

Lemma 7.8 *Suppose that the genus of B_2 is strictly smaller than that of F_1. Then there exist classes $x, y \in H^1(E)$ annihilating $p_2^* H^1 B_2$ (ie $x \smile p_2^* z = y \smile p_2^* z = 0$ for all $z \in H^1(B)$), such that $x \smile y = \Phi_1$, where $\Phi_1 \in H^2(F_1)$ is a generator.*

Proof The cohomological formulation of Lemma 7.4 shows that

$$p_2^* H^1(B_2) \leq H^1(F_1).$$

By (the cohomological reformulation of) Lemma 7.1, $p_2^* H^1(B_2)$ is in fact a *symplectic* subspace of $H^1(F)$, so there exists a symplectic complement. We can then take the desired x and y to be suitable elements of this complement. □

To finish the proof of Theorem 1.1, we will examine where x and y must live, relative to coordinates on $H^*(E)$ coming from the fibering p_2. At this point, the results of Section 6 come into play. In particular, (8) endows $H^1(E, \mathbb{Q})$ with a splitting

$$H^1(E, \mathbb{Q}) = \operatorname{Im} p^* \oplus \ker \widehat{\sigma}^*.$$

For the remainder of the proof, we will assume that all of our cohomology groups have rational coefficients.

Lemma 7.9 *Let $p\colon E \to B$ be any surface bundle over a surface with multisection σ. Suppose that there exists $x \in H^1(E)$ annihilating $p^* H^1(B)$. Then $x \in \ker \widehat{\sigma}^*$.*

Proof Write

$$x = v + p^* b$$

with $v \in \ker \widehat{\sigma}^*$ and $b \in H^1(B)$. If $b \neq 0$, then there exists $c \in H^1(B)$ with $b \smile c \neq 0$. On the one hand, $x \smile p^* c = 0$, by assumption. On the other, letting $[B] \in H_2(B)$ denote the fundamental class, we have by Proposition 6.4 that

$$\begin{aligned}\langle x \smile p^* c, \widehat{\sigma}_*[B]\rangle &= \langle (v + p^* b) \smile p^* c, \widehat{\sigma}_*[B]\rangle \\ &= \langle v \smile p^* c, \widehat{\sigma}_*[B]\rangle + \langle p^*(b \smile c), \widehat{\sigma}_*[B]\rangle \\ &= 0 + \langle \widehat{\sigma}^* p^*(b \smile c), [B]\rangle \\ &= \langle b \smile c, [B]\rangle \neq 0,\end{aligned}$$

since $v \in \ker \widehat{\sigma}^*$. In this case we have reached a contradiction, so $b = 0$ as desired. □

Lemma 7.10 *Let $F_1 \to E \to B_1$ be a surface bundle over a surface with monodromy in $\mathcal{K}_g$ and suppose there is a second fibering $p_2\colon E \to B_2$. Let g denote the genus of F_1 and h denote the genus of B_2. Then $g = h$.*

Proof We have already established (see Lemma 7.5) that

$$\deg(p_2|_{F_1}) = 1.$$

As p_2 has positive degree, we conclude immediately that $g \geq h$. Suppose $g > h$. Then there exist classes $x,\ y \in H^1(E)$ as in the statement of Lemma 7.8. We will make use of the existence of a multisection σ of $p_2\colon E \to B_2$ so that, by Lemma 7.9, we must have $x,\ y \in \ker \widehat{\sigma}^*$. So, by Proposition 6.4,

$$\langle x \smile y, \widehat{\sigma}_*[B_2] \rangle = 0.$$

In the notation of Proposition 7.7, both $p_2^* H^1(B_2)$ and the classes x and y are contained in $H^1(F_1)$ and, as the image of

$$\smile\colon \bigwedge^2 H^1(F_1) \to H^2(F_1)$$

is one-dimensional (since F_1 is a surface), we conclude that $x \smile y = p_2^*(\eta_2)$, where $\eta_2 \in H_2(B_2)$ is a generator. So, then

$$\langle x \smile y, \widehat{\sigma}_*[B_2] \rangle = \langle p_2^*(\eta_2), \widehat{\sigma}^*[B_2] \rangle = \langle \eta_2, [B_2] \rangle = 1.$$

This is a contradiction; necessarily $g = h$. □

This shows that $p_2|_{F_1}$ is a map of degree one between surfaces of the same genus and thus, as is well known,

$$(p_2)_*\colon \pi_1 F_1 \to \pi_1 B_2$$

must be an isomorphism.

End of proof of Theorem 1.1 At this point, we turn to an analysis of the fundamental group. Via the long exact sequence in homotopy for a fibration, there is an exact sequence

$$1 \to \pi_1 F_i \to \pi_1 E \to \pi_1 B_i \to 1$$

for $i = 1, 2$. Consequently, the kernel of

$$(p_1 \times p_2)_*\colon \pi_1 E \to \pi_1 B_1 \times \pi_1 B_2$$

is given by $\pi_1 F_1 \cap \pi_1 F_2$. On the other hand, this is also the kernel of the cross-projection

$$\pi_1 F_1 \to \pi_1 B_2,$$

which was just shown to be an isomorphism. We conclude that $(p_1 \times p_2)_*$ is an isomorphism.

The monodromy of the bundle E can be read off from the fundamental group as the map $\pi_1 B_1 \to \mathrm{Out}(\pi_1 F_1) \approx \mathrm{Mod}(\Sigma_g)$ (the latter isomorphism coming from the theorem of Dehn, Nielsen, and Baer). Since $\pi_1 E$ is a product, this map is trivial. The correspondence (1) then shows that E, being a surface bundle with trivial monodromy, is diffeomorphic to $B_1 \times B_2$. This completes the proof of Theorem 1.1. □

References

[1] **M F Atiyah**, *The signature of fibre-bundles*, from: "Global analysis (papers in honor of K Kodaira)", (D C Spencer, S Iyanaga, editors), Univ. Tokyo Press (1969) 73–84 MR0254864

[2] **A Cavicchioli**, **F Hegenbarth**, **D Repovš**, *On four-manifolds fibering over surfaces*, Tsukuba J. Math. 22 (1998) 333–342 MR1650733

[3] **B Farb**, **D Margalit**, *A primer on mapping class groups*, Princeton Mathematical Series 49, Princeton Univ. Press (2012) MR2850125

[4] **V Guillemin**, **A Pollack**, *Differential topology*, Prentice-Hall, Englewood Cliffs, NJ (1974) MR0348781

[5] **U Hamenstädt**, *On surface subgroups of mapping class groups*, talk at the workshop "Hot topics: Surface subgroups and cube complexes", MSRI (2013) Available at `https://www.msri.org/workshops/723/schedules/16639`

[6] **A Hatcher**, *Algebraic topology*, Cambridge Univ. Press (2002) MR1867354

[7] **J A Hillman**, *Four-manifolds, geometries and knots*, Geometry & Topology Monographs 5 (2002) MR1943724

[8] **F E A Johnson**, *A rigidity theorem for group extensions*, Arch. Math. (Basel) 73 (1999) 81–89 MR1703674

[9] **K Kodaira**, *A certain type of irregular algebraic surfaces*, J. Analyse Math. 19 (1967) 207–215 MR0216521

[10] **S Morita**, *Characteristic classes of surface bundles*, Invent. Math. 90 (1987) 551–577 MR914849

[11] **S Morita**, *Geometry of characteristic classes*, Translations of Mathematical Monographs 199, Amer. Math. Soc. (2001) MR1826571

[12] **I Rivin**, *Rigidity of fibering*, preprint (2011) `arXiv:1106.4595`

[13] **I Rivin**, *Statistics of* 3*–manifolds occsionally fibering over the circle*, preprint (2014) `arXiv:1401.5736`

[14] **N Salter**, *Surface bundles over surfaces with arbitrarily many fiberings*, Geom. Topol. 19 (2015) 2901–2923 MR3416116

Department of Mathematics, University of Chicago
5734 S University Ave, Chicago, IL 60637, USA

nks@math.uchicago.edu

http://math.uchicago.edu/~nks/

Received: 7 December 2014 Revised: 29 March 2015

Algebraic & Geometric Topology 15 (2015) 3653–3705

The algebraic duality resolution at $p = 2$

AGNÈS BEAUDRY

The goal of this paper is to develop some of the machinery necessary for doing $K(2)$–local computations in the stable homotopy category using duality resolutions at the prime $p=2$. The Morava stabilizer group $\mathbb{S}_2$ admits a surjective homomorphism to $\mathbb{Z}_2$ whose kernel we denote by $\mathbb{S}_2^1$. The algebraic duality resolution is a finite resolution of the trivial $\mathbb{Z}_2[\![\mathbb{S}_2^1]\!]$–module $\mathbb{Z}_2$ by modules induced from representations of finite subgroups of $\mathbb{S}_2^1$. Its construction is due to Goerss, Henn, Mahowald and Rezk. It is an analogue of their finite resolution of the trivial $\mathbb{Z}_3[\![\mathbb{G}_2^1]\!]$–module $\mathbb{Z}_3$ at the prime $p=3$. The construction was never published and it is the main result in this paper. In the process, we give a detailed description of the structure of Morava stabilizer group $\mathbb{S}_2$ at the prime 2. We also describe the maps in the algebraic duality resolution with the precision necessary for explicit computations.

55Q45; 55T99, 55P60

1 Introduction

Fix a prime p and recall that L_nS is the Bousfield localization of the sphere spectrum S with respect to the wedge $K(0)\vee\cdots\vee K(n)$, where $K(m)$ is the m^{th} Morava K–theory at the prime p. The chromatic convergence theorem of Hopkins and Ravenel [27, Section 8.6] states that the p–local sphere spectrum is the homotopy limit of its localizations L_nS. Further, there is a homotopy pull-back square:

$$\begin{tikzcd} L_nS \arrow[r] \arrow[d] & L_{K(n)}S \arrow[d] \\ L_{n-1}S \arrow[r] & L_{n-1}L_{K(n)}S \end{tikzcd}$$

In theory, the homotopy groups of S can be recovered from those of its Morava K–theory localizations $L_{K(n)}S$. For this reason, computing $\pi_*L_{K(n)}S$ is one of the central problems in stable homotopy theory. A detailed historical account of chromatic homotopy theory can be found in Goerss, Henn, Mahowald and Rezk [14, Section 1].

The difficulty of computing $\pi_*L_{K(n)}S$ varies with p and n. The computation of $\pi_*L_{K(1)}S$ is related to K–theory and is now well understood. For $n\geq 3$, almost

Published: 12 January 2016 DOI: 10.2140/agt.2015.15.3653

nothing is known, which leaves the case $n = 2$. For $p \geq 5$, $\pi_* L_{K(2)} S$ was computed by Shimomura and Yabe in [33]. Behrens gives an illuminating reconstruction of their results in [4]. The case when $p = 3$ proved much more difficult than the problem for $p \geq 5$. It is now largely understood due to the work of Shimomura, Wang, Goerss, Henn, Karamanov, Mahowald and Rezk (see Goerss and Henn [12], Goerss, Henn and Mahowald [13], Goerss, Henn, Mahowald and Rezk [14; 15], Henn, Karamanov and Mahowald [18] and Shimomura and Wang [32]).

The major breakthrough in understanding the case of $n = 2$ and $p = 3$ was the construction by Goerss, Henn, Mahowald and Rezk [14] of a finite resolution of the $K(2)$–local sphere called the *duality resolution*. The duality resolution comes in two flavors. The *algebraic* duality resolution is a finite resolution of the trivial $\mathbb{Z}_3[\![\mathbb{G}_2]\!]$–module $\mathbb{Z}_3$ by permutation modules induced from representations of finite subgroups G of the extended Morava stabilizer group $\mathbb{G}_2$. Its topological counterpart, the *topological* duality resolution, is a finite resolution of $E_2^{h\mathbb{G}_2}$, where E_2 denotes Morava E–theory. It is composed of spectra of the form $\Sigma^k E_2^{hG}$, and realizes the algebraic duality resolution. Both the algebraic duality resolution and the topological duality resolution give rise to spectral sequences which can be used to study $\pi_* L_{K(2)} S$ at $p = 3$.

The existence of a resolution analogous to that of [14, Theorem 4.1] at the prime $p = 2$ was conjectured by Mahowald using the computations of Shimomura [30] and of Shimomura and Wang [31]. The central result of this paper is its construction, which is due to Goerss, Henn, Mahowald and Rezk. The author is grateful for their blessing to record it here.

More precisely, for the norm-one subgroup $\mathbb{S}_2^1$ defined in (2.3.6), we construct a resolution of $\mathbb{Z}_2$ by modules which are induced from representations of finite subgroups of $\mathbb{S}_2^1$. We add a detailed description of the maps in the resolution, which will be used in later computations. However, we do not construct a full algebraic duality resolution of $\mathbb{Z}_2$ by $\mathbb{Z}_2[\![\mathbb{G}_2]\!]$–modules as in [14, Corollary 4.2] (see Remark 1.2.3), nor do we realize the algebraic resolution topologically as in [14, Section 5]. For work on the topological realization of the algebraic duality resolution, we refer the reader to Bobkova's thesis [6].

The algebraic duality resolution has already proved to be a useful tool for computations. We use the results of this paper in [3] to compute an associated graded for $H^*(\mathbb{S}_2^1, (E_2)_* V(0))$, where $V(0)$ is the mod 2 Moore spectrum. The computations of [3] play a crucial role in [2], where we prove that the strongest form of the chromatic splitting conjecture, as stated by Hovey in [21, Conjecture 4.2(v)], cannot hold when $n = p = 2$.

1.1 Background and notation

As in Goerss, Henn, Mahowald and Rezk [14, page 779], "we unapologetically focus on the case [$p = 2$] and $n = 2$ because this is at the edge of our current knowledge."

We let $K(2)$ be Morava K–theory, so that

$$K(2)_* = \mathbb{F}_2[v_2^{\pm 1}]$$

for v_2 of degree 6, and whose formal group law is the Honda formal group law of height two, which we denote by F_2. The Morava stabilizer group $\mathbb{S}_2$ is the group of automorphisms of F_2 over $\mathbb{F}_4$. It admits an action of the Galois group $\mathrm{Gal}(\mathbb{F}_4/\mathbb{F}_2)$. The extended Morava stabilizer group $\mathbb{G}_2$ is

$$\mathbb{G}_2 = \mathbb{S}_2 \rtimes \mathrm{Gal}(\mathbb{F}_4/\mathbb{F}_2).$$

By the Goerss–Hopkins–Miller theorem (see Goerss and Hopkins [16]), the group $\mathbb{G}_2$ acts on Morava E–theory E_2 by maps of E_∞–ring spectra and, for X a finite spectrum,

$$L_{K(2)}X \simeq E_2^{h\mathbb{G}_2} \wedge X.$$

In fact, for any closed subgroup G of $\mathbb{G}_2$, one can form the homotopy fixed point spectrum E_2^{hG}; see the work of Hopkins and Devinatz [10] and of Davis [9]. For a spectrum X, the action of $\mathbb{G}_2$ on $(E_2)_*$ induces an action on

$$(E_2)_*X := \pi_* L_{K(2)}(E_2 \wedge X).$$

For a closed subgroup G of $\mathbb{G}_2$ and a finite spectrum X, there is a convergent descent spectral sequence

$$E_2^{s,t} := H^s(G, (E_2)_t X) \Longrightarrow \pi_{t-s}(E_2^{hG} \wedge X).$$

We describe the most relevant example for this paper here. There is a norm on the group $\mathbb{S}_2$ whose kernel is denoted $\mathbb{S}_2^1$ (see Goerss, Henn, Mahowald and Rezk [14, Section 1.3]). Further,

$$\mathbb{S}_2 \cong \mathbb{S}_2^1 \rtimes \mathbb{Z}_2. \tag{1.1.1}$$

Similarly, the norm on $\mathbb{S}_2$ induces a norm on $\mathbb{G}_2$. The kernel is denoted $\mathbb{G}_2^1$ and

$$\mathbb{G}_2 \cong \mathbb{G}_2^1 \rtimes \mathbb{Z}_2. \tag{1.1.2}$$

Let π be a topological generator of $\mathbb{Z}_2$ in $\mathbb{G}_2$ and ϕ_π be its action on E_2. If X is finite, there is a fiber sequence

$$L_{K(2)}X \to E_2^{h\mathbb{G}_2^1} \wedge X \xrightarrow{\phi_\pi - \mathrm{id}} E_2^{h\mathbb{G}_2^1} \wedge X. \tag{1.1.3}$$

For this reason, the spectrum $E_2^{h\mathbb{G}_2^1}$ is often called the *half sphere*. One approach for computing $\pi_* L_{K(2)} X$ is to compute the spectral sequence

$$H^s(\mathbb{G}_2^1, (E_2)_t X) \cong H^s(\mathbb{S}_2^1, (E_2)_t X)^{\mathrm{Gal}(\mathbb{F}_4/\mathbb{F}_2)} \Longrightarrow \pi_{t-s} E_2^{h\mathbb{G}_2^1} \wedge X \tag{1.1.4}$$

and then use the fiber sequence (1.1.3) to pass from $\pi_*(E_2^{h\mathbb{G}_2^1} \wedge X)$ to $\pi_* L_{K(2)} X$.

Computing the E_2–term of (1.1.4) can be difficult. At the prime 3, the algebraic duality resolution of Goerss, Henn, Mahowald and Rezk constructed in [14, Theorem 4.1] gives rise to a first quadrant spectral sequence computing the E_2–term of the analogue of (1.1.4). One of the most important consequences of this paper is the existence of such a spectral sequence at the prime $p = 2$ (Theorem 1.2.4 below).

To state the main results and describe this spectral sequence, we must introduce some subgroups of $\mathbb{S}_2$. The group $\mathbb{S}_2$ has a unique conjugacy class of maximal finite subgroups of order 24. Fix a representative and call it G_{24}. The group G_{24} is isomorphic to the semidirect product of a quaternion group denoted Q_8 and a cyclic group of order 3 denoted C_3 (see Section 2.4); that is,

$$G_{24} \cong Q_8 \rtimes C_3.$$

However, there are two conjugacy classes of maximal finite subgroups in $\mathbb{S}_2^1$. If π is as above (1.1.3), the groups G_{24} and

$$G_{24}' := \pi G_{24} \pi^{-1}$$

are representatives for the distinct conjugacy classes. The group $\mathbb{S}_2$ also contains a central subgroup C_2 of order 2 generated by the automorphism $[-1](x)$ of the formal group law F_2 of $K(2)$. Therefore, $\mathbb{S}_2^1$ contains a cyclic subgroup

$$C_6 := C_2 \times C_3.$$

Choose a generator ω of C_3 and an element i in G_{24} such that G_{24} is generated by i and ω. That is,

$$G_{24} = \langle i, \omega \rangle.$$

Let $j = \omega i \omega^2$ and $k = \omega^2 j \omega$. The group $\mathbb{S}_2$ can be decomposed as a semidirect product

$$\mathbb{S}_2 \cong K \rtimes G_{24}$$

for a Poincaré duality group K of dimension 4. Similarly,

$$\mathbb{S}_2^1 \cong K^1 \rtimes G_{24}$$

for a Poincaré duality group K^1 of dimension 3; see Section 2.5. The homology of the groups K and K^1 play a central role in the construction of the duality resolution; see Section 3.1.

The group $\mathbb{S}_2^1$ is a profinite group and one can define the completed group ring

$$\mathbb{Z}_2[\![\mathbb{S}_2^1]\!] = \lim_{i,j} \mathbb{Z}/(2^i)[\mathbb{S}_2^1/U_j]$$

where $\{U_j\}$ forms a system of open subgroups such that $\bigcap_j U_j = \{e\}$. For any closed subgroup G of $\mathbb{S}_2^1$, we let

$$\mathbb{Z}_2[\![\mathbb{S}_2^1/G]\!] := \mathbb{Z}_2[\![\mathbb{S}_2^1]\!] \otimes_{\mathbb{Z}_2[\![G]\!]} \mathbb{Z}_2.$$

1.2 Statement of the results

The main result of this paper is the following theorem.

Theorem 1.2.1 (Goerss, Henn, Mahowald and Rezk, unpublished) *Let $\mathbb{Z}_2$ be the trivial $\mathbb{Z}_2[\![\mathbb{S}_2^1]\!]$–module. There is an exact sequence of complete left $\mathbb{Z}_2[\![\mathbb{S}_2^1]\!]$–modules*

$$0 \to \mathscr{C}_3 \xrightarrow{\partial_3} \mathscr{C}_2 \xrightarrow{\partial_2} \mathscr{C}_1 \xrightarrow{\partial_1} \mathscr{C}_0 \to \mathbb{Z}_2 \to 0$$

where

$$\mathscr{C}_p = \begin{cases} \mathbb{Z}_2[\![\mathbb{S}_2^1/G_{24}]\!] & \text{if } p = 0, \\ \mathbb{Z}_2[\![\mathbb{S}_2^1/C_6]\!] & \text{if } p = 1, 2, \\ \mathbb{Z}_2[\![\mathbb{S}_2^1/G_{24}']\!] & \text{if } p = 3. \end{cases}$$

The resolution of Theorem 1.2.1 is called the *algebraic duality resolution*. The name is motivated by the fact that the exact sequence of Theorem 1.2.1 exhibits a certain twisted duality. To make this precise, let $\mathrm{Mod}(\mathbb{S}_2^1)$ denote the category of finitely generated left $\mathbb{Z}_2[\![\mathbb{S}_2^1]\!]$–modules. As above, let π be a topological generator of $\mathbb{Z}_2$ in $\mathbb{S}_2 \cong \mathbb{S}_2^1 \rtimes \mathbb{Z}_2$. For a module M in $\mathrm{Mod}(\mathbb{S}_2^1)$, let $c_\pi(M)$ denote the left $\mathbb{Z}_2[\![\mathbb{S}_2^1]\!]$–module whose underlying $\mathbb{Z}_2$–module is M, but whose $\mathbb{Z}_2[\![\mathbb{S}_2^1]\!]$–module structure is twisted by the element π.

Theorem 1.2.2 (Henn, Karamanov and Mahowald, unpublished) *Let*

$$\mathscr{C}_p^* = \mathrm{Hom}_{\mathbb{Z}_2[\![\mathbb{S}_2^1]\!]}(\mathscr{C}_p, \mathbb{Z}_2[\![\mathbb{S}_2^1]\!]).$$

There is an isomorphism of complexes of left $\mathbb{Z}_2[\![\mathbb{S}_2^1]\!]$–modules:

$$\begin{tikzcd}
0 \arrow[r] & \mathscr{C}_3 \arrow[r, "\partial_3"] \arrow[d, "f_3"] & \mathscr{C}_2 \arrow[r, "\partial_2"] \arrow[d, "f_2"] & \mathscr{C}_1 \arrow[r, "\partial_1"] \arrow[d, "f_1"] & \mathscr{C}_0 \arrow[r] \arrow[d, "f_0"] & \mathbb{Z}_2 \arrow[r] \arrow[d, equal] & 0 \\
0 \arrow[r] & c_\pi(\mathscr{C}_0^*) \arrow[r, "c_\pi(\partial_1^*)"] & c_\pi(\mathscr{C}_1^*) \arrow[r, "c_\pi(\partial_2^*)"] & c_\pi(\mathscr{C}_2^*) \arrow[r, "c_\pi(\partial_3^*)"] & c_\pi(\mathscr{C}_3^*) \arrow[r] & \mathbb{Z}_2 \arrow[r] & 0
\end{tikzcd}$$

Remark 1.2.3 The resolution of Theorem 1.2.1 has the following shortcoming: it does not extend to a resolution of the group $\mathbb{G}_2$ or of the group $\mathbb{S}_2$ as in [14, Corollary 4.2]. This is due to the fact that (1.1.1) is a nontrivial extension when $n = p = 2$. For $n = 2$ and $p \geq 3$, $\mathbb{S}_2 \cong \mathbb{S}_2^1 \times \mathbb{Z}_p$.

One application of the algebraic duality resolution is given by the following theorem.

Theorem 1.2.4 *Let M be a profinite left $\mathbb{Z}_2[\![\mathbb{S}_2^1]\!]$–module. There is a first quadrant spectral sequence*

$$E_1^{p,q} = \mathrm{Ext}^q_{\mathbb{Z}_2[\![\mathbb{S}_2^1]\!]}(\mathscr{C}_p, M) \Longrightarrow H^{p+q}(\mathbb{S}_2^1, M)$$

with differentials $d_r\colon E_r^{p,q} \to E_r^{p+r,q-r+1}$. Further, there are isomorphisms

$$E_1^{p,q} \cong \begin{cases} H^q(G_{24}, M) & \text{if } p = 0, \\ H^q(C_6, M) & \text{if } p = 1, 2, \\ H^q(G'_{24}, M) & \text{if } p = 3. \end{cases}$$

The spectral sequence of Theorem 1.2.4 is called the algebraic duality resolution spectral sequence. Its computational appeal is twofold. The E_1–term is composed of the cohomology of finite groups. Further, it collapses at the E_4–term. The d_1 differentials are induced by the maps of the exact sequence in Theorem 1.2.1. In order to compute the spectral sequence, it is necessary to have a detailed description of these maps, which we do in Theorem 1.2.6 below.

The following result introduces some important elements in $\mathbb{S}_2^1$.

Theorem 1.2.5 *There is an element α in K^1 such that $\mathbb{S}_2$ is topologically generated by the elements π, α, i and ω. The group $\mathbb{S}_2^1$ is topologically generated by the elements α, i and ω.*

To state the next result, for any element τ in G_{24}, let

$$\alpha_\tau = \tau\alpha\tau^{-1}\alpha^{-1}.$$

Let S_2^1 be the 2–Sylow subgroup of $\mathbb{S}_2^1$. The group S_2^1 admits a decreasing filtration, denoted $F_{n/2}S_2^1$ which will be defined in Section 2.2.

Theorem 1.2.6 *Let e be the canonical generator of $\mathbb{Z}_2[\![\mathbb{S}_2^1]\!]$ and e_p be the canonical generator of $\mathscr{C}_p$. For a subgroup G of S_2^1, let IG be the kernel of the augmentation $\varepsilon\colon \mathbb{Z}_2[\![G]\!] \to \mathbb{Z}_2$. The maps $\partial_p\colon \mathscr{C}_p \to \mathscr{C}_{p-1}$ of Theorem 1.2.1 can be chosen so that:*

(a) $\partial_1(e_1) = (e-\alpha)e_0$.

(b) $\partial_2(e_2) = \Theta e_1$ *for an element Θ in $\mathbb{Z}_2[\![\mathbb{S}_2^1]\!]^{C_3}$ such that*

$$\Theta \equiv e+\alpha+i+j+k-\alpha_i-\alpha_j-\alpha_k \quad \mathrm{mod}\ \mathcal{J},$$

where $\mathcal{J}$ is the left ideal

$$\mathcal{J} = (IF_{4/2}K^1, (IF_{3/2}K^1)(IS_2^1), (IK^1)^7, 2(IK^1)^3, 4IK^1, 8).$$

In particular, $\Theta \equiv e+\alpha$ modulo $(2, (IS_2^1)^2)$.

(c) *There are isomorphisms of $\mathbb{Z}_2[\![\mathbb{S}_2^1]\!]$–modules $g_p\colon \mathscr{C}_p \to \mathscr{C}_p$ and differentials*

$$\partial'_{p+1}\colon \mathscr{C}_{p+1} \to \mathscr{C}_p$$

such that

$$\begin{tikzcd}
0 \arrow[r] & \mathscr{C}_3 \arrow[r, "\partial_3"] \arrow[d, "g_3"] & \mathscr{C}_2 \arrow[r, "\partial_2"] \arrow[d, "g_2"] & \mathscr{C}_1 \arrow[r, "\partial_1"] \arrow[d, "g_1"] & \mathscr{C}_0 \arrow[r] \arrow[d, "g_0"] & \mathbb{Z}_2 \arrow[r] \arrow[d, equal] & 0 \\
0 \arrow[r] & \mathscr{C}_3 \arrow[r, "\partial'_3"] & \mathscr{C}_2 \arrow[r, "\partial'_2"] & \mathscr{C}_1 \arrow[r, "\partial'_1"] & \mathscr{C}_0 \arrow[r] & \mathbb{Z}_2 \arrow[r] & 0
\end{tikzcd}$$

is an isomorphism of complexes. The map $\partial'_3\colon \mathscr{C}_3 \to \mathscr{C}_2$ is given by

$$\partial'_3(e_3) = \pi(e+i+j+k)(e-\alpha^{-1})\pi^{-1}e_2.$$

Theorem 1.2.6 is the key to doing computations using the duality resolution spectral sequence. The most difficult part of Theorem 1.2.6 is giving a good estimate for $\partial_2\colon \mathscr{C}_2 \to \mathscr{C}_1$. A detailed description of the map ∂_2 is given in Section 3.4. Though such precision is not needed for our computations in [3], the hope is that it will be sufficient for most future computations using the duality resolution spectral sequence.

1.3 Organization of the paper

Section 2 is dedicated to the description of the Morava stabilizer group in the special case of $p=2$ and $n=2$. (A more general account of the structure of $\mathbb{S}_n$ can be found in Goerss, Henn, Mahowald and Rezk [14, Section 1].) We begin by recalling the standard filtration on $\mathbb{S}_2$ and defining the norm. This allows us to define the unit norm subgroup $\mathbb{S}_2^1$ and describe its finite subgroups. In particular, we give an explicit choice of maximal finite subgroup G_{24} in Lemma 2.4.3. In Section 2.5, we introduce a subgroup K such that $\mathbb{S}_2 \cong K \rtimes G_{24}$ and compute the cohomology of K and of its norm-one subgroup K^1. These results are due to Goerss and Henn but are not published. We finish the section with a proof of Theorem 1.2.5.

In Section 3, we introduce the finite resolution of the trivial $\mathbb{Z}_2[\![\mathbb{S}_2^1]\!]$–module $\mathbb{Z}_2$. We construct the algebraic duality resolution spectral sequence. We describe the duality properties of the resolution and give a proof of Theorem 1.2.2. We end this section by giving a detailed description of the maps in the resolution, proving Theorem 1.2.6.

The appendix, contains the results on the cohomology of profinite p–adic analytic groups used in this paper.

1.4 Acknowledgements

I thank Mark Behrens, Irina Bobkova, Daniel G Davis, Matt Emerton, Niko Naumann, Doug Ravenel and Dylan Wilson for useful conversations. I thank the referee for a thorough reading and helpful comments. I thank Peter May for his support and for reading everything I have written multiple times. I thank Paul Goerss, Hans-Werner Henn and Charles Rezk for their generosity in sharing their ideas. In particular, I thank Paul and Hans-Werner for spending countless hours and hundreds of emails helping me understand their work and supporting mine. Finally, I will forever be indebted to Mark Mahowald for his incredible intuition and for the time he spent sharing it with me. I hope this paper and those to come will honor his tradition.

