Algebraic & Geometric Topology

Volume 23 (2023)

Issue 4 (pages 1463–1934)

msp

ALGEBRAIC & GEOMETRIC TOPOLOGY

msp.org/agt

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

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

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

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

PUBLISHED BY

mathematical sciences publishers

nonprofit scientific publishing

http://msp.org/

Algebraic & Geometric Topology **23**:4 (2023) 1463–1500
DOI: 10.2140/agt.2023.23.1463
Published: 14 June 2023

The Heisenberg plane

STEVE TRETTEL

The geometry of the Heisenberg group acting on the plane arises naturally in geometric topology as a degeneration of the familiar spaces $\mathbb{S}^2$, $\mathbb{H}^2$ and $\mathbb{E}^2$ via *conjugacy limit* as defined by Cooper, Danciger and Wienhard. This paper considers the deformation and regeneration of Heisenberg structures on orbifolds, adding a carefully worked low-dimensional example to the existing literature on geometric transitions. In particular, the closed orbifolds admitting Heisenberg structures are classified, and their deformation spaces are computed. Considering the regeneration problem, which Heisenberg tori arise as rescaled limits of collapsing paths of constant curvature cone tori is completely determined in the case of a single cone point.

57M50

1 Introduction

Heisenberg geometry is a geometry on the plane given by all translations together with shears parallel to a fixed line. Viewing this fixed line as "space" and any line intersecting it transversely as "time", this is the geometry of $1+1$–dimensional Galilean relativity.

Definition 1.1 Heisenberg geometry is the (G, X) geometry $\mathbb{H}\mathrm{s}^2 := (\mathsf{Heis}, \mathbb{A}^2)$, where

$$\mathsf{Heis} = \left\{ \begin{pmatrix} \pm 1 & a & c \\ 0 & \pm 1 & b \\ 0 & 0 & 1 \end{pmatrix} \middle|\, a, b, c \in \mathbb{R} \right\} \quad \text{and} \quad \mathbb{A}^2 = \{[x : y : 1] \in \mathbb{R}P^2 \mid x, y, \in \mathbb{R}\}.$$

The identity component $\mathsf{Heis}_0 < \mathsf{Heis}$ is the real Heisenberg group, and the index two subgroup of orientation-preserving transformations is denoted Heis_+.

The Heisenberg plane represents a particularly simple example of a non-Riemannian degeneration of Riemannian symmetric spaces via conjugacy limit, as studied by Cooper, Danciger and Wienhard [7], Danciger [9] and Fillastre and Seppi [13]. The semi-Riemannian geometries with automorphism groups $\mathsf{O}(p, q)$ and their degenerations

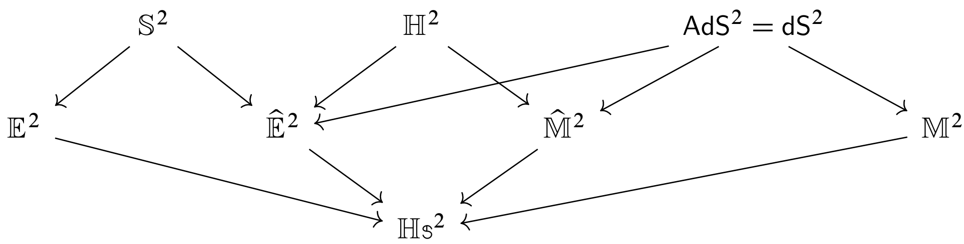


Figure 1: The poset of subgeometries of $\mathbb{R}P^2$ with automorphism groups $\mathsf{PO}(p,q)$ (spherical, hyperbolic and (anti-)de Sitter space) and their degenerations (adapted from [7]). The first degenerations are geometries of Euclidean and Minkowski space together with their contragredient dual representations ($\widehat{\mathbb{M}}^2$ is the *half-pipe* geometry of [9]). The Heisenberg plane is a degeneration of all of these.

form a poset[1] with a minimum element in each dimension [7]. This "most degenerate" geometry has the property that no nontrivial orthogonal group of any dimension appears as a subgroup of its automorphisms, and in dimension two is the Heisenberg plane.

We attempt to provide a detailed exploration of Heisenberg geometry to add to the literature describing explicit geometric transitions. We pay particular attention to aspects of interest to geometric topology, namely classifying Heisenberg orbifolds, calculating deformation their spaces and constructing regenerations of Heisenberg structures into familiar geometries. In order to lower the prerequisites, when some result for the Heisenberg plane is a consequence of more general geometric theorems we mention this, but attempt to also provide self-contained proofs when possible and succinct.

1.1 Heisenberg orbifolds

The first main result concerns the moduli problem for Heisenberg orbifolds. As a subgeometry of the affine plane, all Heisenberg orbifolds are finitely covered by a torus, so computing the deformation space $\mathcal{D}_{\mathbb{H}\mathbb{s}^2}(T^2)$ is the natural starting point. Geometric structures on tori generalize elliptic curves (the conformal structures), especially in the presence of a compatible group operation. As in the complete affine case studied by Baues and Goldman [3], each Heisenberg torus admits a group structure with Heisenberg maps realizing the group operation, which we explicitly describe. As a first

[1]The fact that spheres of increasing radius limit to their tangent plane can be used to produce a degeneration of spherical geometry to Euclidean showing that $\mathbb{E}^2 \prec \mathbb{S}^2$, for example.

step to determining these structures we compute the representation variety of potential holonomies:

Theorem 1.2 *The representation variety* $\mathrm{Hom}(\mathbb{Z}^2, \mathsf{Heis}_0)$ *is isomorphic, as a real algebraic variety, to the product* $V \times \mathbb{R}^2$*, where* V *is the* 3*–dimensional variety* $V = \{(x,y,z,w) \in \mathbb{R}^4 \mid xy = zw\}$*. Topologically, this is homeomorphic to the product of a plane with the cone on a torus.*

The Heisenberg plane admits no invariant Riemannian metric, so the possibility of incomplete structures must be taken seriously. In contrast to the affine case (see Baues [2] and Nagano and Yagi [19]) however, a geometric argument shows all Heisenberg structures are complete, and the deformation space $\mathcal{D}_{\mathbb{H}\mathbb{s}^2}(T^2)$ of tori identifies with the conjugacy classes of faithful representations $\mathbb{Z}^2 \to \mathsf{Heis}_+$ acting properly discontinuously on $\mathbb{R}^2$. The projection onto conjugacy classes admits a section allowing us to select a preferred holonomy (and construct the corresponding developing map) for each point in deformation space.

Theorem 1.3 *All Heisenberg tori are complete, and the projection onto holonomy* $\mathcal{D}_{\mathbb{H}\mathbb{s}^2}(T^2) \to \mathsf{Heis}(\mathbb{Z}^2, \mathsf{Heis}_+)/\mathsf{Heis}_+$ *is an embedding. The deformation space identifies with the classes of faithful representations acting properly discontinuously, and is homeomorphic to* $\mathbb{R}^3 \times \mathbb{S}^1$.

An explicit description of the deformation space of tori greatly simplifies the calculation of the remaining deformation spaces, which is relegated to the appendix. As all structures are complete, the problem of determining Heisenberg structures on an orbifold $\mathcal{O}$ finitely covered by T^2 is equivalent to the following algebraic *extension problem*: when does a representation $\rho\colon \pi_1(T)^2 \to \mathsf{Heis}$ extend to a representation of $\pi_1(\mathcal{O}) > \pi_1(T^2)$?

Theorem 1.4 *There are nine closed Heisenberg orbifolds, namely the quotients of the torus with at most order two cone points and right angled reflector corners. All Heisenberg orbifolds are complete, and the holonomy map*

$$\mathrm{hol}\colon \mathcal{D}_{\mathbb{H}\mathbb{s}^2}(\mathcal{O}) \to \mathrm{Hom}(\pi_1(\mathcal{O}), \mathsf{Heis})/\mathsf{Heis}_+$$

is an embedding.

1.2 Regenerating Heisenberg tori

Our second main result concerns the regeneration of Heisenberg structures to constant curvature ones, adding a detailed example to the collection of regenerations studied in

$\mathcal{O}$	$\mathcal{D}_{\mathbb{Hs}^2}(\mathcal{O})$
$\mathbb{S}^1 \times \mathbb{S}^1$	$\mathbb{R}^3 \times \mathbb{S}^1$
$\mathbb{S}^1 \tilde{\times} \mathbb{S}^1$, $\mathbb{S}^1 \times I$, $\mathbb{S}^1 \tilde{\times} I$	$\mathbb{R}^3 \sqcup \mathbb{R}^2$
$\mathbb{S}^2(2,2,2,2)$	$\mathbb{R}^2 \times \mathbb{S}^1$
$\mathbb{D}^2(2,2;\varnothing), \mathbb{D}^2(\varnothing;2,2,2,2)$	$\mathbb{R}^2 \sqcup \mathbb{R}^2$
$\mathbb{R}P^2(2,2), \mathbb{D}^2(2;2,2)$	$\mathbb{R}^2 \sqcup \mathbb{R}^2$

Table 1: The Heisenberg orbifolds and the homeomorphism type of their deformation spaces.

Danciger [11], Danciger, Guéritaud and Kassel [12], Heusener, Porti and Suárez [16], Leitner [18] and Porti [20]. Understanding the behavior of geometric structures along a transition is in general difficult, as one cannot directly use techniques from either geometry involved. Suitably constructing degenerations of $\mathbb{S}^2$, $\mathbb{E}^2$ and $\mathbb{H}^2$ to the Heisenberg plane within the projective plane allows us to use constructions in projective geometry to bridge the gap and overcome the additional difficulty posed by the lack of an invariant metric on $\mathbb{Hs}^2$.

As the Heisenberg plane is a common degeneration of the familiar constant curvature geometries, focusing on tori we ask when a given Heisenberg torus is the rescaled limit of a sequence of constant curvature cone manifold structures. Restricting to structures with at most one cone point, this has a clean resolution, illustrating a stark dichotomy between two "flavors" of Heisenberg tori: *translation tori* with holonomy images intersecting Heis_0 only in translations, and *shear tori* with holonomy images containing a nontrivial shear.

Theorem 1.5 *Let T be a Heisenberg torus, and $\mathbb{X} \in \{\mathbb{S}^2, \mathbb{E}^2, \mathbb{H}^2\}$. Then if $\mathbb{X}_t$ is a sequence of conjugate models of $\mathbb{X}$ limiting to the Heisenberg plane within $\mathbb{R}P^2$, there is a sequence of $\mathbb{X}_t$–cone tori T_t with a single cone point limiting to T if and only if T is a translation torus.*

A constructive argument for the "if" direction builds a fundamental domain $Q \subset \mathbb{R}^2$ for each translation torus $\mathbb{R}^2/(\mathbb{Z}\vec{v} \oplus \mathbb{Z}\vec{w})$, and a sequence of collapsing $\mathbb{X}$ cone tori such that under rescaling $\mathbb{X}$ degenerates to $\mathbb{Hs}^2$ and the rescaled fundamental domains converge to Q. This construction is analogous to the regeneration of Euclidean tori as hyperbolic cone tori. The "only if" direction follows from a geometric characterization of Heisenberg tori, relating shears in the image of the holonomy homomorphism to the distribution of simple geodesics on the surface.

Theorem 1.6 *A Heisenberg orbifold $\mathcal{O}$ has a nontrivial shear in its holonomy if and only if all simple geodesics on $\mathcal{O}$ are parallel.*

This provides a clear obstruction to regenerating shear tori. Any two simple geodesics on a shear torus are disjoint, but constant curvature cone tori with a single cone point have geodesic representatives of each homotopy class. In particular, any generating set for $H_1(T^2)$ can be pulled tight to give intersecting simple geodesics. An argument in projective geometry shows that any limit of $\mathbb{X} \in \{\mathbb{S}^2, \mathbb{H}^2, \mathbb{E}^2\}$–cone tori as $\mathbb{R}P^2$ structures inherits a collection of intersecting simple geodesics, finishing the proof.

Acknowledgements

I am immensely grateful to Darren Long, my advisor, and Daryl Cooper for many helpful discussions. I have learned a great deal from their helpful suggestions, and appreciate their patience as I worked on this manuscript. Thanks also to Gordon Kirby for listening to me develop these ideas, and much thanks to the anonymous reviewer whose detailed and constructive feedback substantially improved this paper.

2 Background

We list some terminology and notations used throughout the paper for quick reference. We denote by Heis_0 the real Heisenberg group of upper triangular unipotent 3×3 matrices, and by $\mathsf{Heis} = (\mathbb{Z}_2)^2 \rtimes \mathsf{Heis}_0$ the group generated by this together with reflections $\operatorname{diag}(\pm 1, \pm 1, 1)$. Heis_+ is the index two orientation-preserving subgroup of Heis, and Tr the subgroup acting by translations on the plane. The Lie algebra $\mathfrak{heis}$ consists of the strictly upper triangular 3×3 matrices, and provides useful coordinates for the representation varieties. For ease of inline typesetting we will often denote the element

$$\begin{pmatrix} 0 & x & z \\ 0 & 0 & y \\ 0 & 0 & 0 \end{pmatrix} \in \mathfrak{heis}$$

by the shorthand notation $\left(\begin{smallmatrix} x & z \\ & y \end{smallmatrix}\right)$.

We denote a closed two-dimensional orbifold $\mathcal{O}$ with underlying topological space X by $X(\vec{c})$ if $\mathcal{O}$ has cone points of order $\vec{c} = (c_1, \dots, c_m)$, and by $X(\vec{c}; \vec{r})$ if in addition $\partial X \neq \varnothing$ and $\mathcal{O}$ has corner reflectors of order $\vec{r} = (r_1, \dots, r_n)$.

The algebraic variety cut out by $f \in \mathbb{R}[x_1, \dots, x_n]$ is denoted by $V(f)$. A finite presentation for a group $\Gamma = \langle s_1, \dots, s_n \mid r_1, \dots, r_m \rangle$ gives an injection

$$\mathrm{ev}\colon \mathrm{Hom}(\Gamma, \mathsf{Heis}) \hookrightarrow \mathsf{Heis}^n$$

by evaluation on generators: $\mathrm{ev}(\rho) = (\rho(s_1), \dots, \rho(s_n))$. The image is an algebraic variety cut out by the polynomials $\{r_i - I\}$. Pulling this structure back via the evaluation map equips the set of homomorphisms with a variety structure which is independent of the original choice of presentation (see for example [14]).

2.1 Klein geometry

A *geometry* in the sense of Klein is a pair (G, X) consisting of a Lie group G acting analytically and transitively on a smooth manifold X. Examples of Kleinian geometries abound in geometric topology, from spherical geometry as in the sphere with an $\mathsf{SO}(3)$ action, to the hyperbolic plane as a disk in $\mathbb{C}$ together with the Möbius transformations preserving it, and even non-Riemannian examples such as projective space, $(\mathsf{SL}(n+1;\mathbb{R}), \mathbb{R}P^n)$. Consult [7; 15; 23] for additional reference and examples.

For convenience we often work with pointed geometries $(G, (X, x))$ selecting a particular point stabilizer $G_x = \mathrm{stab}_G(x)$. As G acts transitively, the particular choice of basepoint is immaterial and often notationally suppressed. A *morphism of geometries* $(G, X) \to (H, Y)$ is a pair (Φ, F) consisting of a group homomorphism $\Phi\colon G \to H$ with $\Phi(G_x) < H_y$, together with a Φ–equivariant smooth map $F\colon (X, x) \to (Y, y)$. A *subgeometry* of (G, X) is the image of a monomorphism $(H, Y) \hookrightarrow (G, X)$, namely a subset $Y \subset X$ together with a subgroup $H < G$ preserving and acting transitively on Y. An *open subgeometry* is a subgeometry with $Y \subset X$ open. One may alternatively build the theory of Klein geometries abstractly as pairs (G, G_x) of a Lie group and closed subgroup, recovering the space X as $X = G/G_x$ with basepoint G_x. This *automorphism-stabilizer* perspective is equivalent to the *group-space* definitions above, with the map $(G, (X, x)) \mapsto (G, G_x)$ defining an equivalence of categories.