2 The structure of the Morava stabilizer group

2.1 A presentation of $\mathbb{S}_2$

Let F_2 be the Honda formal group law of height 2 at the prime 2. It is the 2–typical formal group law defined over $\mathbb{F}_2$ specified by the 2–series

$$[2]_{F_2}(x) = x^4.$$

The ring of endomorphisms of F_2 over $\mathbb{F}_4$ is isomorphic to the maximal order $\mathcal{O}_2$ in the central division algebra over $\mathbb{Q}_2$ of valuation $\frac{1}{2}$, which we denote by

$$\mathbb{D}_2 = D(\mathbb{Q}_2, \tfrac{1}{2}).$$

We begin by describing this isomorphism. More details can be found in Ravenel [26, A2.2; 27, Chapter 4].

Let $\mathbb{W} = W(\mathbb{F}_4)$ denote the ring of Witt vectors on $\mathbb{F}_4$. The ring $\mathbb{W}$ is isomorphic to the ring of integers of the unique unramified degree 2 extension of $\mathbb{Q}_2$. It is a complete local ring with residue field $\mathbb{F}_4$. The Teichmüller character defines a group homomorphism

$$\tau\colon \mathbb{F}_4^\times \to \mathbb{W}^\times.$$

Let ω be a choice of primitive third root of unity in $\mathbb{F}_4^\times$, and identify ω with its Teichmüller lift $\tau(\omega)$. Given such a choice, there is an isomorphism

$$\mathbb{W} \cong \mathbb{Z}_2[\omega]/(1+\omega+\omega^2).$$

The Galois group $\mathrm{Gal}(\mathbb{F}_4/\mathbb{F}_2)$ is generated by the Frobenius σ. It is the $\mathbb{Z}_2$–linear automorphism of $\mathbb{W}$ determined by

$$\omega^\sigma = \omega^2.$$

The ring $\mathcal{O}_2$ is a noncommutative extension of $\mathbb{W}$. It is given by

$$\mathcal{O}_2 \cong \mathbb{W}\langle S\rangle/(S^2 = 2, aS = Sa^\sigma)$$

for a in $\mathbb{W}$. Note that any element of $\mathcal{O}_2$ can be expressed uniquely as a linear combination $a + bS$ for a and b in $\mathbb{W}$. The division algebra $\mathbb{D}_2$ is given by

$$\mathbb{D}_2 \cong \mathcal{O}_2 \otimes_{\mathbb{Z}_2} \mathbb{Q}_2.$$

The 2–adic valuation v on $\mathbb{Q}_2$ extends uniquely to a valuation v on $\mathbb{D}_2$ such that $v(S) = \frac{1}{2}$. Further, $\mathcal{O}_2 = \{x \in \mathbb{D} \mid v(x) \geq 0\}$ and $\mathcal{O}_2^\times = \{x \in \mathbb{D} \mid v(x) = 0\}$. Therefore, any finite subgroup $G \subseteq \mathbb{D}_2^\times$ is contained in $\mathcal{O}_2^\times$.

Next, we describe the ring of endomorphisms of F_2 and give an explicit isomorphism $\mathrm{End}(F_2) \cong \mathcal{O}_2$. A complete proof can be found in Ravenel [26, Section A2]. First, note that

$$\mathrm{End}(F_2) \subseteq \mathbb{F}_4[\![x]\!].$$

To avoid confusion with the elements $\mathbb{W} \subseteq \mathcal{O}_2$, let ζ be a choice of primitive third root of unity in the field of coefficients $\mathbb{F}_4[\![x]\!]$. Let $S(x)$ correspond to the endomorphism

$$S(x) = x^2$$

so that

$$[2]_{F_2}(x) = x^4 = S(S(x)) = S^2(x).$$

Define $\omega^i(x) = \zeta^i x$ and $0(x) = 0$. Given an element a in $\mathbb{W}$, one can write it uniquely as $a = \sum_{i=0}^{\infty} a_i 2^i$ where a_i in $\mathbb{W}$ satisfies the equation

$$x^4 - x = 0.$$

That is, a_i is in $\{0,1,\omega,\omega^2\}$. Let $\gamma = a+bS$ be an element of $\mathcal{O}_2$. Let $a=\sum_{i\geq 0} a_{2i}2^i$ and $b=\sum_{i\geq 0} a_{2i+1}2^i$. Using the fact that $S^2=2$, the element γ can be expressed uniquely as a power series

$$\gamma = a_0 + 2a_2 + 4a_4 + \cdots + (a_1 + 2a_3 + 4a_6 + \cdots)S = \sum_{i\geq 0} a_i S^i. \tag{2.1.1}$$

One can show that

$$\gamma(x) = a_0(x) +_{F_2} a_1(x^2) +_{F_2} a_2(x^4) +_{F_2} \cdots +_{F_2} a_i(x^{2^i}) +_{F_2} \cdots$$

is a well-defined power series in $\mathbb{F}_4[\![x]\!]$. This determines a ring isomorphism from $\mathcal{O}_2$ to $\mathrm{End}(F_2)$.

The Morava stabilizer group $\mathbb{S}_2$ is the group of automorphisms of F_2. Thus,

$$\mathbb{S}_2 \cong \mathcal{O}_2^\times.$$

Any element of $\mathbb{S}_2$ can be expressed uniquely as a linear combination $a+bS$ for a in $\mathbb{W}^\times$ and b in $\mathbb{W}$. The center of $\mathbb{S}_2$ is given by the Galois invariant elements in $\mathbb{W}^\times$,

$$Z(\mathbb{S}_2) \cong \mathbb{Z}_2^\times.$$

Further, the element ω in $\mathbb{W}^\times$ generates a cyclic group of order 3 in $\mathbb{S}_2$, denoted C_3. The reduction of $\mathbb{W}$ modulo 2 induces an isomorphism $C_3 \cong \mathbb{F}_4^\times$.

The Galois group acts on $\mathbb{S}_2$ by

$$(a+bS)^\sigma = a^\sigma + b^\sigma S.$$

The extended Morava stabilizer group $\mathbb{G}_2$ is defined by

$$\mathbb{G}_2 := \mathbb{S}_2 \rtimes \mathrm{Gal}(\mathbb{F}_4/\mathbb{F}_2).$$

2.2 The filtration

In what follows, we use the presentation of $\mathbb{S}_2$ induced by the isomorphism $\mathbb{S}_2 \cong \mathcal{O}_2^\times$ that was described in Section 2.1. That is,

$$\mathbb{S}_2 \cong \big(\mathbb{W}\langle S\rangle/(S^2=2, aS=Sa^\sigma)\big)^\times$$

for a in $\mathbb{W}$. As described in Henn [17, Section 3], the group $\mathbb{S}_2$ admits the following filtration.

Recall from Section 2.1 that there is a valuation v on $\mathcal{O}_2$ such that

$$v(S) = \tfrac{1}{2}.$$

Regard $\mathbb{S}_2$ as the units in $\mathcal{O}_2$. For all $n \geq 0$, define

$$F_{n/2}\mathbb{S}_2 = \{x \in \mathbb{S}_2 \mid v(x-1) \geq n/2\}.$$

This filtration corresponds to the filtration of $\mathbb{S}_2$ by powers of S; that is, for $n \geq 1$,

$$F_{n/2}\mathbb{S}_2 = \{\gamma \in \mathbb{S}_2 \mid \gamma = 1 + a_n S^n + \cdots\}. \tag{2.2.1}$$

The motivation for indexing the filtration by half integers is that the induced filtration on $\mathbb{Z}_2^\times \subseteq \mathbb{S}_2$ is the usual 2–adic filtration by powers of 2.

Let

$$\mathrm{gr}_{n/2}\,\mathbb{S}_2 := F_{n/2}\mathbb{S}_2 / F_{(n+1)/2}\mathbb{S}_2$$

and

$$\mathrm{gr}\,\mathbb{S}_2 = \bigoplus_{n \geq 0} \mathrm{gr}_{n/2}\,\mathbb{S}_2.$$

Define $S_2 := F_{1/2}\mathbb{S}_2$. The group S_2 is the 2–Sylow subgroup of $\mathbb{S}_2$. The map $\mathbb{S}_2 \to \mathbb{F}_4^\times$ which sends γ to a_0 has kernel S_2. It induces an isomorphism

$$\mathrm{gr}_{0/2}\,\mathbb{S}_2 \cong \mathbb{F}_4^\times.$$

Suppose that $n > 0$ and that γ is an element of $F_{n/2}\mathbb{S}_2$, so that

$$\gamma = 1 + a_n S^n + \cdots$$

for a_i as in (2.1.1). Let $\overline{\gamma}$ denote the image of γ in $\mathrm{gr}_{n/2}\,\mathbb{S}_2$. For $n \geq 1$, the map defined by $\overline{\gamma} \mapsto a_n$ gives a group isomorphism

$$\mathrm{gr}_{n/2}\,\mathbb{S}_2 \cong \mathbb{F}_4.$$

It follows from these observations that the subgroups $F_{n/2}\mathbb{S}_2$ form a basis of open neighborhoods for the identity, so that $\mathbb{S}_2$ is a profinite topological group.

Given any subgroup G of $\mathbb{S}_2$, the filtration on $\mathbb{S}_2$ induces a filtration on G, defined by $F_{n/2}G = F_{n/2}\mathbb{S}_2 \cap G$. Let

$$\mathrm{gr}\,G = \bigoplus_{n \geq 0} \mathrm{gr}_{n/2}\,G \tag{2.2.2}$$

be the associated graded for this filtration.

The associated graded $\mathrm{gr}\,S_2$ has the structure of a restricted Lie algebra. The bracket is induced by the commutator in S_2 and the restriction is induced by squaring. In [17, Lemma 3.1.4], Henn gives an explicit description of the structure of this Lie algebra. We record this result in the case when $p = 2$ and $n = 2$.

Lemma 2.2.1 (Henn) *For $n, m > 0$, let $a \in F_{n/2}S_2$ and $b \in F_{m/2}S_2$. Let $\overline{a}$ be the image of a in $\mathrm{gr}_{n/2}\, S_2$ and $\overline{b}$ be the image of b in $\mathrm{gr}_{m/2}\, S_2$. If $[a,b]$ denotes the commutator $aba^{-1}b^{-1}$, then*

$$\overline{[a,b]} \equiv \overline{a}\overline{b}^{2^n} + \overline{a}^{2^m}\overline{b} \ \in \ \mathrm{gr}_{(n+m)/2}\, S_2.$$

If $P(a) = a^2$, then

$$\overline{P(a)} \equiv \begin{cases} \overline{a}^3 \in \mathrm{gr}_{2/2}\, S_2 & \text{if } n = 1, \\ \overline{a} + \overline{a}^2 \in \mathrm{gr}_{4/2}\, S_2 & \text{if } n = 2, \\ \overline{a} \ \in \mathrm{gr}_{n/2+1}\, S_2 & \text{if } n > 2. \end{cases}$$

2.3 The norm and the subgroups $\mathbb{S}_2^1$ and $\mathbb{G}_2^1$

The group $\mathbb{S}_2 \cong \mathcal{O}_2^\times$ acts on $\mathcal{O}_2$ by right multiplication. This gives rise to a representation $\rho\colon \mathbb{S}_2 \to \mathrm{GL}_2(\mathbb{W})$:

$$\rho(a + bS) = \begin{pmatrix} a & 2b^\sigma \\ b & a^\sigma \end{pmatrix}. \tag{2.3.1}$$

The restriction of the determinant to $\mathbb{S}_2$ is given by

$$\det(a + bS) = aa^\sigma - 2bb^\sigma. \tag{2.3.2}$$

This defines a group homomorphism $\det\colon \mathbb{S}_2 \to \mathbb{Z}_2^\times$.

Lemma 2.3.1 *The determinant* $\det\colon \mathbb{S}_2 \to \mathbb{Z}_2^\times$ *is surjective.*

Before proving this lemma, we introduce elements of $\mathbb{S}_2$ that will play a key role in the remainder of this paper and in future computations. First, let

$$\pi := 1 + 2\omega. \tag{2.3.3}$$

By Hensel's lemma, $\mathbb{Z}_2$ contains two solutions of $f(x) = x^2 + 7$. One of them satisfies

$$\sqrt{-7} \equiv 1 + 4 \mod 8.$$

This allows us to define

$$\alpha := \frac{1 - 2\omega}{\sqrt{-7}}. \tag{2.3.4}$$

Note that the elements π and α are in $\mathbb{W}^\times \subseteq \mathbb{S}_2$.

Proof of Lemma 2.3.1. The group $\mathbb{Z}_2^\times$ is topologically generated by -1 and 3. It suffices to show that these values are in the image of the determinant. A direct computation shows that $\det(\pi) = 3$ and that $\det(\alpha) = -1$. □

Definition 2.3.2 The norm $N\colon \mathbb{S}_2 \to \mathbb{Z}_2^\times/\{\pm 1\}$ is the composite

$$\mathbb{S}_2 \xrightarrow{\det} \mathbb{Z}_2^\times \to \mathbb{Z}_2^\times/\{\pm 1\}.$$

For a prime p and an integer $i \geq 1$, let $U_i = \{x \in \mathbb{Z}_p^\times \mid x \equiv 1 \bmod p^i\}$. For $p = 2$, there is a canonical identification

$$\mathbb{Z}_2^\times \cong \{\pm 1\} \times U_2. \tag{2.3.5}$$

Therefore, the image of the norm is canonically isomorphic to the group U_2. Further, the group U_2 is noncanonically isomorphic to the additive group $\mathbb{Z}_2$.

The subgroup $\mathbb{S}_2^1$ is defined by the short exact sequence

$$1 \to \mathbb{S}_2^1 \to \mathbb{S}_2 \xrightarrow{N} \mathbb{Z}_2^\times/\{\pm 1\} \to 1. \tag{2.3.6}$$

Any element γ such that $N(\gamma)$ is a topological generator of $\mathbb{Z}_2^\times/\{\pm 1\}$ determines a splitting. The element π defined in (2.3.3) is an example. This gives a decomposition

$$\mathbb{S}_2 \cong \mathbb{S}_2^1 \rtimes \mathbb{Z}_2^\times/\{\pm 1\} \cong \mathbb{S}_2^1 \rtimes \mathbb{Z}_2. \tag{2.3.7}$$

Note that the group $\mathbb{S}_2^1$ is closed in $\mathbb{S}_2$ as it is the intersection of the finite index subgroups which are the kernels of the norm followed by the projections $U_2 \to \mathbb{Z}/2^n$ for $n \geq 0$.

The norm N extends to a homomorphism

$$N\colon \mathbb{G}_2 \to \mathbb{Z}_2^\times/\{\pm 1\} \times \mathrm{Gal}(\mathbb{F}_4/\mathbb{F}_2) \to \mathbb{Z}_2^\times/\{\pm 1\},$$

where the second map is the projection. The subgroup $\mathbb{G}_2^1$ is the kernel of the extended norm and

$$\mathbb{G}_2 \cong \mathbb{G}_2^1 \rtimes \mathbb{Z}_2^\times/\{\pm 1\} \cong \mathbb{G}_2^1 \rtimes \mathbb{Z}_2. \tag{2.3.8}$$

We note that there is no splitting which is equivariant with respect to the action of the Galois group.

The filtration on $\mathbb{S}_2$ induces a filtration on $\mathbb{S}_2^1$ and

$$S_2^1 := F_{1/2}\mathbb{S}_2^1 \tag{2.3.9}$$

is the 2–Sylow subgroup of $\mathbb{S}_2^1$.

Remark 2.3.3 Note that for odd primes p, there is a canonical isomorphism

$$\mathbb{Z}_p^\times \cong C_{p-1} \times U_1,$$

where C_{p-1} is a cyclic group of order $p-1$. The exact sequence analogous to (2.3.6) is given by

$$1 \to \mathbb{S}_2^1 \to \mathbb{S}_2 \to \mathbb{Z}_p^\times / C_{p-1} \to 1.$$

Further, it has a central splitting. Therefore, when p is odd, the Morava stabilizer group is a product

$$\mathbb{G}_2 \cong \mathbb{G}_2^1 \times \mathbb{Z}_p^\times / C_{p-1} \cong \mathbb{G}_2^1 \times \mathbb{Z}_p.$$

There is no central splitting at the prime $p = 2$ and the extensions (2.3.7) and (2.3.8) are nontrivial.

We will need the following result in Section 2.5 to prove Theorem 2.5.7.

Lemma 2.3.4 *For $n \geq 1$,*

$$\mathrm{gr}_{n/2}\, S_2^1 = \begin{cases} \mathbb{F}_2 & \text{if } n \text{ is even,} \\ \mathbb{F}_4 & \text{if } n \text{ is odd.} \end{cases}$$

Proof Let $F_{0/2}\mathbb{Z}_2^\times = \mathbb{Z}_2^\times$ and, for $n \geq 2$ even,

$$F_{n/2}\mathbb{Z}_2^\times = F_{(n-1)/2}\mathbb{Z}_2^\times := U_{n/2} = \{\gamma \mid \gamma \equiv 1 \mod 2^{n/2}\}.$$

Let γ be in $\mathbb{S}_2$. Let $n \geq 2$ be even and suppose that γ has an expansion of the form

$$\gamma \equiv 1 + a_{n-1}S^{n-1} + a_n S^n \mod S^{n+1}.$$

By (2.3.2),

$$\det(\gamma) \equiv 1 + 2^{n/2}(a_n + a_n^\sigma) + a_{n-1}a_{n-1}^\sigma 2^{n-1} \mod 2^{n/2+1},$$

which is in $F_{n/2}\mathbb{Z}_2^\times$. Therefore, the determinant preserves this filtration. In fact, it induces short exact sequences of $\mathbb{F}_2$–vector spaces:

$$0 \to \mathrm{gr}_{n/2}\, \mathbb{S}_2^1 \to \mathrm{gr}_{n/2}\, \mathbb{S}_2 \to \mathrm{gr}_{n/2}\, \mathbb{Z}_2^\times \to 0.$$

The result then follows from the fact that

$$\mathrm{gr}_{n/2}\, \mathbb{Z}_2^\times = \begin{cases} \mathbb{F}_2 & \text{if } n \text{ is even,} \\ 0 & \text{if } n \text{ is odd,} \end{cases}$$

and $\mathrm{gr}_{n/2}\, \mathbb{S}_2 \cong (\mathbb{F}_2)^2$ for $n \geq 1$. □

2.4 Finite subgroups of $\mathbb{S}_2$

In this section, we describe the finite subgroups of $\mathbb{S}_2$ that will be used in the construction of the resolution of Theorem 1.2.1. We also prove that there are two conjugacy classes of maximal finite subgroups in $\mathbb{S}_2^1$. This will be used in the proof of Theorem 1.2.1.

Proposition 2.4.1 is a special case of Hewett [19, Theorem 1.4].

Proposition 2.4.1 (Hewett) *Any maximal finite nonabelian subgroup of $\mathbb{S}_2$ is isomorphic to a binary tetrahedral group*

$$G_{24} \cong Q_8 \rtimes C_3.$$

Here, Q_8 is the quaternion group

$$Q_8 \cong \langle i, j \mid i^2 = j^2, iji = j \rangle,$$

and the action of C_3 permutes i, j and ij.

Our next goal is to prove that there are two conjugacy classes of maximal finite subgroups in $\mathbb{S}_2^1$. To do this, we will need some preliminary results. Note that the classification of conjugacy classes of maximal finite subgroups of $\mathbb{S}_2$ is addressed in Hewett [20] and in Bujard [8]. According to Bujard [8, Remark 1.36], Hewett's [20, Theorem 5.3] is incorrect. However, [8, Theorem 1.35] in the case $n = p = 2$ is also stated incorrectly. A correct statement can be found in [8, Theorem 4.30]. To avoid confusion, we restate the results we need.

Proposition 2.4.2 (Bujard) *There is a unique conjugacy class of groups isomorphic to Q_8, and one of groups isomorphic to G_{24}, in $\mathbb{S}_2$.*

Proof For Q_8, this is [8, Lemma 1.25]. For G_{24}, this is [8, Theorem 1.28]. □

It will be useful to have explicit choices of subgroups Q_8 and G_{24}. The proof of the following lemma is a direct computation.

Lemma 2.4.3 (Henn) *Let*

$$i := \frac{1}{1+2\omega}(1 - \alpha S).$$

Define $j = \omega i \omega^2$ and $k = \omega^2 i \omega = ij$. The elements i and j generate a quaternion subgroup of $\mathbb{S}_2$, denoted Q_8. The elements i and ω generate a subgroup isomorphic to G_{24}. Further, in $\mathbb{D}_2$,

$$\omega = -\frac{1 + i + j + k}{2}.$$

For H a subgroup of G, let $N_G(H)$ be the normalizer of H in G. Let $C_G(H)$ be the centralizer of H in G. Note that the element $1+i$ in $\mathbb{D}_2^\times$ is in $N_{\mathbb{D}_2^\times}(Q_8)$. Since the valuation $v(1+i)=\frac{1}{2}$, the restriction of the valuation to the normalizer is surjective. Therefore, there is an exact sequence

$$1 \to N_{\mathbb{S}_2}(Q_8) \to N_{\mathbb{D}_2^\times}(Q_8) \to \tfrac{1}{2}\mathbb{Z} \to 0.$$

Since $\mathbb{D}_2 \cong \mathbb{Q}_2(i,j)$, it follows by the Skolem–Noether theorem that $\mathrm{Aut}(Q_8)$ can be realized by inner conjugation in $\mathbb{D}_2^\times$. There is an exact sequence

$$1 \to C_{\mathbb{D}_2^\times}(Q_8) \to N_{\mathbb{D}_2^\times}(Q_8) \to \mathrm{Aut}(Q_8) \to 0.$$

The next proposition describes which of these automorphisms can be realized by conjugation in $\mathbb{S}_2$.

Proposition 2.4.4 (Henn) *The subgroup of $\mathrm{Aut}(Q_8)$ that can be realized by conjugation by an element of $\mathbb{S}_2$ is isomorphic to the alternating group A_4. It is generated by conjugation by the elements i, j and ω.*

Proof The group $\mathrm{Aut}(Q_8)$ is isomorphic to the symmetric group $\mathfrak{S}_4$. One verifies by a direct computation that conjugation by i, j and ω generates a subgroup of $\mathrm{Aut}(Q_8)$ isomorphic to A_4. Let $\mathrm{Out}_{\mathbb{S}_2}(Q_8)$ be the group of automorphisms of Q_8 that can be realized by conjugation in $\mathbb{S}_2$. Since $C_{\mathbb{D}_2^\times}(Q_8) \cong \mathbb{Q}_2^\times$ and $C_{\mathbb{S}_2}(Q_8) \cong \mathbb{Z}_2^\times$, there is a commutative diagram

$$\begin{tikzcd}
\mathbb{Z}_2^\times \arrow[r] \arrow[d] & \mathbb{Q}_2^\times \arrow[r, "v"] \arrow[d] & \mathbb{Z} \arrow[d] \\
N_{\mathbb{S}_2}(Q_8) \arrow[r] \arrow[d] & N_{\mathbb{D}_2^\times}(Q_8) \arrow[r, "v"] \arrow[d] & \frac{1}{2}\mathbb{Z} \arrow[d] \\
\mathrm{Out}_{\mathbb{S}_2}(Q_8) \arrow[r] & \mathfrak{S}_4 \arrow[r] & \mathbb{Z}/2
\end{tikzcd}$$

where the columns and rows are short exact. Therefore, $\mathrm{Out}_{\mathbb{S}_2}(Q_8) \cong A_4$. □

Lemma 2.4.5 *Let $G_{24} = Q_8 \rtimes C_3$. The normalizer of Q_8 in $\mathbb{S}_2$ is given by*

$$N_{\mathbb{S}_2}(Q_8) \cong U_2 \times G_{24}.$$

Proof By Proposition 2.4.4, there is a short exact sequence

$$1 \to C_{\mathbb{S}_2}(Q_8) \to N_{\mathbb{S}_2}(Q_8) \to A_4 \to 1.$$

The centralizer is the subgroup $\mathbb{Z}_2^\times \cong C_2 \times U_2$ of $\mathbb{S}_2$. Since G_{24} is defined by the extension

$$1 \to C_2 \to G_{24} \to A_4 \to 1,$$

and the elements of U_2 are in the centralizer of G_{24}, it follows that $N_{\mathbb{S}_2}(Q_8)$ is isomorphic to $U_2 \times G_{24}$. □

Since the image of the norm is torsion free, any finite subgroup G of $\mathbb{S}_2$ is contained in the kernel $\mathbb{S}_2^1$. Therefore, $\mathbb{S}_2^1$ has the same maximal finite subgroups as $\mathbb{S}_2$. However, there are more conjugacy classes in $\mathbb{S}_2^1$.

Proposition 2.4.6 *There are two conjugacy classes of maximal finite subgroups in $\mathbb{S}_2^1$. One is the conjugacy class of G_{24} defined in Lemma 2.4.3. The other is $\xi G_{24}\xi^{-1}$, where ξ is any element such that $N(\xi)$ is a topological generator of U_2.*

Proof Let $\mathbb{Z}_2^\times \subseteq \mathbb{S}_2$ be the center. Define

$$\mathbb{S}_2^0 := \mathbb{S}_2^1 \times U_2,$$

where U_2 is as in (2.3.5). The restriction of the determinant to U_2 surjects onto $(\mathbb{Z}_2^\times)^2$. Therefore, there is an exact sequence

$$1 \to \mathbb{S}_2^0 \to \mathbb{S}_2 \to \mathbb{Z}_2^\times/\big(\{\pm 1\}, (\mathbb{Z}_2^\times)^2\big) \to 1,$$

and $\mathbb{S}_2/\mathbb{S}_2^0 \cong \mathbb{Z}/2$. If $N(\xi)$ is a topological generator for $\mathbb{Z}_2^\times/\{\pm 1\}$, then ξ is a representative for the nontrivial coset in $\mathbb{S}_2/\mathbb{S}_2^0$.

By Proposition 2.4.2, there is a unique conjugacy class of subgroups isomorphic to G_{24} in $\mathbb{S}_2$. Since conjugation by any element of the center $\mathbb{Z}_2^\times$ is trivial, any two conjugacy classes in $\mathbb{S}_2^1$ differ by conjugation by an element of $\mathbb{S}_2/\mathbb{S}_2^0 \cong \mathbb{Z}/2$. Therefore, there are at most two conjugacy classes.

Next, we show that the conjugacy classes of G_{24} and $\xi G_{24}\xi^{-1}$ are distinct in $\mathbb{S}_2^1$. Conjugation acts on the 2–Sylow subgroups; hence, it suffices to prove the claim for the subgroup Q_8 of G_{24}. Suppose that there exists an element γ in $\mathbb{S}_2^1$ such that

$$\xi Q_8 \xi^{-1} = \gamma Q_8 \gamma^{-1}.$$

This would imply that $\gamma^{-1}\xi$ is in $N_{\mathbb{S}_2}(Q_8)$. By Lemma 2.4.5, $\gamma^{-1}\xi$ is a product $z\tau$ for z in U_2 and τ in G_{24}. This implies that $\xi = \gamma z \tau$. However, $\gamma z\tau$ is in $\mathbb{S}_2^0$. This is a contradiction, since the residue class of ξ in $\mathbb{S}_2/\mathbb{S}_2^0$ is nontrivial. Therefore, G_{24} and $\xi G_{24}\xi^{-1}$ represent distinct conjugacy classes in $\mathbb{S}_2^1$. □

A choice for ξ is the element π defined in (2.3.3). For the remainder of this paper, G'_{24} will denote

$$G'_{24} := \pi G_{24} \pi^{-1}, \tag{2.4.1}$$

so that G_{24} and G'_{24} are representatives for the two conjugacy classes of maximal finite subgroups in $\mathbb{S}^1_2$.

2.5 The Poincaré duality subgroups

In this section, we introduce the subgroups K and K^1 and we describe their continuous cohomology rings $H^*(K, \mathbb{F}_2)$ and $H^*(K^1, \mathbb{F}_2)$ as G_{24}–modules. The author learned the results of this section from Paul Goerss and Hans-Werner Henn. We refer the reader to the appendix for details on the cohomology of a profinite group.

Let K be the closure of the subgroup of $\mathbb{S}_2$ generated by α (as defined in (2.3.4)) and $F_{3/2}\mathbb{S}_2$. That is,

$$K = \overline{\langle \alpha, F_{3/2}\mathbb{S}_2 \rangle}.$$

Proposition 2.5.1 *The subgroup K is normal in $\mathbb{S}_2$. Further, $S_2 \cong K \rtimes Q_8$ and $\mathbb{S}_2 \cong K \rtimes G_{24}$.*

Proof There is an isomorphism $\mathbb{S}_2 \cong S_2 \rtimes C_3$ and α commutes with the group C_3. Further, for any element γ in S_2, it follows from Lemma 2.2.1 that the commutator $[\gamma, \alpha]$ is in $F_{3/2}\mathbb{S}_2$. Since $\mathbb{S}_2 \cong S_2 \rtimes C_3$, and $F_{3/2}\mathbb{S}_2$ is normal, K is also normal. The quotient S_2/K is a group of order 8 generated by the image of the elements i and j defined in Lemma 2.4.3. The inclusion of Q_8 followed by the projection to S_2/K is an isomorphism. This defines a splitting. Similarly, the group $\mathbb{S}_2/K$ is a group of order 24 generated by the image of ω and i, and this defines a splitting. □

Corollary 2.5.2 *If K^1 is the kernel of the norm restricted to K, then $\mathbb{S}^1_2 \cong K^1 \rtimes G_{24}$.*

Proof The elements α and π are in the group K since $\alpha^{-1}\pi$ is in $F_{3/2}\mathbb{S}_2$. Therefore, the norm restricted to K is surjective and $\mathbb{S}^1_2/K^1 \cong \mathbb{S}_2/K$. □

Our next goal is to compute the group cohomology of K and K^1. We will need a few preliminary results.

Proposition 2.5.3 *Any open subgroup of S_2 or of S^1_2 is a profinite 2–adic analytic group.*

Proof According to Dixon, Du Sautoy, Mann and Segal [11, Theorem 8.1], a topological group is 2–adic analytic if and only if it has an open subgroup which is a finitely generated powerful pro-2 group. (By [11, Definition 3.1], a pro-2 group H is powerful if the quotient $H/\overline{H^4}$ is abelian, where $H^4 = \langle h^4 \mid h \in H \rangle$.) By Lemma 2.2.1, if $n \geq 3$, then $F_{n/2}S_2$ is topologically generated by any finite set of elements that surjects onto $F_{n/2}S_2/F_{(n+2)/2}S_2$. Further, the image of $P^2\colon F_{n/2}S_2 \to F_{(n+4)/2}S_2$ is dense by Lemma 2.2.1. If $n \geq 4$, then

$$[F_{n/2}S_2, F_{n/2}S_2] \subseteq F_{(2n)/2}S_2 \subseteq F_{(n+4)/2}S_2.$$

This implies that $F_{n/2}S_2$ is powerful for $n \geq 4$. Since any open subgroup G of S_2 contains $F_{n/2}S_2$ for some large n, it is a profinite 2–adic analytic group.

The proof for open subgroups of S_2^1 is similar, using $F_{n/2}S_2^1$ instead of $F_{n/2}S_2$. □

By Proposition 2.5.3, open subgroups of S_2 and S_2^1 are compact 2–adic analytic groups. This motivates our use of the following definition, which can be found in Symonds and Weigel [35, Section 4].

Definition 2.5.4 Let G be a compact p–adic analytic group. Then G is a Poincaré duality group of dimension n if G has cohomological dimension n and

$$H^s(G, \mathbb{Z}_p[\![G]\!]) \cong \begin{cases} \mathbb{Z}_p & \text{if } s = n, \\ 0 & \text{if } s \neq n \end{cases}$$

as abelian groups. The right $\mathbb{Z}_p[\![G]\!]$–module $H^n(G, \mathbb{Z}_p[\![G]\!])$ is denoted $D_p(G)$ and called the *compact dualizing module*. If the action of $\mathbb{Z}_p[\![G]\!]$ on $D_p(G)$ is trivial, the group G is called *orientable*.

Remark 2.5.5 For a Poincaré duality group G of dimension n, one can show that $H_n(G, D_p(G))$ is isomorphic to $\mathbb{Z}_p$; see Symonds and Weigel [35, Theorem 4.4.3]. Given a choice of generator $[G]$ for $H_n(G, D_p(G))$, the cap product induces a natural isomorphism

$$H^{n-*}(G, -) \xrightarrow{\cap[G]} H_*(G, D_p(G) \otimes_{\mathbb{Z}_p} -).$$

The following observations are useful to compute $D_p(G)$. Let $\phi_{D_p(G)}\colon G \to \mathbb{Z}_p^*$ be the representation associated to the action of G on $D_p(G)$. Let $L(G)$ be the $\mathbb{Q}_p$–Lie algebra associated to G, as defined in Lazard [24, Definition V.2.4.2.5]. The right conjugation action of G on itself induces a natural right action on $L(G)$, and thus a homomorphism Ad: $G \to \mathrm{Aut}(L(G))$. By [35, Corollary 5.2.5], if G is p–torsion free,

$$\phi_{D_p(G)}(g) = \det(\mathrm{Ad}(g)).$$

Proposition 2.5.6 *If an open subgroup U of $\mathbb{S}_n$ is a Poincaré duality group, then it is orientable.*

Proof This is the argument given by Strickland in the proof of [34, Proposition 5]. For any open subgroup U of $\mathbb{S}_n$, $L(U)$ is isomorphic to the central division algebra $\mathbb{D}_n$ over $\mathbb{Q}_p$ of valuation $1/n$. For g in U, the action $\mathrm{Ad}(g)$ is given by conjugation in $\mathbb{D}_n$, which has determinant one. $\square$

The next result relies on Lazard's theory of groups which are *équi-p–valué*. We refer the reader who is unfamiliar with the theory of Lazard to Huber, Kings and Naumann [22, Section 2] for an overview of the terminology.

Theorem 2.5.7 *For $n \geq 3$, the group $F_{n/2}S_2$ is a Poincaré duality group of dimension 4. The continuous group cohomology $H^*(F_{n/2}S_2, \mathbb{F}_2)$ is the exterior algebra generated by*

$$H^1(F_{n/2}S_2, \mathbb{F}_2) \cong \mathrm{Hom}_{\mathbb{F}_2}(\mathrm{gr}_{n/2} S_2 \oplus \mathrm{gr}_{(n+1)/2} S_2, \mathbb{F}_2) \cong \mathbb{F}_2^4.$$

Similarly, $F_{n/2}S_2^1$ is a Poincaré duality group of dimension 3 and $H^*(F_{n/2}S_2^1, \mathbb{F}_2)$ is the exterior algebra generated by

$$H^1(F_{n/2}S_2^1, \mathbb{F}_2) = \mathrm{Hom}_{\mathbb{F}_2}(\mathrm{gr}_{n/2} S_2^1 \oplus \mathrm{gr}_{(n+1)/2} S_2^1, \mathbb{F}_2) \cong \mathbb{F}_2^3.$$

Proof We define a filtration $w\colon F_{n/2}S_2 \to \mathbb{R}_+^* \cup \{\infty\}$ in the sense of Lazard [24, Definition II.1.1.1]. Let $w(1) = \infty$. For $k \geq 0$ and $x \in F_{(n+2k)/2}S_2 \backslash F_{(n+2k+2)/2}S_2$, let $w(x) = (n+2k)/2$. With this filtration, $F_{n/2}S_2$ is équi-p–valué of rank 4 in the sense of Lazard [24, V.2.2.7], with $\mathrm{gr}\, F_{n/2}S_2$ generated by

$$F_{n/2}S_2/F_{(n+2)/2}S_2 \cong \mathrm{gr}_{n/2} S_2 \oplus \mathrm{gr}_{(n+1)/2} S_2.$$

To verify that w is a filtration and that $F_{n/2}S_2$ is équi-p–valué with respect to w, one uses the formulas of Lemma 2.2.1, noting that the squaring map

$$P\colon F_{(n+2k)/2}S_2/F_{(n+2k+2)/2}S_2 \to F_{(n+2k+2)/2}S_2/F_{(n+2k+4)/2}S_2$$

is an isomorphism if and only if $n \geq 3$. The result then follows from [24, Proposition V.2.5.7.1], which states that $H^*(F_{n/2}S_2, \mathbb{F}_2)$ is an exterior algebra on the $\mathbb{F}_2$–linear dual of

$$F_{n/2}S_2/P(F_{n/2}S_2) \cong F_{n/2}S_2/F_{(n+2)/2}S_2 \cong \mathbb{F}_4^2 \cong \mathbb{F}_2^4.$$

According to Symonds and Weigel [35, Theorem 5.1.5], this also implies that $F_{n/2}S_2$ is a Poincaré duality group of dimension 4 (note that in [35], the authors imply in the third footnote that they use the terms *uniformly powerful pro-p* and *équi-p–valué* interchangeably.)

To prove the second claim, we use the same filtration, $F_{n/2}S_2^1$. By Lemma 2.3.4,

$$F_{n/2}S_2^1/F_{(n+2)/2}S_2^1 \cong \mathbb{F}_4 \oplus \mathbb{F}_2 \cong \mathbb{F}_2^3. \qquad \square$$

Recall that we use the convention that

$$\alpha_\tau = [\tau, \alpha] = \tau\alpha\tau^{-1}\alpha^{-1}.$$

The following congruences will be used in the computations of this section:

$$\begin{aligned} i &\equiv 1 + S \quad \bmod S^2, & j &\equiv 1 + \omega^2 S \quad \bmod S^2, \\ -1 &\equiv 1 + S^2 \quad \bmod S^4, & \alpha &\equiv 1 + \omega S^2 \quad \bmod S^4, \\ \alpha_i &\equiv 1 + S^3 \quad \bmod S^4, & \alpha_j &\equiv 1 + \omega^2 S^3 \quad \bmod S^4, \\ \alpha^2 &\equiv 1 + S^4 \quad \bmod S^5, & \alpha\pi &\equiv 1 + \omega S^4 \quad \bmod S^5. \end{aligned}$$

They are obtained by a direct computation using the definitions of π, α, i and j, which were given in (2.3.3), (2.3.4) and Lemma 2.4.3.

Definition 2.5.8 Let

$$\alpha_0 = \alpha, \quad \alpha_1 = \alpha_i, \quad \alpha_2 = \alpha_j, \quad \alpha_3 = \alpha^2, \quad \alpha_4 = \alpha\pi,$$

and let x_s in $\mathrm{Hom}_{\mathbb{F}_2}(\mathrm{gr}\,\mathbb{S}_2, \mathbb{F}_2)$ be the function dual to the image of α_s in $\mathrm{gr}\,\mathbb{S}_2$. The action of conjugation by an element τ on an element g is denoted by $\tau_*(g)$.

Remark 2.5.9 The action of conjugation by τ can be computed using Lemma 2.2.1 and the formula

$$[\tau, \gamma]\gamma = \tau_*(\gamma).$$

Corollary 2.5.10 *The continuous group cohomology $H^*(F_{3/2}S_2, \mathbb{F}_2)$ is the exterior algebra generated by*

$$H^1(F_{3/2}S_2, \mathbb{F}_2) \cong \mathbb{F}_2\{x_1, x_2, x_3, x_4\}$$

for x_s as in Definition 2.5.8. The action of α on $H^1(F_{3/2}S_2, \mathbb{F}_2)$ is trivial.

Similarly, $H^*(F_{3/2}S_2^1, \mathbb{F}_2)$ is the exterior algebra generated by

$$H^1(F_{3/2}S_2^1, \mathbb{F}_2) \cong \mathbb{F}_2\{x_1, x_2, x_3\}$$

with a trivial action by α.

Proof By Theorem 2.5.7, $H^*(F_{3/2}S_2,\mathbb{F}_2)$ is an exterior algebra generated by the $\mathbb{F}_2$–linear dual of $F_{3/2}S_2/F_{5/2}S_2$. This group is generated by the image of α_1, α_2, α_3 and α_4 of Definition 2.5.8. Therefore, $H^*(F_{3/2}S_2,\mathbb{F}_2)$ is the exterior algebra generated by $\mathbb{F}_2\{x_1,x_2,x_3,x_4\}$. Since α is in $F_{2/2}S_2$, if g is in $F_{3/2}S_2$, the commutator $[\alpha,g]$ is in $F_{5/2}S_2$. Using Remark 2.5.9, we conclude that the action of α on $H^1(F_{3/2}S_2,\mathbb{F}_2)$ is trivial.

The second claim follows in the same way from the fact that $F_{3/2}S_2^1/F_{5/2}S_2^1$ is generated by the image of α_1, α_2 and α_3. □

Lemma 2.5.11 *For α_i as defined in Definition 2.5.8, and $\overline{\alpha}_i$ its image in $H_1(K,\mathbb{Z}_2)$, there is an isomorphism*

$$H_1(K,\mathbb{Z}_2)\cong\mathbb{Z}/4\{\overline{\alpha}_0\}\oplus\mathbb{Z}/2\{\overline{\alpha}_1,\overline{\alpha}_2\}\oplus\mathbb{Z}_2\{\overline{\alpha}_4\},$$

where $2\overline{\alpha}_0$ is the image of $\alpha_3=\alpha^2$. Similarly,

$$H_1(K^1,\mathbb{Z}_2)\cong\mathbb{Z}/4\{\overline{\alpha}_0\}\oplus\mathbb{Z}/2\{\overline{\alpha}_1,\overline{\alpha}_2\}.$$

The conjugation action of Q_8 on K factors through the quotient of Q_8 by the central subgroup C_2. The induced action on $H_1(K,\mathbb{Z}_2)$ is trivial on $\overline{\alpha}_4$. On the other generators, it is given by

$$\begin{aligned} i_*(\overline{\alpha}_0)&=\overline{\alpha}_0+\overline{\alpha}_1, & j_*(\overline{\alpha}_0)&=\overline{\alpha}_0+\overline{\alpha}_2,\\ i_*(\overline{\alpha}_1)&=\overline{\alpha}_1, & j_*(\overline{\alpha}_1)&=\overline{\alpha}_1+2\overline{\alpha}_0,\\ i_*(\overline{\alpha}_2)&=\overline{\alpha}_2+2\overline{\alpha}_0, & j_*(\overline{\alpha}_2)&=\overline{\alpha}_2.\end{aligned}$$

Hence, $H_1(K^1,\mathbb{Z}_2)$ is generated by the image of α as a G_{24}–module.

Proof First, we prove that the group $[K,K]$ is dense in $F_{5/2}S_2^1$. Note that K is contained in $F_{2/2}S_2$. Let a and b be in K. For $\overline{a}$ and $\overline{b}$ as in Lemma 2.2.1,

$$\overline{[a,b]}=\overline{a}\overline{b}^4+\overline{a}^4\overline{b}\in\mathrm{gr}_{4/2}\,S_2.$$

Since $\overline{a}$ and $\overline{b}$ are in $\mathbb{F}_4$ and $x^4=x$ for all x in $\mathbb{F}_4$, this implies that $\overline{[a,b]}=0$ in $\mathrm{gr}_{4/2}\,S_2$. Therefore, $[a,b]\in F_{5/2}S_2$.

Since the norm is multiplicative, the elements of $[K,K]$ have norm one. Hence, $[K,K]$ is contained in $F_{5/2}S_2^1$. Further, the map from $[K,K]$ to $F_{5/2}S_2^1/F_{7/2}S_2^1$ induced by the inclusion is surjective. Indeed, $F_{5/2}S_2^1/F_{7/2}S_2^1$ is generated by the images of the elements $[\alpha,[i,\alpha]]$, $[\alpha,[j,\alpha]]$ and $[[i,\alpha],[j,\alpha]]$, all of which are in $[K^1,K^1]$. By Corollary 2.5.10, this implies that the composite

$$[K^1,K^1]\hookrightarrow[K,K]\to H_1(F_{5/2}S_2^1,\mathbb{F}_2)$$

is surjective. According to Behrens and Lawson [5, Theorem 2.1], it then follows from results of Koch and Serre that $[K^1, K^1]$ and $[K, K]$ are dense in $F_{5/2}S_2^1$. Therefore,

$$\overline{[K^1, K^1]} = \overline{[K, K]} = F_{5/2}S_2^1.$$

Hence, $H_1(K, \mathbb{Z}_2) \cong K/F_{5/2}S_2^1$ and $H_1(K^1, \mathbb{Z}_2) \cong K^1/F_{5/2}S_2^1$. Since α and π are in K, the norm $N\colon K \to \mathbb{Z}_2^\times/\{\pm 1\}$ is split surjective. The image of $\alpha_4 = \alpha\pi$ generates $\mathbb{Z}_2^\times/\{\pm 1\} \cong \mathbb{Z}_2$. Therefore,

$$H_1(K, \mathbb{Z}_2) \cong H_1(K^1, \mathbb{Z}_2) \oplus \mathbb{Z}_2\{\overline{\alpha}_4\}.$$

Finally, $H_1(K^1, \mathbb{Z}_2)$ is generated by the image of $\alpha_0 = \alpha$, $\alpha_1 = \alpha_i$ and $\alpha_2 = \alpha_j$. Since α_i and α_j are in $F_{3/2}S_2$, it follows from Lemma 2.2.1 that α_i^2 and α_j^2 are in

$$F_{5/2}S_2 = \overline{[K^1, K^1]}.$$

Therefore, the images of α_i and α_j have order 2 in $K^1/\overline{[K^1, K^1]}$. Finally, $\alpha^2 \equiv 1 + S^4$ modulo S^5, so that the image of α has order 4 in $K^1/\overline{[K^1, K^1]}$. We conclude that

$$H_1(K^1, \mathbb{Z}_2) \cong \mathbb{Z}/4\{\overline{\alpha}_0\} \oplus \mathbb{Z}/2\{\overline{\alpha}_1, \overline{\alpha}_2\}.$$

The action of Q_8 by conjugation factors through $C_2 = \{\pm 1\}$ since C_2 is in the center of $\mathbb{S}_2$. The action of the generators i and j is computed using Remark 2.5.9 and the following relations, which hold modulo S^5:

$$\begin{aligned} &[i, \alpha] \equiv \alpha_i, \quad &&[i, \alpha_i] \equiv 1, \quad &&[i, \alpha_j] \equiv \alpha^2, \quad &&[i, \alpha\pi] \equiv 1, \\ &[j, \alpha] \equiv \alpha_j, \quad &&[j, \alpha_i] \equiv \alpha^2, \quad &&[j, \alpha_j] \equiv 1, \quad &&[j, \alpha\pi] \equiv 1. \end{aligned}$$

These relations are obtained from Lemma 2.2.1. □

Corollary 2.5.12 *The group K is an orientable Poincaré duality group of dimension 4 and the group K^1 is an orientable Poincaré duality group of dimension 3.*

Proof It is a theorem of Serre [29, Section 1] that the cohomological dimension of a p–torsion free profinite group G is equal to the cohomological dimension of any of its open subgroups. The group K is a 2–group, and by Lemma 2.2.1, the squaring operation P on K has a trivial kernel. Therefore, K is torsion free. The group $F_{3/2}S_2$ is an open subgroup of K. Hence, the cohomological dimension of K is equal to the cohomological dimension of $F_{3/2}S_2$, so that K has cohomological dimension 4. Similarly, K^1 has cohomological dimension 3 since it contains $F_{3/2}S_2^1$ as an open subgroup. According to Symonds and Weigel [35, Proposition 4.4.1], a profinite group G of finite cohomological dimension is a Poincaré duality group if and only if it contains an open subgroup which is a Poincaré duality group. Therefore, both K and K^1 are Poincaré duality groups.

Since K is an open subgroup of $\mathbb{S}_2$, it follows from Proposition 2.5.6 that it is orientable. It remains to prove that K^1 is orientable. Let $\mathbb{Z}_2^\times$ be the center of $\mathbb{S}_2$. Let $U_2 \subseteq \mathbb{Z}_2^\times$ be as in (2.3.5). The group $H = K^1 \times U_2$ is an open subgroup of $\mathbb{S}_2$, and hence H is orientable. Further, $L(H) \cong L(K^1) \oplus L(U_2)$, and the action of H preserves the summands. Recall from Remark 2.5.5 that the representation $\phi_{D_2(H)}\colon H \to \mathbb{Z}_2^*$ is given by the determinant of the adjoint action of H on $L(H)$. For g in H,

$$\det(\mathrm{Ad}(g)) = \det(\mathrm{Ad}(g)|_{L(K^1)}) \det(\mathrm{Ad}(g)|_{L(U_2)}).$$

Since U_2 is abelian, $\det(\mathrm{Ad}(g)|_{L(U_2)}) = 1$. It follows from the orientability of H that $\det(\mathrm{Ad}(g)|_{L(K^1)}) = 1$. In particular, this holds for any g in K^1, and the representation $\phi_{D_2(K^1)}$ is trivial. □

Theorem 2.5.13 (Goerss and Henn, unpublished) *As an $\mathbb{F}_2$–algebra,*

$$H^*(K, \mathbb{F}_2) \cong \mathbb{F}_2[x_0, x_1, x_2, x_4]/(x_0^2, x_1^2 + x_0x_1, x_2^2 + x_0x_2, x_4^2),$$

where x_s has degree one and is as in Definition 2.5.8. Further,

$$H^*(K^1, \mathbb{F}_2) \cong \mathbb{F}_2[x_0, x_1, x_2]/(x_0^2, x_1^2 + x_0x_1, x_2^2 + x_0x_2).$$

The conjugation action of Q_8 factors through $Q_8/C_2 \cong C_2 \times C_2$. It is trivial on x_0 and x_4. On x_1 and x_2, it is described by

$$\begin{aligned} i_*(x_1) &= x_0 + x_1, & j_*(x_1) &= x_1, \\ i_*(x_2) &= x_2, & j_*(x_2) &= x_0 + x_2, \end{aligned}$$

so that the induced representation on $H^1(K^1, \mathbb{F}_2)$ is isomorphic to the augmentation ideal $I(Q_8/C_2)$, and $H^2(K^1, \mathbb{F}_2)$ is isomorphic to the coaugmentation ideal $I(Q_8/C_2)^*$.

Proof The spectral sequence for the group extension

$$1 \to F_{3/2}\mathbb{S}_2 \to K \to \mathbb{Z}/2\{\overline{\alpha}_0\} \to 1$$

has E_2–term given by

$$\mathbb{F}_2[x_0] \otimes E(x_1, x_2, x_3, x_4).$$

It follows from Lemma 2.5.11 and Lemma A.1.5 of the appendix that $x_0^2 = 0$. Since x_3 is the function dual to the image of α^2 in $\mathrm{gr}\,\mathbb{S}_2$, we have that $d_2(x_3) = x_0^2$. Using the isomorphism $H^1(K, \mathbb{F}_2) \cong \mathrm{Hom}(K, \mathbb{F}_2)$ and Lemma 2.5.11, one computes that

$$H^1(K, \mathbb{F}_2) \cong \mathbb{F}_2\{x_0, x_1, x_2, x_4\}.$$

Hence, $d_r(x_i) = 0$ for $i \neq 3$. All other differentials are determined by these differentials, and

$$E_3 \cong E_\infty \cong E(x_0, x_1, x_2, x_4).$$

Similarly, the E_2–term for the extension

$$1 \to F_{3/2}\mathbb{S}_2^1 \to K^1 \to \mathbb{Z}/2\{\overline{\alpha}_0\} \to 1$$

is given by $\mathbb{F}_2[x_0] \otimes E(x_1, x_2, x_3)$, and $E_3 \cong E_\infty \cong E(x_0, x_1, x_2)$.

Now we determine the multiplicative extensions. First, note that it follows from Lemma A.1.5 that $x_4^2 = 0$ since x_4 is dual to a class that lifts to the free class $\overline{\alpha}_4$ in $H_1(K, \mathbb{Z}_2)$. Similarly, x_1^2 and x_2^2 are nonzero since they lift to 2–torsion classes $\overline{\alpha}_1$ and $\overline{\alpha}_2$ in $H_1(K, \mathbb{Z}_2)$. Therefore, x_1^2 and x_2^2 are linear combinations of x_0x_1 and x_0x_2. We will show that $x_1^2 = x_0x_1$. The proof that $x_2^2 = x_0x_2$ is similar.

Let N be the closure of the normal subgroup of K^1 generated by $F_{6/2}K^1$, α^2 and α_j. That is,

$$N = \overline{\langle F_{6/2}K^1, \alpha^2, \alpha_j \rangle}.$$

Since $[K^1, F_{3/2}K^1] \subseteq F_{6/2}K^1$, and $[\alpha, \alpha_j] = \alpha_j^2$, the group K^1/N is a group of order 8 generated by the image a of α and the image b of α_i. The order of a is 2 and the order of b is 4. Further, since $[\alpha, \alpha_i] = \alpha_i^2$, the group K^1/N is isomorphic to the dihedral group D_8. Now, note that

$$H_1(D_8, \mathbb{F}_2) \cong \mathbb{F}_2\{a, b\}.$$

It is proved in Adem and Milgram [1, Chapter IV, Theorem 2.7] that

$$H^*(D_8, \mathbb{F}_2) \cong \mathbb{F}_2[x, y, w]/(xy),$$

where x is the function dual to a and y is the function dual to $a + b$. Changing the basis of $H_1(D_8, \mathbb{F}_2)$ from $\langle a, a+b \rangle$ to $\langle a, b \rangle$ sends the basis $\langle x, y \rangle$ of $H^1(D_8, \mathbb{F}_2)$ to the basis $\langle y_0, y_1 \rangle = \langle x + y, y \rangle$. We obtain the following presentation

$$H^*(K^1/N, \mathbb{F}_2) \cong \mathbb{F}_2[y_0, y_1, w]/(y_1^2 + y_0y_1).$$

The projection induces a map

$$f\colon H^*(K^1/N, \mathbb{F}_2) \to H^*(K^1, \mathbb{F}_2)$$

with $f(y_0) = x_0$ and $f(y_1) = x_1$. Therefore, $x_1^2 + x_0x_1 = 0$ in $H^2(K^1/N, \mathbb{F}_2)$.

The action of Q_8 follows from Lemma 2.5.11. The isomorphism between the representation $H^1(K^1, \mathbb{F}_2)$ and the representation $I(Q_8/C_2)$ defined by

$$0 \to I(Q_8/C_2) \to \mathbb{F}_2[Q_8/C_2] \xrightarrow{\varepsilon} \mathbb{F}_2 \to 0$$

is given by sending x_0 to the invariant $e+i+j+ij$, x_1 to $e+j$ and x_2 to $e+i$. □

The following description of the integral homology of K^1 will be used heavily in the proof of Theorem 1.2.1.

Corollary 2.5.14 (Goerss and Henn, unpublished) *The integral homology of K^1 is given by*

$$H_n(K^1, \mathbb{Z}_2) = \begin{cases} \mathbb{Z}_2 & \text{if } n = 0, 3, \\ \mathbb{Z}/4 \oplus (\mathbb{Z}/2)^2 & \text{if } n = 1, \\ 0 & \text{if } n = 2. \end{cases}$$

Proof The result for $n = 1$ is Lemma 2.5.11. The homology $H_*(K^1, \mathbb{F}_2)$ is dual to $H^*(K^1, \mathbb{F}_2)$, computed in Theorem 2.5.13. The groups $H_n(K^1, \mathbb{Z}_2)$ for $n = 2, 3$ are computed from the long exact sequence associated to

$$0 \to \mathbb{Z}_2 \xrightarrow{2} \mathbb{Z}_2 \to \mathbb{F}_2 \to 0,$$

using the fact that $H_n(K^1, \mathbb{F}_2)$ and $H_1(K^1, \mathbb{Z}_2)$ are known. □

We finish this section by proving Theorem 1.2.5.

Proof of Theorem 1.2.5 Since $\mathbb{S}_2 \cong S_2 \rtimes C_3$ and ω generates C_3, it suffices to show that S_2 is generated by π, α, i and $j = \omega i \omega^{-1}$. Further, according to Behrens and Lawson [5, Theorem 2.1], it suffices to prove that the inclusion $\langle \pi, \alpha, i, j \rangle \to S_2$ induces a surjective map

$$H_1(S_2, \mathbb{F}_2) \cong S_2/\overline{S_2^*},$$

where S_2^* is the group $S_2^2[S_2, S_2]$. The claim then follows from the isomorphism $S_2 \cong K \rtimes Q_8$, the surjectivity of the map

$$\langle \pi, \alpha, \alpha_i, \alpha_j \rangle \to K/\overline{K^*}$$

and the fact that i and j generate Q_8. The argument for $\mathbb{S}_2^1$ is similar. □

3 The algebraic duality resolution

This section is devoted to the construction of the algebraic duality resolution and the description of its properties. We refer the reader to the appendix for background on the cohomology of profinite groups.

3.1 The resolution

From now on, we fix $p=2$. The goal of this section is to prove Theorem 1.2.1, which was stated in Section 1.2, and is restated as Theorem 3.1.7 below. The proof is broken into a series of results given in Lemma 3.1.1, Lemma 3.1.2, Lemma 3.1.3 and Theorem 3.1.6. All results in this section are due to Goerss, Henn, Mahowald and Rezk.

Let G_{24} be the maximal finite subgroup of $\mathbb{S}_2$ defined in Lemma 2.4.3. Recall that $G'_{24}=\pi G_{24}\pi^{-1}$ for $\pi=1+2\omega$ in $\mathbb{S}_2$. It was shown in Proposition 2.4.6 that there are two conjugacy classes of maximal finite subgroups in $\mathbb{S}_2^1$, and that G_{24} and G'_{24} are representatives. Recall that $C_2=\{\pm 1\}$ is the subgroup generated by $[-1](x)$ and $C_6=C_2\times C_3$. The group K^1 is the Poincaré duality subgroup of $\mathbb{S}_2^1$ which was defined in Section 2.5.

Lemma 3.1.1 *Let* $\mathscr{C}_0=\mathbb{Z}_2[\![\mathbb{S}_2^1/G_{24}]\!]$ *with canonical generator* e_0. *Let* $\varepsilon\colon \mathscr{C}_0\to\mathbb{Z}_2$ *be the augmentation*

$$\varepsilon\colon \mathbb{Z}_2[\![\mathbb{S}_2^1/G_{24}]\!]\to \mathbb{Z}_2\otimes_{\mathbb{Z}_2[\![\mathbb{S}_2^1]\!]}\mathbb{Z}_2[\![\mathbb{S}_2^1]\!]\otimes_{\mathbb{Z}_2[G_{24}]}\mathbb{Z}_2\cong\mathbb{Z}_2.$$

Let N_0 *be defined by the short exact sequence*

$$0\to N_0\to\mathscr{C}_0\xrightarrow{\varepsilon}\mathbb{Z}_2\to 0. \tag{3.1.1}$$

Then N_0 *is the left* $\mathbb{Z}_2[\![\mathbb{S}_2^1]\!]$*–submodule of* $\mathscr{C}_0$ *generated by* $(e-\alpha)e_0$, *for* e *the unit in* $\mathbb{S}_2^1$ *and* α *as defined in* (2.3.4).

Proof Since $\mathbb{S}_2^1\cong K^1\rtimes G_{24}$, $\mathscr{C}_0\cong\mathbb{Z}_2[\![K^1]\!]$ as a $\mathbb{Z}_2[\![K^1]\!]$–module. Therefore, N_0 is isomorphic to the augmentation ideal IK^1. Lemma A.1.4 of the appendix implies that $H_1(K^1,\mathbb{Z}_2)\cong H_0(K^1,N_0)$, where an isomorphism sends the image of g in $K^1/\overline{[K^1,K^1]}$ to the image of $e-g$ in $IK^1/\overline{(IK^1)^2}$. It was shown in Lemma 2.5.11 that $K^1/\overline{[K^1,K^1]}$ is generated by α as a G_{24}–module. This implies that, as a G_{24}–module, $H_0(K^1,N_0)$ is generated by the image of $(e-\alpha)e_0$. Therefore, the map $F\colon\mathbb{Z}_2[\![\mathbb{S}_2^1]\!]\to N_0$ defined by $F(\gamma)=\gamma(e-\alpha)e_0$ induces a surjective map

$$\mathbb{F}_2\otimes_{\mathbb{Z}_2[\![K^1]\!]}F\colon \mathbb{F}_2\otimes_{\mathbb{Z}_2[\![K^1]\!]}\mathbb{Z}_2[\![\mathbb{S}_2^1]\!]\to\mathbb{F}_2\otimes_{\mathbb{Z}_2[\![K^1]\!]}N_0.$$

By Lemma A.1.3 of the appendix, F itself is surjective, and $(e-\alpha)e_0$ generates N_0 as a $\mathbb{Z}_2[\![\mathbb{S}_2^1]\!]$–module. □

Lemma 3.1.2 *Let* N_0 *be as in Lemma 3.1.1. Let* $\mathscr{C}_1=\mathbb{Z}_2[\![\mathbb{S}_2^1/C_6]\!]$ *with canonical generator* e_1. *There is a map* $\partial_1\colon\mathscr{C}_1\to N_0$ *defined by*

$$\partial_1(\gamma e_1)=\gamma(e-\alpha)e_0 \tag{3.1.2}$$

for γ in $\mathbb{Z}_2[\![\mathbb{S}_2^1]\!]$. Further, let N_1 be defined by the short exact sequence

$$0 \to N_1 \to \mathscr{C}_1 \xrightarrow{\partial_1} N_0 \to 0, \tag{3.1.3}$$

and let Θ_0 in $\mathbb{Z}_2[\![\mathbb{S}_2^1]\!]$ be any element such that $\Theta_0 e_1$ is in the kernel of ∂_1 and

$$\Theta_0 e_1 \equiv (3+i+j+k)e_1 \mod (4, IK^1).$$

Then $\Theta_0 e_1$ generates N_1 over $\mathbb{S}_2^1$.

Proof The element α satisfies $\tau\alpha = \alpha\tau$ for $\tau \in C_6$. Therefore, the map ∂_1 given by (3.1.2) is well defined.