A geometry is said to be *effective* if the only automorphism acting trivially is the identity. Its failure to be effective is measured by the intersection of all point stabilizers $K_G = \bigcap_{x \in X} \mathrm{stab}_G(x)$, and a geometry is *locally effective* if K_G is discrete. The assignment $E\colon (G, X) \mapsto (G/K_G, X)$ induces an equivalence of categories onto the subcategory of effective geometries; we say two geometries are *effectively equivalent* if their images under $E(\cdot)$ are isomorphic. As is commonplace, we switch between effectively equivalent geometries when convenient.

2.2 Geometric structures and collapse

A (G,X) structure on a manifold M is defined by a maximal atlas of X–valued charts on M with transition maps in G. The set of such structures is denoted by $\mathcal{S}_{(G,X)}(M)$. Pulling an atlas back to the universal cover and analytically continuing a chosen base chart provides an alternative definition via a *developing pair*: an immersion $f\colon \widetilde{M} \to X$ called the developing map, which is equivariant with respect to the holonomy homomorphism $\rho\colon \pi_1(M) \to G$. A (G,X) structure on M, only determines such a developing pair up to the action of G by $g\,.\,(f,\rho) = (g\,.\,f, g\rho g^{-1})$, identifying $\mathcal{S}_{(G,X)}(M)$ with the set of G orbits of developing pairs under this action. $\mathcal{S}_{(G,X)}(M)$ inherits the quotient topology from the space of developing pairs topologized by uniform convergence on compact sets. The *deformation space* of (G,X) structures $\mathcal{D}_{(G,X)}(M)$ is the result of further identifying isotopy classes of structures. More precisely, let $\mathrm{Diff}_0(M)$ denote the diffeomorphisms of M isotopic to the identity, and $\widetilde{\mathrm{Diff}}_0(M)$ their lifts to $\pi_1(M)$–equivariant maps $\widetilde{M} \to \widetilde{M}$. Then $\mathcal{D}_{(G,X)}(M)$ is the quotient of $\mathcal{S}_{(G,X)}(M)$ by the action of $\widetilde{\mathrm{Diff}}_0(M)$ by precomposition on the developing map factor. More detailed accounts of deformation space can be found in [2; 14; 15]. A subgeometry $(H,Y) < (G,X)$ induces a map $\mathcal{D}_{(H,Y)}(M) \to \mathcal{D}_{(G,X)}(M)$ by viewing an (H,Y) structure up to (G,X) equivalence, called *weakening*. Note that this map is rarely injective. For example, weakening Euclidean to affine structures collapses the entirety of $\mathcal{D}_{\mathbb{E}^2}(T^2)$ to a point. Dually, a developing pair for a (G,X) structure with holonomy image in (H,Y) can be *strengthened* to an (H,Y) structure by only considering equivalence up to H–conjugacy.

A sequence of geometric structures *degenerates* if the developing maps fail to converge to an immersion even after adjusting by diffeomorphisms of M and coordinate changes in G. Of particular interest are *collapsing degenerations*, with developing maps converging to a submersion into a lower-dimensional submanifold and holonomies limiting to a representation into the subgroup preserving this submanifold. A trivial example is given by the collapse of Euclidean manifolds under volume rescaling. Given a Euclidean structure (f,ρ) on a manifold M^n and any $r \in \mathbb{R}_+$, the developing pair $(rf, r\rho)$ describes the rescaled manifold with volume r^n times that of the original. As $r \to 0$ these structures collapse to a constant map and the trivial holonomy. More interesting examples include the collapse of hyperbolic structures onto a codimension-1 hyperbolic space as studied by Danciger [9; 11; 10] and the collapse of hyperbolic and spherical structures in [20; 22].

Collapsing geometric structures can often be "saved" by allowing more flexible coordinate changes. If a geometry (H,Y) can be realized as an open subgeometry of (G,X), then a sequence (f_n,ρ_n) of collapsing (H,Y) structures may actually converge *as* (G,X) *structures*, meaning there are $g_n\in G$ such that the developing pairs $g_n\,.\,(f_n,\rho_n)$ converge to a (G,X) developing pair (f_∞,ρ_∞). When f_∞ has image in an open subset $Z\subset X$, and ρ_∞ maps into the subgroup $L<G$ of Z–preserving transformations, this (G,X) developing pair strengthens to an (L,Z) structure. It is tempting to say that *within* (G,X) these (H,Y) structures converge to an (L,Z) structure. Formalizing this notion motivates the field of *transitional geometry*.

2.3 Geometric transitions

A *geometric transition* is a continuous path of geometries (H_t,Y_t), each isomorphic to a fixed geometry (H,Y), which converge to a geometry $(L,Z)\not\cong(H,Y)$. This is difficult to define in full generality, but here it suffices to formalize geometric transitions occurring *as subgeometries of a fixed ambient geometry*. Subgeometries (H,H_x) of (G,G_x) correspond directly to closed subgroups $H<G$ (with $H_x=H\cap G_x$), providing a natural topology on the space of subgeometries of (G,X). The hyperspace $\mathfrak{C}_G$ of closed subgroups of a compact Lie group G admits the Hausdorff metric, inducing a topology in which $\{Z_n\}$ converges to the set of all subsequential limits of sequences $\{z_n\}\in Z_n$. This generalizes to all Lie groups G by equipping $\mathfrak{C}_G$ with the topology of Hausdorff convergence on compact sets, otherwise known as the Chabauty topology [6].

Definition 2.1 Given a geometry $(G,(X,x))$, the space of open subgeometries $\mathfrak{S}_{(G,X)}$ is defined by

$$\mathfrak{S}_{(G,X)}=\{(H,H\cap G_x)\mid H<G \text{ and } \dim H-\dim(H\cap G_x)=\dim G-\dim G_x\}$$

equipped with the subspace topology from $\mathfrak{C}_G\times\mathfrak{C}_{G_x}$.

Definition 2.2 A *continuous path of subgeometries* of (G,X) is a continuous map $I\to\mathfrak{S}_{(G,X)}$. A geometry (L,Z) is a *degeneration of* (H,Y) *in* (G,X) if there is a continuous path $\gamma\colon[0,1]\to\mathfrak{S}_{(G,X)}$ with $\gamma(t)\cong(H,Y)$ for $t\neq 0$ and $\gamma(0)\cong(L,Z)$. A geometry (L,Z) is a *transitional geometry from* (H,Y) *to* (H',Y') *in* (G,X) if it is a degeneration of both (H,Y) and (H',Y').

The automorphisms G of the ambient geometry act on the space of subgeometries by $g\,.\,(H,Y)=(gHg^{-1},g\,.\,Y)$. A degeneration which occurs as the limit of a sequence

$g_t \,.\, (H, Y)$ for $g_t \in G$ is called a *conjugacy limit* of (H, Y) in (G, X). This provides the necessary background to formally consider the degeneration and regeneration of geometric structures.

Definition 2.3 Fix an ambient geometry (G, X) and a subgeometry (H, Y). Then a collapsing sequence of (H, Y) structures (f_t, ρ_t) on a manifold M *degenerates to an* (L, Z) *structure* if there is a path $g_t \in G$ with $g_t \,.\, (H, Y) \to (L, Z)$ such that $g_t \,.\, (f_t, \rho_t)$ converges as developing pairs. Dually, an (L, Z) structure on M is said to *regenerate into* (H, Y) if such a collapsing path of (H, Y) structures exists.

Danciger develops *half-pipe geometry* [9], as half-pipe structures are the limits of the aforementioned collapse of hyperbolic cone manifolds onto codimension-1 hyperbolic space, and together with Guéritaud and Kassel studies regenerations of AdS spacetimes from flat spacetimes [12]. Hodgson [17] and Porti [20] analyze Euclidean limits resulting from hyperbolic cone manifolds collapsing to a point, which plays an important role in the orbifold theorem of Cooper, Hodgson, and Kerckhoff [8] and Boileau, Leeb and Porti [5]. Further work of Porti studies the nonuniform collapse of hyperbolic structures and regenerations of Nil [21] and Sol [16], and the work of Ballas, Cooper and Leitner concerns the degeneration of cusps in projective space [1; 18].

2.4 An example: the spherical-to-hyperbolic transition

As a final installment of introductory material, we introduce models of the constant curvature geometries $\mathbb{S}^2$, $\mathbb{E}^2$ and $\mathbb{H}^2$ as subgeometries of projective space, and then construct a geometric transition from spherical to hyperbolic space via conjugacy limit.

Definition 2.4 As subgeometries of projective space, the constant curvature geometries are realized by the following three models:

- $\mathbb{S}^2 = (\mathsf{SO}(3), \mathbb{R}P^2)$. This twofold quotient of the unit sphere is often called the *elliptic plane* in older literature.
- $\mathbb{E}^2 = (\mathsf{Euc}(2), \mathbb{A}^2)$ with $\mathsf{Euc}(2) = \left(\begin{smallmatrix} \mathsf{SO}(2) & \mathbb{R}^2 \\ 0 & 1 \end{smallmatrix}\right)$, the Euclidean group acting transitively on the affine patch $\mathbb{A}^2 = \{[x : y : 1]\} \subset \mathbb{R}P^2$.
- $\mathbb{H}^2 = (\mathsf{SO}(2,1), \mathbb{D}^2)$ with $\mathbb{D}^2 = \{[x : y : 1] \mid x^2 + y^2 < 1\}$, the unit disk in the affine patch $\mathbb{A}^2$.

Note that the projective point $p = [0 : 0 : 1]$ lies in each of the above models, and the stabilizing subgroup of p is equal in all three geometries to $S = \left(\begin{smallmatrix} \mathsf{SO}(2) & 0 \\ 0 & 1 \end{smallmatrix}\right)$.

The underlying spaces of these geometries will often be denoted by $\mathbb{S}^2$, $\mathbb{E}^2$ and $\mathbb{H}^2$ as well to remind us of the inherent geometric structure. On the level of curvature one can easily imagine producing a transition from (a small patch of) spherical space to (a small patch of) hyperbolic space through Euclidean geometry by appropriately varying the Riemannian metric. Below we give an example realizing this transition as a *conjugacy limit* connecting the three specific models above within an ambient copy of $\mathbb{R}P^2$.

From the group stabilizer perspective, the models above are given by the points $\mathbb{S}^2 = (\mathsf{SO}(3), S)$, $\mathbb{E}^2 = (\mathsf{Euc}(2), S)$ and $\mathbb{H}^2 = (\mathsf{SO}(2,1), S)$ in the space $\mathfrak{S}_{\mathbb{R}P^2}$ of subgeometries of the projective plane. Let $C_t = \operatorname{diag}(1,1,t)$ and define the path $\gamma\colon[-1,1]\to\mathfrak{C}_{\mathsf{GL}(3;\mathbb{R})}$:

$$\gamma(t) = \begin{cases} C_t\,.\,(\mathsf{SO}(2,1), S) & \text{if } t<0,\\ (\mathsf{Euc}(2), S) & \text{if } t=0,\\ C_t\,.\,(\mathsf{SO}(3), S) & \text{if } t>0.\end{cases}$$

The point stabilizer subgroup S is invariant under C_t conjugacy, thus checking the continuity of γ reduces to considering the limits of $C_t.\mathsf{SO}(3)$ and $C_t.\mathsf{SO}(2,1)$ in $\mathfrak{C}_{\mathsf{GL}(3;\mathbb{R})}$. The fact that each of these paths has limit $\mathsf{Euc}(2)$ as $t\to 0$ is a straightforward computation in the Lie algebra, a reduction which is justified by [7, Proposition 3.1] as both are conjugacy limits of algebraic groups. Thus γ realizes a continuous transition as subgeometries of $\mathbb{R}P^2$ from $\gamma(-1)=\mathbb{H}^2$ to $\gamma(1)=\mathbb{S}^2$ through $\gamma(0)=\mathbb{E}^2$.

3 Heisenberg geometry

The Heisenberg plane is not a metric geometry but supports other familiar geometric quantities. The standard area form $dA = dx\wedge dy$ on $\mathbb{R}^2$ is invariant under the action of Heis_+, furnishing $\mathbb{H}\mathbb{s}^2$ with a well defined notion of area. The 1–form dy is Heis_0 invariant, and induces a Heis–invariant foliation of $\mathbb{H}\mathbb{s}^2$ by horizontal lines together with a transverse measure. As a subgeometry of the affine plane, $\mathbb{H}\mathbb{s}^2$ inherits an affine connection and notion of geodesic. A curve γ is a geodesic if $\gamma''=0$, tracing out a constant speed straight line in $\mathbb{H}\mathbb{s}^2$.

Heisenberg geometry arises as a limit of the constant curvature spaces $\mathbb{S}^2$, $\mathbb{H}^2$ and $\mathbb{E}^2$ by "zooming into while unequally stretching" a projective model. Details can be reconstructed from [7], and the precise characterization is reviewed in Section 4. Here we briefly explore one degeneration of hyperbolic space to the Heisenberg plane as subgeometries of $\mathbb{R}P^2$. Acting on $\mathbb{H}^2\in\mathfrak{S}_{\mathbb{R}P^2}$ by the path $A_t = \operatorname{diag}(t^2, t, 1)$ results in a path of subgeometries $A_t\mathbb{H}^2$ isomorphic to the hyperbolic plane with underlying space the origin-centered ellipsoid in $\mathbb{A}^2$ with semimajor and semiminor axes of lengths t^2 and t parallel to the x and y axes, respectively. As t tends to infinity, the limit of these domains is $\mathbb{A}^2$ and the groups $A_t\mathsf{O}(2,1)A_t^{-1}$ limit to Heis. The aforementioned invariant foliation on $\mathbb{H}\mathbb{s}^2$ is a remnant of this stretching, and is parallel to the limiting direction of the major axes of $A_t\mathbb{H}^2$.

Unlike the degeneration of $\mathbb{S}^2$ and $\mathbb{H}^2$ to Euclidean space, the uneven stretching required to produce a Heisenberg limit distorts even the point stabilizer subgroups, which become noncompact in the limit. Conjugation by A_t stretches the circle

$$S = \begin{pmatrix} \mathsf{SO}(2) & 0 \\ 0 & 1 \end{pmatrix} \subset \mathsf{M}(3;\mathbb{R})$$

into ellipses of increasing eccentricity limiting to the parallel lines $\left(\begin{smallmatrix} 1 & \pm x \\ 0 & 1 \end{smallmatrix}\right)$ in the upper 2×2 block. As a consequence, the role of the unit tangent bundle in the constant curvature geometries is replaced for the Heisenberg plane by an appropriate space of based lines. Indeed let $\mathcal{L} = \mathbb{P}\mathsf{T}(\mathbb{H}\mathbb{s}^2)$ be the space of pointed lines in the Heisenberg plane, and $\mathcal{H} \subset \mathcal{L}$ those belonging to the invariant horizontal foliation. The action of Heis_0 on the plane extends to a simple transitive action on $\mathcal{L} \smallsetminus \mathcal{H}$, analogous to the action of $\mathrm{Isom}(\mathbb{X})$ on the unit tangent bundle $\mathsf{UT}(\mathbb{X})$ for $\mathbb{X} \in \{\mathbb{H}^2, \mathbb{E}^2, \mathbb{S}^2\}$. The noncompactness of point stabilizers is sufficient to preclude an invariant Riemannian metric, but moreover the existence of shears in the automorphism group of Heis forces any continuous Heis–invariant map $d\colon \mathbb{R}^2\times\mathbb{R}^2 \to \mathbb{R}$ to be constant along the lines $\{x\}\times\mathbb{R}$ in both factors of the domain, so there are no continuous Heis–invariant distance functions at all.

3.1 Heisenberg structures on orbifolds

As a subgeometry of the affine plane, every Heisenberg structure on an orbifold $\mathcal{O}$ canonically weakens to an affine structure. This provides strong restrictions on which orbifolds can possibly admit Heisenberg structures. It follows from a result of Benzécri that closed affine orbifolds have Euler characteristic zero [4]; an additional self contained proof appears in [2]. The deformation space of affine tori has been computed [2], and weakening Heisenberg structures to affine structures provides a (noninjective) map $\omega\colon \mathcal{D}_{\mathbb{H}\mathbb{s}^2}(T^2) \to \mathcal{D}_{\mathbb{A}^2}(T^2)$. Each Heisenberg orbifold inherits an area form from $\mathbb{H}\mathbb{s}^2$ and has a well defined finite total area. The group $\mathbb{R}_+$ of homotheties of the plane acts on $\mathcal{D}_{\mathbb{H}\mathbb{s}^2}(\mathcal{O})$, sending an orbifold $\mathcal{O}$ with total area α to an orbifold $r\,.\,\mathcal{O}$ with area $r^2\alpha$, allowing the deformation space to be easily recovered from the space of unit area structures.