Let N_1 be the kernel of ∂_1. Note that $\mathbb{Z}_2[\![\mathbb{S}_2^1/C_6]\!] \cong \mathbb{Z}_2[\![K^1]\!]^4$ as $\mathbb{Z}_2[\![K^1]\!]$–modules, generated by e_1, ie_1, je_1 and ke_1. Therefore, there is an isomorphism of G_{24}–modules

$$H_0(K^1, \mathscr{C}_1) \cong \mathbb{Z}_2[G_{24}/C_6].$$

As $\mathbb{Z}_2[\![K^1]\!]$–modules, $H_0(K^1, \mathscr{C}_1) \cong \mathbb{Z}_2^4$ generated by the image of the classes e_1, ie_1, je_1 and ke_1. Since $N_0 \cong IK^1$, Lemma A.1.4 of the appendix and Corollary 2.5.14 imply that

$$H_1(K^1, N_0) \cong H_2(K^1, \mathbb{Z}_2) = 0.$$

Therefore, the long exact sequence on cohomology gives rise to a short exact sequence

$$0 \to H_0(K^1, N_1) \to H_0(K^1, \mathscr{C}_1) \to H_0(K^1, N_0) \to 0.$$

By Lemma A.1.4 of the appendix and Lemma 2.5.11,

$$H_0(K^1, N_0) \cong H_1(K^1, \mathbb{Z}_2) \cong \mathbb{Z}/4 \oplus (\mathbb{Z}/2)^2,$$

which is all torsion. Thus, we can identify $H_0(K^1, N_1)$ with a free submodule of $H_0(K^1, \mathscr{C}_1)$. Further, it must have rank 4 over $\mathbb{Z}_2$. This can be made explicit as follows.

The map $H_0(K^1, \partial_1)$ sends the residue class of τe_1 to that of $\tau(e-\alpha)e_0$. For τ in G_{24}, $\tau^{-1}e_0 = e_0$, hence $\tau(e-\alpha)e_0 = (e-\tau_*(\alpha))e_0$, where $\tau_*(\alpha) = \tau\alpha\tau^{-1}$. Again, using the boundary isomorphism $H_1(K^1, \mathbb{Z}_2) \cong H_0(K^1, N_0)$ of Lemma A.1.4, the formulas of Lemma 2.5.11 together with the fact that $k = ij$ can be used to compute

$$\partial_1(e_1) \equiv \overline{\alpha}, \quad \partial_1(ie_1) \equiv \overline{\alpha} + \overline{\alpha}_i, \quad \partial_1(je_1) \equiv \overline{\alpha} + \overline{\alpha}_j, \quad \partial_1(ke_1) \equiv 3\overline{\alpha} + \overline{\alpha}_i + \overline{\alpha}_j.$$

Here, $\overline{a}$ is the image of a in $H_1(K^1, \mathbb{Z}_2)$. As α generates a group isomorphic to $\mathbb{Z}/4$, and α_i and α_j both generate groups isomorphic to $\mathbb{Z}/2$, a set of $\mathbb{Z}_2$ generators for the kernel of $H_0(K^1,\partial_1)$ is given by the elements

$$f_1=-4e_1,\quad f_2=2(i-e)e_1,\quad f_3=2(j-e)e_1,\quad f_4=(k-i-j-e)e_1.$$

Let

$$f=(3e+i+j+k)e_1\in H_0(K^1,N_1).$$

Then f generates $H_0(K^1,N_1)\cong\mathbb{Z}_2[G_{24}/C_6]$ as a G_{24}–module. Indeed, using the fact that $G_{24}/C_6\cong Q_8/C_2$, one computes

$$f_1=1/3(i+j+k-5)f,\quad f_2=if-f,\quad f_3=jf-f,\quad f_4=-k(f+f_1).$$

(Note that $-\tau$ denotes $(-1)\cdot\tau$ for the coefficient -1 in $\mathbb{Z}_2$, as opposed to the generator of the central C_2 in Q_8.)

Next, we show that if

$$f'\equiv f \mod (4,IK^1),$$

then f' also generates $H_0(K^1,N_1)$ as a G_{24}–module. To do this, note that $\mathbb{Z}_2[Q_8/C_2]$ is a complete local ring with maximal ideal $\mathfrak{m}=(2,IQ_8/C_2)$. Hence, any element congruent to 1 modulo $\mathfrak{m}$ is invertible. Therefore, if $f'=f+\epsilon f$ for ϵ in $\mathfrak{m}$, then f' is also a generator. However, for a in $H_0(K_1,\mathscr{C}_1)$,

$$4ae_1=a\tfrac{1}{3}((e-i)+(e-j)+(e-k)+2e)f.$$

Hence, $a\frac{1}{3}((e-i)+(e-j)+(e-k)+2e)$ is in $\mathfrak{m}$. Therefore, $4H_0(K_1,\mathscr{C}_1)$ is contained in $\mathfrak{m}f$.

Let Θ_0 in $\mathbb{Z}_2[\![\mathbb{S}_2^1]\!]$ be such that

$$\Theta_0e_1\equiv(3+i+j+k)e_1 \mod (4,IK^1).$$

Let $F\colon\mathbb{Z}_2[\![\mathbb{S}_2^1]\!]\to N_1$ be the map defined by $F(\gamma)=\gamma\Theta_0e_1$. It induces a surjective map

$$\mathbb{F}_2\otimes_{\mathbb{Z}_2[\![K^1]\!]}F\colon\mathbb{F}_2\otimes_{\mathbb{Z}_2[\![K^1]\!]}\mathbb{Z}_2[\![\mathbb{S}_2^1]\!]\to\mathbb{F}_2\otimes_{\mathbb{Z}_2[\![K^1]\!]}N_1.$$

By Lemma A.1.3 of the appendix, F itself is surjective, and Θ_0e_1 generates N_1 as a $\mathbb{Z}_2[\![\mathbb{S}_2^1]\!]$–module. □

Define $\mathrm{tr}_{C_3}\colon\mathbb{Z}_2[\![\mathbb{S}_2^1]\!]\to\mathbb{Z}_2[\![\mathbb{S}_2^1]\!]$ to be the $\mathbb{Z}_2$–linear map induced by

(3.1.4) $$\mathrm{tr}_{C_3}(g)=g+\omega g\omega^{-1}+\omega^{-1}g\omega$$

for g in $\mathbb{S}_2^1$ and ω our chosen generator of C_3.

Lemma 3.1.3 *Let $\mathscr{C}_2=\mathbb{Z}_2[\![\mathbb{S}_2^1/C_6]\!]$ with canonical generator e_2. Let Θ in $\mathbb{Z}_2[\![\mathbb{S}_2^1]\!]$ satisfy:*

(1) *$\tau\Theta = \Theta\tau$ for τ in C_6,*

(2) *Θe_1 is in the kernel of $\partial_1\colon \mathscr{C}_1 \to \mathscr{C}_0$,*

(3) *$\Theta e_1 \equiv (3+i+j+k)e_1$ modulo $(4, IK^1)$.*

Then the map of $\mathbb{Z}_2[\![\mathbb{S}_2^1]\!]$–modules $\partial_2\colon \mathscr{C}_2 \to \mathscr{C}_1$ defined by

$$\partial_2(\gamma e_2) = \gamma\Theta e_2 \tag{3.1.5}$$

surjects onto $N_1 = \ker(\partial_1)$. Further, if N_2 is defined by the exact sequence

$$0 \to N_2 \to \mathscr{C}_2 \xrightarrow{\partial_2} N_1 \to 0, \tag{3.1.6}$$

then $N_2 \cong \mathbb{Z}_2[\![K^1]\!]$ as $\mathbb{Z}_2[\![K^1]\!]$–modules.

Proof Choose an element Θ_0 which generates N_1 as in Lemma 3.1.2. Recall that $C_6 \cong C_2 \times C_3$ and that C_2 is in the center of $\mathbb{S}_2$. Therefore, for tr_{C_3} as defined by (3.1.4),

$$\Theta = \tfrac{1}{3}\,\mathrm{tr}_{C_3}(\Theta_0)$$

satisfies properties (1), (2) and (3). The map ∂_2 given by (3.1.5) is well defined and surjects onto N_1 by Lemma A.1.3.

Let $N_2 \subseteq \mathscr{C}_2$ be the kernel of ∂_2 as in the statement of the result. The map ∂_2 induces an isomorphism $H_0(K^1, \mathscr{C}_2) \cong H_0(K^1, N_1)$. Hence, for all n,

$$H_n(K^1, N_2) \cong H_{n+1}(K^1, N_1) \cong H_{n+2}(K^1, N_0) \cong H_{n+3}(K^1, \mathbb{Z}_2).$$

This implies:

$$H_n(K^1, N_2) \cong \begin{cases} \mathbb{Z}_2 & \text{if } n = 0, \\ 0 & \text{if } n > 0. \end{cases}$$

Choose an element e' in N_2 such that e' reduces to a generator of $\mathbb{Z}_2$ in $H_0(K^1, N_2)$. Define $\phi\colon \mathbb{Z}_2[\![K^1]\!] \to N_2$ by $\phi(k) = ke'$. Then

$$\mathrm{Tor}_0^{\mathbb{Z}_2[\![K^1]\!]}(\mathbb{F}_2, \phi)$$

is an isomorphism, and

$$\mathrm{Tor}_1^{\mathbb{Z}_2[\![K^1]\!]}(\mathbb{F}_2, \phi)$$

is surjective. By Lemma A.1.3 of the appendix, ϕ is an isomorphism of $\mathbb{Z}_2[\![K^1]\!]$–modules. $\square$

Splicing the exact sequences (3.1.1), (3.1.3) and (3.1.6) gives an exact sequence

$$0 \to N_2 \to \mathscr{C}_2 \to \mathscr{C}_1 \to \mathscr{C}_0 \to \mathbb{Z}_2 \to 0, \tag{3.1.7}$$

which is a free resolution of $\mathbb{Z}_2$ as a trivial $\mathbb{Z}_2[\![K^1]\!]$–module. The next goal is to find an isomorphism $N_2 \cong \mathbb{Z}_2[\![\mathbb{S}_2^1/G_{24}']\!]$, where $G_{24}' = \pi G_{24} \pi^{-1}$ represents the other conjugacy class of maximal finite subgroups in $\mathbb{S}_2^1$. To prove this, we will need a few results. Before stating these, we introduce some notation.

Let G be a subgroup of $\mathbb{S}_2$ which contains the central subgroup C_2. We define

$$PG := G/C_2.$$

We let

$$A_4 := PG_{24}, \tag{3.1.8}$$

$$A_4' := PG_{24}'. \tag{3.1.9}$$

The choice of notation is justified by the fact that both of these groups are isomorphic to the alternating group on four letters. Note also that, since C_2 is central, $PC_6 \cong C_3$ and $P\mathbb{S}_2^1 \cong K^1 \rtimes A_4$. Therefore, for any G which contains C_2,

$$\mathbb{Z}_2[\![\mathbb{S}_2^1/G]\!] \cong \mathbb{Z}_2[\![P\mathbb{S}_2^1/PG]\!]$$

as $\mathbb{Z}_2[\![\mathbb{S}_2^1]\!]$–modules. To prove that $N_2 \cong \mathbb{Z}_2[\![\mathbb{S}_2^1/G_{24}']\!]$, it will thus be sufficient to prove that

$$N_2 \cong \mathbb{Z}_2[\![P\mathbb{S}_2^1/A_4']\!]$$

as $\mathbb{Z}_2[\![P\mathbb{S}_2^1]\!]$–modules.

We showed in Corollary 2.5.12 that K^1 is a Poincaré duality group (see Definition 2.5.4). Further, there is an isomorphism of $\mathbb{Z}_2[\![K^1]\!]$–modules

$$\mathbb{Z}_2[\![P\mathbb{S}_2^1/A_4']\!] \cong \mathbb{Z}_2[\![K^1]\!].$$

Hence,

$$H^n(K^1, \mathbb{Z}_2[\![P\mathbb{S}_2^1/A_4']\!]) \cong \begin{cases} \mathbb{Z}_2 & \text{if } n = 3, \\ 0 & \text{otherwise.} \end{cases} \tag{3.1.10}$$

Lemma 3.1.4 *The inclusion $\iota\colon K^1 \to P\mathbb{S}_2^1$ induces an isomorphism*

$$\iota^*\colon H^3(P\mathbb{S}_2^1, \mathbb{Z}_2[\![P\mathbb{S}_2^1/A_4']\!]) \to H^3(K^1, \mathbb{Z}_2[\![P\mathbb{S}_2^1/A_4']\!]).$$

Proof The action of A_4 on $H^3(K^1, \mathbb{Z}_2[\![P\mathbb{S}_2^1/A_4']\!])$ is trivial. This follows from the fact that there are no nontrivial one-dimensional representations of A_4. Indeed,

$$\mathrm{Hom}(A_4, Gl_1(\mathbb{Z}_2)) = H^1(A_4, \mathbb{Z}_2^\times).$$

and $H^1(A_4, \mathbb{Z}_2^\times) = 0$. Since $P\mathbb{S}_2^1 \cong K^1 \rtimes A_4$, there is a spectral sequence

$$H^p(A_4, H^q(K^1, \mathbb{Z}_2[\![P\mathbb{S}_2^1/A_4']\!])) \Longrightarrow H^{p+q}(P\mathbb{S}_2^1, \mathbb{Z}_2[\![P\mathbb{S}_2^1/A_4']\!]).$$

Because the action of A_4 on $H^3(K^1, \mathbb{Z}_2[\![P\mathbb{S}_2^1/A_4']\!])$ is trivial, (3.1.10) implies that the edge homomorphism

$$H^3(P\mathbb{S}_2^1, \mathbb{Z}_2[\![P\mathbb{S}_2^1/A_4']\!]) \to H^0(A_4, H^3(K^1, \mathbb{Z}_2[\![P\mathbb{S}_2^1/A_4']\!]))$$

induced by the inclusion $\iota\colon K^1 \to P\mathbb{S}_2^1$ is an isomorphism. □

Lemma 3.1.5 *There are surjections*

$$\eta\colon \mathrm{Hom}_{\mathbb{Z}_2[\![P\mathbb{S}_2^1]\!]}(N_2, \mathbb{Z}_2[\![P\mathbb{S}_2^1/A_4']\!]) \to H^3(P\mathbb{S}_2^1, \mathbb{Z}_2[\![P\mathbb{S}_2^1/A_4']\!]),$$
$$\eta'\colon \mathrm{Hom}_{\mathbb{Z}_2[\![K^1]\!]}(N_2, \mathbb{Z}_2[\![P\mathbb{S}_2^1/A_4']\!]) \to H^3(K^1, \mathbb{Z}_2[\![P\mathbb{S}_2^1/A_4']\!])$$

making the following diagram commute

$$\begin{tikzcd}
\mathrm{Hom}_{\mathbb{Z}_2[\![P\mathbb{S}_2^1]\!]}(N_2,\ \mathbb{Z}_2[\![P\mathbb{S}_2^1/A_4']\!]) \arrow[r, "\eta"] \arrow[d, "\iota^*"'] & H^3(P\mathbb{S}_2^1, \mathbb{Z}_2[\![P\mathbb{S}_2^1/A_4']\!]) \arrow[d, "\iota^*"] \\
\mathrm{Hom}_{\mathbb{Z}_2[\![K^1]\!]}(N_2, \mathbb{Z}_2[\![P\mathbb{S}_2^1/A_4']\!]) \arrow[r, "\eta'"] & H^3(K^1, \mathbb{Z}_2[\![P\mathbb{S}_2^1/A_4']\!])
\end{tikzcd} \tag{3.1.11}$$

where ι^* is the map induced by the inclusion $\iota\colon K^1 \to P\mathbb{S}_2^1$.

Proof Let $\mathscr{B}_p = \mathscr{C}_p$ for $0 \le p < 3$ and $\mathscr{B}_3 = N_2$. Resolving $\mathscr{B}_p$ by projective $\mathbb{Z}_2[\![P\mathbb{S}_2^1]\!]$–modules gives rise to spectral sequences

$$E_1^{p,q} \cong \mathrm{Ext}^q_{\mathbb{Z}_2[\![P\mathbb{S}_2^1]\!]}(\mathscr{B}_p, \mathbb{Z}_2[\![P\mathbb{S}_2^1/A_4']\!]) \Longrightarrow H^{p+q}(P\mathbb{S}_2^1, \mathbb{Z}_2[\![P\mathbb{S}_2^1/A_4']\!])$$

and

$$F_1^{p,q} \cong \mathrm{Ext}^q_{\mathbb{Z}_2[\![K^1]\!]}(\mathscr{B}_p, \mathbb{Z}_2[\![P\mathbb{S}_2^1/A_4']\!]) \Longrightarrow H^{p+q}(K^1, \mathbb{Z}_2[\![P\mathbb{S}_2^1/A_4']\!]).$$

These are first quadrant cohomology spectral sequences, with differentials

$$d_r\colon E_r^{p,q} \to E_r^{p+r,q-r+1}$$

and

$$d_r\colon F_r^{p,q} \to F_r^{p+r,q-r+1}.$$

Further, $\iota\colon K^1 \to P\mathbb{S}_2^1$ induces a map of spectral sequences

$$\iota^*\colon E_r^{p,q} \to F_r^{p,q}.$$

Let η be the edge homomorphism

$$\eta\colon E_1^{3,0} \to H^3(P\mathbb{S}_2^1, \mathbb{Z}_2[\![P\mathbb{S}_2^1/A_4']\!]),$$

and let η' be the edge homomorphism

$$\eta'\colon F_1^{3,0} \to H^3(K^1, \mathbb{Z}_2[\![P\mathbb{S}_2^1/A_4']\!]).$$

First, note that since the modules $\mathscr{B}_p$ are projective $\mathbb{Z}_2[\![K^1]\!]$–modules, $F_r^{p,q}$ collapses with $F_\infty^{p,q} = 0$ for $q > 0$ so that

$$F_\infty^{3,0} \to H^3(K^1; \mathbb{Z}_2[\![P\mathbb{S}_2^1/A_4']\!])$$

is surjective. Hence, η' is surjective.

In order to show that η is surjective, it is sufficient to show that $E_1^{3-q,q} = 0$ for $q > 0$. For $q = 1$ and $q = 2$, this follows from the fact that $\mathbb{Z}_2[\![P\mathbb{S}_2^1/C_3]\!]$ is a projective $\mathbb{Z}_2[\![P\mathbb{S}_2^1]\!]$–module. Hence, if $q > 0$, then

$$\operatorname{Ext}^q_{\mathbb{Z}_2[\![P\mathbb{S}_2^1]\!]}(\mathbb{Z}_2[\![P\mathbb{S}_2^1/C_3]\!], \mathbb{Z}_2[\![P\mathbb{S}_2^1/A_4']\!]) = 0.$$

It remains to show that

$$E_1^{0,3} = \operatorname{Ext}^3_{\mathbb{Z}_2[\![P\mathbb{S}_2^1]\!]}(\mathscr{B}_0, \mathbb{Z}_2[\![P\mathbb{S}_2^1/A_4']\!])$$

is zero, where $\mathscr{B}_0 = \mathbb{Z}_2[\![P\mathbb{S}_2^1/A_4]\!]$.

Let $V \cong C_2 \times C_2$ be the 2–Sylow subgroup of A_4. Then

$$\begin{aligned} E_1^{0,3} = \operatorname{Ext}^3_{\mathbb{Z}_2[\![P\mathbb{S}_2^1]\!]}(\mathscr{B}_0, \mathbb{Z}_2[\![P\mathbb{S}_2^1/A_4']\!]) &\cong H^3(A_4, \mathbb{Z}_2[\![P\mathbb{S}_2^1/A_4']\!]) \\ &\cong H^3(V, \mathbb{Z}_2[\![P\mathbb{S}_2^1/A_4']\!])^{C_3}. \end{aligned}$$

Let $G_n = PF_{n/2}S_2^1 \rtimes A_4'$ and $X_n = P\mathbb{S}_2^1/G_n$. The profinite A_4–set $P\mathbb{S}_2^1/A_4'$ is isomorphic to the inverse limit of the finite A_4–sets X_n. There is an exact sequence

$$0 \to \lim{}^1 H^2(V, \mathbb{Z}_2[X_n]) \to H^3(V, \mathbb{Z}_2[\![P\mathbb{S}_2^1/A_4']\!]) \to \lim_n H^3(V, \mathbb{Z}_2[X_n]) \to 0.$$

Since the groups $H^2(V, \mathbb{Z}_2[X_n])$ are finite, the Mittag–Leffler condition is satisfied and $\lim{}^1 H^2(V, \mathbb{Z}_2[X_n]) = 0$. Hence,

$$H^3(V, \mathbb{Z}_2[\![P\mathbb{S}_2^1/A_4']\!]) \cong \lim_n H^3(V, \mathbb{Z}_2[X_n]).$$

We will show that there is an integer N such that $H^3(V, \mathbb{Z}_2[X_n]) = 0$ for all $n \geq N$. This implies that $H^3(V, \mathbb{Z}_2[\![P\mathbb{S}_2^1/A_4']\!])$ is zero, so that $E_1^{0,3} = 0$.

Note that

$$H^3(V, \mathbb{Z}_2[X_n]) \cong \bigoplus_{x \in V \backslash X_n / G_n} H^3(V, \mathbb{Z}_2[V/V_x]) \cong \bigoplus_{x \in V \backslash X_n / G_n} H^3(V_x, \mathbb{Z}_2)$$

for $V_x = \{g \in V \mid gxG_n = xG_n\}$. If the inclusion $V_x \subseteq V$ is an equality, then $x^{-1}Vx$ is a subgroup of G_n. We show that there exists an integer N such that, for all $n \geq N$, there is no element x in $P\mathbb{S}_2^1$ such that $x^{-1}Vx \subseteq G_n$. This implies that, for $n \geq N$, for all choices of coset representatives $x \in V \backslash X_n / G_n$, the group V_x is either trivial or it has order 2. In both cases, $H^3(V_x, \mathbb{Z}_2) = 0$. Hence, for $n \geq N$, $H^3(V, \mathbb{Z}_2[X_n]) = 0$.

Suppose that there is a sequence of integers n_m and elements x_{n_m} such that $x_{n_m}^{-1} V x_{n_m} \subseteq G_{n_m}$. Since $P\mathbb{S}_2^1$ is compact, we can choose the sequence (x_{n_m}) to converge to some element y. The groups G_n are closed and nested, so the continuity of the group multiplication implies that $y^{-1}Vy \subseteq G_n$ for all $n \in \mathbb{N}$. Therefore,

$$y^{-1}Vy \subseteq \bigcap_n G_n = A_4',$$

and hence $y^{-1}Vy = V'$, where V' is the 2–Sylow subgroup of A_4'. However, it follows from Proposition 2.4.6 that V and V' are not conjugate in $P\mathbb{S}_2^1$. Therefore, such a sequence cannot exist, and there must be some integer N such that, for all $n \geq N$, there is no x in $P\mathbb{S}_2^1$ such that $x^{-1}Vx \subseteq G_n$. □

Theorem 3.1.6 *There is an isomorphism of left $\mathbb{Z}_2[\![\mathbb{S}_2^1]\!]$–modules*

$$\phi\colon \mathbb{Z}_2[\![\mathbb{S}_2^1/G_{24}']\!] \to N_2,$$

where $G_{24}' = \pi G_{24} \pi^{-1}$.

Proof It suffices to construct an isomorphism $\varphi\colon N_2 \to \mathbb{Z}_2[\![P\mathbb{S}_2^1/PG_{24}']\!]$ of left $\mathbb{Z}_2[\![P\mathbb{S}_2^1]\!]$–modules. The result then follows by letting $\phi = \varphi^{-1}$, considered as a map of $\mathbb{Z}_2[\![\mathbb{S}_2^1]\!]$–modules.

Recall from Corollary 2.5.12 that K^1 is an orientable Poincaré duality group of dimension 3, as in Definition 2.5.4. That is, the compact dualizing module $D_2(K^1)$ is isomorphic to the trivial $\mathbb{Z}_2[\![K^1]\!]$–module $\mathbb{Z}_2$ and $H_3(K^1, \mathbb{Z}_2) \cong \mathbb{Z}_2$. Choose a generator $[K^1]$ of $H_3(K^1, \mathbb{Z}_2)$. As in Remark 2.5.5, there is a natural isomorphism

$$H^{3-*}(K^1, -) \xrightarrow{\cap [K^1]} H_*(K^1, -).$$

Let

$$\nu\colon H_3(K^1,\mathbb{Z}_2)\to \mathbb{Z}_2\otimes_{\mathbb{Z}_2[\![K^1]\!]} N_2$$

be the edge homomorphism for the homology spectral sequence obtained from (3.1.7). Define

$$\mathrm{ev}\colon \mathrm{Hom}_{\mathbb{Z}_2}(\mathbb{Z}_2\otimes_{\mathbb{Z}_2[\![K^1]\!]} N_2, \mathbb{Z}_2\otimes_{\mathbb{Z}_2[\![K^1]\!]} \mathbb{Z}_2[\![P\mathbb{S}_2^1/A_4']\!])\to H_0(K^1,\mathbb{Z}_2[\![P\mathbb{S}_2^1/A_4']\!])$$

by

$$\mathrm{ev}(f)=f(\nu([K^1])).$$

Let $\iota\colon K^1\to P\mathbb{S}_2^1$ be the inclusion. Let η and η' be the edge homomorphisms of Lemma 3.1.5. We obtain the following commutative diagram:

$$\begin{tikzcd}
\mathrm{Hom}_{\mathbb{Z}_2[\![P\mathbb{S}_2^1]\!]}(N_2,\mathbb{Z}_2[\![P\mathbb{S}_2^1/A_4']\!]) \arrow[r, "\eta"] \arrow[d, "\iota^*"'] & H^3(P\mathbb{S}_2^1,\mathbb{Z}_2[\![P\mathbb{S}_2^1/A_4']\!]) \arrow[d, "\iota^*"] \\
\mathrm{Hom}_{\mathbb{Z}_2[\![K^1]\!]}(N_2,\mathbb{Z}_2[\![P\mathbb{S}_2^1/A_4']\!]) \arrow[r, "\eta'"] \arrow[d, "\mathbb{Z}_2\otimes_{\mathbb{Z}_2[\![K^1]\!]}-"'] & H^3(K^1,\mathbb{Z}_2[\![P\mathbb{S}_2^1/A_4']\!]) \arrow[d, "\cap[K^1]"] \\
\mathrm{Hom}_{\mathbb{Z}_2}(\mathbb{Z}_2\otimes_{\mathbb{Z}_2[\![K^1]\!]}N_2,\mathbb{Z}_2\otimes_{\mathbb{Z}_2[\![K^1]\!]}\mathbb{Z}_2[\![P\mathbb{S}_2^1/A_4']\!]) \arrow[r, "\mathrm{ev}"] & H_0(K^1,\mathbb{Z}_2[\![P\mathbb{S}_2^1/A_4']\!])
\end{tikzcd}$$

Since $\cap[K^1]\circ\eta'$ is surjective, so is the map ev. Both N_2 and $\mathbb{Z}_2[\![P\mathbb{S}_2^1/A_4']\!]$ are free of rank one over $\mathbb{Z}_2[\![K^1]\!]$. Hence, $\mathbb{Z}_2\otimes_{\mathbb{Z}_2[\![K^1]\!]} N_2$ and $\mathbb{Z}_2\otimes_{\mathbb{Z}_2[\![K^1]\!]}\mathbb{Z}_2[\![P\mathbb{S}_2^1/A_4']\!]$ are abstractly isomorphic to $\mathbb{Z}_2$. Since ev is a surjective group homomorphism from $\mathbb{Z}_2$ to itself, it is an isomorphism. It follows from Lemma A.1.3 that any element of $\mathrm{Hom}_{\mathbb{Z}_2[\![K^1]\!]}(N_2,\mathbb{Z}_2[\![P\mathbb{S}_2^1/A_4']\!])$ that becomes a unit after applying $\mathbb{Z}_2\otimes_{\mathbb{Z}_2[\![K^1]\!]}-$ is an isomorphism. By Lemma 3.1.5, the composite $\cap[K^1]\circ\iota^*\circ\eta$ is surjective. Therefore, we can choose φ in $\mathrm{Hom}_{\mathbb{Z}_2[\![P\mathbb{S}_2^1]\!]}(N_2,\mathbb{Z}_2[\![P\mathbb{S}_2^1/A_4']\!])$ such that $\cap[K^1]\circ\iota^*\circ\eta(\varphi)$ is a generator of $H_0(K^1,\mathbb{Z}_2[\![P\mathbb{S}_2^1/A_4']\!])$. Then $\iota^*(\varphi)$ in $\mathrm{Hom}_{\mathbb{Z}_2[\![K^1]\!]}(N_2,\mathbb{Z}_2[\![P\mathbb{S}_2^1/A_4']\!])$ is an isomorphism, and hence φ must be an isomorphism. □

Combining the previous results, we can finally prove Theorem 1.2.1. We restate it here for convenience.

Theorem 3.1.7 *Let $\mathbb{Z}_2$ be the trivial $\mathbb{Z}_2[\![\mathbb{S}_2^1]\!]$–module. There is an exact sequence of complete $\mathbb{Z}_2[\![\mathbb{S}_2^1]\!]$–modules*

$$0\to\mathscr{C}_3\xrightarrow{\partial_3}\mathscr{C}_2\xrightarrow{\partial_2}\mathscr{C}_1\xrightarrow{\partial_1}\mathscr{C}_0\xrightarrow{\varepsilon}\mathbb{Z}_2\to 0,$$

where $\mathscr{C}_0\cong\mathbb{Z}_2[\![\mathbb{S}_2^1/G_{24}]\!]$ and $\mathscr{C}_1\cong\mathscr{C}_2\cong\mathbb{Z}_2[\![\mathbb{S}_2^1/C_6]\!]$ and $\mathscr{C}_3=\mathbb{Z}_2[\![\mathbb{S}_2^1/G_{24}']\!]$. Further, this is a free resolution of the trivial $\mathbb{Z}_2[\![K^1]\!]$–module $\mathbb{Z}_2$.

Proof Let
$$\mathscr{C}_3 := \mathbb{Z}_2[\![\mathbb{S}_2^1/G'_{24}]\!].$$
Let $\phi\colon \mathscr{C}_3 \to N_2$ be the isomorphism of Theorem 3.1.6. Let $\partial_3\colon \mathscr{C}_3 \to \mathscr{C}_2$ be the isomorphism ϕ followed by the inclusion of N_2 in $\mathscr{C}_2$. This gives an exact sequence

$$0 \to \mathscr{C}_3 \to \mathscr{C}_2 \xrightarrow{\partial_2} N_1 \to 0. \tag{3.1.12}$$

Splicing the exact sequences of (3.1.1), (3.1.3) and (3.1.12) finishes the proof. □

The exact sequence of Theorem 3.1.7 is called the *algebraic duality resolution*. The duality properties it satisfies will be described in Section 3.3.

3.2 The algebraic duality resolution spectral sequence

The algebraic duality resolution gives rise to a spectral sequence called the *algebraic duality resolution spectral sequence*, which we describe here. The following result is a refinement of Theorem 1.2.4, which was stated in Section 1.2. We define
$$Q'_8 := \pi Q_8 \pi^{-1}.$$
We also let V be the 2–Sylow subgroup of A_4 and V' be the 2–Sylow subgroup of A'_4, where $A_4 \cong PG_{24}$ and $A'_4 = PG'_{24}$ as defined in (3.1.8) and (3.1.9).

Theorem 3.2.1 *Let M be a profinite $\mathbb{Z}_2[\![\mathbb{S}_2^1]\!]$–module. There is a first quadrant spectral sequence*
$$E_1^{p,q} = \mathrm{Ext}^q_{\mathbb{Z}_2[\![\mathbb{S}_2^1]\!]}(\mathscr{C}_p, M) \Longrightarrow H^{p+q}(\mathbb{S}_2^1, M)$$
with differentials $d_r\colon E_r^{p,q} \to E_r^{p+r,q-r+1}$. Further,
$$E_1^{p,q} \cong \begin{cases} H^q(G_{24}, M) & \text{if } p = 0, \\ H^q(C_6, M) & \text{if } p = 1, 2, \\ H^q(G'_{24}, M) & \text{if } p = 3. \end{cases}$$
Similarly, there are first quadrant spectral sequences
$$E_1^{p,q} = \mathrm{Ext}^q_{\mathbb{Z}_2[\![G]\!]}(\mathscr{C}_p, M) \Longrightarrow H^{p+q}(G, M),$$
where G is S_2^1, $P\mathbb{S}_2^1$ or PS_2^1. The E_1–term is
$$E_1^{p,q} \cong \begin{cases} H^p(Q_8; M) & \text{if } q = 0, \\ H^p(C_2; M) & \text{if } q = 1, 2, \\ H^p(Q'_8; M) & \text{if } q = 3 \end{cases}$$

when G is S_2^1,

$$E_1^{p,q} \cong \begin{cases} H^p(A_4; M) & \text{if } q = 0, \\ H^p(C_3; M) & \text{if } q = 1, 2, \\ H^p(A_4'; M) & \text{if } q = 3 \end{cases}$$

when G is $P\mathbb{S}_2^1$ and

$$E_1^{p,q} \cong \begin{cases} H^p(V; M) & \text{if } q = 0, \\ H^p(\{e\}; M) & \text{if } q = 1, 2, \\ H^p(V'; M) & \text{if } q = 3 \end{cases}$$

when G is PS_2^1.

Proof There are two equivalent constructions. First, recall that the algebraic duality resolution is spliced from the exact sequences

$$0 \to N_i \to \mathscr{C}_i \to N_{i-1} \to 0, \tag{3.2.1}$$

with $\mathscr{C}_3 = N_2$ and $N_{-1} = \mathbb{Z}_2$. The exact couple

$$\begin{tikzcd} \mathrm{Ext}(N_*, M) \arrow[rr, "\delta_*"] & & \mathrm{Ext}(N_{*-1}, M) \arrow[dl, "r_*"] \\ & \mathrm{Ext}(\mathscr{C}_*, M) \arrow[ul, "i^*"] & \end{tikzcd}$$

gives rise to the algebraic duality resolution spectral sequence.

Alternatively, one can resolve each $\mathscr{C}_\bullet \to \mathbb{Z}_2$ into a double complex of projective finitely generated $\mathbb{Z}_2[\![\mathbb{S}_2^1]\!]$–modules. The total complex $\mathrm{Tot}(P_{p,q})$ for $p \geq 0$ is a projective resolution of $\mathbb{Z}_2$ as a $\mathbb{Z}_2[\![\mathbb{S}_2^1]\!]$–module. The homology of the double complex $\mathrm{Hom}_{\mathbb{Z}_2[\![\mathbb{S}_2^1]\!]}(\mathrm{Tot}(P_{p,q}), M)$ is

$$\mathrm{Ext}^{p+q}_{\mathbb{Z}_2[\![\mathbb{S}_2^1]\!]}(\mathbb{Z}_2, M) \cong H^{p+q}(\mathbb{S}_2^1, M).$$

The identification of the E_1–term follows from Shapiro's Lemma A.1.2 of the appendix. Indeed, any finite subgroup H of $\mathbb{S}_2^1$ is closed. Further, since $\mathbb{S}_2^1 \cong S_2^1 \rtimes C_3$,

$$\begin{aligned} \mathrm{Ext}^q_{\mathbb{Z}_2[\![\mathbb{S}_2^1]\!]}(\mathbb{Z}_2[\![\mathbb{S}_2^1]\!] \otimes_{\mathbb{Z}_2[H]} \mathbb{Z}_2, M) &\cong \big(\mathrm{Ext}^q_{\mathbb{Z}_2[\![S_2^1]\!]}(\mathbb{Z}_2[\![S_2^1]\!] \otimes_{\mathbb{Z}_2[\mathrm{Syl}_2(H)]} \mathbb{Z}_2, M)\big)^{C_3} \\ &\cong \big(\mathrm{Ext}^q_{\mathbb{Z}_2[\mathrm{Syl}_2(H)]}(\mathbb{Z}_2, M)\big)^{C_3} \cong H^q(H, M). \end{aligned}$$

For the groups S_2^1, $P\mathbb{S}_2$ and PS_2^1, one applies the same construction, keeping the following isomorphisms in mind. Let $H \subseteq \mathbb{S}_2^1$ be a finite subgroup which contains C_6

and let $PH = H/C_2$. There are isomorphisms

$$\mathbb{Z}_2[\![\mathbb{S}_2^1/H]\!] \cong \mathbb{Z}_2[\![P\mathbb{S}_2^1]\!] \otimes_{\mathbb{Z}_2[PH]} \mathbb{Z}_2 \cong \mathbb{Z}_2[\![P\mathbb{S}_2^1/PH]\!]$$

and

$$\mathbb{Z}_2[\![\mathbb{S}_2^1/H]\!] \cong \mathbb{Z}_2[\![PS_2^1]\!] \otimes_{\mathbb{Z}_2[\mathrm{Syl}_2(PH)]} \mathbb{Z}_2 \cong \mathbb{Z}_2[\![P\mathbb{S}_2^1/\,\mathrm{Syl}_2(PH)]\!]$$

as $\mathbb{Z}_2[\![P\mathbb{S}_2^1]\!]$ and $\mathbb{Z}_2[\![PS_2^1]\!]$–modules, respectively. □

3.3 The duality

The algebraic duality resolution of Theorem 3.1.7 owes its name to the fact that it satisfies a certain twisted duality. This duality is crucial for computations as it allows us to understand the map $\partial_3\colon \mathscr{C}_3 \to \mathscr{C}_2$.

Let $\mathrm{Mod}(\mathbb{S}_2^1)$ denote the category of finitely generated left $\mathbb{Z}_2[\![\mathbb{S}_2^1]\!]$–modules. Let $\pi = 1+2\omega$ in $\mathbb{S}_2$ be as defined in (2.3.3). For M in $\mathrm{Mod}(\mathbb{S}_2^1)$, let $c_\pi(M)$ denote the left $\mathbb{Z}_2[\![\mathbb{S}_2^1]\!]$–module whose underlying $\mathbb{Z}_2$–module is M, but for which the action of γ in $\mathbb{S}_2^1$ on an element m in $c_\pi(M)$ is given by

$$\gamma\cdot m = \pi\gamma\pi^{-1}m.$$

If $\phi\colon M \to N$ is a morphism of left $\mathbb{Z}_2[\![\mathbb{S}_2^1]\!]$–modules, let $c_\pi(\phi)\colon c_\pi(M) \to c_\pi(N)$ be given by

$$c_\pi(\phi)(m) = \phi(m).$$

Then $c_\pi\colon \mathrm{Mod}(\mathbb{S}_2^1) \to \mathrm{Mod}(\mathbb{S}_2^1)$ is a functor. In fact, c_π is an involution, since $\pi^2 = -3$ is in the center of $\mathbb{S}_2$. We can now prove Theorem 1.2.2, which is restated here for convenience.

Theorem 3.3.1 (Henn, Karamanov and Mahowald, unpublished) *There exists an isomorphism of complexes of left $\mathbb{Z}_2[\![\mathbb{S}_2^1]\!]$–modules:*

$$\begin{tikzcd}
0 \arrow[r] & \mathscr{C}_3 \arrow[r, "\partial_3"] \arrow[d, "f_3"] & \mathscr{C}_2 \arrow[r, "\partial_2"] \arrow[d, "f_2"] & \mathscr{C}_1 \arrow[r, "\partial_1"] \arrow[d, "f_1"] & \mathscr{C}_0 \arrow[r, "\varepsilon"] \arrow[d, "f_0"] & \mathbb{Z}_2 \arrow[r] \arrow[d, equal] & 0 \\
0 \arrow[r] & c_\pi(\mathscr{C}_0^*) \arrow[r, "c_\pi(\partial_1^*)"] & c_\pi(\mathscr{C}_1^*) \arrow[r, "c_\pi(\partial_2^*)"] & c_\pi(\mathscr{C}_2^*) \arrow[r, "c_\pi(\partial_3^*)"] & c_\pi(\mathscr{C}_3^*) \arrow[r, "\overline{\varepsilon}"] & \mathbb{Z}_2 \arrow[r] & 0
\end{tikzcd}$$

Proof The proof is similar to the proof of Henn, Karamanov, Mahowald [18, Proposition 3.8]. Let $\mathscr{C}_p^* = \mathrm{Hom}_{\mathbb{Z}_2[\![\mathbb{S}_2^1]\!]}(\mathscr{C}_p, \mathbb{Z}_2[\![\mathbb{S}_2^1]\!])$ and $\partial_p^* = \mathrm{Hom}_{\mathbb{Z}_2[\![\mathbb{S}_2^1]\!]}(\partial_p, \mathbb{Z}_2[\![\mathbb{S}_2^1]\!])$ be the $\mathbb{Z}_2[\![\mathbb{S}_2^1]\!]$–duals of $\mathscr{C}_p$ and ∂_p in the sense of Equation (A.1.1). The resolution of Theorem 3.1.7 gives rise to a complex

$$0 \to \mathscr{C}_0^* \xrightarrow{\partial_1^*} \mathscr{C}_1^* \xrightarrow{\partial_2^*} \mathscr{C}_2^* \xrightarrow{\partial_3^*} \mathscr{C}_3^* \to 0. \tag{3.3.1}$$

Because K^1 has finite index in $\mathbb{S}_2^1$, the induced and coinduced modules of $\mathbb{Z}_2[\![K^1]\!]$ are isomorphic; see Symonds and Weigel [35, Section 3.3]. Therefore

$$\mathrm{Hom}_{\mathbb{Z}_2[\![\mathbb{S}_2^1]\!]}(\mathscr{C}_p, \mathbb{Z}_2[\![\mathbb{S}_2^1]\!]) \cong \mathrm{Hom}_{\mathbb{Z}_2[\![K^1]\!]}(\mathscr{C}_p, \mathbb{Z}_2[\![K^1]\!])$$

and the homology of the complex (3.3.1) is $H^n(K^1, \mathbb{Z}_2[\![K^1]\!])$. By Corollary 2.5.12, $H^n(K^1, \mathbb{Z}_2[\![K^1]\!])$ is 0 for $n \neq 3$ and $\mathbb{Z}_2$ for $n = 3$. Further, the action of G_{24} on $H^3(K^1, \mathbb{Z}_2[\![K^1]\!]) \cong \mathbb{Z}_2$ is trivial, as there are no nontrivial one dimensional 2–adic representations of G_{24}. Hence, (3.3.1) is a resolution of $\mathbb{Z}_2$ as a trivial $\mathbb{Z}_2[\![\mathbb{S}_2^1]\!]$–module.

The module $\mathscr{C}_p^*$ is of the form $\mathbb{Z}_2[\![\mathbb{S}_2^1/H]\!]$ via the isomorphism t defined in (A.1.2). Let $\overline{\varepsilon}$ be the augmentation

$$\overline{\varepsilon}\colon \mathscr{C}_3^* \to \mathbb{Z}_2.$$

Because the augmentation $\varepsilon\colon \mathbb{Z}_2[\![K^1]\!] \to \mathbb{Z}_2$ induces an isomorphism

$$\mathrm{Hom}_{\mathbb{Z}_2[\![K^1]\!]}(\mathbb{Z}_2, \mathbb{Z}_2) \cong \mathrm{Hom}_{\mathbb{Z}_2[\![K^1]\!]}(\mathbb{Z}_2[\![K^1]\!], \mathbb{Z}_2),$$

one can choose an isomorphism $H^3(K, \mathbb{Z}_2[\![K^1]\!]) \to \mathbb{Z}_2$ making the following diagram commute:

$$\begin{tikzcd} \mathscr{C}_3^* \arrow[r] \arrow[d, "\overline{\varepsilon}"] & H^3(K^1, \mathbb{Z}_2[\![K^1]\!]) \arrow[dl, dotted] \\ \mathbb{Z}_2 & \end{tikzcd}$$

Therefore, the dual resolution is given by

$$0 \to \mathscr{C}_0^* \xrightarrow{\partial_1^*} \mathscr{C}_1^* \xrightarrow{\partial_2^*} \mathscr{C}_2^* \xrightarrow{\partial_3^*} \mathscr{C}_3^* \xrightarrow{\overline{\varepsilon}} \mathbb{Z}_2 \to 0.$$

Take the image of this resolution in $\mathrm{Mod}(\mathbb{S}_2^1)$ under the involution c_π. Let e_3^π be the canonical generator of $c_\pi(\mathscr{C}_3^*)$. The map $f_0\colon \mathscr{C}_0 \to c_\pi(\mathscr{C}_3^*)$ defined by

$$f_0(e_0) = e_3^\pi$$

is an isomorphism of $\mathbb{Z}_2[\![\mathbb{S}_2^1]\!]$–modules, and the following diagram is commutative:

$$\begin{tikzcd}
\mathscr{C}_0 \arrow[r, "\varepsilon"] \arrow[d, "f_0"] & \mathbb{Z}_2 \arrow[r] \arrow[d, equal] & 0 \\
c_\pi(\mathscr{C}_3^*) \arrow[r, "\overline{\varepsilon}"] & \mathbb{Z}_2 \arrow[r] & 0
\end{tikzcd}$$

Therefore, f_0 induces an isomorphism $\ker\varepsilon \cong \ker\overline{\varepsilon}$. As both

$$\mathscr{C}_2 \xrightarrow{\partial_2} \mathscr{C}_1 \xrightarrow{\partial_1} \ker\varepsilon$$

and

$$c_\pi(\mathscr{C}_2^*) \xrightarrow{c_\pi(\partial_2^*)} \mathscr{C}_3^* \xrightarrow{c_\pi(\partial_3^*)} \ker\overline{\varepsilon}$$

are the beginning of projective resolutions of $\ker\varepsilon$ and $\ker\overline{\varepsilon}$ as $\mathbb{Z}_2[\![P\mathbb{S}_2^1]\!]$–modules, f_0 lifts to a chain map:

$$\begin{tikzcd}
0 \arrow[r] & \mathscr{C}_3 \arrow[r, "\partial_3"] \arrow[d, "f_3"] & \mathscr{C}_2 \arrow[r, "\partial_2"] \arrow[d, "f_2"] & \mathscr{C}_1 \arrow[r, "\partial_1"] \arrow[d, "f_1"] & \ker\varepsilon \arrow[r] \arrow[d, "f_0"] & 0 \\
0 \arrow[r] & c_\pi(\mathscr{C}_0^*) \arrow[r, "c_\pi(\partial_1^*)"] & c_\pi(\mathscr{C}_1^*) \arrow[r, "c_\pi(\partial_2^*)"] & c_\pi(\mathscr{C}_2^*) \arrow[r, "c_\pi(\partial_3^*)"] & \ker\overline{\varepsilon} \arrow[r] & 0
\end{tikzcd}$$

Let $PS_2^1 = S_2^1/C_2$, where S_2^1 denotes the 2–Sylow subgroup of $\mathbb{S}_2^1$. By construction, f_0 is an isomorphism, which implies that $\mathbb{F}_2 \otimes_{\mathbb{Z}_2[\![PS_2^1]\!]} f_1$ and $\mathbb{F}_2 \otimes_{\mathbb{Z}_2[\![PS_2^1]\!]} f_2$ are isomorphisms. As $\mathscr{C}_p$ and $c_\pi(\mathscr{C}_p^*)$ are projective $\mathbb{Z}_2[\![P\mathbb{S}_2^1]\!]$–modules for $p = 1, 2$, Lemma A.1.3 of the appendix implies that f_1 and f_2 are isomorphisms. Finally, f_3 must be an isomorphism by the five lemma. □

3.4 A description of the maps

This section is dedicated to proving the statements in Theorem 1.2.6. The first statement of Theorem 1.2.6 is that

$$\partial_1(e_1) = (e-\alpha)e_0.$$

This was shown in Theorem 3.1.7. In this section, we prove the remaining statements of that theorem.

The following result provides a description of the maps $\partial_3\colon \mathscr{C}_3 \to \mathscr{C}_2$ and proves the last part of Theorem 1.2.6. It is a consequence of Theorem 3.3.1.

Theorem 3.4.1 *There are isomorphisms of $\mathbb{Z}_2[\![\mathbb{S}_2^1]\!]$–modules $g_p\colon \mathscr{C}_p \to \mathscr{C}_p$ and differentials*

$$\partial'_{p+1}\colon \mathscr{C}_{p+1} \to \mathscr{C}_p$$

such that

$$\begin{tikzcd}
0 \arrow[r] & \mathscr{C}_3 \arrow[r, "\partial_3"] \arrow[d, "g_3"] & \mathscr{C}_2 \arrow[r, "\partial_2"] \arrow[d, "g_2"] & \mathscr{C}_1 \arrow[r, "\partial_1"] \arrow[d, "g_1"] & \mathscr{C}_0 \arrow[r, "\varepsilon"] \arrow[d, "g_0"] & \mathbb{Z}_2 \arrow[r] \arrow[d, equal] & 0 \\
0 \arrow[r] & \mathscr{C}_3 \arrow[r, "\partial'_3"] & \mathscr{C}_2 \arrow[r, "\partial'_2"] & \mathscr{C}_1 \arrow[r, "\partial'_1"] & \mathscr{C}_0 \arrow[r, "\varepsilon"] & \mathbb{Z}_2 \arrow[r] & 0
\end{tikzcd} \tag{3.4.1}$$

is an isomorphism of complexes. The map $\partial'_3\colon \mathscr{C}_3 \to \mathscr{C}_2$ is given by

$$\partial'_3(e_3) = \pi(e+i+j+k)(e-\alpha^{-1})\pi^{-1}e_2. \tag{3.4.2}$$

Proof We will construct a commutative diagram:

$$\begin{tikzcd}
0 \arrow[r] & \mathscr{C}_3 \arrow[r, "\partial_3"] \arrow[d, "f_3"] & \mathscr{C}_2 \arrow[r, "\partial_2"] \arrow[d, "f_2"] & \mathscr{C}_1 \arrow[r, "\partial_1"] \arrow[d, "f_1"] & \mathscr{C}_0 \arrow[r, "\varepsilon"] \arrow[d, "f_0"] & \mathbb{Z}_2 \arrow[r] \arrow[d, equal] & 0 \\
0 \arrow[r] & c_\pi(\mathscr{C}_0^*) \arrow[r, "c_\pi(\partial_1^*)"] \arrow[d, "q_3"] & c_\pi(\mathscr{C}_1^*) \arrow[r, "c_\pi(\partial_2^*)"] \arrow[d, "q_2"] & c_\pi(\mathscr{C}_2^*) \arrow[r, "c_\pi(\partial_3^*)"] \arrow[d, "q_1"] & c_\pi(\mathscr{C}_3^*) \arrow[r, "\overline{\varepsilon}"] \arrow[d, "q_0"] & \mathbb{Z}_2 \arrow[r] \arrow[d, equal] & 0 \\
0 \arrow[r] & \mathscr{C}_3 \arrow[r, "\partial'_3"] & \mathscr{C}_2 \arrow[r, "\partial'_2"] & \mathscr{C}_1 \arrow[r, "\partial'_1"] & \mathscr{C}_0 \arrow[r, "\varepsilon"] & \mathbb{Z}_2 \arrow[r] & 0
\end{tikzcd}$$

The maps g_p will be the composites of the vertical maps. First, let $e_p^\pi \in c_\pi(\mathscr{C}_p^*)$ be the canonical generator. Define isomorphisms $q_p\colon c_\pi(M_{3-p}^*) \to M_p$ by

$$q_p(e_{3-p}^\pi) = e_p.$$

Define $g_p\colon \mathscr{C}_p \to \mathscr{C}_p$ by

$$g_p = q_p f_p$$

and $\partial'_{p+1}\colon \mathscr{C}_{p+1} \to \mathscr{C}_p$ by

$$\partial'_{p+1} = q_p c_\pi(\partial_{3-p}^*) q_{p+1}^{-1}.$$

By construction, (3.4.1) is commutative.

In order to compute ∂'_3, it is necessary to understand ∂_1^*. By definition,

$$\partial_1^*(e_0^*)(e_1) = e_0^*((e-\alpha)e_1) = (e-\alpha)\sum_{h\in G_{24}} h.$$

However,

$$\begin{aligned}(e-\alpha)\sum_{h\in G_{24}} h &= (e-\alpha)\sum_{h\in C_6} h(e+i^{-1}+j^{-1}+k^{-1})\\ &= \sum_{h\in C_6} h(e-\alpha)(e+i^{-1}+j^{-1}+k^{-1})\\ &= ((e+i+j+k)(e-\alpha^{-1})e_1^*)(e_1).\end{aligned}$$

Hence,

$$\partial_1^*(e_0^*) = (e+i+j+k)(e-\alpha^{-1})e_1^*.$$

A diagram chase shows that ∂_3' is given by (3.4.2). □

The maps $\partial_1\colon \mathscr{C}_1 \to \mathscr{C}_0$ and $\partial_3\colon \mathscr{C}_3 \to \mathscr{C}_2$ now have explicit descriptions up to isomorphisms. The map $\partial_2\colon \mathscr{C}_2 \to \mathscr{C}_1$ is harder to describe. Theorem 3.4.5 and Corollary 3.4.6 below give an estimate for this map. These are technical results which will be used in our computations in [3]. Note that Theorem 3.1.7, Theorem 3.4.1 and Corollary 3.4.6 below prove Theorem 1.2.6, which was stated in Section 1.2.

Recall that

$$\alpha_\tau = [\tau,\alpha] = \tau\alpha\tau^{-1}\alpha^{-1}.$$

We will need the following result to describe the element Θ of Lemma 3.1.3.

Lemma 3.4.2 *Let $n \geq 2$ and x be in $IF_{n/2}K^1$. There exist h_0, h_1 and h_2 in $\mathbb{Z}_2[\![F_{n/2}K^1]\!]$ such that*

$$x = \begin{cases} h_0(e-\alpha^{2^{m-1}}) + h_1(e-\alpha_i^{2^{m-1}}) + h_2(e-\alpha_j^{2^{m-1}}) & \text{if } n=2m,\\ h_0(e-\alpha^{2^{m}}) + h_1(e-\alpha_i^{2^{m-1}}) + h_2(e-\alpha_j^{2^{m-1}}) & \text{if } n=2m+1.\end{cases} \tag{3.4.3}$$

Proof Define a map of $\mathbb{Z}_2[\![F_{n/2}K^1]\!]$–modules

$$p\colon \bigoplus_{i=0}^{2} \mathbb{Z}_2[\![F_{n/2}K^1]\!]_i \to IF_{n/2}K^1$$

by sending (h_0,h_1,h_2) to the element given by (3.4.3). It is sufficient to show that the map induced by p surjects onto

$$H_1(F_{n/2}K^1,\mathbb{F}_2) \cong \mathbb{F}_2 \otimes_{\mathbb{Z}_2[\![F_{n/2}K^1]\!]} IF_{n/2}K^1.$$

By Lemma 2.2.1, $H_1(F_{n/2}K^1, \mathbb{F}_2)$ is generated by the classes

$$\begin{aligned} \alpha^{2^{m-1}},\ \alpha_i^{2^{m-1}},\ \alpha_j^{2^{m-1}} &\quad \text{if } n = 2m, \\ \alpha^{2^{m}},\ \alpha_i^{2^{m-1}},\ \alpha_j^{2^{m-1}} &\quad \text{if } n = 2m+1. \end{aligned}$$

Therefore, $\mathbb{F}_2 \otimes_{\mathbb{Z}_2[\![F_{n/2}K^1]\!]} p$ is surjective, and hence so is p. □

The ideal

$$\mathcal{I} = ((IK^1)^7, 2(IK^1)^3, 4(IK^1), 8)$$

will play a crucial role in the following estimates.

Corollary 3.4.3 *Let e_0 be the canonical generator of $\mathscr{C}_0$ and g be in $F_{8/2}K^1$. There exists h in $\mathbb{Z}_2[\![\mathbb{S}_2^1]\!]$ such that $(e-g)e_0 = h(e-\alpha)e_0$ with $h \equiv 0 \mod \mathcal{I}$.*

Proof By Lemma 3.4.2, there exist h_0, h_1 and h_2 in $\mathbb{Z}_2[\![F_{8/2}K^1]\!]$ such that

$$e - g = h_0(e-\alpha^8) + h_1(e-\alpha_i^8) + h_2(e-\alpha_j^8).$$

Since

$$(e - x^8) = \Big(\sum_{s=0}^{7} x^s\Big)(e-x),$$

this implies that

$$e-g = h_0\Big(\sum_{s=0}^{7}\alpha^s\Big)(e-\alpha) + h_1\Big(\sum_{s=0}^{7}\alpha_i^s\Big)(e-\alpha_i) + h_2\Big(\sum_{s=0}^{7}\alpha_j^s\Big)(e-\alpha_j).$$

Let

$$h = h_0\Big(\sum_{s=0}^{7}\alpha^s\Big) + h_1\Big(\sum_{s=0}^{7}\alpha_i^s\Big)(i-\alpha_i) + h_2\Big(\sum_{s=0}^{7}\alpha_j^s\Big)(j-\alpha_j).$$

If $\tau \in G_{24}$, then $\tau e_0 = e_0$. Hence,

$$(\tau - \alpha_\tau)(e-\alpha)e_0 = (e-\alpha_\tau)e_0.$$

Using this fact, one verifies that $(e-g)e_0 = h(e-\alpha)e_0$. Further,

$$\sum_{s=0}^{7} x^s \equiv (1-x)^7 + 2x^4(x-1)^3 + 4x^2(x-1) \mod (8).$$

Since α, α_i and α_j are in K^1 and K^1 is a normal subgroup, this implies that

$$h \equiv 0 \mod ((IK^1)^7, 2(IK^1)^3, 4(IK^1), 8).$$ □

We will use the following result.

Lemma 3.4.4 *The element $\alpha_i\alpha_j\alpha_k$ is in $F_{4/2}K^1$. The element $\alpha_i\alpha_j\alpha_k\alpha^2$ is in $F_{8/2}K^1$.*

Proof Let $T=\alpha S$ in $\mathcal{O}_2\cong \mathrm{End}(F_2)$. Then $T^2=-2$, and $aT=Ta^{\sigma}$ for a in $\mathbb{W}$. As defined in (2.3.4) and Lemma 2.4.3, we have

$$\alpha=\frac{1}{\sqrt{-7}}(1-2\omega),\qquad i=-\tfrac{1}{3}(1+2\omega)(1-T),$$
$$j=-\tfrac{1}{3}(1+2\omega)(1-\omega^2T),\quad k=-\tfrac{1}{3}(1+2\omega)(1-\omega T).$$

Further,

$$\alpha^{-1}=-\frac{1}{\sqrt{-7}}(1-2\omega^2).$$

We use the fact that $\frac{1}{3}$ and $\frac{1}{\sqrt{-7}}$ are in $Z(\mathbb{S}_2)$. We also use the fact $\tau^{-1}=-\tau$ for $\tau=i$, j and k and the fact that $S^4=4$ and $S^8=16$.

First, note that

$$\begin{aligned} i\alpha&=-\frac{1}{3\sqrt{-7}}(1+2\omega)(1-T)(1-2\omega)\\ &=-\frac{1}{3\sqrt{-7}}(1+2\omega)((1-2\omega)-(1-2\omega^2)T)\\ &=-\frac{1}{3\sqrt{-7}}((5+4\omega)+(1-4\omega)T).\end{aligned}$$

Further,

$$\begin{aligned} i^{-1}\alpha^{-1}&=-\frac{1}{3\sqrt{-7}}(1+2\omega)(1-T)(1-2\omega^2)\\ &=-\frac{1}{3\sqrt{-7}}(1+2\omega)((1-2\omega^2)-(1-2\omega)T)\\ &=-\frac{1}{3\sqrt{-7}}((-1+4\omega)-(5+4\omega)T).\end{aligned}$$

Therefore,

$$\begin{aligned} \alpha_i=i\alpha i^{-1}\alpha^{-1}&=-\tfrac{1}{63}((5+4\omega)+(1-4\omega)T)((-1+4\omega)-(5+4\omega)T)\\ &\equiv 13+(2+8\omega)T \mod S^8.\end{aligned}$$

Using the fact that $\alpha_j=\omega\alpha_i\omega^2$ and $\alpha_k=\omega^2\alpha_i\omega$, this implies that

$$\alpha_j\equiv 13+\omega^2(2+8\omega)T \mod S^8,\quad \alpha_k\equiv 13+\omega(2+8\omega)T \mod S^8.$$

Hence,

$$\begin{aligned}\alpha_i\alpha_j &\equiv (13+(2+8\omega)T)(13+\omega^2(2+8\omega)T)\\ &\equiv \big(9+\omega^2(10+8\omega)T+(10+8\omega)T+(2+8\omega)(\omega(2+8\omega^2))T^2\big)\\ &\equiv 9+8\omega+(8+14\omega)T \mod S^8,\end{aligned}$$

so that

$$\begin{aligned}\alpha_i\alpha_j\alpha_k &\equiv (9+8\omega+(8+14\omega)T)(13+\omega(2+8\omega)T)\\ &\equiv (5+8\omega+(8+6\omega)T+(9+8\omega)\omega(2+8\omega)T+(8+14\omega)\omega^2(2+8\omega^2)T^2)\\ &\equiv 13+8\omega \mod S^8.\end{aligned}$$

This shows that $\alpha_i\alpha_j\alpha_k \equiv 1$ modulo S^4. Finally, note that

$$\begin{aligned}\alpha_i\alpha_j\alpha_k\alpha^2 \equiv (13+8\omega)\Big(\frac{1}{\sqrt{-7}}(1-2\omega)\Big)^2 \equiv -\tfrac{1}{7}(13+8\omega)(1-2\omega)^2 \equiv -\tfrac{9}{7}\\ \equiv 1 \bmod S^8,\end{aligned}$$

which shows that $\alpha_i\alpha_j\alpha_k\alpha^2$ is in $F_{8/2}K^1$. □

Theorem 3.4.5 *There exists Θ in $\mathbb{Z}_2[\![\mathbb{S}_2^1]\!]$ satisfying the conditions of Lemma 3.1.3 such that*

$$\begin{aligned}\Theta \equiv e+\alpha+i+j+k-\alpha_i-\alpha_j-\alpha_k\\ -\tfrac{1}{3}\,\mathrm{tr}_{C_3}((e-\alpha_i)(j-\alpha_j)+(e-\alpha_i\alpha_j)(k-\alpha_k)+(e-\alpha_i\alpha_j\alpha_k)(e+\alpha))\end{aligned}$$

modulo $\mathcal{I}=((IK^1)^7, 2(IK^1)^3, 4(IK^1), 8)$, where tr_{C_3} is defined by (3.1.4).