Observation The action of $\mathbb{R}_+$ by homotheties on the plane induces an action on $\mathcal{D}_{\mathbb{H}\mathrm{s}^2}(\mathcal{O})$ defined by $r.[f,\rho]=[rf,r\rho]$. This gives a homeomorphism

$$\mathcal{D}_{\mathbb{H}\mathrm{s}^2}(\mathcal{O})=\mathbb{R}_+\times\mathcal{T}_{\mathbb{H}\mathrm{s}^2}(\mathcal{O})$$

for $\mathcal{T}_{\mathbb{H}\mathrm{s}^2}(\mathcal{O})$, the subspace of unit area structures, analogous to the Teichmüller space for Euclidean tori.

As dy is invariant under the action of Heis_0, any Heisenberg surface with holonomy into Heis_0 inherits a closed nondegenerate 1–form and corresponding foliation. This observation leads to a self-contained proof that every Heisenberg orbifold has vanishing Euler characteristic, simple enough that we include it for completeness.

Proposition 3.1 *Every closed Heisenberg orbifold is finitely covered by a torus with holonomy in* Heis_0*.*

Proof Let $\mathcal{O}$ be a Heisenberg orbifold, with developing map $f\colon\widetilde{\mathcal{O}}\to\mathbb{H}\mathrm{s}^2$ and holonomy $\rho\colon\pi_1(\mathcal{O})\to\mathsf{Heis}$. As f immerses $\widetilde{\mathcal{O}}$ in the plane it has no singular locus, thus $\widetilde{\mathcal{O}}$ is a manifold and $\mathcal{O}$ is good. Then by the classification of two-dimensional orbifolds $\mathcal{O}$ is not the spindle or teardrop, and is finitely covered by some surface $\Sigma\to\mathcal{O}$. The Heisenberg structure on $\mathcal{O}$ pulls back to Σ with developing pair $(f,\rho|_{\pi_1(\Sigma)})$. Passing to an at most 4–sheeted cover, we may assume the holonomy of Σ takes values in Heis_0. Thus Σ inherits a nondegenerate 1–form $\omega\in\Omega^1(\Sigma)$ from dy on $\mathbb{H}\mathrm{s}^2$. Choose a Riemannian metric g on Σ. Then ω defines a nonvanishing vector field X_ω by $\omega(\cdot)=g(X_\omega,\cdot)$, and so $\chi(\Sigma)=0$. As Heis_0 acts by orientation-preserving transformations, Σ is a torus. $\square$

Thus Heisenberg tori with holonomy in Heis_0 play a fundamental role in the classification of Heisenberg orbifolds, and it is natural to study them first. By the previous observation, in particular it suffices to study the Teichmüller space of unit area structures, whose holonomy are determined up to conjugacy and homotheties of the plane.

3.2 Representations of $\mathbb{Z}^2$ into Heis

To classify tori with holonomy into Heis_0, we compute the representation variety $\mathcal{R}=\mathrm{Hom}(\mathbb{Z}^2,\mathsf{Heis}_0)$. The quotients of $\mathcal{R}$ by homothety and Heisenberg conjugacy are denoted by $\mathcal{H}=\mathcal{R}/\mathbb{R}_+$ and $\mathcal{X}=\mathcal{R}/\mathsf{Heis}_0$, respectively. The holonomies of unit area structures lie in the double quotient $\mathcal{U}=\mathcal{X}/\mathbb{R}_+\cong\mathcal{H}/\mathsf{Heis}_0$. Representations into the center of Heis_0 act by collinear translations on $\mathbb{H}\mathbb{s}^2$, and a simple argument of Section 3.3 precludes these from being the holonomy of any Heisenberg structure. Thus, we are primarily concerned with the subset $\mathcal{R}^\star \subset \mathcal{R}$ of representations not into the center, and its quotients $\mathcal{X}^\star \subset \mathcal{X}$, $\mathcal{H}^\star \subset \mathcal{H}$ and $\mathcal{U}^\star \subset \mathcal{U}$. Explicitly dealing with these representation spaces is easiest using coordinates from the Lie algebra, introduced below.

Proposition 3.2 *The map* $\log\colon \mathsf{Heis}_0 \to \mathfrak{heis}$ *induces an isomorphism of varieties* $\mathrm{Hom}(\mathbb{Z}^2, \mathsf{Heis}_0) \cong \mathrm{Hom}(\mathbb{R}^2, \mathfrak{heis})$.

Proof Both Heis_0 and $\mathfrak{heis}$ inherit their structure as algebraic varieties from their inclusion in the affine space $\mathsf{M}(3,\mathbb{R})$ of 3×3 real matrices. As $\mathfrak{heis}$ is nilpotent, the power series $\exp\colon \mathfrak{heis} \to \mathsf{Heis}_0$ terminates, and thus is algebraic. Indeed, exp is an isomorphism of varieties with polynomial inverse $\log\colon \mathsf{Heis}_0 \to \mathfrak{heis}$. Recall that evaluation on the generators $e_1, e_2 \in \mathbb{Z}^2 \subset \mathbb{R}^2$ identifies the collections of representations with subvarieties of $\mathsf{Heis}_0 \times \mathsf{Heis}_0$ and $\mathfrak{heis}\times\mathfrak{heis}$, respectively. Applying the exponential/logarithm coordinatewise provides the required algebraic isomorphism $\mathrm{Hom}(\mathbb{Z}^2, \mathsf{Heis}_0) \cong \mathrm{Hom}(\mathbb{R}^2, \mathfrak{heis})$.

$$\begin{tikzcd}
\mathrm{Hom}(\mathbb{Z}^2, \mathsf{Heis}_0) \arrow[r, "\mathrm{ev}"] \arrow[d, shift left, "\log"] & \mathsf{Heis}_0 \times \mathsf{Heis}_0 \arrow[d, shift left, "\log\times\log"] \\
\mathrm{Hom}(\mathbb{R}^2, \mathfrak{heis}) \arrow[r, "\mathrm{ev}"] \arrow[u, shift left, "\exp"] & \mathfrak{heis}\times\mathfrak{heis} \arrow[u, shift left, "\exp\times\exp"]
\end{tikzcd}$$ □

We continue to denote the induced isomorphisms $\mathcal{R} \cong \mathrm{Hom}(\mathbb{R}^2, \mathfrak{heis})$ by exp and log, and call the vector $(\vec{x}, \vec{y}, \vec{z}) \in \mathbb{R}^6$ the *Lie algebra coordinates* for the representation $\rho \in \mathcal{R}$ when $\mathrm{ev}(\log\rho) = \left(\begin{pmatrix} x_1 & z_1 \\ & y_1\end{pmatrix}, \begin{pmatrix} x_2 & z_2 \\ & y_2\end{pmatrix}\right)$.

Proposition 3.3 $\mathcal{R}$ *is isomorphic to* $V(x_1y_2 - x_2y_1) \times \mathbb{R}^2$.

Proof Evaluation on the generators identifies the representation variety $\mathrm{Hom}(\mathbb{R}^2, \mathfrak{heis})$ with the kernel of the Lie bracket $[\cdot,\cdot]\colon \mathfrak{heis}^2 \to \mathfrak{heis}$. Indeed

$$\left[\begin{pmatrix} x_1 & z_1 \\ & y_1\end{pmatrix}, \begin{pmatrix} x_2 & z_2 \\ & y_2\end{pmatrix}\right] = \begin{pmatrix} 0 & x_1y_2 - x_2y_1 \\ & 0\end{pmatrix},$$

so $\ker[\cdot,\cdot]$ is cut out precisely by $x_1y_2 = x_2y_1$ in $\mathfrak{heis}^2$ and $(\vec{x}, \vec{y}, \vec{z}) \in \mathbb{R}^6$ is the Lie algebra coordinates of a representation $\rho \in \mathcal{R}$ if and only if $(\vec{x}, \vec{y}) \in V(x_1y_2 - x_2y_1)$ and $(z_1, z_2) \in \mathbb{R}^2$. □

Proposition 3.4 *The space $\mathcal{H}^\star = \mathcal{R}^\star/\mathbb{R}_+$ of representations modulo homothety with image not contained in the center of* Heis *is homeomorphic to $\mathbb{R}^2 \times T^2$.*

Proof Denote by $\mathbb{R}^2_{(\vec{x},\vec{y})}$ the fiber above $(\vec{x},\vec{y})$ under the projection $(\vec{x},\vec{y},\vec{z}) \mapsto (\vec{x},\vec{y})$. The hypersurface $V = V(x_1y_2 - x_2y_1)$ has one singularity at zero, above which $\mathbb{R}^2_{(0,0)}$ consists of the representations into the center. Homotheties of $\mathbb{H}\mathrm{s}^2$ induce the $\mathbb{R}_+$ action $t\,.\,(\vec{x},\vec{y},\vec{z}) = (t\vec{x}, t\vec{y}, t\vec{z})$ on $\mathcal{R}$, thus $V \subset \mathbb{R}^4$ is a cone and $\mathcal{H}^\star$ identifies with the product of $\mathbb{R}^2$ with the intersection $V \cap \mathbb{S}^3$. The change of coordinates on $\mathbb{R}^4$ given by $(x_1, x_2, y_1, y_2) = (u_1 + v_1, v_2 + u_2, v_2 - u_2, u_1 - v_1)$ provides an isomorphism $V \cong V(u_1^2 + u_2^2 - v_1^2 - v_2^2)$ identifying $V \cap \mathbb{S}^3$ with the Clifford torus $T = \{(u,v) \in \mathbb{C}^2 : \|\vec{u}\| = \|\vec{v}\| = 1/\sqrt{2}\}$, verifying the claim. □

Corollary 3.5 *The section of $\mathcal{R}^\star \to \mathcal{H}^\star$ sending each homothety class*

$$[\rho]_{\mathbb{R}_+} = [(\vec{x},\vec{y},\vec{z})]_{\mathbb{R}_+}$$

to the representative with $(\vec{x},\vec{y}) \in T^2 \subset \mathbb{S}^3$ is a diffeomorphism of $\mathcal{H}^\star$ onto its image. This identifies $\mathcal{H}^\star$ with the algebraic variety $V(x_2y_1 - x_1y_2, \|x\|^2 + \|y\|^2 - 1) \subset \mathbb{R}^6$.

We have identified the space $\mathcal{R}$ of all representations as a product $\mathbb{R}^2 \times V$ of a plane with a cone on the torus, with representations into the center parametrized by the plane above the cone point of V. Restricting to representations *not* into the center, it proves useful to remove this cone point and consider the space $V \smallsetminus \{0\} \cong \mathbb{R}_+ \times T^2$, which we denote by $V^\star$ to remain consistent with other notations.

Proposition 3.6 *Let $\mathcal{X}^\star$ be the conjugacy quotient $\mathcal{X}^\star = \mathcal{R}^\star/\mathrm{Heis}_0$. Then the function $\pi : \mathcal{X}^\star \to V^\star$ defined by sending the Heis_0 orbit of $\rho = (\vec{x},\vec{y},\vec{z}) \in \mathcal{R}^\star$ to $(\vec{x},\vec{y}) \in V^\star$ equips $\mathcal{X}^\star$ with the structure of a line bundle over $V^\star$. Topologically we can identify this line bundle up to isomorphism by noting that it is once-twisted above each generator of $\pi_1(V^\star) = \mathbb{Z}^2$.*

Proof A computation reveals the conjugation action of Heis_0 on $\mathcal{R}$ in Lie algebra coordinates is expressed as

$$\begin{pmatrix} 1 & g & k \\ & 1 & h \\ & & 1 \end{pmatrix} . (\vec{x},\vec{y},\vec{z}) = (\vec{x},\vec{y},\vec{z} + g\vec{y} - h\vec{x}).$$

Thus Heis_0 acts trivially on the first factor of $\mathcal{R} = V \times \mathbb{R}^2$ and the orbit of a point $\vec{z} \in \mathbb{R}^2_{(\vec{x},\vec{y})}$ is the coset of $\mathrm{span}\{\vec{x},\vec{y}\} \subset \mathbb{R}^2_{(\vec{x},\vec{y})}$ containing it. In the subset $\mathcal{R}^\star$ at least one of $\vec{x}$ or $\vec{y}$ is nonzero, and the condition that $(\vec{x},\vec{y}) \in V(x_1y_2 - x_2y_1) = V\big(\det\big(\begin{smallmatrix} x_1 & y_1 \\ x_2 & y_2 \end{smallmatrix}\big)\big)$

implies $\vec{x}$ and $\vec{y}$ are linearly dependent. It follows that the Heis_0 orbits on $\mathcal{R}^\star$ are lines, foliating each $\mathbb{R}^2_{(\vec{x},\vec{y})}$ over $V^\star$, and the leaf space is a line bundle over $V^\star$.

Equipping each $\mathbb{R}^2_{(\vec{x},\vec{y})}$ with the standard Euclidean inner product, the orthogonal line $\ell^\perp_{(\vec{x},\vec{y})} \subset \mathbb{R}^2_{(\vec{x},\vec{y})}$ gives canonical representatives for cosets of $\ell_{(\vec{x},\vec{y})} = \operatorname{span}\{\vec{x},\vec{y}\}$. This defines a section $\mathcal{X}^\star \to \mathcal{R}^\star$ sending a conjugacy class $[\rho]_{\mathsf{Heis}_0} = [(\vec{x},\vec{y},\vec{z})]_{\mathsf{Heis}_0}$ to its representation with $\vec{z}$–coordinate on $\ell^\perp_{(\vec{x},\vec{y})}$, and identifies

$$\mathcal{X}^\star = \{(\vec{x},\vec{y},\vec{z}) \mid (\vec{x},\vec{y}) \in V^\star, \vec{z} \in \ell^\perp_{(\vec{x},\vec{y})}\}$$

with a subbundle of $V^\star \times \mathbb{R}^2 \to V^\star$.

Line bundles over $V^\star \cong \mathbb{R}_+ \times T^2$ are in bijection with $H^1(T^2,\mathbb{Z}_2) \cong \mathbb{Z}_2^2$, determined up to isomorphism by whether pulling back along generators of $\pi_1(T)^2$ gives cylinders or Möbius bands. A convenient choice of generators in the $(\vec{u},\vec{v})$ coordinates introduced above is $\alpha(\theta) = (\vec{e_1},\vec{p_\theta})$ and $\beta(\theta) = (\vec{p_\theta},\vec{e_1})$ for $e_1 = \left(\begin{smallmatrix}1\\0\end{smallmatrix}\right)$ and $\vec{p_\theta} = \left(\begin{smallmatrix}\cos\theta\\ \sin\theta\end{smallmatrix}\right)$. An explicit computation using the description of $\mathcal{X}^\star$ above shows the bundle restricts to a Möbius band above each of α and β, so $\mathcal{X}^\star$ is the line bundle over $\mathbb{R}_+ \times T^2$ represented by $(1,1) \in H^1(T^2,\mathbb{Z}_2)$. $\square$

The choice of explicit sections has identified $\mathcal{H}^\star$ and $\mathcal{X}^\star$ with subsets of $\mathcal{R}$. The space of interest $\mathcal{U}^\star$ identifies with their intersection, $\mathcal{X}^\star \cap \mathcal{H}^\star$, which is the restriction of $\mathcal{X}^\star \to V^\star$ to the base $T^2 \subset \mathbb{S}^3$.