Proof We will use the following facts. First, note that

$$\tau e_q = e_q$$

for $\tau\in G_{24}$ and $q=0$, or for $\tau\in C_6$ and $q=1$. This implies that

$$\tau(e-\alpha)e_0 = (e-\alpha_\tau\alpha)e_0.$$

Since $j=\omega i\omega^{-1}$ and $k=\omega^{-1}i\omega$, it also implies that

$$\omega i e_q = j e_q, \quad \omega^2 i e_q = k e_q.$$

The element $\alpha\in\mathbb{W}^\times\subseteq\mathbb{S}_2$ commutes with ω. This implies that

$$\omega\alpha_i e_q = \alpha_j e_q.$$

We will use the fact that for $\tau \in G_{24}$,

$$(\tau - \alpha_\tau)(e-\alpha)e_0 = (e-\alpha_\tau)e_0.$$

We will also use the identity

$$e - gh = (e-g) + (e-h) - (e-g)(e-h).$$

Let $\Theta_0 = e+i$. Then $\mathrm{tr}_{C_3}(\Theta_0)e_1 = (3+i+j+k)e_1$ and

$$\begin{aligned}
\partial_1(\Theta_0 e_1) &= (e+i)(e-\alpha)e_0 \\
&= (e-\alpha)e_0 + (e-\alpha_i\alpha)e_0 \\
&= 2(e-\alpha)e_0 + (e-\alpha_i)e_0 - (e-\alpha_i)(e-\alpha)e_0 \\
&= (e-\alpha^2)e_0 + (e-\alpha)^2 e_0 + (e-\alpha_i)e_0 - (e-\alpha_i)(e-\alpha)e_0.
\end{aligned}$$

Let $\Theta_1 = e + i - (e-\alpha) + (e-\alpha_i)$. Then,

$$\partial_1(\Theta_1 e_1) = (e-\alpha^2)e_0 + (e-\alpha_i)e_0.$$

Therefore,

$$\begin{aligned}
&\partial_1(\mathrm{tr}_{C_3}(\Theta_1)e_1) \\
&\quad = 3(e-\alpha^2)e_0 + (e-\alpha_i)e_0 + (e-\alpha_j)e_0 + (e-\alpha_k)e_0 \\
&\quad = 3(e-\alpha^2)e_0 + (e-\alpha_i)(e-\alpha_j)e_0 + (e-\alpha_i\alpha_j)e_0 + (e-\alpha_k)e_0 \\
&\quad = 3(e-\alpha^2)e_0 + (e-\alpha_i)(e-\alpha_j)e_0 + (e-\alpha_i\alpha_j)(e-\alpha_k)e_0 + (e-\alpha_i\alpha_j\alpha_k)e_0 \\
&\quad = 2(e-\alpha^2)e_0 + (e-\alpha_i)(e-\alpha_j)e_0 + (e-\alpha_i\alpha_j)(e-\alpha_k)e_0 \\
&\qquad\qquad + (e-\alpha_i\alpha_j\alpha_k)(e-\alpha^2)e_0 + (e-\alpha_i\alpha_j\alpha_k\alpha^2)e_0.
\end{aligned}$$

Let

$$\begin{aligned}
\Theta_2 = \mathrm{tr}_{C_3}\big(e+i-(e-\alpha)+(e-\alpha_i)\big) - 2(e+\alpha) \\
-(e-\alpha_i)(j-\alpha_j) - (e-\alpha_i\alpha_j)(k-\alpha_k) - (e-\alpha_i\alpha_j\alpha_k)(e+\alpha).
\end{aligned}$$

Then $\Theta_2 \equiv 3+i+j+k \mod (4, IK^1)$. Further,

$$\partial_1(\Theta_2 e_1) = (e-\alpha_i\alpha_j\alpha_k\alpha^2)e_0.$$

By Lemma 3.4.4, $\alpha_i\alpha_j\alpha_k\alpha^2 \in F_{8/2}K^1$. By Corollary 3.4.3, there exists h such that

$$(e-\alpha_i\alpha_j\alpha_k\alpha^2)e_0 = h(e-\alpha)e_0$$

and $h \equiv 0$ modulo $\mathcal{I}$, where $\mathcal{I} = ((IK^1)^7, 2(IK^1)^3, 4IK^1, 8)$. Therefore,

$$\partial_1((\Theta_2 - h)e_1) = (e-\alpha_i\alpha_j\alpha_k\alpha^2)e_0 - h(e-\alpha)e_0 = 0.$$

Define

$$\Theta = \tfrac{1}{3}\,\mathrm{tr}_{C_3}(\Theta_2 - h).$$

Then Θ satisfies the conditions of Lemma 3.1.3. Further,

$$\Theta \equiv e + \alpha + i + j + k - \alpha_i - \alpha_j - \alpha_k \\ - \tfrac{1}{3}\,\mathrm{tr}_{C_3}((e-\alpha_i)(j-\alpha_j) + (e-\alpha_i\alpha_j)(k-\alpha_k) + (e-\alpha_i\alpha_j\alpha_k)(e+\alpha))$$

modulo $\mathcal{I}$. □

Corollary 3.4.6 *Let $\mathcal{J} = (IF_{4/2}K^1, (IF_{3/2}K^1)(IS_2^1), \mathcal{I})$. The element Θ from Theorem 3.4.5 satisfies*

$$\Theta \equiv e + \alpha + i + j + k - \alpha_i - \alpha_j - \alpha_k \mod \mathcal{J},$$

and $\Theta \equiv e + \alpha$ modulo $(2, (IS_2^1)^2)$.

Proof First, note that $\alpha_\tau \in F_{3/2}K^1$ for $\tau \in G_{24}$. Further, by Lemma 3.4.4, $\alpha_i\alpha_j\alpha_k$ is in $F_{4/2}K^1$. Hence, it follows from Theorem 3.4.5 that

$$\Theta \equiv e + \alpha + i + j + k - \alpha_i - \alpha_j - \alpha_k \mod \mathcal{J}.$$

For the second claim, we first prove that $\mathcal{J} \subseteq (2, (IS_2^1)^2)$. It is clear that

$$((IF_{3/2}K^1)(IS_2^1), \mathcal{I}) \subseteq (2, (IS_2^1)^2).$$

Further, it follows from Lemma 3.4.2 and the fact that $(e - x^{2^k}) \equiv (e-x)^{2^k}$ modulo (2) that

$$IF_{4/2}K^1 \subseteq (2, (IS_2^1)^2).$$

Therefore, $\mathcal{J} \subseteq (2, (IS_2^1)^2)$. Hence,

$$\Theta \equiv e + \alpha + i + j + k - \alpha_i - \alpha_j - \alpha_k \mod (2, (IS_2^1)^2).$$

Further, $(e-i)(e-j) \equiv e + i + j + k$ modulo (2) and

$$e - \alpha_i = i\alpha((e-\alpha^{-1})(e-i^{-1}) - (e-i^{-1})(e-\alpha^{-1})).$$

Therefore, $e + i + j + k$ and $e - \alpha_\tau$ are in $(2, (IS_2^1)^2)$. We conclude that

$$\Theta \equiv e + \alpha \mod (2, (IS_2^1)^2).$$ □

Appendix: Background on profinite groups

We use the terminology of Ribes and Zalesskii [28, Section 5]. Let G be a profinite p–adic analytic group and $\{U_k\}$ be a system of open normal subgroups of G such that $\bigcap_k U_k = \{e\}$. The *completed group ring of* G is

$$\mathbb{Z}_p[\![G]\!] := \lim_{n,k} \mathbb{Z}/(p^n)[G/U_k].$$

The *augmentation* is the continuous homomorphism of $\mathbb{Z}_p$–modules $\varepsilon\colon \mathbb{Z}_p[\![G]\!] \to \mathbb{Z}_p$ defined by $\varepsilon(g) = 1$ for $g \in G$. The *augmentation ideal* IG is the kernel of ε.

A *left* $\mathbb{Z}_p[\![G]\!]$*–module* is a $\mathbb{Z}_p[\![G]\!]$–module M which is a Hausdorff topological abelian group with a continuous structure map $\mathbb{Z}_p[\![G]\!] \times M \to M$. The module M is *finitely generated* if it is the closure of the $\mathbb{Z}_p[\![G]\!]$–module generated by a finite subset of M. It is *discrete* if it is the union of its finite $\mathbb{Z}_p[\![G]\!]$–submodules and *profinite* if it is the inverse limit of its finite $\mathbb{Z}_p[\![G]\!]$–submodule quotients; see [28, Lemma 5.1.1]. The module M is complete with respect to the IG–adic topology if

$$M \cong \lim_{n,k} \mathbb{Z}_p/(p^n)[G/U_k] \otimes_{\mathbb{Z}_p[\![G]\!]} M.$$

It is a theorem of Lazard that $\mathbb{Z}_p[\![G]\!]$ is Noetherian; see Symonds and Weigel [35, Theorem 5.1.2]. Finitely generated $\mathbb{Z}_p[\![G]\!]$–modules are thus both profinite and complete with respect to the IG–adic topology.

Let $M = \lim_i M_i$ be a profinite $\mathbb{Z}_p[\![G]\!]$–bimodule and $N = \lim_j N_j$ a profinite left $\mathbb{Z}_p[\![G]\!]$–module. Then

$$M \otimes_{\mathbb{Z}_p[\![G]\!]} N = \lim_{i,j} M_i \otimes_{\mathbb{Z}_p[\![G]\!]} N_j$$

denotes the completed tensor product, which is itself a profinite left $\mathbb{Z}_p[\![G]\!]$–module [28, Section 5.5]. The abelian group of continuous $\mathbb{Z}_p[\![G]\!]$–homomorphisms is denoted by

$$\mathrm{Hom}_{\mathbb{Z}_p[\![G]\!]}(M, N).$$

This is a topological space with the compact open topology. If M is finitely generated, then it is compact; see [35, Section 3.7].

Lazard also proves in [24, V.3.2.7] that the trivial $\mathbb{Z}_p[\![G]\!]$–module $\mathbb{Z}_p$ admits a resolution by finitely generated $\mathbb{Z}_p[\![G]\!]$–modules. A $\mathbb{Z}_p[\![G]\!]$–module M which admits a projective resolution $P_\bullet \to M$ by finitely generated $\mathbb{Z}_p[\![G]\!]$–modules is said to be of type $\mathbf{FP}^\infty$; see [35, Section 3.7]. For such M, we let

$$\mathrm{Ext}^n_{\mathbb{Z}_p[\![G]\!]}(M, N) = H^n(\mathrm{Hom}_{\mathbb{Z}_p[\![G]\!]}(P_\bullet, N))$$

and

$$\mathrm{Tor}_n^{\mathbb{Z}_p[\![G]\!]}(M, N) = H_n(P_\bullet \otimes_{\mathbb{Z}_p[\![G]\!]} N).$$

These functors are studied by Symonds and Weigel in [35, Section 3.7]. There are isomorphisms

$$H^n(G, N) \cong \mathrm{Ext}^n_{\mathbb{Z}_p[\![G]\!]}(M, N),$$

where $H^n(G, N)$ is the cohomology computed with continuous cochains and

$$H_n(G, N) \cong \mathrm{Tor}_n^{\mathbb{Z}_p[\![G]\!]}(\mathbb{Z}_p, N);$$

see Neukirch, Schmidt and Wingberg [25, Propositions 5.2.6, 5.2.14] or the discussion in Kohlhaase [23, Section 3]. Therefore, these functors satisfy the usual properties of group cohomology; see Ribes and Zalesskii [28, Section 6]. In particular, for $[G, G]$ the commutator subgroup, $G^* = G^p[G, G]$, and $\mathbb{Z}_p$ and $\mathbb{F}_p$ the trivial modules, we have

$$H_1(G, \mathbb{Z}_p) \cong G/\overline{[G, G]}, \qquad H_1(G, \mathbb{F}_p) \cong G/\overline{G^*},$$
$$H^1(G, \mathbb{Z}_p) \cong \mathrm{Hom}(G, \mathbb{Z}_p), \quad H^1(G, \mathbb{F}_p) \cong \mathrm{Hom}(G, \mathbb{F}_p).$$

Examples A.1.1 We give examples, which we use in this paper, in [3] and in [2].

(a) The modules

$$\mathbb{Z}_p[\![G/H]\!] := \mathbb{Z}_p[\![G]\!] \otimes_{\mathbb{Z}_p[H]} \mathbb{Z}_p$$

for H a finite subgroup of G and $\mathbb{Z}_p$ the trivial $\mathbb{Z}_p[H]$–module are finitely generated, and thus profinite and complete.

(b) The $\mathbb{Z}_p[\![G]\!]$–dual of a finitely generated $\mathbb{Z}_p[\![G]\!]$–module M is defined as

$$M^* := \mathrm{Hom}_{\mathbb{Z}_p[\![G]\!]}(M, \mathbb{Z}_p[\![G]\!]), \tag{A.1.1}$$

with the action of $g \in G$ on $\phi \in M^*$ defined by

$$(g\phi)(m) = \phi(m)g^{-1}.$$

This gives M^* the structure of a finitely generated left $\mathbb{Z}_p[\![G]\!]$–module; see Symonds and Weigel [35, 3.7.1] and Henn, Karamanov and Mahowald [18, Section 3.4]. For example, if $H \subseteq G$ is a finite subgroup and $[g]$ denotes the coset gH, there is a canonical isomorphism

$$t\colon \mathbb{Z}_p[\![G/H]\!] \to \mathbb{Z}_p[\![G/H]\!]^* \tag{A.1.2}$$

which sends $[g]$ to the map $[g]^*\colon \mathbb{Z}_p[\![G/H]\!] \to \mathbb{Z}_p[\![G]\!]$ defined by

$$[g]^*([x]) = x \sum_{h \in H} hg^{-1}.$$

We refer the reader to [18, Section 3.4] for a detailed discussion of $\mathbb{Z}_p[\![G]\!]$–duals.

(c) In the case when $G = S_n$ is the p–Sylow subgroup of $\mathbb{S}_n$, an important example is the continuous $\mathbb{Z}_p[\![S_n]\!]$–module $(E_n)_*X = \pi_* L_{K(n)}(E_n \wedge X)$ for a spectrum X; see Goerss, Henn, Mahowald and Rezk [14, Section 2]. In the case when X is a finite spectrum, $(E_n)_*X$ is profinite, although it is not known if, in general, it is finitely generated over $\mathbb{Z}_p[\![S_n]\!]$. For a more extensive discussion, see the work of Kohlhaase in [23].

Lemma A.1.2 (Shapiro's Lemma) *Let G be a profinite p–analytic group and let H be a closed subgroup. Let M be a $\mathbb{Z}_p[\![H]\!]$–module of type $\mathbf{FP}^\infty$ and let $N = \lim_i N_i$ be a profinite $\mathbb{Z}_p[\![G]\!]$–module, which is also a $\mathbb{Z}_p[\![H]\!]$–module via restriction. Then*

$$\operatorname{Ext}^*_{\mathbb{Z}_p[\![G]\!]}(\mathbb{Z}_p[\![G]\!] \otimes_{\mathbb{Z}_p[\![H]\!]} M, N) \cong \operatorname{Ext}^*_{\mathbb{Z}_p[\![H]\!]}(M, N).$$

Proof Let $P_\bullet \to M$ be a projective resolution of M by finitely generated $\mathbb{Z}_p[\![H]\!]$–modules. According to Brumer [7, Lemma 4.5], $\mathbb{Z}_p[\![G]\!]$ is a projective $\mathbb{Z}_p[\![H]\!]$–module. Hence, the functor $\mathbb{Z}_p[\![G]\!] \otimes_{\mathbb{Z}_p[\![H]\!]} (-)$ is exact, and $\mathbb{Z}_p[\![G]\!] \otimes_{\mathbb{Z}_p[\![H]\!]} P_\bullet$ is a projective resolution of $\mathbb{Z}_p[\![G]\!] \otimes_{\mathbb{Z}_p[\![H]\!]} M$ by finitely generated $\mathbb{Z}_p[\![G]\!]$–modules. Finally, note that

$$\begin{aligned}
\operatorname{Hom}_{\mathbb{Z}_p[\![G]\!]}(\mathbb{Z}_p[\![G]\!] \otimes_{\mathbb{Z}_p[\![H]\!]} P_\bullet, N) &\cong \lim_i \operatorname{Hom}_{\mathbb{Z}_p[\![G]\!]}(\mathbb{Z}_p[\![G]\!] \otimes_{\mathbb{Z}_p[\![H]\!]} P_\bullet, N_i) \\
&\cong \lim_i \operatorname{Hom}_{\mathbb{Z}_p[\![H]\!]}(P_\bullet, N_i) \\
&\cong \operatorname{Hom}_{\mathbb{Z}_p[\![H]\!]}(P_\bullet, N),
\end{aligned}$$

where the first isomorphism is proved by Symonds and Weigel [35, (3.7.1)] and the second follows from Ribes and Zalesskii [28, Proposition 5.5.4(c)]. □

The following result is Lemma 4.3 of Goerss, Henn, Mahowald and Rezk [14]. It is a version of Nakayama's lemma in this setting.

Lemma A.1.3 *Let G be a finitely generated profinite p–group. Let M and N be finitely generated complete $\mathbb{Z}_p[\![G]\!]$–modules and $f\colon M \to N$ be a map of complete $\mathbb{Z}_p[\![G]\!]$–modules. If the induced map*

$$\mathbb{F}_p \otimes_{\mathbb{Z}_p[\![G]\!]} f\colon \mathbb{F}_p \otimes_{\mathbb{Z}_p[\![G]\!]} M \to \mathbb{F}_p \otimes_{\mathbb{Z}_p[\![G]\!]} N$$

is surjective, then so is f. If the map

$$\operatorname{Tor}_q^{\mathbb{Z}_p[\![G]\!]}(\mathbb{F}_p, f)\colon \operatorname{Tor}_q^{\mathbb{Z}_p[\![G]\!]}(\mathbb{F}_p, M) \to \operatorname{Tor}_q^{\mathbb{Z}_p[\![G]\!]}(\mathbb{F}_p, N)$$

is an isomorphism for $q = 0$ and surjective for $q = 1$, then f is an isomorphism.

The following is a restatement of some of the results which can be found in Ribes and Zalesskii [28, Lemma 6.8.6].

Lemma A.1.4 *Let G be a profinite group and let IG be the augmentation ideal. For a profinite $\mathbb{Z}_p[\![G]\!]$–module M, the boundary map for the short exact sequence*

$$0 \to IG \to \mathbb{Z}_p[\![G]\!] \xrightarrow{\varepsilon} \mathbb{Z}_p \to 0$$

induces an isomorphism

$$H_{n+1}(G, M) \cong \mathrm{Tor}_n^{\mathbb{Z}_p[\![G]\!]}(IG, M).$$

For the trivial module $M = \mathbb{Z}_p$, this isomorphism sends g in $G/\overline{[G,G]}$ to the residue class of $e - g$ in $H_0(G, IG) \cong IG/\overline{IG^2}$. Let G^* be the subgroup generated by $[G,G]$ and G^p. For $M = \mathbb{F}_p$, it sends g in $G/\overline{G^*}$ to the residue class of $e - g$ in $\mathbb{F}_p \otimes_{\mathbb{Z}_p} IG/\overline{IG^2}$.

Finally, we note the following classical result.

Lemma A.1.5 *Let G be a profinite 2–analytic group and suppose that $H_1(G, \mathbb{Z}_2) \cong G/\overline{[G,G]}$ is a finitely generated 2–group. Suppose that the residue class of an element g in $G/\overline{[G,G]}$ generates a summand isomorphic to $\mathbb{Z}/2^k$. Let x in $H^1(G, \mathbb{Z}/2) \cong \mathrm{Hom}(G, \mathbb{Z}/2)$ be the homomorphism dual to g. Then x^2 is nonzero in $H^2(G, \mathbb{Z}/2)$ if and only if $k = 1$.*

Proof This follows from the fact that x in $H^1(G, \mathbb{Z}/2)$ has a nonzero Bockstein in the long exact sequence associated to the extension of trivial modules

$$1 \to \mathbb{Z}/2 \to \mathbb{Z}/4 \to \mathbb{Z}/2 \to 1$$

if and only if g generates a $\mathbb{Z}/2$ summand. □

References

[1] **A Adem**, **R J Milgram**, *Cohomology of finite groups*, 2nd edition, Grundl. Math. Wissen. 309, Springer, Berlin (2004) MR2035696

[2] **A Beaudry**, *The chromatic splitting conjecture at $n = p = 2$*, preprint (2015) `arXiv:1502.02190v2`

[3] **A Beaudry**, *Towards $\pi_* L_{K(2)} V(0)$ at $p = 2$*, preprint (2015) `arXiv:1501.06082`

[4] **M Behrens**, *The homotopy groups of $S_{E(2)}$ at $p \geq 5$ revisited*, Adv. Math. 230 (2012) 458–492 MR2914955

[5] **M Behrens**, **T Lawson**, *Isogenies of elliptic curves and the Morava stabilizer group*, J. Pure Appl. Algebra 207 (2006) 37–49 MR2244259

[6] **I Bobkova**, *Resolutions in the $K(2)$–local category at the prime 2*, PhD thesis, Northwestern University, Ann Arbor, MI (2014) MR3251316 Available at `http://search.proquest.com/docview/1558126694`

[7] **A Brumer**, *Pseudocompact algebras, profinite groups and class formations*, J. Algebra 4 (1966) 442–470 MR0202790

[8] **C Bujard**, *Finite subgroups of extended Morava stabilizer groups*, preprint (2012) `arXiv:1206.1951v2`

[9] **D G Davis**, *Homotopy fixed points for $L_{K(n)}(E_n \wedge X)$ using the continuous action*, J. Pure Appl. Algebra 206 (2006) 322–354 MR2235364

[10] **E S Devinatz**, **M J Hopkins**, *Homotopy fixed point spectra for closed subgroups of the Morava stabilizer groups*, Topology 43 (2004) 1–47 MR2030586

[11] **J D Dixon**, **M P F du Sautoy**, **A Mann**, **D Segal**, *Analytic pro-p groups*, 2nd edition, Cambridge Studies Adv. Math. 61, Cambridge Univ. Press (1999) MR1720368

[12] **P G Goerss**, **H-W Henn**, *The Brown–Comenetz dual of the $K(2)$–local sphere at the prime 3*, Adv. Math. 288 (2016) 648–678

[13] **P G Goerss**, **H-W Henn**, **M Mahowald**, *The rational homotopy of the $K(2)$–local sphere and the chromatic splitting conjecture for the prime 3 and level 2*, Doc. Math. 19 (2014) 1271–1290 MR3312144

[14] **P Goerss**, **H-W Henn**, **M Mahowald**, **C Rezk**, *A resolution of the $K(2)$–local sphere at the prime 3*, Ann. of Math. 162 (2005) 777–822 MR2183282

[15] **P Goerss**, **H-W Henn**, **M Mahowald**, **C Rezk**, *On Hopkins' Picard groups for the prime 3 and chromatic level 2*, J. Topol. 8 (2015) 267–294 MR3335255

[16] **P G Goerss**, **M J Hopkins**, *Moduli spaces of commutative ring spectra*, from: "Structured ring spectra", (Cambridge, editor), London Math. Soc. Lecture Note Ser. 315, Cambridge Univ. Press (2004) 151–200 MR2125040

[17] **H-W Henn**, *Centralizers of elementary abelian p–subgroups and mod-p cohomology of profinite groups*, Duke Math. J. 91 (1998) 561–585 MR1604171

[18] **H-W Henn**, **N Karamanov**, **M Mahowald**, *The homotopy of the $K(2)$–local Moore spectrum at the prime 3 revisited*, Math. Z. 275 (2013) 953–1004 MR3127044

[19] **T Hewett**, *Finite subgroups of division algebras over local fields*, J. Algebra 173 (1995) 518–548 MR1327867

[20] **T Hewett**, *Normalizers of finite subgroups of division algebras over local fields*, Math. Res. Lett. 6 (1999) 271–286 MR1713129

[21] **M Hovey**, *Bousfield localization functors and Hopkins' chromatic splitting conjecture*, from: "The Čech centennial", (M Cenkl, H Miller, editors), Contemp. Math. 181, Amer. Math. Soc. (1995) 225–250 MR1320994

[22] **A Huber**, **G Kings**, **N Naumann**, *Some complements to the Lazard isomorphism*, Compos. Math. 147 (2011) 235–262 MR2771131

[23] **J Kohlhaase**, *On the Iwasawa theory of the Lubin–Tate moduli space*, Compos. Math. 149 (2013) 793–839 MR3069363

[24] **M Lazard**, *Groupes analytiques p–adiques*, Inst. Hautes Études Sci. Publ. Math. 26 (1965) 389–603 MR0209286

[25] **J Neukirch**, **A Schmidt**, **K Wingberg**, *Cohomology of number fields*, Grundl. Math. Wissen. 323, Springer, Berlin (2000) MR1737196

[26] **D C Ravenel**, *Complex cobordism and stable homotopy groups of spheres*, Pure and Applied Mathematics 121, Academic Press, Orlando, FL (1986) MR860042

[27] **D C Ravenel**, *Nilpotence and periodicity in stable homotopy theory*, Annals of Mathematics Studies 128, Princeton Univ. Press (1992) MR1192553

[28] **L Ribes**, **P Zalesskii**, *Profinite groups*, 2nd edition, Ergeb. Math. Grenzgeb. 40, Springer, Berlin (2010) MR2599132

[29] **J-P Serre**, *Sur la dimension cohomologique des groupes profinis*, Topology 3 (1965) 413–420 MR0180619

[30] **K Shimomura**, *The Adams–Novikov E_2–term for computing $\pi_*(L_2V(0))$ at the prime* 2, Topology Appl. 96 (1999) 133–152 MR1702307

[31] **K Shimomura**, **X Wang**, *The Adams–Novikov E_2–term for $\pi_*(L_2S^0)$ at the prime* 2, Math. Z. 241 (2002) 271–311 MR1935487

[32] **K Shimomura**, **X Wang**, *The homotopy groups $\pi_*(L_2S^0)$ at the prime* 3, Topology 41 (2002) 1183–1198 MR1923218

[33] **K Shimomura**, **A Yabe**, *The homotopy groups $\pi_*(L_2S^0)$*, Topology 34 (1995) 261–289 MR1318877

[34] **N P Strickland**, *Gross–Hopkins duality*, Topology 39 (2000) 1021–1033 MR1763961

[35] **P Symonds**, **T Weigel**, *Cohomology of p–adic analytic groups*, from: "New horizons in pro-p groups", (M du Sautoy, D Segal, A Shalev, editors), Progr. Math. 184, Birkhäuser, Boston (2000) 349–410 MR1765127

Department of Mathematics, University of Chicago
1118 East 58th Street, Chicago, IL 60637, USA

beaudry@math.uchicago.edu

Received: 19 December 2014 Revised: 30 March 2015

msp

Algebraic & Geometric Topology 15 (2015) 3707–3729

Braiding link cobordisms and non-ribbon surfaces

MARK C HUGHES

We define the notion of a braided link cobordism in $S^3 \times [0,1]$, which generalizes Viro's closed surface braids in $\mathbb{R}^4$. We prove that any properly embedded oriented surface $W \subset S^3 \times [0,1]$ is isotopic to a surface in this special position, and that the isotopy can be taken rel boundary when ∂W already consists of closed braids. These surfaces are closely related to another notion of surface braiding in $D^2 \times D^2$, called braided surfaces with caps, which are a generalization of Rudolph's braided surfaces. We mention several applications of braided surfaces with caps, including using them to apply algebraic techniques from braid groups to studying surfaces in 4–space, as well as constructing singular fibrations on smooth 4–manifolds from a given handle decomposition.

57M12; 57M25, 57R52

1 Introduction

Two of the most useful and foundational results in knot theory and low-dimensional topology are the classical theorems of Alexander and Markov. These theorems allow us to study knots entirely within the realm of braids and braid closures, where we can exploit either the algebraic structure of the braid group, the special position of a closed braid in S^3 or the fact that braids with isotopic closures can be related by special braid moves. These results have been used in numerous applications, examples of which include the construction and categorification of quantum link invariants due to Freyd, Yetter, Hoste, Lickorish, Millett and Ocneanu [9], Jones [13] and Khovanov and Rozansky [20], the construction of open book decompositions on 3–manifolds of Alexander [2], and studying the slice and ribbon genera of knots of Rudolph [25; 27].

The notion of a closed braid as a specially positioned 1–dimensional submanifold of 3–dimensional space has been generalized by different authors to certain classes of surfaces in 4–space. One such generalization is due to Rudolph [25], who considered surfaces $S \subset D^2 \times D^2$ on which the projection to the second factor $\mathrm{pr}_2\colon D^2 \times D^2 \to D^2$ restricts as a branched covering. These *braided surfaces* generalize the classical notion of a (geometric) braid as a 1–dimensional submanifold of $D^2 \times [0,1]$ on which the projection $\mathrm{pr}_{[0,1]}\colon D^2 \times [0,1] \to [0,1]$ restricts as an ordinary covering. Any braided

Published: 12 January 2016 DOI: 10.2140/agt.2015.15.3707

surface is necessarily ribbon and Rudolph showed that every orientable ribbon surface with boundary properly embedded in $D^2 \times D^2$ is isotopic to a braided surface.

Braided surfaces are closely related to a similar notion due to Viro [29], called 2*–braids*. Analogous to classical braids, 2–braids admit a closure operation yielding closed surfaces in S^4. Viro [29] and Kamada [16] independently proved a 4–dimensional Alexander theorem by showing that every closed oriented surface in S^4 is isotopic to the closure of a 2–braid. Kamada [15; 19] additionally proved a 4–dimensional Markov theorem which relates any pair of 2–braids with isotopic closures.

Like their lower-dimensional counterparts, braided ribbon surfaces have found use in various applications, including finding obstructions to sliceness in knot theory [27], the study of Stein fillings of contact 3–manifolds and the construction of Lefschetz fibrations on 4–dimensional 2–handlebodies (ie 4–manifolds admitting handle decompositions with no 3– or 4–handles). Indeed, using the fact that any oriented 4–dimensional 2–handlebody X admits a covering over $D^2 \times D^2$ branched along an orientable ribbon surface, Loi and Piergallini [22] were able to construct Lefschetz fibrations on X and used them to give a topological characterization of Stein surfaces with boundary.

As Rudolph's braided surfaces do not include non-ribbon surfaces, the above techniques were not sufficient for studying smooth 4–manifolds with 3– or 4–handles. Indeed, the branched coverings of such manifolds over $D^2 \times D^2$ do not have ribbon branch loci. Expanding these applications thus requires a more general notion of braided surface.

In this paper we generalize these notions further, by defining *braided link cobordisms* (or simply *braided cobordisms*). These are surfaces $W \subset S^3 \times [0,1]$, smoothly and properly embedded, on which the projection $\mathrm{pr}_2\colon S^3 \times [0,1] \to [0,1]$ restricts as a Morse function with each regular level set $W \cap (S^3 \times \{t\})$ a closed braid in $S^3 \times \{t\}$. Braided cobordisms generalize Viro's closed 2–braids to oriented surfaces with boundary. We prove the following:

Theorem 1.1 *Let $W \subset S^3 \times [0,1]$ be an oriented surface smoothly and properly embedded. Then W is isotopic to a braided cobordism. If the boundary links of ∂W are already closed braids, then this isotopy can be chosen rel ∂W.*

Theorem 1.1 can be thought of as a cobordism analogue to the classical Alexander's theorem and will be proven in Section 3. Our construction will be similar to Kamada's construction [16], which implies our result in the case that W is a closed surface. The bulk of the additional work here will be in carrying out the construction in a way that allows us to keep ∂W fixed during the required ambient isotopies. This boundary-fixing requirement is considered with an eye toward applications (see either Jacobsson [12] for a construction using Khovanov homology, which is not invariant under general isotopies of W, or below for other applications).