Corollary 3.7 *Let $\mathcal{U}^\star$ denote the quotient of $\mathcal{R}^\star$ by homothety and conjugacy (equivalently, the quotient of $\mathcal{X}^\star$ by homothety). Then the map $\mathcal{U}^\star \to T^2$ defined by sending the orbit of $\rho = (\vec{x},\vec{y},\vec{z})$ to $(\vec{x}/\|\vec{x}\|,\vec{y}/\|\vec{y}\|) \in V \cap \mathbb{S}^3 \cong T^2$ equips $\mathcal{U}^\star$ with the structure of a line bundle over the torus. We may realize $\mathcal{U}^\star$ explicitly as the subvariety of $\mathcal{U}^\star \subset \mathbb{R}^6$ consisting of triples of vectors $(\vec{x},\vec{y},\vec{z})$ such that $\vec{x}$ and $\vec{y}$ are collinear, and $\vec{z}$ is orthogonal to their span:*

$$\mathcal{U}^\star = V\begin{pmatrix}\|x\|^2 + \|y\|^2 = 1, & \vec{z}\cdot\vec{x} = 0\\ x_1y_2 - x_2y_1 = 0, & \vec{z}\cdot\vec{y} = 0\end{pmatrix} \subset \mathbb{R}^6.$$

As with $\mathcal{X}^\star$, we may characterize the bundle $\mathcal{U}^\star \to T^2$ topologically by noting that its restriction to each standard generator of T^2 is a Möbius band.

The developing pair of a Heisenberg torus is only well defined up to orientation-preserving transformations, so potential holonomies lie in the space $\mathcal{R}/\mathsf{Heis}_+$, a twofold quotient of $\mathcal{U}^\star$ computed here. We will deal with this $\mathbb{Z}_2 = \mathsf{Heis}_+/\mathsf{Heis}_0$ ambiguity after determining which points of $\mathcal{U}^\star$ are in fact holonomies.

3.3 The deformation space of tori

As a warm-up to computing the deformation space of Heisenberg tori, we review the analogous problem for both Euclidean and affine structures. Euclidean tori are complete metric spaces, and so are determined by their holonomy, which is necessarily discrete and faithful (for instance, by Thurston [23, Proposition 3.4.10]). Discrete subgroups $\mathbb{Z}^2 < \mathrm{Isom}(\mathbb{E}^2)$ act by translations, and thus the deformation space of Euclidean tori identifies with the $\mathrm{Isom}(\mathbb{E}^2)$–conjugacy classes of marked planar lattices, $\mathcal{D}_{\mathbb{E}^2}(T^2) \cong \mathsf{GL}(2;\mathbb{R})/\mathsf{O}(2)$. The unit area structures are parametrized by the familiar Teichmüller space $\mathbb{H}^2 = \mathsf{SL}(2;\mathbb{R})/\mathsf{SO}(2)$.

The affine plane admits no invariant metric, which complicates the story significantly. Complete affine structures have universal cover affinely diffeomorphic to $\mathbb{A}^2$, but in contrast to the Euclidean case incomplete structures abound. The work of Baues [2] provides a remarkably comprehensive description of the classification of affine tori, in particular containing the following classification theorem:

Theorem 3.8 [2, Theorem 5.1] *The universal cover of an affine torus is affinely diffeomorphic to one of the following spaces: the affine plane $\mathbb{A}^2$, the half plane $\mathcal{H} = \{(x,y) \mid y > 0\}$, the quarter plane $\mathcal{Q} = \{(x,y) \in \mathbb{A}^2 \mid x, y > 0\}$ or the universal cover of the punctured plane $\mathcal{P} = \widetilde{\mathbb{A}^2 \smallsetminus 0}$. Furthermore the developing maps of affine structures are covering projections onto their images.*

As $\mathbb{H}\mathrm{s}^2$ admits no invariant metric, we must be prepared for complications similar to the affine case. Such difficulties do not materialize however, as canonically weakening Heisenberg structures to affine ones, we may use the classification above to show all Heisenberg tori are complete.

Corollary 3.9 *All Heisenberg structures on the torus are complete.*

Proof Let (f,ρ) be the developing pair for a Heisenberg torus T, considered as an affine structure. If T is not complete, there is an affine transformation A with $A\,.\,f(\widetilde{T}) \in \{\mathcal{H}, \mathcal{Q}, \mathbb{A}^2 \smallsetminus 0\}$ and holonomy $A\rho A^{-1}$ preserving this developing image. But by the classification of affine tori, holonomies of these tori contain elements of $\det \neq 1$, whereas Heis is unipotent so $\det A\rho(\mathbb{Z}^2)A^{-1} = \{1\}$. Thus T is in fact complete, with developing map a diffeomorphism $f\colon \widetilde{T} \to \mathbb{A}^2$. □

Here we pursue a self-contained computation of the deformation space $\mathcal{D}_{\mathbb{H}\mathrm{s}^2}(T^2)$, using the understanding of representations $\mathbb{Z}^2 \to \mathsf{Heis}_0$ up to conjugacy developed in

Section 3.1. Specifically, for $\rho \in \mathrm{Hom}(\mathbb{Z}^2, \mathsf{Heis})$ we either construct a corresponding developing map f giving a Heisenberg structure (f, ρ) on T^2 (and prove its uniqueness), or we show no developing map for ρ can exist.

A developing map for $\rho\colon \mathbb{Z}^2 \to \mathsf{Heis}$ is a ρ–equivariant immersion $f\colon \mathbb{R}^2 \to \mathbb{H}\mathrm{s}^2$. A natural ρ–equivariant self map of the plane can be constructed directly from ρ, relying on the fact that each representation of $\mathbb{Z}^2$ extends uniquely to a representation $\hat{\rho}\colon \mathbb{R}^2 \to \mathsf{Heis}_0$ via $\hat{\rho}(x, y) = \rho(e_1)^x \rho(e_2)^y$. The orbit map $f_\rho\colon \mathbb{R}^2 \to \mathbb{H}\mathrm{s}^2$ defined by $(x, y) \mapsto \hat{\rho}(x, y) \,.\, \vec{0}$ for this extended representation is ρ–equivariant, and thus a developing map for a Heisenberg structure when it is an immersion. As the following two propositions show, this construction actually produces developing maps for all complete Heisenberg tori (and thus by Corollary 3.9 for all Heisenberg tori, although with the aim of producing a self-contained proof we do not presume that here).

Proposition 3.10 *Let $\mathcal{F} \subset \mathcal{U}$ be the subset of representations ρ with extensions $\hat{\rho}$ acting freely on $\mathbb{H}\mathrm{s}^2$. Then each $\rho \in \mathcal{F}$ determines a unique Heisenberg structure on T^2 which is complete, and all complete structures with holonomy in Heis_0 arise this way.*

Proof If $\hat{\rho}$ acts freely, the orbit map $f_\rho\colon \mathbb{R}^2 \to \mathbb{H}\mathrm{s}^2$ is injective, and a computation reveals $(df_\rho)_0\colon T_0\mathbb{R}^2 \to T_0\mathbb{H}\mathrm{s}^2$ is injective. Furthermore $(df_\rho)_x = \hat{\rho}(x) \,.\, (df_\rho)_0$, so f_ρ is an immersion of $\mathbb{R}^2$ and (f_ρ, ρ) is a developing pair for a Heisenberg torus. Similarly, the other orbit maps $\vec{u} \mapsto \hat{\rho}(\vec{u}) \,.\, q$ are immersions (thus open maps) for any $q \in \mathbb{H}\mathrm{s}^2$, and distinct $\hat{\rho}(\mathbb{R}^2)$ orbits partition $\mathbb{H}\mathrm{s}^2$ into a disjoint union of open sets. Then by connectedness f_ρ is onto, hence a diffeomorphism, so the corresponding Heisenberg structure is complete.

Alternatively, let $\rho\colon \mathbb{Z}^2 \to \mathsf{Heis}_0$ be the holonomy of a complete torus, but assume $\hat{\rho}\colon \mathbb{R}^2 \to \mathsf{Heis}_0$ fails to act freely. Then some element, and hence some 1–parameter subgroup $L < \mathbb{R}^2$, fixes a point under the action induced by $\hat{\rho}$. This line L intersects $\mathbb{Z}^2$ only in $\vec{0}$ (as ρ acts freely by completeness), and so is dense in the quotient $\mathbb{R}^2/\mathbb{Z}^2$. Thus there are sequences $\vec{v}_n \in \mathbb{Z}^2$ with $\rho(v_n)$ coming arbitrarily close to stabilizing a point, and $\hat{\rho}$ does not act properly discontinuously, contradicting completeness.

Finally, let (f, ρ) be a complete structure and (ϕ, ρ) another structure with the same holonomy. Then $f^{-1}\phi\colon \widetilde{T} \to \widetilde{T}$ is $\pi_1(T)$–equivariant and descends to a diffeomorphism $\psi\colon T \to T$. But ψ_* is the identity on fundamental groups, and as the torus is $K(\pi, 1)$, ψ is isotopic to the identity. Thus (f, ρ) and (ϕ, ρ) are developing pairs for the same Heisenberg structure. □

Constructing developing maps from the extensions $\hat{\rho}$ provides endows these tori with the structure of a commutative group via the identification $\hat{\rho}(\mathbb{R}^2)/\rho(\mathbb{Z}^2) \cong f_\rho(\mathbb{R}^2)/\rho(\mathbb{Z}^2)$. The existence of this group structure can more generally be deduced from the similar observation of Baues and Goldman concerning affine structures [3].

Corollary 3.11 *Complete Heisenberg tori are the group objects in the category of Heisenberg manifolds, analogous to elliptic curves in the category of Riemann surfaces.*

Proposition 3.12 *The subset $\mathcal{F} \subset \mathcal{U}$ of conjugacy classes with freely acting extensions $\hat{\rho}\colon \mathbb{R}^2 \to \mathsf{Heis}_0$ is a trivial $\mathbb{R}^\times$ bundle over the cylinder $\mathsf{Cyl} = T^2 \smallsetminus S$ for S, the circle defined by the intersection of $T^2 = V(x_1y_2 - x_2y_1) \cap \mathbb{S}^3$ with the plane $V(y_1, y_2)$.*

Proof A representation $\hat{\rho} \in \mathcal{U}$ is faithful if and only if the logarithm of its generators $\begin{pmatrix} x_1 & z_1 \\ & y_1 \end{pmatrix}$ and $\begin{pmatrix} x_2 & z_2 \\ & y_2 \end{pmatrix}$ are linearly independent in $\mathfrak{heis}$. In Lie algebra coordinates, linearly dependent elements of $\mathfrak{heis}^2$ form the variety $\mathsf{Rk}_1 \subset \mathsf{M}_{3\times 2}(\mathbb{R})$ of rank one matrices $(\vec{x}, \vec{y}, \vec{z}) = \begin{pmatrix} x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \end{pmatrix}$, alternatively described as triples of simultaneously collinear vectors $\vec{x} \parallel \vec{y} \parallel \vec{z} \in \mathbb{R}^2$. There are no faithful $\mathbb{R}^2$ representations into the 1–dimensional center of Heis, so it suffices to consider the representations in $\mathcal{U}^\star$. Recalling Corollary 3.7, points $(\vec{x}, \vec{y}, \vec{z})$ of $\mathcal{U}^\star$ satisfy $\vec{x} \parallel \vec{y}$, and $\vec{z}$ is perpendicular to their span. Thus any $(\vec{x}, \vec{y}, \vec{z}) \in \mathcal{U}^\star \cap \mathsf{Rk}_1$ necessarily has $\vec{z} = 0$, so the intersection $\mathcal{U}^\star \cap \mathsf{Rk}_1$ is the torus $(\vec{x}, \vec{y}, 0) \subset \mathcal{X}^\star$. The conjugacy classes of faithful representations constitute the complement of this zero section of $\mathcal{U}^\star \to T^2$.

A nonidentity element of Heis_0 stabilizes a point of $\mathbb{H}\mathbb{s}^2$ if and only if it acts trivially on the leaf space of the invariant foliation and has nontrivial shear. In Lie algebra coordinates this forms the set $\mathcal{S} = \left\{\begin{pmatrix} x & z \\ & 0 \end{pmatrix} \mid x \neq 0\right\} \subset \mathfrak{heis}$. The extension $\hat{\rho}$ acts freely if and only if, in Lie algebra coordinates, each generator misses $\mathcal{S}$. All faithful representations $(\vec{x}, \vec{y}, \vec{z})$ with $y_1, y_2 \neq 0$ act freely, and all with $\vec{y} = 0$ fail to. If $\vec{y} = (0, y_2)$, then $\rho \in \mathcal{R}$ implies $x_1 = 0$ so ρ acts freely, and similarly for $\vec{y} = (y_1, 0)$. Thus faithful representations fail to act freely if and only if $\vec{y} = 0$, and the space of freely acting representations is $\mathcal{F} = \mathcal{U}^\star \smallsetminus V(z_1, z_2) \cup V(y_1, y_2)$.

The intersection $S = T^2 \cap V(y_1, y_2)$ is a $(1, 1)$ curve with respect to the $(\vec{u}, \vec{v})$ coordinates, and $\mathcal{U}^\star \smallsetminus V(y_1, y_2)$ is an $\mathbb{R}$–bundle over $\mathsf{Cyl} = T^2 \smallsetminus S$. This bundle is trivial as the generator of $\pi_1(\mathsf{Cyl})$ is parallel to $V(y_1, y_2)$ and the restriction of the doubly twisted bundle $\mathcal{X}$ to a $(1, 1)$ curve in the base is a cylinder. The subvariety $V(z_1, z_2)$ is the zero section of this bundle, thus its complement is the trivial $\mathbb{R}^\times$ bundle over Cyl. □

This classification gives a simple, self contained argument that no incomplete structures exist. An incomplete structure must have holonomy in $\mathcal{U} \smallsetminus \mathcal{F}$, but geometric reasons preclude these from being the holonomy of Heisenberg tori. This completes the classification of tori with Heis_0 holonomy, and a quick observation implies there can be no others.

Proposition 3.13 *Representations $\rho \in \mathcal{U} \smallsetminus \mathcal{F}$ are not the holonomy of any Heisenberg torus. Consequently all Heisenberg tori are complete, with holonomy into* Heis_0.

Proof There are three classes of elements in $\mathcal{U} \smallsetminus \mathcal{F}$: representations into the center, representations $(\vec{x}, \vec{y}, \vec{z})$ with $\vec{z} = 0$ and representations with $\vec{y} = 0$. These classes are all topologically conjugate, and preserve a fibration of the plane $\mathbb{H}\mathbb{s}^2 \twoheadrightarrow \mathbb{R}$. Representations into the center act by translations parallel to the x axis, preserving the invariant foliation of $\mathbb{H}\mathbb{s}^2$, and similarly for those with $\vec{y} = 0$. Representations with $\vec{z} = 0$ are not faithful, and factor through a representation $\mathbb{R} \to \mathsf{Heis}$ with orbits foliating the plane by parabolas.

To see that these cannot be the holonomy of tori, let $\rho \in \mathcal{U} \smallsetminus \mathcal{F}$ preserve the fibration $\pi\colon \mathbb{H}\mathbb{s}^2 \twoheadrightarrow \mathbb{R}$, and assume (f, ρ) is a developing pair for some Heisenberg torus. Let $\Omega = f(\widetilde{T})$ be the developing image, and note that $\pi(\Omega) \subset \mathbb{R}$ is open as f is a local diffeomorphism and π is a bundle projection. Let $Q \subset \widetilde{T}$ be a compact fundamental domain for the action of $\mathbb{Z}^2$ by covering transformations, and note that $\pi(f(Q)) = \pi(f(\Omega))$ as ρ is fiber preserving. But $\pi(f(Q))$ is compact, and thus not open in $\mathbb{R}$, a contradiction.

It follows from this that all Heisenberg tori are complete, and have holonomy in Heis_0. Indeed let T be any Heisenberg torus with developing pair (f, ρ) and $\widetilde{T} \to T$ the cover corresponding to the subgroup $\rho(\mathbb{Z}^2) \cap \mathsf{Heis}_0$. Then $\widetilde{T}$ is complete so T is also, and $\rho(\mathbb{Z}^2)$ acts freely and properly discontinuously on $\mathbb{H}\mathbb{s}^2$. As T^2 is orientable, the holonomy takes values in Heis_+, but every element of $\mathsf{Heis}_+ \smallsetminus \mathsf{Heis}_0$ fixes a point in $\mathbb{H}\mathbb{s}^2$ so in fact ρ is Heis_0–valued and $T = \widetilde{T}$. □

Thus a representation $\rho\colon \mathbb{Z}^2 \to \mathsf{Heis}$ is either the holonomy of a unique complete structure on T^2, or is not the holonomy of any geometric structure at all. After dealing with the slight annoyance of Heis_0 vs Heis_+ conjugacy, this directly provides a description of the Teichmüller space $\mathcal{T}_{\mathbb{H}\mathbb{s}^2}(T^2)$ of unit area structures and the corresponding deformation space $\mathcal{D}_{\mathbb{H}\mathbb{s}^2}(T^2) = \mathbb{R}_+ \times \mathcal{T}_{\mathbb{H}\mathbb{s}^2}(T^2)$.