We also define a related class of surfaces in $D^2 \times D^2$, called *braided surfaces with caps*, which generalize Rudolph's braided surfaces (see Section 2.4), and which are closely related to braided cobordisms. Theorem 1.1 then gives us the following:

Corollary 1.2 *Let S be a smooth oriented properly embedded surface in $D^2 \times D^2$. Then S is isotopic to a braided surface with caps. If ∂S is already a closed braid, then the isotopy can be chosen rel ∂S.*

These generalized surface braiding results make it possible to extend applications which rely on Rudolph's braiding algorithm. Here we outline one such application, which involves extending Loi and Piergallini's techniques to construct broken Lefschetz fibrations on oriented smooth 4–manifolds. Let X be a smooth, oriented, compact 4–manifold and Σ a compact oriented surface. Then a surjective map $f\colon X \to \Sigma$ is called a *Lefschetz fibration* if around every critical point the map f can be modeled in orientation-preserving complex coordinates locally as $f(u,v) = u^2 + v^2$. It is called a *broken Lefschetz fibration* if, along with these isolated critical points, it also contains embedded circles of critical points near which f is locally modeled by $f(\theta, x, y, z) = (\theta, x^2 + y^2 - z^2)$.

Lefschetz fibrations are closely related to symplectic structures on X — see Donaldson [8] and Gompf and Stipsicz [11] — and allow us to express the 4–manifold X combinatorially in terms of the monodromy of a regular fiber (see [11]). Broken Lefschetz fibrations exist more generally, but share a similar relation to near-symplectic structures — see Auroux, Donaldson and Katzarkov [3] — and can be used to define invariants of smooth 4–manifolds and finitely presented groups; see Baykur [5]. They were introduced in [3], which constructed a broken Lefschetz fibration on S^4. Later, it was shown independently by Akbulut and Karakurt [1], Baykur [4] and Lekili [21] that any oriented smooth 4–manifold admits a broken Lefschetz fibration over S^2. Although their approaches differ, none of them build the desired fibration directly from a given handle decomposition of X, instead relying on the modification of critical points of generic maps or deep classification results from contact topology.

Corollary 1.2 allows us to extend Loi and Piergallini's techniques to construct broken Lefschetz fibrations from handle decompositions on a wide class of 4–manifolds. Indeed, given a handle decomposition of X with $\partial X \neq \varnothing$, we can construct a branched covering $h\colon X \to D^2 \times D^2$ one handle at a time, so that the branch locus is a surface with only cusp and node singularities. In many cases this branch locus can be made to be orientable and hence, by Corollary 1.2, can be isotoped to a braided surface with caps in $D^2 \times D^2$. The desired fibration on X is then obtained as the composition $\mathrm{pr}_2 \circ h\colon X \to D^2$. This construction yields fibrations directly from the handle decomposition of X and can be combined with techniques of Gay and Kirby [10] to give broken Lefschetz fibrations on closed 4–manifolds.

Another avenue of application lies in using algebraic information from a braid to answer geometric questions about its closure. Indeed, Rudolph used braided ribbon surfaces to study quasipositive links [26; 27; 28] (links which bound braided ribbon surfaces with only positive branch points), as well as to find bounds on the ribbon genus of a link in terms of algebraic information from the braid group [25]. Using braided (non-ribbon) surfaces with caps, this latter approach can be extended further to look for bounds on the genus of an arbitrary surface bounded by a link in terms of algebraic information from its boundary. Furthermore, there are a number of link invariants whose definitions require they be computed on closed braid diagrams (see eg [20]). By examining links that are joined by a given braided cobordism W, one could attempt to extend these invariants across W and uncover interesting relationships between the invariants along ∂W and the surface W. The author intends to pursue these questions further in upcoming work.

The remainder of this paper will be organized as follows. In Section 2 we define various notions of surface braidings in $D^2 \times D^2$ and $S^3 \times [0,1]$, as well as outline the relationship between them. In Section 3 we present diagrammatic methods for studying 1–dimensional braids and surfaces in 4–space and use them to prove Theorem 1.1 and Corollary 1.2.

Acknowledgements The author would like to thank Oleg Viro for many helpful comments and suggestions.

2 Braided surfaces in 4–space

2.1 Links as braid closures

Let $D^2 \subset \mathbb{C}$ be the closed unit disk, $S^1 = \partial D^2$ and $S^3 = \{(z,w) \mid |z|^2 + |w|^2 = 1\} \subset \mathbb{C}^2$ the unit 3–sphere. We set $T_1 = S^3 \cap \{|z| \leq 1/\sqrt{2}\}$ and $T_2 = S^3 \cap \{|w| \leq 1/\sqrt{2}\}$, which are both tori, and let $U = S^3 \cap \{w = 0\}$ (ie the core of T_2). We say that an oriented link L in S^3 is a *closed braid* if $L \subset S^3 \backslash U$ and $\arg(w)$ is strictly increasing as we traverse the components of L in the positively oriented direction. We call U the *axis* of the closed braid.

Alexander's theorem then says that any oriented link in S^3 is isotopic to a closed braid. Markov's theorem says that any two closed braids which are isotopic as links can be joined by a sequence of isotopies through closed braids as well as stabilization and destabilizations moves which increase and decrease the braid index, respectively.

2.2 Movie presentations of braided cobordisms

Recall from Section 1 that a braided cobordism is a surface $W \subset S^3 \times [0,1]$, smoothly and properly embedded, on which the projection $\mathrm{pr}_2\colon S^3 \times [0,1] \to [0,1]$ restricts as a Morse function with each regular level set $W_t = W \cap (S^3 \times \{t\})$ a closed braid in $S^3 \times \{t\}$. We will assume in what follows that $\mathrm{pr}_2|_W$ is injective on its set of critical points. Each regular W_t with $t < 1$ is oriented as the boundary of $W \cap (S^3 \times [t,1])$.

We now establish a diagrammatic method for describing braided cobordisms. Choose a point $p \in U \subset S^3$ with $\{p\} \times [0,1]$ disjoint from W, and identify the complement of p in (S^3, U) with $(\mathbb{R}^3, z\text{–axis})$. Choose the identification so that $\arg(w)$ corresponds to the angular cylindrical coordinate on $\mathbb{R}^3$. Here we let (x,y,z) denote the usual coordinates on $\mathbb{R}^3$, while t denotes the coordinate on $[0,1]$.

Let $\pi\colon \mathbb{R}^3 \to \mathbb{R}^2$ denote the orthogonal projection to the xy–plane. After perturbing W slightly if necessary, we can assume that $\pi \times \mathrm{id}\colon \mathbb{R}^3 \times [0,1] \to \mathbb{R}^2 \times [0,1]$ restricts to a family of regular link projections $W_t \to \mathbb{R}^2 \times [0,1]$ for all but finitely many $t \in [0,1]$. After decorating with over- and under-crossing information, we obtain a continuous family of link diagrams with finitely many singular diagrams. As each regular W_t is a closed braid, each regular diagram will be the diagram of a closed braid, while passing a singular still will change the diagram by one of the following:

(1) Addition or deletion of a single loop around $0 \in \mathbb{R}^2$ disjoint from the rest of the diagram (corresponding to local maximum and minimum points of W).

(2) Addition or deletion of a single crossing between adjacent strands in the braid diagram by a band surgery (corresponding to saddle points of W).

(3) A single braid-like Reidemeister move of type II or III, where each strand involved in the move is oriented in the positive direction.

We refer to this family of link diagrams as the *movie presentation* of W. Note that, because we are not assuming W is in general position with respect to the z– and t–projections, our definition of movie presentation differs slightly from that used by other authors (see eg Carter, Kamada and Saito [7]). During the proof of Theorem 1.1, we will also consider movie presentations using projections other than the orthogonal projection $\pi\colon \mathbb{R}^3 \to \mathbb{R}^2$ to the xy–plane.

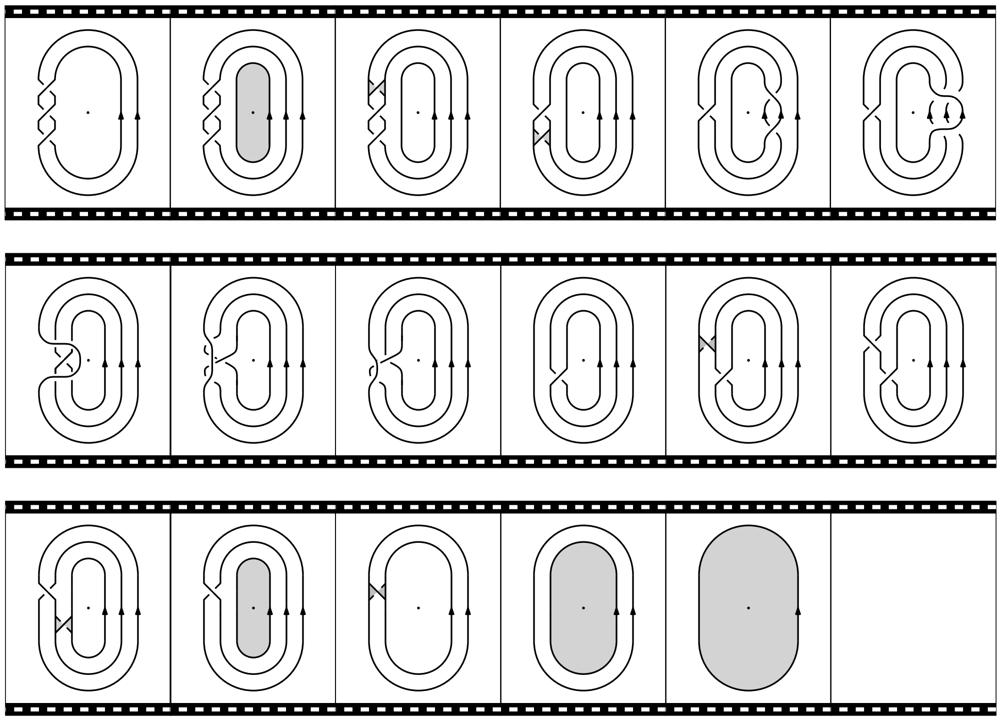


Figure 1: Braided movie presentation

The surface W can then be described by taking a finite number of the nonsingular stills, where each one differs from the previous still by a single modification as described above, or by a planar isotopy preserving the closed braid structure. Some caution is needed in using such descriptions, as different choices of planar isotopies linking two adjacent diagrams can result in non-isotopic embeddings (see eg Jacobsson [12]). See Figure 1 for a genus-1 example of a braided movie presentation between the trefoil and the empty knot (the stills are read as lines of text, from left to right).

2.3 Braided surfaces in $D^2 \times D^2$

Rudolph [25] defined a *braided surface* to be a smooth, properly embedded, oriented surface $S \subset D^2 \times D^2$ on which the projection to the second factor $\mathrm{pr}_2\colon D^2 \times D^2 \to D^2$ restricts as a simple branched covering. Examples of these braided surfaces can be obtained by taking intersections of nonsingular complex plane curves with 4–balls in $\mathbb{C}^2$ and they can be used to study the links that arise as their boundaries in $S^3 = \partial D^4$ (see eg [26; 27; 28]).

Let S be a braided surface. In a neighborhood of any branch point p of the covering $\mathrm{pr}_2|_S$, there are local complex coordinates u and v on D^2 such that S is given by the equation $u^2 = v$ in the coordinates (u, v) on $D^2 \times D^2$.

The boundary of $D^2 \times D^2$ decomposes as $\partial(D^2 \times D^2) = (D^2 \times S^1) \cup (S^1 \times D^2)$ in the obvious way and we set $\partial_1 = D^2 \times S^1$ and $\partial_2 = S^1 \times D^2$. We then define closed braids in $\partial(D^2 \times D^2)$ as links in ∂_1 on which the projection $\mathrm{pr}_2\colon \partial_1 \to S^1$ restricts to a covering map. Note that the boundary of a braided surface is a closed braid in $\partial(D^2 \times D^2)$.

One feature of Rudolph's braided surfaces is that they are all necessarily *ribbon*. A properly embedded surface S in $D^4 = \{(z, w) \mid |z|^2 + |w|^2 \leq 1\}$ is said to be *ribbon embedded* if the function $|z|^2 + |w|^2$ restricts to S as a Morse function with no local maximal points on $\mathrm{int}\, S$. A properly embedded surface in D^4 is said to be *ribbon* if it is isotopic to a surface which is ribbon embedded. By fixing an identification of $D^2 \times D^2$ with D^4, we can similarly consider ribbon surfaces in $D^2 \times D^2$ (the definition of ribbon embeddings in $D^2 \times D^2$ will depend on our choice of identification, though the resulting class of ribbon surfaces will not).

Rudolph proved that any orientable ribbon surface in $D^2 \times D^2$ is isotopic to a braided surface, though in general this isotopy cannot be chosen to fix ∂S, even if ∂S is already a closed braid.

Viro defined a similar notion, which he called a 2–*braid*, by additionally requiring that $\partial S \subset \partial_1 = D^2 \times S^1$ be a trivial closed braid (ie $\partial S = P \times S^1$ for some finite subset $P \subset D^2$). These 2–braids come equipped with a closure operation, yielding closed surfaces in S^4. Viro [29] and Kamada [16] independently proved that every closed oriented surface in S^4 is isotopic to the closure of a 2–braid. These 2–braids were studied further by Kamada [14; 15; 17; 18; 19], who also proved a 4–dimensional Markov theorem relating any two 2–braids with isotopic closures.

2.4 Braided surfaces with caps

The embedded surfaces in $D^2 \times D^2$ we consider in this paper will not in general be ribbon and hence cannot be braided via Rudolph's algorithm. We thus consider a less restrictive notion of braiding, which we define now.

Let $\phi\colon F \to \Sigma$ be a smooth map of oriented surfaces. Then a *cap of F with respect to ϕ* is an embedded disk $D \subset F$ such that

(1) ϕ restricts to embeddings on $\mathrm{int}\, D$ and on ∂D,

(2) F and Σ admit coordinate charts of the form $S^1 \times [-1, 1]$ around $\partial D = S^1 \times \{0\}$ and $\phi(\partial D) = S^1 \times \{0\}$, on which ϕ is given by $(\theta, t) \mapsto (\theta, t^2)$,

(3) in the above coordinate chart around $\phi(\partial D)$, the curve $S^1 \times \{1\}$ lies in $\phi(\mathrm{int}\, D)$.

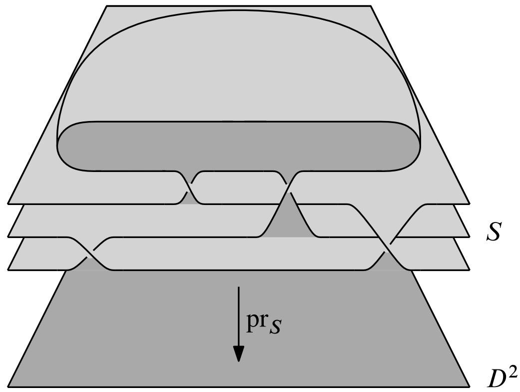


Figure 2: Cross-section of a braided surface with caps

Now let $S \subset D^2 \times D^2$ and let pr_S denote the restriction of pr_2 to S. We say that S is a *braided surface with caps* if the critical points of pr_S all correspond either to isolated simple branch points or boundaries of caps of S with respect to pr_S. Moreover, we will often assume that the critical values in D^2 form a set of embedded concentric circles (corresponding to the boundaries of caps), with isolated critical values lying inside the innermost circle. See Figure 2 for a cross-sectional diagram of a braided surface with a single cap.

2.5 Braided surfaces with caps from braided cobordisms

Braided cobordisms are closely related to braided surfaces with caps, a fact which we illuminate here. We begin by defining a smooth map $\rho\colon S^3 \to D^2$ as follows. Let $\lambda\colon [0,1] \to [0,1]$ be a smooth function with $\lambda(t) = t$ on $[0, \frac{1}{4}]$, $\lambda \equiv 1/\sqrt{2}$ on $[1/\sqrt{2}, 1]$ and so that $d\lambda/dt > 0$ on $[0, 1/\sqrt{2})$. Then we define $\rho\colon S^3 \to D^2$ as

$$\rho(z,w) = \frac{\sqrt{2}w\lambda(|w|)}{|w|}$$

for $w \neq 0$ and $\rho(z,0) = 0$. Clearly ρ is smooth, with $T_1 = \rho^{-1}(\partial D^2)$ and $T_2 = \overline{\rho^{-1}(\operatorname{int} D^2)}$. Furthermore, using ρ we can fix a fibering of T_1 over S^1 with fiber D^2 and a fibering of T_2 over D^2 with fiber S^1. A link $L \subset T_1$ is a closed braid if and only if $\rho|_L\colon L \to S^1$ is a covering map. We call the degree of the covering map $\rho|_L$ the *index* of the closed braid L.

We now identify $\partial(D^2 \times D^2)$ with S^3 by a smooth homeomorphism κ, which smooths the corners of $\partial(D^2 \times D^2)$ and identifies ∂_1 with T_1 and ∂_2 with T_2. Furthermore, we assume that κ is a diffeomorphism away from the corners of $\partial(D^2 \times D^2)$ and maps the fibers of pr_2 diffeomorphically onto the fibers of ρ.

For $0 \leq t \leq 1$, we can multiply $\partial(D^2 \times D^2) \subset \mathbb{C}^2$ by a factor of $\frac{1}{2}(t+1)$ and use κ to identify the resulting set with $S^3 \times \{t\}$. We thus obtain an identification of $S^3 \times [0,1]$ with a collar neighborhood ν of $\partial(D^2 \times D^2)$ in $D^2 \times D^2$, which we denote by $\kappa' \colon \nu \to S^3 \times [0,1]$.

As any properly embedded surface S in $D^2 \times D^2$ can easily be arranged to lie in the collar neighborhood ν, we see that after smoothing corners any such surface gives rise to a smooth properly embedded surface in $S^3 \times [0,1]$ whose boundary lies in $S^3 \times \{1\}$, and vice versa.

Lemma 2.1 *Suppose $W \subset S^3 \times [0,1]$ is a braided cobordism with $W \cap (S^3 \times \{0\}) = \varnothing$. Then $(\kappa')^{-1}(W)$ will be a braided surface with caps in $D^2 \times D^2$ (after a small isotopy smoothing corners around the boundaries of the caps).*

Proof Let $S = (\kappa')^{-1}(W)$ and let pr_S denote the restriction of pr_2 to S. Each local maximum or minimum point of $W \subset S^3 \times [0,1]$ with respect to the height function will lie in $T_2 \times [0,1]$ and we can arrange that each saddle point of W lies in $T_1 \times [0,1]$. Furthermore, by flattening a neighborhood of each local maximum and minimum point, we can isotope W so that it intersects $T_2 \times [0,1] = S^1 \times D^2 \times [0,1]$ in a collection of disks of the form $\{p\} \times D^2 \times \{t\}$. The image of any such disk under $(\kappa')^{-1}$ will be a disk in $\frac{1}{2}(t+1) \cdot \partial_2$ and the restriction of pr_S to its interior will be free of critical points.

Now $W'_t = W \cap (T_1 \times \{t\})$ will be a (possibly singular) closed braid in $T_1 \times \{t\}$ for each $0 \leq t \leq 1$. Each singular braid W'_t will consist of a closed braid with a pair of strands intersecting at a point, with distinct tangent lines. These self-intersections correspond to saddle points of the surface W. Each $(\kappa')^{-1}(W'_t)$ will thus also be a possibly singular closed braid in $\frac{1}{2}(t+1) \cdot \partial_1$, where each singular point gives rise to a simple branch point of the projection pr_S. The nonsingular points of these closed braids all correspond to regular points of pr_S.

Finally, it remains to consider what happens along the boundaries of the disks in $W \cap (T_2 \times [0,1])$. For any disk D corresponding to a local minimum of W, the boundary of $(\kappa')^{-1}(D)$ can be smoothed in such a way that the resulting points are all regular points of the map pr_S. If D instead corresponds to a local maximum, then the boundary of $(\kappa')^{-1}(D)$ is instead smoothed in such a way that $(\kappa')^{-1}(D)$ becomes a cap of S with respect to pr_S. Since all critical points of pr_S are either isolated simple branch points or lie along the boundary of a cap, $S \subset D^2 \times D^2$ is a braided surface with caps. □

3 Braiding link cobordisms

We start the proof of Theorem 1.1. For the duration of the proof, it will be convenient to think of our cobordisms as lying in $\mathbb{R}^3 \times [0,1]$, so that we can use the diagrammatic approach described in Section 2.2. Suppose that $W \subset \mathbb{R}^3 \times [0,1]$ is a properly embedded oriented link cobordism between closed braids $B_0 \subset \mathbb{R}^3 \times \{0\}$ and $B_1 \subset \mathbb{R}^3 \times \{1\}$. Assume furthermore that the restriction of the projection $\mathrm{pr}_2\colon \mathbb{R}^3 \times [0,1] \to [0,1]$ to W is a Morse function. For any such surface $W \subset \mathbb{R}^3 \times [0,1]$ and any $[a,b] \subset [0,1]$, let $W_{[a,b]} = W \cap (\mathbb{R}^3 \times [a,b])$ and $W_t = W \cap (\mathbb{R}^3 \times \{t\})$.

3.1 Braiding around critical points

We begin by proving that W can be "braided" in a neighborhood of the critical points of $\mathrm{pr}_2|_W$. This will reduce the problem of proving Theorem 1.1 to proving it for cobordisms W without critical points.

Lemma 3.1 *There is an isotopy of W rel ∂W taking W to a surface W' such that $W'_{[a,b]}$ is a braided cobordism for $[a,b] \in \{[0,\frac{1}{6}],[\frac{1}{3},\frac{2}{3}],[\frac{5}{6},1]\}$ and is free of critical points for $[a,b] \in \{[\frac{1}{6},\frac{1}{3}],[\frac{2}{3},\frac{5}{6}]\}$.*

Proof As both B_0 and B_1 are closed braids, W_t will also be a closed braid for t close to 0 and 1, so we can assume that W_t is a closed braid for all $t \in [0,\frac{1}{6}] \cup [\frac{5}{6},1]$. Push all minimal points into $\mathbb{R}^3 \times [0,\frac{1}{6}]$, all maximal points into $\mathbb{R}^3 \times [\frac{5}{6},1]$ and all saddle points into $\mathbb{R}^3 \times \{\frac{1}{2}\}$ (see [19] for details). The maximal and minimal points can easily be positioned in such a way that $W'_{[0,1/6]}$ and $W'_{[5/6,1]}$ remain braided.

Now, passing each saddle point changes the level set W_t by surgery along a 2–dimensional 1–handle. After a small perturbation in a neighborhood of each saddle point, we can assume that these 1–handles all lie in $\mathbb{R}^3 \times \{\frac{1}{2}\}$. By adding a half-twist in each band, we can arrange that each segment of $W_{1/2+\varepsilon}$ and $W_{1/2-\varepsilon}$ involved in the surgeries are oriented in the positive direction (see Figure 3, where $W_{1/2}$ is shown). Keeping these bands in place, the remaining strands of $W_{1/2}$ can be braided using the standard proof of the classical Alexander's theorem. Thus we can arrange $W_{1/2}$ so that it is a closed braid both before and after the surgeries, and can extend the closed braid structure to the rest of $W'_{[1/3,2/3]}$. □

The above argument is due to Kamada [19].

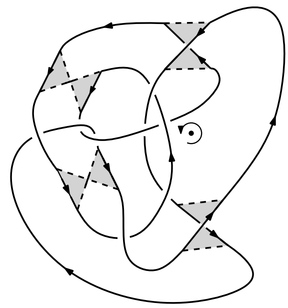


Figure 3: Arranging 1–handles

3.2 Braiding critical point-free cobordisms

Any cobordism W which is free of critical points is topologically just a union of cylinders, and is isotopic to a product cobordism. In general, however, the isotopy taking W to a product cobordism cannot be chosen to fix the boundary. Consider, for example, the movie presentation of the critical point-free cobordism W depicted in Figure 4 (where the middle still is meant to imply that the bottom strand is given a non-zero number of full twists as we look at the level sets moving down). Here, W is isotopic to a product cobordism, but there is no such isotopy fixing ∂W.

The movie presentations of a critical point-free cobordism is described entirely by its starting diagram and a sequences of Reidemeister moves and planar isotopies. We will complete the proof of Theorem 1.1 in two stages, first by proving it for critical point-free cobordisms whose movie presentation is described entirely by a planar isotopy (ie no Reidemeister moves take place between nearby stills) before proving it for the general case. Before doing this however, we must first recall a geometric set of Markov moves for classical links used by Morton [24], as well as his threading construction, which gives a diagrammatic approach to studying isotopies of closed braids. The proof of Theorem 1.1 relies on enhancements of the arguments used in his proof of Markov's theorem.

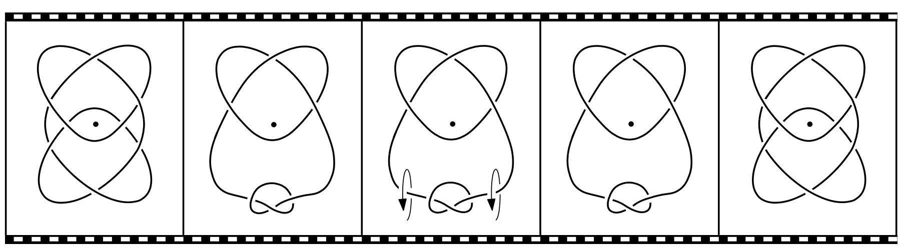


Figure 4: Critical point-free cobordism not isotopic rel boundary to product cobordism

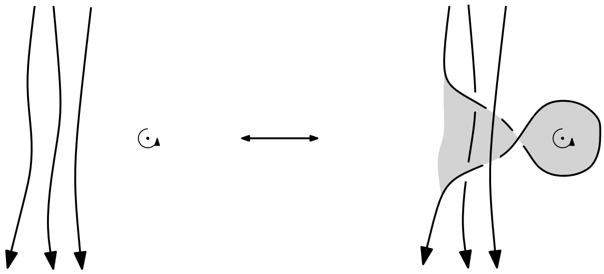


Figure 5: Simple Markov equivalence

3.3 Geometric Markov moves for closed braids in $\mathbb{R}^3$

Morton's geometric formulation of Markov's theorem states that two closed braids which are isotopic as links can be joined by a sequence of braid isotopies and simple Markov equivalences. A *braid isotopy* between two closed braids L_0 and L_1 in $\mathbb{R}^3$ is an isotopy ϕ_α of $\mathbb{R}^3$, ie a continuous family of maps $\phi_\alpha\colon \mathbb{R}^3 \to \mathbb{R}^3$ parametrized by $\alpha \in [0,1]$ with $\phi_0 = \mathrm{id}_{\mathbb{R}^3}$ such that $\phi_\alpha(L_0)$ is a closed braid for all α and $\phi_1(L_0) = L_1$.

The second move on closed braids is a geometric version of braid stabilization. Let B and B' be closed braids and suppose there is an oriented embedded disk $R \subset \mathbb{R}^3$ intersecting the z–axis transversely in a single point. Suppose also that $\partial R = c \cup c'$, where $c = B \cap R$ and $c' = B' \cap R$ are connected and where the boundary orientation of ∂R is winding clockwise along c and counterclockwise along c'. Suppose further that $B \backslash c = B' \backslash c'$. Then B and B' are said to be *simply Markov equivalent* (see Figure 5, where the disk R is shaded).

The projections of such B and B' to the xy–plane differ by a sequence of Reidemeister moves which includes precisely one move of type I creating an extra loop around the origin.

3.4 Threading construction

Let P be the xz–plane and let $\pi'\colon \mathbb{R}^3 \to P$ be the orthogonal projection. Let $h \subset P$ be the image of the z–axis under π'. Suppose D is the diagram in P of an oriented link L. Let $C \subset D$ denote the double points (crossings) of L under the projection π'.

A *choice of overpasses* for D is a pair of disjoint finite subsets S, $F \subset D \backslash C$ such that each link component contains points from $S \cup F$ and points of S alternate with points of F when traveling along any component. Furthermore, when traveling in the positively oriented direction, each arc of the form $[s, f]$ contains no under-crossings and each arc $[f, s]$ contain no over-crossings.

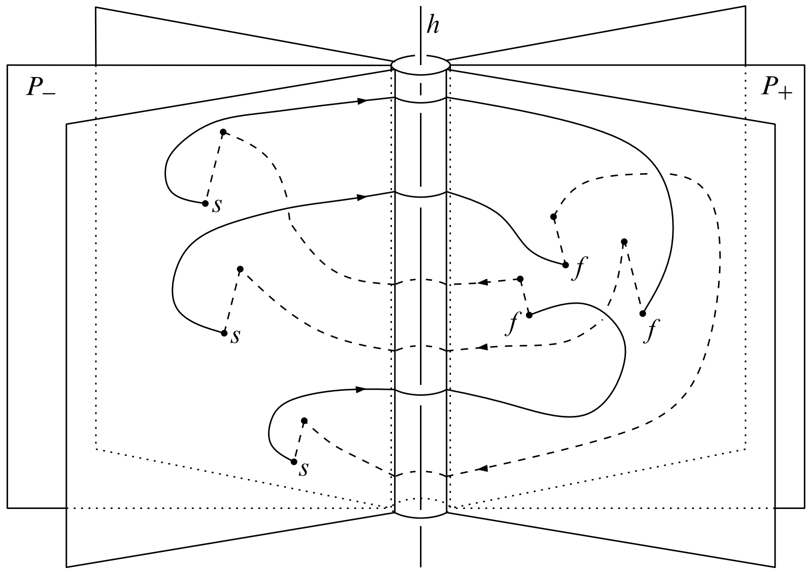


Figure 6: Trefoil as a closed braid given by a threading

Now let $P_+ = P \cap \{x > 0\}$ and $P_- = P \cap \{x < 0\}$ be the right- and left-hand regions of P separated by h, respectively. Although h is not a component of L, we can enhance the diagram D by assigning a crossing choice whenever D intersects h transversely.

Given such an enhanced diagram, h is said to *thread* the diagram D for some choice of overpasses (S, F) if h intersects D transversely, $S \subset P_-$, $F \subset P_+$ and

(1) D crosses over h when traveling from P_- to P_+,

(2) D crosses under h when traveling from P_+ to P_-.

Threadings of link diagrams allow us to study closed braids on the level of link diagrams. The following lemma is due to Morton (see [24]):

Lemma 3.2 *Suppose D is a diagram that is threaded by h for some choice of overpasses. Then there is a closed braid L with diagram D.*

The idea behind the proof of the lemma is summarized in Figure 6. Note that, even if the over- and under-crossing information of D with h has not been specified, there is a unique assignment to each such crossing so that the resulting diagram lifts to a closed braid. Conversely, it is also easy to show that any closed braid is braid-isotopic to one whose diagram is threaded by h for some choice of overpasses.

3.5 Braiding movie presentations without Reidemeister moves

Now suppose that $W \subset \mathbb{R}^3 \times [0,1]$ is a critical point-free cobordism between two closed braids and consider the movie presentation of W, this time projecting each $W_t \subset \mathbb{R}^3 \times \{t\} = \mathbb{R}^3$ to the plane P via the projection π'. We let D_t denote the (possibly singular) diagram of W_t in P for each $t \in [0,1]$. As W is free of critical points, nearby diagrams will differ by either a planar isotopy or Reidemeister move. If the movie presentation of W does not involve any Reidemeister moves, then it can be described completely by specifying the initial diagram D_0 and a planar isotopy ϕ_α of P, with $\phi_\alpha(D_0) = D_\alpha$ for all α. In what follows it will be convenient to specify the movie presentations of such surfaces in this way.

We prove Theorem 1.1 first in the special case when D_0 and D_1 are threaded and the movie presentation of W does not involve any Reidemeister moves:

Proposition 3.3 *Suppose W has no critical points and that its movie presentation does not involve any Reidemeister moves. Suppose further that W_0 and W_1 are closed braids with diagrams D_0 and D_1 threaded by h for some choices of overpasses. Then W is isotopic relative its boundary to a braided cobordism.*

In order to prove the above proposition we will need to lift the planar isotopy joining D_0 and D_1 to a sequence of braid isotopies and simple Markov equivalences in $\mathbb{R}^3$. For the rest of this section we assume W is as described in the statement of Proposition 3.3. The first lemma we will need is the following:

Lemma 3.4 *Let ψ_α be a planar isotopy of P taking D_0 to D_1 which fixes h setwise. Suppose further that $\psi_\alpha \equiv \psi_0$ and $\psi_{1-\alpha} \equiv \psi_1$ for α in a small neighborhood of 0. Then there is a braid isotopy ϕ_α taking W_0 to W_1 such that $\pi' \circ \phi_\alpha(W_0) = \psi_\alpha(D_0)$ for all $\alpha \in [0,1]$.*

Proof For any $p \in W_0$ and $\alpha \in [0,1]$, the $x-$ and z–coordinates of $\phi_\alpha(p)$ are determined by ψ_α. The y–coordinate of $\phi_\alpha(p)$ can then be chosen uniquely so that the radial coordinate of $\phi_\alpha(p)$ remains constant for all α. It thus suffices to note that any two closed braids with the same diagram are also braid-isotopic, via a straight line isotopy. □

Let (S_0, F_0), $(S_1, F_1) \subset P$ denote the overpasses chosen for the threadings of D_0 and D_1, respectively, and let ψ_α denote a planar isotopy of P associated to the movie presentation of W, ie $\psi_\alpha(D_0) = D_\alpha$ for all $\alpha \in [0,1]$. We can assume that

$$S_0 \cap \psi_1^{-1}(S_1) = F_0 \cap \psi_1^{-1}(F_1) = \varnothing.$$

The following lemma will allow us to assume that the choices of overpasses for both D_0 and D_1 coincide and that they can be assumed to be fixed by the planar isotopy ψ_α.

Lemma 3.5 *W is isotopic relative its boundary to a cobordism whose movie presentation is determined by the diagram D_0 and a planar isotopy φ_α, where $\varphi_\alpha(S_0) = S_0$ and $\varphi_\alpha(F_0) = F_0$ for $0 \le \alpha \le \frac{1}{2}$, and $\varphi_\alpha(S_1) = S_1$ and $\varphi_\alpha(F_1) = F_1$ for $\frac{1}{2} \le \alpha \le 1$.*

Proof We can assume that, for all $q \in S_1 \cup F_1$, the sets $\{\psi_\alpha^{-1}(q) \mid 0 \le \alpha \le 1\}$ are disjoint embedded arcs in P which do not intersect $S_0 \cup F_0$ (see for example [6, Lemma 10.4]). For each $q \in S_1 \cup F_1$ choose a small regular neighborhood A_q of $\{\psi_\alpha^{-1}(q) \mid 0 \le \alpha \le 1\}$ so that the A_q are pairwise disjoint and also do not intersect $S_0 \cup F_0$.

Now let ξ_α be a planar isotopy of P which restricts to the identity on the complement of $\bigcup A_q$ and is such that for all $\alpha \in [0,1]$ and all $p \in \psi_1^{-1}(S_1 \cup F_1)$ we have $\xi_\alpha(p) = \psi_{1-\alpha}^{-1} \circ \psi_1(p)$. Let $\Gamma_{\tau,\alpha}$ be the one-parameter family of planar isotopies of P, with $\tau \in [0,1]$, defined by

$$\Gamma_{\tau,\alpha} = \begin{cases} \xi_{2\tau\alpha} & \text{if } 0 \le \alpha \le \frac{1}{2}, \\ \xi_{\tau(2-2\alpha)} & \text{if } \frac{1}{2} \le \alpha \le 1. \end{cases}$$

After an isotopy of W which rescales the t–coordinate, we can arrange that the movie presentation of W is instead described by the planar isotopy

$$\Phi_\alpha = \begin{cases} \mathrm{id}_P & \text{if } 0 \le \alpha \le \frac{1}{2}, \\ \psi_{2\alpha-1} & \text{if } \frac{1}{2} \le \alpha \le 1. \end{cases}$$

Now consider the composition $\Phi_\alpha \circ \Gamma_{\tau,\alpha}$. Letting τ range from 0 to 1 shows that the surface W, which is described by the diagram D_0 and the planar isotopy $\Phi_\alpha = \Phi_\alpha \circ \Gamma_{0,\alpha}$, is isotopic to a surface described by D_0 and the planar isotopy

$$\varphi_\alpha := \Phi_\alpha \circ \Gamma_{1,\alpha} = \begin{cases} \xi_{2\alpha} & \text{if } 0 \le \alpha \le \frac{1}{2}, \\ \psi_{2\alpha-1} \circ \xi_{2-2\alpha} & \text{if } \frac{1}{2} \le \alpha \le 1. \end{cases}$$

As the ξ_α is the identity outside of $\bigcup A_q$, for any $p \in S_0 \cup F_0$ and any $\alpha \in [0, \frac{1}{2}]$ we have $\varphi_\alpha(p) = \xi_{2\alpha}(p) = p$. For $\alpha \in [\frac{1}{2}, 1]$ and $q \in S_1 \cup F_1$ we have

$$\varphi_\alpha(q) = \psi_{2\alpha-1} \circ \xi_{2-2\alpha}(q) = \psi_{2\alpha-1} \circ \psi_{1-(2-2\alpha)}^{-1}(q) = q,$$

as required. Note that all the isotopies described above fix $W_0 \cup W_1 = \partial W$. □

By the above lemma it is enough to prove Proposition 3.3 in the case when $S = S_0 = S_1$, $F = F_0 = F_1$ and all points in $S \cup F$ are fixed by ψ_α. Indeed, since the points in $S_0 \cup F_0$ are stationary during the first half of the planar isotopy φ_α and since they form a choice of overpasses for which D_0 is threaded, they must also form a choice

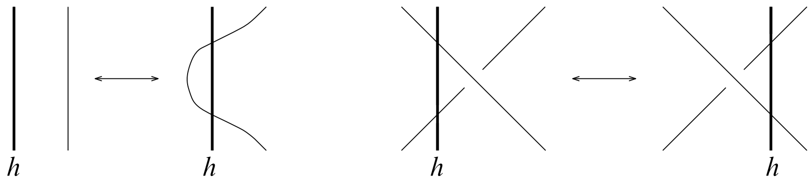


Figure 7: Reidemeister-like moves involving h

of overpasses which give rise to a threading of $D_{1/2}$. Likewise, $D_{1/2}$ is threaded by h with the choice of overpasses (S_1, F_1), since they remain stationary for during the second half of φ_α and give a threading of D_1. By Lemma 3.2 we can arrange W locally near $\mathbb{R}^3 \times \{\frac{1}{2}\}$ so that $W_{1/2}$ is a closed braid with diagram $D_{1/2}$ threaded with either choice of overpasses and prove Proposition 3.3 for $W_{[0,1/2]}$ and $W_{[1/2,1]}$.

Suppose then that W is as above. Although the movie presentation of W does not involve any Reidemeister moves, it will (after perturbing W slightly away from the boundary) contain Reidemeister II– and III–like moves involving components of the diagrams and the z–axis h (see Figure 7). These Reidemeister-like moves are like classical Reidemeister moves, but where no crossing information is specified at double points of the projection involving h. The absence of crossing information with h reflects the fact that the movie presentation of W does not specify the relative position of the links W_t above or below P and that the components of the link are free to pass through the z–axis during isotopies in $\mathbb{R}^3$.

We can thus break the planar isotopy ψ_α determining W into a sequence of transformations that take into account the relative position of the diagrams D_t with h. More precisely, we can divide the interval [0, 1] into smaller subintervals $[t_{j-1}, t_j]$ such that for each j there is either

(1) a planar isotopy ϕ_α^j of P that fixes h setwise with $\phi_\alpha^j(D_{t_{j-1}}) = D_{t_{j-1}+\alpha(t_j - t_{j-1})}$ for all $\alpha \in [0, 1]$, or

(2) a Reidemeister-like move of type II or III taking $D_{t_{j-1}}$ to D_{t_j} involving (but fixing) h.

We will simplify notation and write D^j and W^j instead of D_{t_j} and W_{t_j}, respectively, for each j. Since we are assuming that the points of $S \cup F$ are fixed throughout the planar isotopy ψ_α, we can fix (S, F) as a choice of overpass for each D^j. Furthermore, for each diagram we fix the unique choice of h–crossing information so that D^j is threaded by h.

Before proceeding, we need to eliminate any situations as in Figure 8. Here we have a Reidemeister-like move of type III where the center crossing cannot pass to the

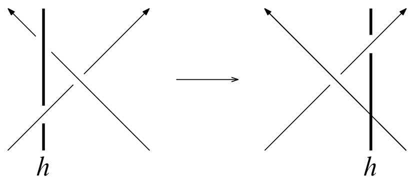


Figure 8: Reidemeister-like move of type III which does not lift to a braid isotopy

other side of h without first introducing crossing changes. These can be eliminated by making a local replacement as in Figure 9, where the offending move has been replaced by a sequence consisting of three Reidemeister-like moves, two of type II and one of type III (which lifts to an isotopy avoiding the z–axis). This local replacement does not change the isotopy class of W rel ∂W.

Lemma 3.6 *If W^{j-1} is a closed braid, then the transformation $D^{j-1} \to D^j$ lifts to $\mathbb{R}^3$ as a sequence of braid isotopies and simple Markov equivalences on W^{j-1}.*

Proof Note first that, since W^{j-1} is a closed braid and D^{j-1} is threaded, the h–crossing information on D^{j-1} will match that coming from the projection of W^{j-1}.

For transformations of type (1) above, Lemma 3.4 shows that the planar isotopy between D^{j-1} and D^j can be lifted to a braid isotopy on W^{j-1}.

Suppose now that D^j is obtained from D^{j-1} by a Reidemeister-like move of type II (or its inverse) as in Figure 7. Then, as D^{j-1} is threaded, locally it must look like either the right- or left-hand side of one of the transformations in Figure 10. Note that by assumption no points of S or F can occur anywhere in these local pictures. Clearly D^j can be lifted to a closed braid W^j which agrees with W^{j-1} away from the Reidemeister-like move of type II, so that W^{j-1} and W^j are simply Markov equivalent.

Now suppose that D^j is obtained from D^{j-1} by a Reidemeister-like move of type III. It is easy to verify that for most configurations of D^{j-1} the move can be lifted to a braid

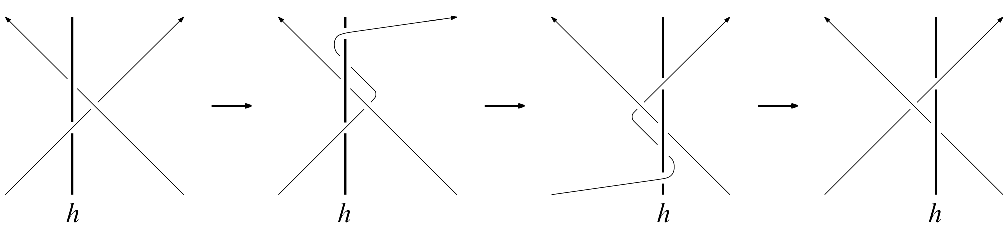


Figure 9: Replacing bad Reidemeister-like moves of type III with a sequence of moves that lift to braid isotopies and simple Markov equivalences

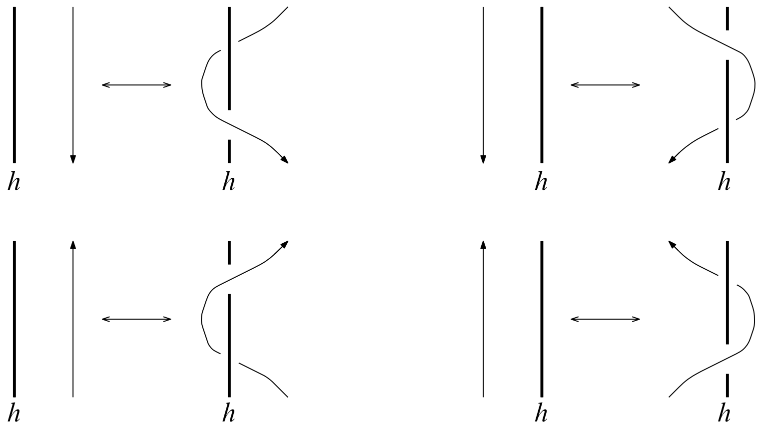


Figure 10: Reidemeister-like moves of type II

isotopy taking W^{j-1} to a closed braid W^j with diagram D^j. The only exceptions arise as in Figure 8, but these were all replaced previously by sequences of moves that can be lifted. □

Starting with the closed braid $W_0 \subset \mathbb{R}^3 \times \{0\}$, we can construct a new surface W' by tracing the path of W_0 in $\mathbb{R}^3 \times [0,1]$ as we apply the sequence of lifted braid isotopies and simple Markov equivalences obtained from the previous lemma. Away from the simple Markov equivalences each level set W'_t will be a closed braid. By construction, the movie presentation of W' will be the same as that of W, hence it will be isotopic to W rel $\partial W'$. To prove Proposition 3.3 it thus remains only to show that W can be braided in neighborhoods of the simple Markov equivalences.

Proof of Proposition 3.3 Suppose that, for some $s \in [0,1]$ and $\varepsilon > 0$, the closed braids $W_{s-\varepsilon}$ and $W_{s+\varepsilon}$ differ by a simple Markov equivalence spanned by a disk R. After a small isotopy in the neighborhood of the hyperplane $\mathbb{R}^3 \times \{s\}$ we can assume that R lies entirely in this hyperplane and that the orthogonal projection of ∂R to the xy–plane yields a figure eight.

Decompose R as the boundary sum of two closed disks R' and R'' (equipped with the orientation of W), where R' intersects the z–axis transversely in a single point and $\partial R'$ is a simple curve which is strictly monotone in the angular direction (see Figure 11). Push R' to either $\mathbb{R}^3 \times \{s+\varepsilon\}$ or $\mathbb{R}^3 \times \{s-\varepsilon\}$ (depending on whether $\partial R'$ is monotone increasing or decreasing, respectively) while keeping R'' fixed. This gives rise to a new maximal disk (minimal disk, respectively) while R'' yields a new saddle band. After a slight local perturbation these new critical disks can be changed to isolated critical points, completing the proof of Proposition 3.3. □

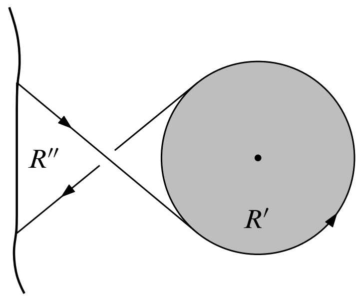


Figure 11: Decomposing R as the boundary sum of R' and R''

3.6 Braiding movie presentations with Reidemeister moves

Now consider an arbitrary critical point-free cobordism W between two closed braids. The movie presentation of W under the projection to P will in general include Reidemeister moves as well as planar isotopies. Recycling notation from above, let D_t denote the diagram of W_t and divide the interval $[0,1]$ into smaller subintervals $[t_{j-1}, t_j]$ such that for each j there is either

(1) a planar isotopy ϕ_α^j of P which has $\phi_\alpha^j(D_{t_{j-1}}) = D_{t_{j-1}+\alpha(t_j - t_{j-1})}$ for all $\alpha \in [0,1]$, or

(2) a Reidemeister move taking $D_{t_{j-1}}$ to D_{t_j}.

As above we will simplify notation and write D^j and W^j instead of D_{t_j} and W_{t_j}, respectively, for each j. To complete the proof of Theorem 1.1 we need:

Lemma 3.7 *Suppose D^j is obtained from D^{j-1} by a Reidemeister move of any type. Then there is a planar isotopy ζ_α of P such that $\zeta_1(D^{j-1})$ and $\zeta_1(D^j)$ are both threaded by h for some choice of overpasses and, if W^{j-1} is a closed braid with diagram $\zeta_1(D^{j-1})$, then the Reidemeister move taking $\zeta_1(D^{j-1})$ to $\zeta_1(D^j)$ lifts to a braid isotopy of W^{j-1}.*

To see that this completes the proof of Theorem 1.1, note first that by [24, Theorem 2] there are braid isotopies taking W_0 and W_1 to closed braids whose diagrams in P are threaded by h for some choices of overpasses. Thus we can assume that the diagrams D_0 and D_1 are both threaded. We also assume that in the movie presentation of W the sequence involved alternates between planar isotopies and Reidemeister moves, beginning and finishing with planar isotopies. Suppose for some j that D^j is obtained from D^{j-1} by a Reidemeister move and let ϕ_α^{j-1} and ϕ_α^{j+1} be the planar isotopies taking D^{j-2} to D^{j-1} and D^j to D^{j+1}, respectively. Then we can replace

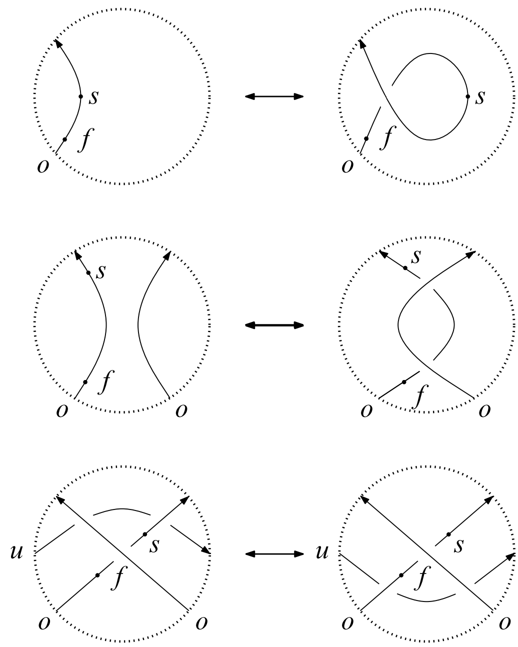


Figure 12: Overpass choices in a neighborhood of type I and II moves

D^{j-1} and D^j with $\zeta_1(D^{j-1})$ and $\zeta_1(D^j)$, respectively, and ϕ_α^{j-1} and ϕ_α^{j+1} with $\zeta_\alpha \circ \phi_\alpha^{j-1}$ and $\zeta_{1-\alpha} \circ \phi_\alpha^{j+1}$, respectively, without changing the isotopy class of W rel ∂W. Performing a similar replacement one by one around all Reidemeister moves in the movie presentation, we see that W is isotopic relative its boundary to a cobordism whose movie presentation involves only Reidemeister moves and planar isotopies between threaded diagrams.

Thus we can assume that each of the D^j are threaded and that the W^j are all closed braids. By Lemma 3.7, the portions of W corresponding to planar isotopies in the movie presentation are then isotopic relative their boundaries to braided cobordisms, while by Proposition 3.3 we see that the same is true for portions of W corresponding to Reidemeister moves. Thus W itself is isotopic relative its boundary to a braided cobordism, completing the proof.

Proof of Lemma 3.7 Begin by making a choice of overpasses for D^{j-1} and D^j which agree outside some small neighborhood of the move in question. In the small neighborhood of the move we choose points which give a valid choice of overpasses both before and after the move. See examples of different possible configurations in Figure 12, where incoming strands are labeled with o if they are part of an overpass or u if they are part of an underpass.

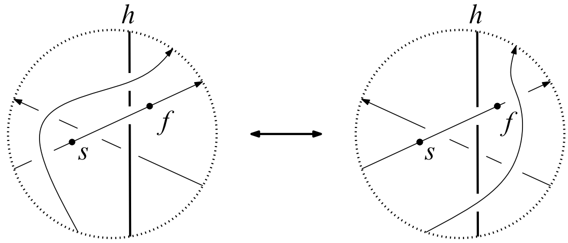


Figure 13: Threading near a Reidemeister move of type III

Now let ζ_α be a planar isotopy which repositions all of the S points to P_- (the left half of the plane P) and all the F points to P_+ (the right half of P). Once positioned in this way, there is a unique way to assign over- and under-crossings of D^{j-1} and D^j with h so that both diagrams are threaded by h.

Note that, in the case of moves of type I and II, we can choose S, F and ζ_α so that the Reidemeister move of interest happens away from h. It is then easy to see that the Reidemeister move of interest lifts to a braid isotopy.

Moves of type III cannot be arranged to take place away from h, however. Of the three strands in this local picture, one strand will cross over the other two, one will pass under the other two, while the third will pass over one and under the other. Choose S and F away from this picture so that the top strand is part of an over-crossing, the bottom strand is part of an under-crossing and place a single point from each of S and F on the third strand to create a valid choice of overpasses.

Now we can arrange the diagrams so that h separates S and F, and so that the uppermost strand crosses over h in a neighborhood of the move (the orientation of this strand determines whether it will cross h at the top or bottom of the local picture). Regardless then of the orientation on the other two strands or their shared crossing, the uppermost strand is free to pass over the crossing and both the nearby S and F points as in Figure 13, a move which can clearly be lifted to a braid isotopy in $\mathbb{R}^3$. This completes the proof of Lemma 3.7 and of Theorem 1.1. □

Corollary 1.2 now follows easily by combining Theorem 1.1 with Lemma 2.1.

Remark Suppose now that the cobordism W we start with is in ribbon position, ie has no local maximal points with respect to the t–coordinate. Although we may hope to preserve this property during the braiding procedure described above, this will not be possible in general. Indeed, Morton [23] gave an example of a 4–strand braid β with unknotted closure which is *irreducible*, meaning any simplification of β using Markov moves necessarily raises the braid index to 5. As noted by Rudolph [26], it is not difficult to see that any braided ribbon cobordism bounded by the closure of β must have genus at least 1, even though it clearly bounds a ribbon embedded disk in $S^3 \times [0,1]$.

References

[1] **S Akbulut**, **Ç Karakurt**, *Every* 4*–manifold is BLF*, J. Gökova Geom. Topol. GGT 2 (2008) 83–106 MR2466002

[2] **J Alexander**, *A lemma on systems of knotted curves*, Proc. Nat. Acad. Sci. USA 9 (1923) 93–95

[3] **D Auroux**, **S K Donaldson**, **L Katzarkov**, *Singular Lefschetz pencils*, Geom. Topol. 9 (2005) 1043–1114 MR2140998

[4] **R İ Baykur**, *Existence of broken Lefschetz fibrations*, Int. Math. Res. Not. 2008 (2008) MR2439543

[5] **R İ Baykur**, *Broken Lefschetz fibrations and smooth structures on* 4*–manifolds*, from: "Proceedings of the Freedman Fest", (R Kirby, V Krushkal, Z Wang, editors), Geom. Topol. Monogr. 18 (2012) 9–34 MR3084230

[6] **G Burde**, **H Zieschang**, *Knots*, de Gruyter Studies in Mathematics 5, de Gruyter, Berlin (1985) MR808776

[7] **S Carter**, **S Kamada**, **M Saito**, *Surfaces in* 4*–space*, Encyclopaedia of Mathematical Sciences 142, Springer, Berlin (2004) MR2060067

[8] **S K Donaldson**, *Lefschetz fibrations in symplectic geometry*, from: "Proceedings of the International Congress of Mathematicians, Extra Vol. II" (1998) 309–314 MR1648081

[9] **P Freyd**, **D Yetter**, **J Hoste**, **W B R Lickorish**, **K Millett**, **A Ocneanu**, *A new polynomial invariant of knots and links*, Bull. Amer. Math. Soc. 12 (1985) 239–246 MR776477

[10] **D T Gay**, **R Kirby**, *Constructing Lefschetz-type fibrations on four-manifolds*, Geom. Topol. 11 (2007) 2075–2115 MR2350472

[11] **R E Gompf**, **A I Stipsicz**, 4*–manifolds and Kirby calculus*, Graduate Studies in Mathematics 20, Amer. Math. Soc. (1999) MR1707327

[12] **M Jacobsson**, *An invariant of link cobordisms from Khovanov homology*, Algebr. Geom. Topol. 4 (2004) 1211–1251 MR2113903

[13] **V F R Jones**, *A polynomial invariant for knots via von Neumann algebras*, from: "Fields medallists' lectures", (M Atiyah, D Iagolnitzer, editors), World Sci. Ser. 20^{th} Century Math. 5, World Sci. Publ. (1997) 448–458 MR1622916

[14] **S Kamada**, 2*–dimensional braids and chart descriptions*, from: "Topics in knot theory", (M E Bozhüyük, editor), NATO Adv. Sci. Inst. Ser. C Math. Phys. Sci. 399, Kluwer, Dordrecht (1993) 277–287 MR1257915

[15] **S Kamada**, *Alexander's and Markov's theorems in dimension four*, Bull. Amer. Math. Soc. 31 (1994) 64–67 MR1254074

[16] **S Kamada**, *A characterization of groups of closed orientable surfaces in* 4*–space*, Topology 33 (1994) 113–122 MR1259518

[17] **S Kamada**, *On braid monodromies of non-simple braided surfaces*, Math. Proc. Cambridge Philos. Soc. 120 (1996) 237–245 MR1384466

[18] **S Kamada**, *Arrangement of Markov moves for* 2*–dimensional braids*, from: "Low-dimensional topology", (H Nencka, editor), Contemp. Math. 233, Amer. Math. Soc. (1999) 197–213 MR1701684

[19] **S Kamada**, *Braid and knot theory in dimension four*, Mathematical Surveys and Monographs 95, Amer. Math. Soc. (2002) MR1900979

[20] **M Khovanov**, **L Rozansky**, *Matrix factorizations and link homology, II*, Geom. Topol. 12 (2008) 1387–1425 MR2421131

[21] **Y Lekili**, *Wrinkled fibrations on near-symplectic manifolds*, Geom. Topol. 13 (2009) 277–318 MR2469519

[22] **A Loi**, **R Piergallini**, *Compact Stein surfaces with boundary as branched covers of* B^4, Invent. Math. 143 (2001) 325–348 MR1835390

[23] **H R Morton**, *An irreducible* 4*–string braid with unknotted closure*, Math. Proc. Cambridge Philos. Soc. 93 (1983) 259–261 MR691995

[24] **H R Morton**, *Threading knot diagrams*, Math. Proc. Cambridge Philos. Soc. 99 (1986) 247–260 MR817666

[25] **L Rudolph**, *Braided surfaces and Seifert ribbons for closed braids*, Comment. Math. Helv. 58 (1983) 1–37 MR699004

[26] **L Rudolph**, *Special positions for surfaces bounded by closed braids*, Rev. Mat. Iberoamericana 1 (1985) 93–133 MR836285

[27] **L Rudolph**, *Quasipositivity as an obstruction to sliceness*, Bull. Amer. Math. Soc. 29 (1993) 51–59 MR1193540

[28] **L Rudolph**, *Knot theory of complex plane curves*, from: "Handbook of knot theory", (W Menasco, M Thistlethwaite, editors), Elsevier, Amsterdam (2005) 349–427 MR2179266

[29] **O Y Viro**, lecture, Osaka City University (1999)

Department of Mathematics, Brigham Young University
312 TMCB, Provo, UT 84602, USA

hughes@mathematics.byu.edu

http://math.byu.edu/~hughes/

Received: 2 March 2015 Revised: 31 March 2015

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.

Preparing your article for Algebraic & Geometric Topology

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.

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

AGT papers are published simultaneously on the `arXiv`. If your paper has previously been deposited there, we shall need the `arXiv` password at acceptance time, to submit the published version to the `arXiv`.

References

Bibliographical references should be listed alphabetically at the end of the paper. All references in the bibliography should be cited at least once in the text. Use of BibTeX is preferred but not required. Any bibliographical citation style may be used, but will be converted to the house style (see a current issue for examples).

Figures

Figures, whether prepared electronically or hand-drawn, must be of publication quality. Figures prepared electronically should be submitted in Encapsulated PostScript (EPS) or in a form that can be converted to EPS, such as GnuPlot, Maple, Mathematica or XFig. Many drawing tools such as Adobe Illustrator and Aldus FreeHand also produce EPS output. Figures containing bitmaps should be generated at the highest possible resolution. If there is doubt whether a particular figure is in an acceptable format, check with production by sending an email to agt@msp.warwick.ac.uk.

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 15 Issue 6 (pages 3107–3729) 2015

Universality of multiplicative infinite loop space machines 3107
DAVID GEPNER, MORITZ GROTH and THOMAS NIKOLAUS

Floer homology and splicing knot complements 3155
EAMAN EFTEKHARY

Higher Hochschild cohomology of the Lubin–Tate ring spectrum 3215
GEOFFROY HOREL

Fixed-point free circle actions on 4–manifolds 3253
WEIMIN CHEN

Exactly fourteen intrinsically knotted graphs have 21 edges 3305
MINJUNG LEE, HYOUNGJUN KIM, HWA JEONG LEE and SEUNGSANG OH

Equivalence classes of augmentations and Morse complex sequences of Legendrian knots 3323
MICHAEL B HENRY and DAN RUTHERFORD

On finite derived quotients of 3–manifold groups 3355
WILL CAVENDISH

A colored operad for string link infection 3371
JOHN BURKE and ROBIN KOYTCHEFF

Systoles and kissing numbers of finite area hyperbolic surfaces 3409
FEDERICA FANONI and HUGO PARLIER

Combinatorial cohomology of the space of long knots 3435
ARNAUD MORTIER

On the K–theory of subgroups of virtually connected Lie groups 3467
DANIEL KASPROWSKI

McCool groups of toral relatively hyperbolic groups 3485
VINCENT GUIRARDEL and GILBERT LEVITT

A generating set for the palindromic Torelli group 3535
NEIL J FULLARTON

Character varieties of double twist links 3569
KATHLEEN L PETERSEN and ANH T TRAN

The L^2–Alexander torsion is symmetric 3599
JÉRÔME DUBOIS, STEFAN FRIEDL and WOLFGANG LÜCK

Cup products, the Johnson homomorphism and surface bundles over surfaces with multiple fiberings 3613
NICK SALTER

The algebraic duality resolution at $p = 2$ 3653
AGNÈS BEAUDRY

Braiding link cobordisms and non-ribbon surfaces 3707
MARK C HUGHES

1472-2747(2015)15:6;1-Y