Theorem 3.14 *The projection onto holonomy identifies the Teichmüller space of unit area Heisenberg tori with the quotient of $\mathcal{F}$ by the free $\mathbb{Z}_2$ action of conjugacy by* $\mathrm{diag}(-1, -1, 1)$ *and* $\mathcal{T}_{\mathbb{H}\mathbb{s}^2}(T^2) \cong \mathcal{F}/\mathbb{Z}^2 \cong \mathbb{R}^2 \times \mathbb{S}^1$.

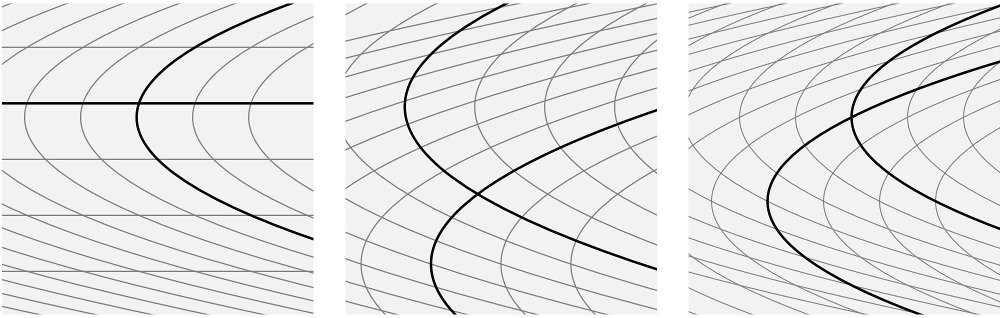


Figure 2: Some examples of developing maps for Heisenberg shear tori.

Proof The map hol: $\mathrm{Dev}_{\mathbb{H}\mathrm{s}^2}(T^2) \to \mathcal{R}$ that projects a developing pair onto its holonomy is a local homeomorphism by the Ehresmann–Thurston principle, which induces a continuous map $\overline{\mathrm{hol}}\colon \mathcal{D}_{\mathbb{H}\mathrm{s}^2}(T^2) \to \mathcal{R}/\mathsf{Heis}_+$. The work above shows the map $\mathrm{dev}\colon \mathcal{F} \to \mathcal{D}_{\mathbb{H}\mathrm{s}^2}(T^2)$ defined by $\rho \mapsto [f_\rho, \rho]$ is a continuous surjection onto the Teichmüller space $\mathcal{T}_{\mathbb{H}\mathrm{s}^2}(T^2)$. As $\mathcal{F} \subset \mathcal{U}$ was defined only up to Heis_0 conjugacy, dev factors through the quotient by $(\mathsf{Heis}_+/\mathsf{Heis}_0) \cong \mathbb{Z}_2$ conjugacy to a continuous bijection $\overline{\mathrm{dev}}\colon \mathcal{F}/\mathbb{Z}_2 \to \mathcal{T}_{\mathbb{H}\mathrm{s}^2}(T^2)$. The composition $\overline{\mathrm{hol}} \circ \overline{\mathrm{dev}}$ is the identity on $\mathcal{F}/\mathbb{Z}_2$, so $\overline{\mathrm{dev}}$ is a homeomorphism.

Thus $\mathcal{T}_{\mathbb{H}\mathrm{s}^2}(T^2) \cong \mathcal{F}/\mathbb{Z}_2$. The quotient $\mathsf{Heis}_+/\mathsf{Heis}_0 \cong \mathbb{Z}_2$ generated by $\mathrm{diag}(-1,-1,1)$ acts by conjugation in Lie algebra coordinates as

$$\mathrm{diag}(-1,-1,1).(\vec{x},\vec{y},\vec{z}) = (\vec{x},-\vec{y},-\vec{z}).$$

This action is free on $\mathcal{F}$ and the quotient $\mathcal{T}_{\mathbb{H}\mathrm{s}^2}(T^2)$ is the trivial $\mathbb{R}_+$ bundle over Cyl, which is homeomorphic to the open solid torus $\mathbb{R}^2 \times \mathbb{S}^1$, and $\mathcal{D}_{\mathbb{H}\mathrm{s}^2}(T^2) \cong \mathbb{R}^3 \times \mathbb{S}^1$. $\square$

The identification $\mathcal{T}_{\mathbb{H}\mathrm{s}^2}(T^2) = \mathcal{F}/\mathbb{Z}_2$ identifies two distinct classes of Heisenberg tori: those containing a shear in their holonomy and those with holonomy into the subgroup of translations of the plane. We will refer to these as *shear tori* and *translation tori*, respectively.

Corollary 3.15 *The space of unit-area translation tori is homeomorphic to* $\mathbb{R} \times \mathbb{S}^1$*, corresponding to the points of* $\mathcal{F} \cap V(x_1, x_2)$*.*

It is notable that the set of developing pairs for Heisenberg translation tori is the same as the set of developing pairs for Euclidean tori, but the corresponding deformation spaces are not homeomorphic, with $\mathcal{T}_{\mathbb{E}^2}(T^2)$ a disk and $\mathcal{T}_{\mathbb{H}\mathrm{s}^2}(T^2)$ a cylinder. This is due to

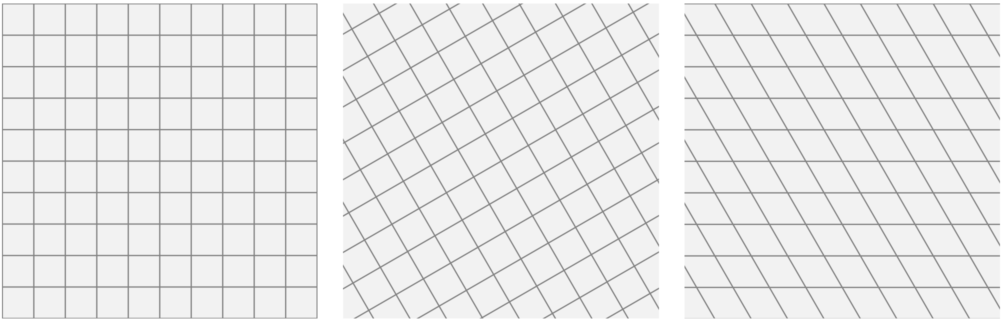


Figure 3: Developing maps for translation tori. The left two are equivalent as Euclidean structures, whereas the right two are as Heisenberg structures. All three represent the same (unique) affine translation torus.

the different notions of equivalence coming from Heis_+ and $\mathrm{Isom}_+(\mathbb{E}^2)$ conjugacy, the former acting by shears and the latter by rotations. The familiar fact that Euclidean torus has a representative holonomy containing horizontal translations is a consequence of this, as is the fact that each Heisenberg translation torus has a representative holonomy translating along (Euclidean) orthogonal lines.

Every Heisenberg structure canonically weakens to an affine structure, defining the map $\omega\colon \mathcal{D}_{\mathbb{H}\mathbb{s}^2}(T^2) \to \mathcal{D}_{\mathbb{A}^2}(T^2)$ with image in the complete structures.

Corollary 3.16 *The space $\omega(\mathcal{D}_{\mathbb{H}\mathbb{s}^2}(T^2))$ of Heisenberg structures up to affine equivalence is one-dimensional, and homeomorphic to $\mathbb{R}$.*

Proof By Goldman and Baues [3], the space of complete affine structures on T^2 is diffeomorphic to the plane, and by completeness we identify this with its projection onto holonomy. This realizes $\omega(\mathcal{D}_{\mathbb{H}\mathbb{s}^2}(T^2))$ as the quotient of $\mathcal{F}$ by affine conjugacy, on which the subgroups of rotations and linearly independent scalings act freely. Thus the $\mathbb{S}^1$ factor and $\mathbb{R}^2_+$ directions of independent scalings collapse in the quotient, and $\omega(\mathcal{D}_{\mathbb{H}\mathbb{s}^2}(T^2)) \cong \mathbb{R}$. □

3.4 Which orbifolds admit Heisenberg structures?

We may use this description of the deformation space of tori to understand all Heisenberg orbifolds. An orbifold covering $\pi : \mathcal{Q} \to \mathcal{O}$ induces a map $\pi^*\colon \mathcal{D}_{\mathbb{H}\mathbb{s}^2}(\mathcal{O}) \to \mathcal{D}_{\mathbb{H}\mathbb{s}^2}(\mathcal{Q})$ by pullback of geometric structures, which is easily expressed on developing pairs as $\pi^*([f, \rho]) = [f, \rho|_{\pi_1(\mathcal{Q})}]$ for $\pi_1(\mathcal{Q}) < \pi_1(\mathcal{O})$, the subgroup corresponding to the cover.

Proposition 3.17 *All Heisenberg structures on orbifolds are complete, and projection onto the holonomy is an embedding* $\mathcal{D}_{\mathbb{H}\mathbb{s}^2}(\mathcal{O}) \hookrightarrow \mathrm{Hom}(\pi_1(\mathcal{O}), \mathsf{Heis})/\mathsf{Heis}_+$. *Under this identification, a finite sheeted covering* $\mathcal{Q} \to \mathcal{O}$ *describes the deformation space* $\mathcal{D}_{\mathbb{H}\mathbb{s}^2}(\mathcal{O})$ *as the preimage of* $\mathcal{D}_{\mathbb{H}\mathbb{s}^2}(\mathcal{Q})$ *under the restriction* $\pi^*\colon \rho \mapsto \rho|_{\pi_1(\mathcal{Q})}$.

Proof Let $\mathcal{O}$ be a Heisenberg orbifold with developing pair $[f,\rho]$, and choose a finite covering $\pi\colon T \to \mathcal{O}$. Then by the completeness of $\pi^*[f,\rho] \in \mathcal{D}_{\mathbb{H}\mathbb{s}^2}(T)$, the developing map f is a diffeomorphism and $\rho|_{\pi_1(T^2)}$ (hence ρ, as $\pi_1(T^2)$ is finite index in $\pi_1(\mathcal{O})$) acts properly discontinuously. As $\pi_1(T^2) < \pi_1(\mathcal{O})$ is an essential subgroup for all orbifolds covered by the torus, the faithfulness of $\rho|_{\pi_1(T^2)}$ implies faithfulness of ρ. Thus the structure $[f,\rho]$ on $\mathcal{O}$ is complete. Let $[\phi,\rho]$ be another Heisenberg structure on $\mathcal{O}$ with the same holonomy, then $\phi f^{-1}\colon \widetilde{\mathcal{O}} \to \widetilde{\mathcal{O}}$ is $\pi_1(\mathcal{O})$ equivariant and descends to a Heisenberg map $\mathcal{O} \to \mathcal{O}$, inducing the identity on fundamental groups. Thus these structures represent the same point in deformation space, so projection onto holonomy is an embedding. $\square$

This further restricts the possible topologies of Heisenberg orbifolds. In particular, any torsion in the fundamental group is represented faithfully by the holonomy, so orbifolds may only have corner reflectors and cone points of order two. In the appendix, we show that all of these actually admit Heisenberg structures, and calculate their deformation spaces.

Corollary 3.18 *If* $\mathcal{O}$ *is a Heisenberg orbifold, necessarily* $\mathcal{O}$ *is* T^2, *the Klein bottle* $\mathbb{S}^1 \tilde{\times} \mathbb{S}^1$, *or the pillowcase* $\mathbb{S}^2(2,2,2,2)$ *or one of their quotients: the cylinder* $\mathbb{S}^1 \times I$, *the Möbius band* $\mathbb{S}^1 \tilde{\times} I$, *the square* $\mathbb{D}^2(\varnothing; 2,2,2,2)$, $\mathbb{D}^2(2,2;\varnothing)$, $\mathbb{D}^2(2;2,2)$ *and* $\mathbb{R}P^2(2,2)$.

4 Collapse and regenerations

Unless otherwise specified, $\mathbb{X}$ denotes any one of the constant curvature geometries $\mathbb{S}^2$, $\mathbb{E}^2$ or $\mathbb{H}^2$ realized as a subgeometry of $\mathbb{R}P^2$ (see Section 2.4) throughout. Conjugate models will be denoted by $C.\mathbb{X}$ for $C \in \mathsf{GL}(3;\mathbb{R})$. Recall, a collapsing path $[f_t,\rho_t]$ of $\mathbb{X}$ structures degenerates to a Heisenberg structure if there is a path $C_t \in \mathsf{GL}(3;\mathbb{R})$ with $C_t.[f_t,\rho_t] = [C_t f_t, C_t\rho_t C_t^{-1}]$ converging in the space of developing pairs to $[f_\infty,\rho_\infty]$ with f_∞ an immersion into the affine patch $\mathbb{H}\mathbb{s}^2 = \{[x:y:1]\}$ and ρ_∞ with image in Heis. We may view these rescaled $\mathbb{X}$ structures as geometric structures modeled on the conjugate subgeometry $C_t.\mathbb{X}$ which converge to a Heisenberg structure as $C_t\,.\,\mathbb{X}$ itself converges to $\mathbb{H}\mathrm{s}^2$. The following proposition, a consequence of [7] (or a straightforward calculation of conjugacy limits of Lie algebras), describes which conjugacies of $\mathbb{X} \in \{\mathbb{S}^2, \mathbb{E}^2, \mathbb{H}^2\}$ limit to the Heisenberg plane:

Proposition 4.1 *Let $\mathbb{X}$ be a projective model of a constant curvature geometry in $\mathbb{R}P^2$, and $C_t : [0,\infty) \to \mathsf{PGL}(3;\mathbb{R})$ be any path of projective transformations. After potentially rescaling the matrix representatives and applying the KBH decomposition* (Theorem 4.1 of [7]) *we write $C_t = K_t D_t H_t$ for $K_t \in \mathsf{O}(3)$, $H_t \in \mathrm{Isom}(\mathbb{X})$ and $D_t = \mathrm{diag}(\lambda_t, \mu_t, 1)$ with $\lambda_t \geq \mu_t \geq 1$. Then for $\mathbb{X} \in \{\mathbb{S}^2, \mathbb{H}^2\}$, the path of geometries $C_t\,.\,\mathbb{X}$ limits to the Heisenberg plane if and only if*

(i) *K_t converges in $\mathsf{O}(3)$, and*

(ii) *λ_t, μ_t and λ_t/μ_t all diverge to ∞.*

For $\mathbb{X} = \mathbb{E}^2$, the divergence $\lambda_t/\mu_t \to \infty$ alone is necessary and sufficient for (ii).

For convenience, we may without loss of generality restrict our attention to conjugacy limits by diagonal matrices $D_t = \mathrm{diag}(\lambda_t, \mu_t, 1)$ with $\lambda_t > \mu_t > 1$. To see this, let $\mathbb{X} \in \{\mathbb{S}^2, \mathbb{E}^2, \mathbb{H}^2\}$ and suppose C_t is any path of projective transformations such that $C_t\mathbb{X} \to \mathbb{H}\mathrm{s}^2$. Writing $C_t = K_t D_t H_t$ as above, we note that $C_t\mathbb{X} = K_t D_t \mathbb{X}$ for all t as $H_t \in \mathrm{Isom}(\mathbb{X})$, and as K_t converges, we see that $K_t^{-1} C_t \mathbb{X} = D_t \mathbb{X}$ is conjugate to the original path, even in the limit.

In this section, we classify which Heisenberg tori arise as rescaled limits of collapsing constant curvature geometric structures. As all constant curvature tori are Euclidean, we consider the natural generalization of *cone manifold structures* on the torus, which exist in both positive and negative curvature.

4.1 Constant curvature cone tori

Definition 4.2 An $\mathbb{X}$ cone surface is a surface Σ with a complete path metric that is the metric completion of an $\mathbb{X}$ structure on the complement of a discrete set.

An $\mathbb{X}$ cone torus T with cone points $C = \{p_1, \dots, p_n\}$ gives an incomplete $\mathbb{X}$ structure on $T^2_\star = T^2 \smallsetminus C$ encoded by a class of developing pairs [8]. The space of all such $\mathbb{X}$ cone tori can be identified with the subset $\mathcal{C}_{\mathbb{X}}(T^2) \subset \mathcal{D}_{\mathbb{X}}(T^2_\star)$ with metric completions T^2, given the subspace topology under this inclusion.

Definition 4.3 A path T_t of $\mathbb{X}$ cone tori converges projectively if the associated incomplete structures $(f_t, \rho_t) \in \mathcal{D}_{\mathbb{X}}(T^2_\star)$ converge in $\mathcal{D}_{\mathbb{R}P^2}(T^2_\star)$ to a projective structure (f_∞, ρ_∞) which can be completed to a projective torus T. Conversely, we say a

Heisenberg torus T *regenerates* to $\mathbb{X}$ structures if there is a sequence of $\mathbb{X}$ cone tori converging to T in $\mathbb{R}P^2$.

In the above definition we always require the limiting projective structure on the torus to be nonsingular and allow only sequences of Riemannian cone tori where the cone point(s) vanish in the limit. Allowing singularities in the limiting structure requires a notion of *real projective cone manifolds*, which is beyond the scope of this work.

In considering the question of regeneration, we further restrict our attention to sequences containing tori with a single cone point. Cone tori with a single cone point admit a convenient combinatorial description via marked parallelograms, which provides us substantial control. A marked $\mathbb{X}$ parallelogram is a convex quadrilateral $Q \subset \mathbb{X}$ with opposing geodesic sides of equal length, equipped with an ordering of the vertices (v_1, v_2, v_3, v_4) proceeding counterclockwise from some initial vertex v_1. Such a marked parallelogram is determined by a vertex $v = v_1$, the geodesic lengths of the sides adjacent to v and the angle of incidence at v. The moduli space $\mathcal{P}(\mathbb{X})$ of marked parallelograms is $\mathbb{R}^2_+ \times (0, \pi)$ in nonpositive curvature and $(0, \pi/(2\kappa))^2 \times (0, \pi)$ in spherical space of radius κ. Just as deformation spaces of Euclidean tori can be identified with isometry classes of marked parallelograms $\mathcal{P}(\mathbb{E}^2)$, so can the deformation spaces of $\mathbb{H}^2$ and $\mathbb{S}^2$ cone structures (with the caveat that in positive curvature we must restrict our interest to sufficiently small cone angle).

Proposition 4.4 *The map* $\mathrm{Glue}\colon \mathcal{P}(\mathbb{X}) \to \mathcal{C}_{\mathbb{X}}(T_\star)$ *induced by isometrically identifying opposing sides of* $Q \in \mathcal{P}(\mathbb{X})$ *is a homeomorphism onto its image. For* $\mathbb{X} = \mathbb{H}^2$ *this image is the entire deformation space* $\mathcal{C}_{\mathbb{X}}(T_\star)$*. For* $\mathbb{X} = \mathbb{S}^2$ *the image contains all cone tori whose marked curves each have length less than* $\frac{1}{2}\pi$*.*

Proof There is a unique orientation-preserving isometry sending any oriented line segment in $\mathbb{X}$ to any other of the same length. Thus a marked quadrilateral $Q \subset \mathbb{X}$ determines unique side pairings $A, B \in \mathrm{Isom}_+(\mathbb{X})$ identifying opposing sides. The quotient is topologically a torus and inherits an $\mathbb{X}$ structure on the complement of $[v]$. If Q' is isometric to Q then there is a $g \in \mathrm{Isom}(\mathbb{X})$ with $g \,.\, Q = Q'$, so the induced structures are isomorphic and Glue is well defined.

We may also define an inverse *cutting map* as follows. A marked $\mathbb{X}$ cone torus T has generators $a, b \in \pi_1(T)$ based at the cone point, which may be pulled tight relative to p to length minimizing representatives α and β as T is a compact path metric space. These are locally length minimizing, and so $\mathbb{X}$–geodesics away from p. As $a \simeq \alpha$ and $b \simeq \beta$ generate $\pi_1(T)$, α and β have algebraic intersection number 1. As each is

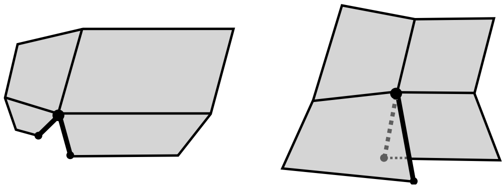


Figure 4: Small portions of the developing map for a hyperbolic and spherical cone torus

globally length minimizing in its pointed homotopy class, the complement $T \smallsetminus \{\alpha \cup \beta\}$ contains no bigons. From this it follows that $\alpha \cap \beta = \{p\}$, and so cutting along α and β gives a simply connected surface locally modeled on $\mathbb{X}$, with four geodesic boundary components, opposing pairs of which have equal length.

For $\mathbb{X} = \mathbb{H}^2$ such a surface always embeds in $\mathbb{H}^2$ as a hyperbolic parallelogram, so this process defines a map $\mathrm{Cut}\colon \mathcal{C}_{\mathbb{X}}(T_\star) \to \mathcal{P}(\mathbb{X})$. When $\mathbb{X} = \mathbb{S}^2$ it is possible that the resulting surface does not embed in $\mathbb{S}^2$ (indeed the area of Q may exceed the area of $\mathbb{S}^2$!). However, if both α and β have length less than $\frac{1}{2}\pi$, then Q certainly embeds in $\mathbb{S}^2$ (in fact it embeds into a hemisphere, and thus into the projective model $\mathbb{R}P^2$ of spherical geometry). These maps $\{\mathrm{Cut}, \mathrm{Glue}\}$ are inverses where their composition is defined, and thus define a pair of homeomorphisms. $\square$

To study regenerations from this combinatorial perspective, we characterize when a collapsing path in $\mathcal{C}_{\mathbb{X}}(T_\star)$ converges in $\mathcal{D}_{\mathbb{R}P^2}(T_\star)$ in terms of marked parallelograms. First, we show such a characterization is possible as all convergent paths of cone tori admit such a representation.

Proposition 4.5 *Let $\mathbb{X}_t$ be a sequence of geometries conjugate to a constant curvature geometry $\mathbb{X}$ which converges to $\mathbb{H}\mathbb{s}^2$ in the space $\mathfrak{S}_{\mathbb{R}P^2}$ of subgeometries of $\mathbb{R}P^2$. If T_t is any convergent sequence of $\mathbb{X}_t$ cone tori, then for all sufficiently large t the structures T_t lie in the image of the gluing map* $\mathrm{Glue}\colon \mathcal{P}(\mathbb{X}) \to \mathcal{C}_{\mathbb{X}}(T_\star)$.

Proof For $\mathbb{X} = \mathbb{H}^2$ the gluing map is surjective by Proposition 4.4 so there is nothing more to prove. For $\mathbb{X} = \mathbb{S}^2$ by the same proposition it is enough to show that eventually all the structures T_t have marked curves of sufficiently short length. Choose a smooth curve γ representing one of the markings on $T_\star$. For each t we pull γ tight, fixing the cone point to a geodesic whose length $\ell_t \le \mathrm{Length}_{T_t}(\gamma)$ defines the length of this marking curve in the T_t structure. As $t \to \infty$, we show that $\mathrm{Length}_{T_t}(\gamma)$, and hence ℓ_t tends to 0.

Let (f_t, ρ_t) be a convergent sequence of developing pairs for the cone tori T_t. Choosing a lift $\tilde{\gamma}$ of γ we note that $\mathrm{Length}_{T_t}(\gamma) = \mathrm{Length}_{\mathbb{X}_t}(f_t \circ \tilde{\gamma})$, allowing computation of lengths in T_t via the geometry $\mathbb{X}_t$. By the assumed convergence of this path of structures the developing maps f_t converge to some $f\colon \widetilde{T}_\star \to \mathbb{R}P^2$ uniformly in the C^∞ topology on compact sets, so fixing any $\epsilon > 0$, for all sufficiently large t, $|\mathrm{Length}_{\mathbb{X}_t}(f_t \circ \tilde{\gamma}) - \mathrm{Length}_{\mathbb{X}_t}(f \circ \tilde{\gamma})| < \varepsilon$.

Therefore we only need to understand the length of the fixed curve $f \circ \tilde{\gamma}$ in the changing geometries $\mathbb{X}_t$. The geometry $\mathbb{X}_t = D_t \mathbb{S}^2$ is a conjugate of spherical geometry by some projective transformation D_t, which, as in the discussion following Proposition 4.1, we may without loss of generality take to be represented by a diagonal matrix $D_t = \mathrm{diag}(\lambda_t, \mu_t, 1)$. Changing perspective (applying D_t^{-1}) we note that $\mathrm{Length}_{D_t \mathbb{S}^2}(f \circ \gamma) = \mathrm{Length}_{\mathbb{S}^2}(D_t^{-1} \circ f \circ \tilde{\gamma})$, allowing us to instead compute the length of a varying curve in a fixed model of $\mathbb{S}^2$.

As $D_t \mathbb{S}^2$ limits to the Heisenberg plane, by Proposition 4.1 the eigenvalues λ_t and μ_t of D_t diverge to ∞, hence the effect of D_t^{-1} on the standard affine patch $\{[x:y:1]\}$ of $\mathbb{R}P^2$ is to collapse everything towards the origin. Thus as $t \to \infty$, the curve $D_t^{-1} \circ f \circ \gamma$ converges to a constant map, and its sequence of lengths converges to 0. All together this implies for any $\varepsilon > 0$, for all sufficiently large t,

$$\begin{aligned} \ell_t \le \mathrm{Length}_{T_t}(\gamma) = \mathrm{Length}_{\mathbb{X}_t}(f \circ \tilde{\gamma}) &< \mathrm{Length}_{\mathbb{X}_t}(f \circ \tilde{\gamma}) + \tfrac{1}{2}\varepsilon \\ &= \mathrm{Length}_{\mathbb{S}^2}(D_t^{-1} f \circ \tilde{\gamma}) + \tfrac{1}{2}\varepsilon < \varepsilon. \end{aligned}$$

This is stronger than we strictly require: taking $\varepsilon = \frac{1}{2}\pi$ and applying this to both marking curves is enough to provide the desired result in light of Proposition 4.4. □

Next, we give a precise description of these convergent sequences of structures in terms of their parallelogram representatives.

Proposition 4.6 *Let $\mathbb{X}_t = D_t \mathbb{X}$ be a sequence of geometries conjugate to $\mathbb{X}$ which converges to $\mathbb{H}\mathrm{s}^2$ in the space $\mathfrak{S}_{\mathbb{R}P^2}$ of subgeometries of $\mathbb{R}P^2$, and T_t a sequence of $\mathbb{X}_t$ cone tori. Then the structures T_t converge to a Heisenberg torus if and only if:*

1. *For all sufficiently large t, there is a choice of embeddings $Q_t \hookrightarrow \mathbb{X}_t \subset \mathbb{R}P^2$ of the fundamental parallelograms for T_t whose images converge in the Hausdorff space of closed subsets of $\mathbb{R}P^2$ to a projective quadrilateral Q.*
2. *The induced side pairings A_t and B_t of Q_t converge in $\mathsf{PGL}(3;\mathbb{R})$ to a commuting pair of projective transformations A and B.*

Proof Again without loss of generality we may assume that the conjugating transformations D_t are represented by diagonal matrices. Let (f_t, ρ_t) be a convergent sequence of developing pairs for the incomplete structures on $T_\star = T^2 \smallsetminus \{*\}$ for $\mathbb{X}_t$ cone tori T_t. Choose a generating set $a, b \in \pi_1(T_\star)$ and a basepoint $q \in \widetilde{T}_\star$. The universal cover $\widetilde{T}_\star$ is tiled by ideal quadrilaterals formed from the lifts of a and b. For each t these can be straightened to geodesics in the $\mathbb{X}_t$ structure; let $\widetilde{Q}_t \subset \widetilde{T}_\star$ be the geodesic quadrilateral containing $q \in \widetilde{T}_\star$.

By Proposition 4.5, for all sufficiently large t, the quadrilateral $\widetilde{Q}_t$ can be embedded as a subset of $\mathbb{X}_t$. For these structures the developing map f_t itself provides such an embedding, and we define $Q_t = f_t(\widetilde{Q}_t) \subseteq \mathbb{X}_t$ together with the side pairings $A_t = \rho_t(a)$ and $B_t = \rho_t(b)$. When $\mathbb{X} = \mathbb{H}^2$ the convergence of developing pairs then implies A_t and B_t are convergent in $\mathsf{PGL}(3;\mathbb{R})$ to A and B and Q_t converges to Q_∞, a fundamental domain for the Heisenberg structure T with sides paired by the commuting transformations A and B.

Conversely let Q_t be a sequence of $\mathbb{X}_t$ parallelograms convergent in the Hausdorff space $\mathfrak{C}_{\mathbb{R}P^2}$ of closed subsets of $\mathbb{R}P^2$ to an affine parallelogram Q. The triples (Q_t, A_t, B_t) of the quadrilateral with side pairings define $\mathbb{X}_t$ cone tori, and hence $\mathbb{R}P^2$ punctured tori for all t. As $t \to \infty$ these converge to a punctured torus T_∞ with holonomy in Heis, and so $T_\infty \in \mathcal{D}_{\mathbb{H}\mathbb{s}^2}(T_\star)$. As $[A, B] = I$, the limiting holonomy factors through $\mathbb{Z} \oplus \mathbb{Z}$ and so the limiting torus can be completed to a torus T_∞. That the limits A and B and in Heis follows from the definition of $\mathbb{X}_t$ converging to $\mathbb{H}\mathbb{s}^2$, so this limiting projective structure canonically strengthens to a Heisenberg structure. □

4.2 Translation tori

This combinatorial description of cone tori with at most one cone point provides enough control to completely understand the regeneration of translation tori.

Theorem 4.7 *Let $\mathbb{X} \in \{\mathbb{S}^2, \mathbb{E}^2, \mathbb{H}^2\}$ and $\mathbb{X}_t = D_t . \mathbb{X}$ be a sequence of diagonal conjugates converging to $\mathbb{H}\mathbb{s}^2$. Given any translation torus T there is a sequence of $\mathbb{X}_t$ cone tori with at most one cone point converging to T.*

Proof (Euclidean case) Heisenberg tori arise as limits of collapsing families of *smooth* Euclidean tori (there are no Euclidean cone tori with a single cone point, per Gauss–Bonnet). Let T be a Heisenberg translation torus and $\mathbb{E}_t = D_t . \mathbb{E}^2$ be a sequence of diagonal conjugates of $\mathbb{E}^2$ converging to the Heisenberg plane. Choose a fundamental domain Q for $T \subset \mathbb{H}\mathbb{s}^2$, together with side pairings A and B by translations for T.

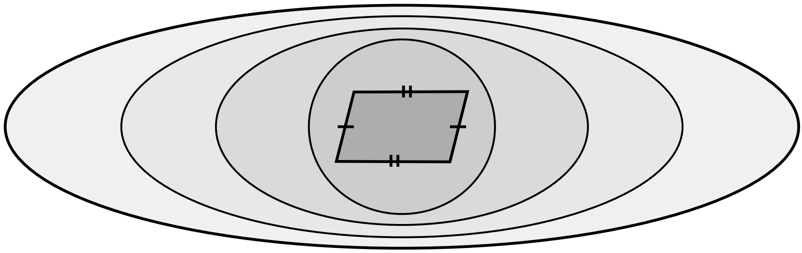


Figure 5: A fixed Quadrilateral and various conjugate models of $\mathbb{H}^2$ containing it.

The underlying spaces for the models $\mathbb{E}^2$, $\mathbb{E}_t$ and $\mathbb{Hs}^2$ in $\mathbb{R}P^2$ are all the entire affine patch $\mathbb{A}^2 = \{[x : y : 1]\}$, and the group Tr of translations acting on this affine patch is contained in each conjugate $D_t \operatorname{Isom}(\mathbb{E}^2) D_t^{-1}$ as well as Heis. Thus (Q, A, B) encodes an $\mathbb{E}_t$ structure $[f, \rho]_{\mathbb{E}_t}$ on T^2 for each $t \in \mathbb{R}_+$. Canonically weakening to projective structures, this is the constant sequence $[f, \rho]_{\mathbb{R}P^2}$, thus clearly convergent. As $\rho(\mathbb{Z}^2) \subset \mathsf{Tr} < \mathsf{Heis}$, the limit canonically strengthens to the original Heisenberg structure $[f, \rho]_{\mathbb{Hs}^2}$. □

When viewed as Euclidean structures in the fixed model $\mathbb{E}^2$, the developing pairs $[D_t^{-1} f, D_t^{-1} \rho D_t]$ encode a collapsing collection of tori with one of the generators of the holonomy shrinking much faster than the other. That is, even after rescaling to unit area structures this path fails to converge in Teichmüller space and limits to a point in the Thurston boundary. The foliation represented by this point can actually be seen in the limiting Heisenberg structure as the invariant foliation pulled back from dy on $\mathbb{Hs}^2$.

The approach for producing translation tori as limits of hyperbolic and spherical cone tori is similar in spirit, but more involved in the details. Again we take a fundamental domain with side pairings (Q, A, B) for the proposed limit, and view Q as a geometric parallelogram in each of the model geometries $\mathbb{X}_t$. Side pairings $A_t, B_t \in \operatorname{Isom}(\mathbb{X}_t)$ are uniquely determined by each $\mathbb{X}_t$ structure on Q, and converge to A and B in the limit.

Proof (hyperbolic and spherical cases) If $\mathbb{X} \in \{\mathbb{S}^2, \mathbb{H}^2\}$, let Q be an origin-centered fundamental domain for T with side pairings $A, B \in \mathsf{Tr}$. The existence of a convergent sequence of $\mathbb{X}_t$ cone tori $T_t \to T$ follows from:

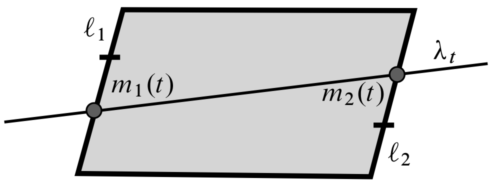
Claim 1 *For large t, the quadrilateral Q defines an $\mathbb{X}_t$ parallelogram.*

Claim 2 *The side pairing A_t preserves the entire projective line through the $\mathbb{X}_t$ midpoints of paired sides.*

Claim 3 *If Q is an $\mathbb{X}_t$ parallelogram for all t and $A_t \in \mathrm{Isom}(\mathbb{X}_t)$ pairs opposing sides, A_t converges as a sequence of projective transformations.*

Claim 4 *The $\mathbb{X}_t$ midpoints of the edges of Q converge to the Euclidean midpoints as $t \to \infty$.*

Given that Q defines an $\mathbb{X}_t$ parallelogram, there are unique side pairing transformations $A_t, B_t \in \mathrm{Isom}(\mathbb{X}_t)$ determining an $\mathbb{X}_t$ cone torus. By the third claim, these sequences of transformations converge in $\mathsf{PGL}(3,\mathbb{R})$, and as $\mathbb{X}_t \to \mathbb{H}\mathbb{s}^2$ in fact $A_\infty, B_\infty \in \mathsf{Heis}_0$. Recalling the discussion in Section 3, Heis_0 acts simply transitively on the subspace $\mathcal{L} \smallsetminus \mathcal{H}$ of pointed lines, so the limiting transformations are completely determined by their action on a pair (p, ℓ) of a point p on a nonhorizontal line ℓ.



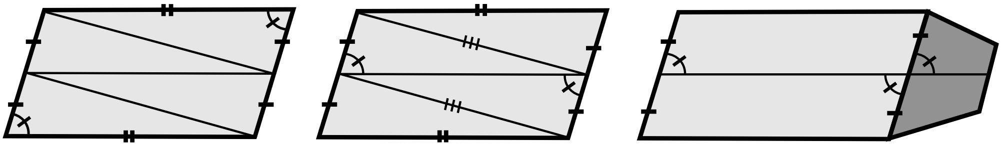
Let ℓ_1 and ℓ_2 be a pair of opposing sides of Q, with Euclidean midpoints m_1 and m_2. For each t, let $m_1(t)$ and $m_2(t)$ be the $\mathbb{X}_t$ corresponding midpoints, and λ_t the projective line connecting them. The second claim implies A_t preserves λ_t, and so the fourth fact above implies that A_∞ preserves $\lambda = \overline{m_1 m_2}$. Thus A_∞ sends the pair (m_1, ℓ_1) to (m_2, ℓ_2), as well as the pair (m_1, λ) to (m_2, λ). At least one of the lines ℓ_1 or λ is nonhorizontal, and so this completely determines the behavior of A_∞. As this agrees precisely with the action of the original transformation A, we have $A_\infty = A$ and similarly for B. Thus the sequence of cone tori corresponding to the triples (Q, A_t, B_t) converges to the original Heisenberg torus T as $t \to \infty$. $\square$

Thus the proof reduces to an argument for the four claims above. Throughout, it's often helpful to switch between the perspectives of a fixed fundamental domain Q in expanding model geometries $\mathbb{X}_t$ and the equivalent picture of shrinking domains Q_t in the fixed model $\mathbb{X}$.

Claim 1 *Let Q be an affine parallelogram centered at $\vec{0} \in \mathbb{A}^2$ and $\mathbb{X}_t \to \mathbb{H}\mathbb{s}^2$ a sequence of diagonal conjugates of $\mathbb{X} \in \{\mathbb{S}^2, \mathbb{H}^2\}$. Then for all $t \gg 0$, Q defines an $\mathbb{X}_t$ parallelogram.*

Proof The π–rotation about $\vec{0} \in \mathbb{A}^2$ represented by $R = \mathrm{diag}(-1,-1,1)$ is in $\mathsf{O}(3) \cap \mathsf{O}(2,1)$ and is invariant under diagonal conjugacy. Thus for each t, $R \in \mathrm{Isom}(\mathbb{X}_t)$. As Q is an affine parallelogram with centroid $\vec{0}$, $RQ = Q$, so there is an $\mathbb{X}_t$ isometry exchanging opposing sides of Q. Thus if $Q \subset \mathbb{X}_t$, it defines an $\mathbb{X}_t$ parallelogram. For $\mathbb{X} = \mathbb{S}^2$ this is always satisfied, and for $\mathbb{X} = \mathbb{H}^2$, the domains $\mathbb{X}_t$ limit to the affine patch and so eventually contain any compact subset. □

Claim 2 *Let $A \in \mathrm{Isom}(\mathbb{X})$ pair opposing sides of the $\mathbb{X}$ parallelogram Q. Then A preserves the projective line through the midpoints of the paired sides.*



Proof We argue in classical axiomatic geometry without assuming the parallel postulate as this applies equally to $\mathbb{S}^2$ and $\mathbb{H}^2$. Opposite angles of a constant curvature parallelogram are congruent. Connect the opposing sides of Q paired by A_t with a line segment λ through their midpoints. This divides Q into two quadrilaterals, subdivided by their diagonals into four triangles. The outer two of these triangles are congruent by side-angle-side, and so the diagonals are congruent. Thus the inner two triangles are congruent by side-side-side, meaning the opposite angles made by the edges with the line connecting their midpoints are equal. Consider Q and its translate $A \,.\, Q$. These share an edge, which meets the segments λ and $A_t\lambda$ at its midpoint m. As A is an isometry, it follows that opposite angles at m are congruent. Thus λ and $A \,.\, \lambda$ are segments of a single projective line, so A preserves the line extending λ as claimed. □

Claim 3 *The side pairings $A_t, B_t \in \mathrm{Isom}(\mathbb{X})$ converge in $\mathsf{PGL}(3,\mathbb{R})$.*

Proof A projective transformation of $\mathbb{R}P^2$ is completely determined by its values on a projective basis (a collection of four points in general position). The vertices (v_i) of Q form a convenient projective basis with images $(A_t v_i)$ completely specifying the transformations A_t. These transformations converge in $\mathsf{PGL}(3;\mathbb{R})$ if and only if $(A_t v_i)$ limits to a projective basis, which, as the images $A_t v_i$ remain in a bounded neighborhood of Q,[2] is equivalent to no triangle $\Delta \subset Q$ formed by three vertices of Q collapsing in the limit. That is, it suffices to show $\mathrm{Area}_{\mathbb{E}^2}(A_t\Delta)/\,\mathrm{Area}_{\mathbb{E}^2}(\Delta) \not\to 0$.

[2]The conjugating path C_t is *expansive* with eigenvalues $\lambda_t > \mu_t$ each monotonic in t. Then it's easy to see for $\mathbb{X} = \mathbb{H}^2$ that $A_t Q \subset AQ$ and for $\mathbb{X} = \mathbb{S}^2$ that $A_t Q < A_0 Q$, for all $t > 0$.

Diagonal transformations act linearly on the affine patch and do not change ratios of areas, thus we may transform this to the fixed model $\mathbb{X}$ with a collapsing sequence of triangles Δ_t being moved by transformations $C_t = D_t A_t D_t^{-1}$. For large t, both Δ_t and $C_t\Delta_t$ are extremely close to the origin $\vec{0} \in \mathbb{A}^2$ and we may estimate their area ratio analytically. By Claim 2, C_t preserves the geodesic through the midpoints of paired sides, thus is either a hyperbolic in $\mathrm{Isom}(\mathbb{H}^2)$ or rotation in $\mathrm{Isom}(\mathbb{S}^2)$ with axis represented by an ideal point relative the affine patch. In each of these cases we may bound the distortion of Euclidean area under these isometries as follows.

Up to conjugation by a rotation we may express any such isometry as

$$C = \begin{pmatrix} c(\tau) & 0 & s(\tau) \\ 0 & 1 & 0 \\ s(\tau) & 0 & c(\tau) \end{pmatrix},$$

where $(c, s) = (\cosh, \sinh)$ for $\mathbb{X} = \mathbb{H}^2$ and $(\cos, \sin)$ for $\mathbb{X} = \mathbb{S}^2$, and τ is the translation length along the preserved geodesic. At any $p = [x : y : 1] \in \mathbb{A}^2$, the infinitesimal area distortion $J(p)$ is given by the Jacobian of the projective action of C on the affine patch. On any region $R \subset \mathbb{A}^2$ then, the overall area distortion $\mathrm{Area}_{\mathbb{E}^2}(C(R))/\,\mathrm{Area}_{\mathbb{E}^2} R$ is bounded below by $J_{\min}(R) = \inf_{p\in R} J(p)$ and above by $J_{\max}(R) = \sup_{p\in R} J(p)$.

Consider again the region Δ_t and the side pairing C_t with side pairing of translation length τ_t. As $t \to \infty$, both Δ_t and $C_t\Delta_t$ collapse to $\vec{0}$. Thus for any $\varepsilon > 0$ and all sufficiently large t, both of these regions are subsets of the ε–ball about 0, and we may bound the overall area distortion by $J_{\min}(B_\varepsilon)$ and $J_{\max}(B_\varepsilon)$. Computing these, we see

$$\frac{1}{(c(\tau) + \varepsilon s(\tau))^3} \le \frac{\mathrm{Area}_{\mathbb{E}^2}(C_t\Delta_t)}{\mathrm{Area}_{\mathbb{E}^2}(\Delta_t)} \le \frac{1}{(c(\tau) - \varepsilon s(\tau))^3}.$$

As $t \to \infty$ the translation length τ_t converges to 0 (as the geometric structure's developing map collapses to the constant map onto the origin). Thus, the above bounds squeeze the limiting area of $C_t\Delta_t$ to Δ_t by 1, so the area of $A_t\Delta$ does not collapse in the limit. □

Claim 4 *Let $\ell \subset \mathbb{A}^2$ be a line segment and $\mathbb{X}_t \to \mathbb{H}\mathbb{s}^2$ as above. Then the $\mathbb{X}_t$ midpoint of ℓ converges to the Euclidean midpoint.*

Proof Let $\ell = \overline{pq}$ and $m \in \ell$ be the Euclidean midpoint. Viewing ℓ in $\mathbb{X}_t$, it has $\mathbb{X}_t$ midpoint y_t, and to show $y_t \to m$ it suffices to see $d_{\mathbb{X}_t}(p, m)/d_{\mathbb{X}_t}(m, q) \to 1$. Ratios of collinear line segment lengths are invariant under linear transformations, so we may choose to view this situation in the fixed model $\mathbb{X}$ for ease of calculation, with a shrinking line segment $\ell_t = \overline{p_t q_t}$ with Euclidean midpoint m_t and $\mathbb{X}$ midpoint x_t.

For $\mathbb{X}=\mathbb{H}^2$ a straightforward computation shows that the length of any segment $\ell\subset B_{\mathbb{E}^2}(0,\varepsilon)$ is bounded by a multiple of its Euclidean length

$$\mathrm{Length}_{\mathbb{E}^2}(\ell)\le \mathrm{Length}_{\mathbb{X}}(\ell)\le K_\varepsilon\,\mathrm{Length}_{\mathbb{E}^2}(\ell),$$

where K_ε may be chosen[3] so that $K_\varepsilon>1$ and $\lim_{\varepsilon\to 0}K_\varepsilon=1$. Similarly, pulling back the spherical metric to the affine patch there is such a $K_\varepsilon>1$ with

$$\frac{\mathrm{Length}_{\mathbb{E}^2}(\ell)}{K_\varepsilon}\le \mathrm{Length}_{\mathbb{X}}(\ell)\le \mathrm{Length}_{\mathbb{E}^2}(\ell).$$

We may use this to bound the difference between the $\mathbb{X}$ and Euclidean midpoints of the shrinking segments ℓ_t:

$$\frac{1}{K_\varepsilon}=\frac{d_{\mathbb{E}^2}(p_t,m_t)}{K_\varepsilon d(m_t,q_t)}\le\frac{d_{\mathbb{X}}(p_t,m_t)}{d_{\mathbb{X}}(m_t,q_t)}=\frac{d_{\mathbb{X}_t}(p,m)}{d_{\mathbb{X}_t}(m,q)}\le\frac{K_\varepsilon d_{\mathbb{E}^2}(p_t,m_t)}{d_{\mathbb{E}^2}(m_t,q_t)}=K_\varepsilon.$$

As $\mathbb{X}_t\to\mathbb{H}\mathbb{s}^2$, ℓ_t collapses to $\vec{0}$ and we may take smaller and smaller ε so this ratio converges to 1. □

4.3 Shear tori

Every translation Heisenberg torus arises as a limit of Euclidean, hyperbolic or spherical cone tori with at most one cone point. Translation structures are rather special Heisenberg tori, compromising a codimension-1 subset of deformation space. Here we investigate the generic case, Heisenberg tori with nontrivial shears in their holonomy, and show none regenerate as cone structures with a single cone point. Shears of the plane fix a single line, and alter the slope of all lines not parallel to this. All shears in Heis are parallel, so the holonomy of any shear torus leaves invariant precisely one slope on $\mathbb{H}\mathbb{s}^2$. This has strong consequences for the distribution of geodesics on Heisenberg orbifolds.

Proposition 4.8 *A Heisenberg orbifold $\mathcal{O}$ has a shear in its holonomy if and only if all simple geodesics on $\mathcal{O}$ are pairwise disjoint.*

Proof Let $\mathcal{O}$ be a shear orbifold and γ a simple geodesic on $\mathcal{O}$. As $\mathcal{O}$ is covered by a complete torus we identify $\widetilde{\mathcal{O}}$ with $\mathbb{H}\mathbb{s}^2$, and the preimage of γ under the covering with a $\pi_1(\mathcal{O})$–invariant collection $\{\tilde{\gamma}\}$ of lines in $\mathbb{H}\mathbb{s}^2$. As γ is simple these are pairwise disjoint and so parallel in $\mathbb{A}^2$. Because $\mathcal{O}$ has a shear structure, some $\alpha\in\pi_1(\mathcal{O})$ acts on $\mathbb{H}\mathbb{s}^2$ by a nontrivial shear, which alters the slope of all nonhorizontal lines. Thus $\{\tilde{\gamma}\}$ is a subset of the horizontal foliation. But this holds for any simple geodesic on $\mathcal{O}$

[3]For hyperbolic space we may choose $K_\varepsilon=1/\sqrt{1-4\varepsilon^2}$ and for the sphere $K_\varepsilon=1/(1+\varepsilon^2)$ with ε measured in the Euclidean metric on the affine patch.

so any two must each lift to a subset of the horizontal foliation, which are then disjoint or (by $\pi_1(\mathcal{O})$ invariance) equal. If the two geodesics lift to disjoint collections then their projections are also disjoint, meaning any two distinct simple geodesics on T cannot intersect.

Conversely, assume $\mathcal{O}$ is an orbifold covered by a translation torus T given by the developing pair (f, ρ), for $\rho\colon \mathbb{Z}^2 \to \mathsf{Tr}$. Then $\rho(e_1)$ and $\rho(e_2)$ are linearly independent translations, each preserving each component of a family of parallel lines descending to closed intersecting geodesics on T and further descend to intersecting geodesics on $\mathcal{O}$. □

Hyperbolic, spherical and Euclidean (cone) tori behave quite differently than this. Recall that any generators $\langle a, b\rangle = \pi_1(T)$ have geodesic representatives through the cone point, and cutting along these gives a constant curvature parallelogram with side pairings. Claim 2 of the previous section shows these side parings must preserve the full projective lines through the midpoints of the paired edges, so these descend to intersecting closed geodesics on T. The following argument shows this property remains true in the limit:

Theorem 4.9 *Let $\mathbb{X} \in \{\mathbb{S}^2, \mathbb{E}^2, \mathbb{H}^2\}$ and $\mathbb{X}_t = D_t\mathbb{X}$ be a sequence of conjugate geometries converging to the Heisenberg plane for D_t diagonal. Let T_t be a sequence of $\mathbb{X}_t$ cone tori with at most one cone point converging to some Heisenberg torus T. Then T is a translation torus.*

Proof By Proposition 4.6 we may represent these structures by a sequence of $\mathbb{X}_t$ parallelograms (Q_t, A_t, B_t) converging to the triple $(Q_\infty, A_\infty, B_\infty)$ describing the Heisenberg torus T.

Claim 2 of the previous section implies that for each t, the side pairing A_t preserves the projective line α_t connecting the $\mathbb{X}_t$ midpoints of the paired sides. As $t \to \infty$ this sequence of lines in $\mathbb{R}P^2$ subconverges to a projective line α_∞. Since $A_t(\alpha_t) = \alpha_t$ for all t, it follows that $A_\infty(\alpha_\infty) = \alpha_\infty$, so this line is preserved by the limiting action. By Claim 3, α_∞ passes through the Euclidean midpoints of opposing sides of Q_∞. Thus α_∞ and β_∞ descend to closed geodesics on T.

As α_t and β_t intersect ∂Q_t in the $\mathbb{X}_t$ midpoints of opposing sides, they divide Q_t into four congruent quadrilaterals. Thus the lines α_t and β_t intersect at the center of mass of Q_t. It follows that in the limit the lines α_∞ and β_∞ intersect at the center of Q_∞, and the closed geodesics on T given by the projections of α_∞ and β_∞ intersect. As T has intersecting geodesics, T cannot have any shears in its holonomy, and thus is a translation torus. □

$$\begin{tikzcd}
 & & & & T^2 \arrow[dlll] \arrow[d] \arrow[drr] \arrow[ddrrrr, bend left] & & & & \\
 & K \arrow[dl] & & & \mathrm{Cyl} \arrow[dllll] \arrow[dll] \arrow[d] & & \mathbb{S}^2(2,2,2,2) \arrow[dllll] \arrow[dll] \arrow[d] \arrow[drr] & & \\
\mathrm{Mob} & & \mathbb{D}^2(\varnothing,2,2,2,2) & & \mathbb{D}^2(2,2;\varnothing) & & \mathbb{D}^2(2;2,2) & & \mathbb{R}\mathrm{P}^2(2,2)
\end{tikzcd}$$

Figure 6: All Heisenberg orbifolds are finitely covered by a Heisenberg torus, and furthermore all with cone points or corner reflectors are covered by the pillowcase $\mathbb{S}^2(2,2,2,2)$.

It would be interesting to consider the regeneration of shear tori without restricting to a single cone point. In particular, whether a sequence of Euclidean cone tori with two cone points, one of cone angle less than π and the other greater than π, could converge to a Heisenberg shear torus provides an intriguing possibility that is yet unknown to the author. Constructing such examples (or proving the nonexistence thereof) likely requires different techniques than those of Section 4.

Appendix Heisenberg orbifolds

Proposition 3.17 provides a strategy for computing the remaining orbifold deformation spaces: given $\mathcal{D}(\mathcal{Q})$ and a covering map $\mathcal{Q}\to\mathcal{O}$ we identify $\mathcal{D}(\mathcal{O})$ with the collection of all extensions of $\rho\in\mathrm{Hom}(\pi_1(\mathcal{Q}),\mathsf{Heis})/\mathsf{Heis}_+$ to $\pi_1(\mathcal{O})$ up to Heis conjugacy fixing ρ. Figure 6 shows all Heisenberg orbifolds, with arrows representing the finite covers used in the calculation of their deformation spaces.

Recall that a *translation torus* has holonomy acting purely by translations. The Teichmüller space of translation tori is homeomorphic to $\mathbb{R}_+\times\mathbb{S}^1$, and is parametrized by rectangular lattices with ratio of generator lengths in $\mathbb{R}_+$ and angle of first vector $\theta\in\mathbb{S}^1$ with the horizontal. A translation torus is called *axis-aligned* if the holonomy contains a translation along the invariant foliation (up to Heis_0 conjugacy such a structure can actually be assumed to have holonomy generated by translations along the coordinate axes). Within the Teichmüller space $\mathcal{T}_{\mathbb{H}\mathrm{s}^2}(T^2)$, the subset of axis-aligned translation tori is homeomorphic to $\mathbb{R}_+\sqcup\mathbb{R}_+$, corresponding to the points of $\mathcal{F}\cap V(x_1,x_2,y_1y_2)$.

Proposition A.1 *Every Heisenberg structure on the pillowcase* $P=\mathbb{S}^2(2,2,2,2)$ *is uniquely covered by a translation torus, and so* $\mathcal{T}_{\mathbb{H}\mathrm{s}^2}(P)\cong\mathbb{R}\times\mathbb{S}^1$.

Proof The twofold branched cover $T \to \mathbb{S}^2(2,2,2,2) = P$ exhibits $\pi_1(P)$ as a $\mathbb{Z}_2 = \langle r \rangle$ extension of $\pi_1(T) = \langle a, b \rangle$ with $rar = a^{-1}$ and $rbr = b^{-1}$. Thus $\mathcal{D}_{\mathbb{H}\mathrm{s}^2}(P)$ is parametrized by pairs $[\rho, R]$ for R conjugating images under ρ to their inverses. Any orientation-preserving element of order two in Heis is a π–rotation about some point $p \in \mathbb{H}\mathrm{s}^2$. Rotations only conjugate translations to their inverses, so ρ is the holonomy of a translation torus. Given any translation torus, the π–rotation about any point in the plane provides an extension of ρ, and any two are conjugate by conjugacies fixing ρ. Thus restriction provides a bijection from $\mathcal{D}_{\mathbb{H}\mathrm{s}^2}(\mathbb{S}^2(2,2,2,2))$ onto translation tori. $\square$

Proposition A.2 *All Heisenberg cylinders are quotients of an axis-aligned translation torus, or a shear torus with one generator of the holonomy a horizontal translation. Thus $\mathcal{T}_{\mathbb{H}\mathrm{s}^2}(\mathsf{Cyl}) \cong \mathbb{R} \sqcup \mathbb{R}^2$.*

Proof The doubling mirror double of a cylinder is a torus, and the corresponding orbifold cover $T \to \mathsf{Cyl}$ exhibits $\pi_1(\mathsf{Cyl})$ as a $\mathbb{Z}_2 = \langle f \rangle$ extension of $\pi_1(T)$ with $faf = a$ and $fbf = b^{-1}$. Thus $\mathcal{D}_{\mathbb{H}\mathrm{s}^2}(\mathsf{Cyl})$ is parametrized by conjugacy classes of pairs $[\rho, F]$ with $\rho \in \mathcal{D}(T)$ and F satisfying the relations above with respect to $\rho(a)$ and $\rho(b)$. For each ρ with $\rho(a)$ a horizontal translation, there is a one-parameter family of solutions F to the system, all conjugate via conjugacies fixing ρ to a reflection across the horizontal, $\operatorname{diag}\{1,-1,1\}$. Thus there is a unique quotient corresponding to each $\rho \in \mathcal{D}_{\mathbb{H}\mathrm{s}^2}(T)$ with $\rho(a)$ a horizontal translation. If $\rho(a)$ is not a horizontal translation, the system of equations above only has solutions when $\rho \in \mathcal{D}(T)$ is an axis-aligned translation torus with $\rho(a)$ vertical, $\rho(b)$ horizontal and $F = \operatorname{diag}\{-1,1,1\}$. Thus the Teichmüller space consists of the union of the space of axis-aligned tori with all tori having $\rho(a)$, a horizontal translation. The space of tori with $\rho(a)$ horizontal identifies with a slice $\mathbb{R}_+ \times \mathbb{R}$ of $\mathcal{T}_{\mathbb{H}\mathrm{s}^2}(T^2) = \mathbb{R}_+ \times \mathbb{R} \times \mathbb{S}^1$ with fixed $\theta = 0 \in \mathbb{S}^1$, intersecting the space $\mathbb{R}_+ \sqcup \mathbb{R}_+$ of axis-aligned translation tori in one copy of $\mathbb{R}_+$. $\square$

Proposition A.3 *All Heisenberg Klein bottles are quotients of an axis-aligned translation torus, or a shear torus with one generator of the holonomy a horizontal translation. Thus $\mathcal{T}_{\mathbb{H}\mathrm{s}^2}(\mathsf{K}) \cong \mathbb{R} \sqcup \mathbb{R}^2$.*

Proof The Klein bottle K has orientation double cover $T \to K$ corresponding to $\pi_1(K) = \langle x, b \mid xbx^{-1} = b^{-1} \rangle$ with $\pi_1(T) = \langle x^2, b \rangle$, so $\mathcal{D}(K)$ is parametrized by pairs $[\rho, X]$ for $\rho \in \mathcal{D}_{\mathbb{H}\mathrm{s}^2}(T)$ and $X^2 = \rho(a)$ satisfying $X\rho(b)X^{-1}\rho(b) = I$. As orientation-reversing elements of Heis square to translations, $\rho(a) \in \mathsf{Tr}$, and we distinguish two cases depending on the component X lies in.

If $X \in \operatorname{diag}\{-1,1,1\}\mathsf{Heis}_0$ reflects across the vertical and conjugates $\rho(b) \in \mathsf{Heis}_0$ to its inverse, $\rho(b)$ cannot have any vertical translation component, and so preserves the horizontal foliation. As $\rho \in \mathcal{D}_{\mathbb{H}_{\mathbb{S}^2}}(K)$, combining with $\rho(a) \in \mathsf{Tr}$ shows ρ is the holonomy of an axis-aligned translation torus, and there is a unique solution for X up to conjugacy:

$$\tilde{\rho}(X) = \begin{pmatrix} -1 & 0 & 0 \\ 0 & 1 & \frac{1}{2}r \\ 0 & 0 & 1 \end{pmatrix}.$$

If $X \in \operatorname{diag}\{1,-1,1\}\mathsf{Heis}_0$ reflects across the horizontal, the only solutions to $X^2 = \rho(a)$ are horizontal translations, and $\rho(b)$ must not have horizontal translational component. There is a one-parameter family of solutions X to the system, all conjugate, via conjugacies fixing ρ, to a glide reflection across the horizontal:

$$\begin{pmatrix} -1 & 0 & -\frac{1}{2}\lambda \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \qquad \square$$

Corollary A.4 *The space of Möbius bands identifies with the space of Klein bottles or cylinders*, $\mathcal{T}_{\mathbb{H}_{\mathbb{S}^2}}(\mathsf{M}) \cong \mathbb{R} \sqcup \mathbb{R}^2$.

Proof A Heisenberg Möbius band has mirror double a Klein bottle and orientation double cover an annulus, so points of $\mathcal{D}_{\mathbb{H}_{\mathbb{S}^2}}(M)$ correspond to triples $[\rho, F, X]$ for $[\rho, X] \in \mathcal{D}(K)$ and $[\rho, F] \in \mathcal{D}(\mathsf{Cyl})$ satisfying $FX = XF$. Every $\rho \in \mathcal{D}_{\mathbb{H}_{\mathbb{S}^2}}(T)$ that extends to a representation of $\pi_1(\mathsf{Cyl})$ does so uniquely, and also uniquely extends to a representation of $\pi_1(K)$, and so there is a unique Möbius band covered by the torus with holonomy ρ. $\square$

Proposition A.5 *Suppose* $\mathcal{O} \in \{D^2(2,2;\varnothing), \mathbb{D}^2(\varnothing,2,2,2,2), \mathbb{R}P^2(2,2)\}$. *Then each Heisenberg structure on* $\mathcal{O}$ *is the quotient of a unique axis-aligned translation torus. Thus* $\mathcal{T}_{\mathbb{H}_{\mathbb{S}^2}}(\mathcal{O}) \cong \mathbb{R}_+ \sqcup \mathbb{R}_+$.

Proof These three orbifolds are twofold covered by $\mathbb{S}^2(2,2,2,2)$, and thus fourfold covered by translation tori. The orbifolds $\mathbb{D}^2(2,2;\varnothing)$ and $\mathbb{D}^2(\varnothing;2,2,2,2)$ are also covered by the annulus, and the only translation annuli are axis-aligned. Each such axis-aligned torus has a unique $\mathbb{D}^2(2,2;\varnothing)$ and $\mathbb{D}^2(\varnothing;2,2,2,2)$ quotient. The orbifold $\mathbb{R}P^2(2,2)$ arises as a fourfold quotient of the torus by glide reflections x and y such that $\pi_1(T^2) = \langle x^2, y^2 \rangle$. As seen in Proposition A.3, each glide reflection squaring to a generator of $\pi_1(T^2)$ is along an axis of $\mathbb{R}^2$, so in this case the torus cover must be an axis-aligned translation torus. Each such cover admits a unique $\mathbb{R}P^2(2,2)$ quotient. $\square$

Proposition A.6 *The orbifold* $\mathbb{D}^2(2;2,2)$ *has Teichmüller space homeomorphic to* $\mathbb{R} \sqcup \mathbb{R}$.

Proof This orbifold is the quotient of the pillowcase by a reflection passing through two opposing cone points, and thus is fourfold covered by a translation torus. Algebraically this is an extension of $\pi_1(P) = \langle a, b, r \rangle$ by $\langle f \rangle = \mathbb{Z}_2$ satisfying $faf = b$, $fbf = a$ and $frf = r^{-1}$. Up to Heis_+ conjugacy we may choose representations for homothety classes of translation tori translating along $v_\theta = \left(\begin{smallmatrix}\cos\theta \\ \sin\theta\end{smallmatrix}\right)$ and $\lambda v_\theta^\perp = \left(\begin{smallmatrix}-\lambda\sin\theta \\ \lambda\cos\theta\end{smallmatrix}\right)$, uniquely defined for $\theta \in [0,\pi)$ and $\lambda > 0$. The only reflections F representing f are parallel to the x or y axes, so the covering torus T cannot be axis-aligned for this to pass through the cone points of the pillow quotient. For $F \in \mathrm{diag}(-1,1,1)\mathsf{Heis}_0$, computing with the relations shows there is a solution if and only if $\theta \in (0,\pi)$ and $\lambda = \tan\theta$. Similarly, for $F \in \mathrm{diag}(1,-1,1)\mathsf{Heis}_0$, a solution exists for $\theta \in \left(\frac{1}{2}\pi, \pi\right)$ and $\lambda = -\tan\theta$. These solutions are unique up to conjugacy and so $\mathsf{T}_{\mathbb{H}\mathrm{s}^2}(\mathbb{D}^2(2;2,2)) \cong \mathbb{R} \sqcup \mathbb{R}$. $\square$

References

[1] **S A Ballas**, **D Cooper**, **A Leitner**, *Generalized cusps in real projective manifolds: classification*, J. Topol. 13 (2020) 1455–1496 MR Zbl

[2] **O Baues**, *The deformation of flat affine structures on the two-torus*, from "Handbook of Teichmüller theory, IV" (A Papadopoulos, editor), IRMA Lect. Math. Theor. Phys. 19, Eur. Math. Soc., Zürich (2014) 461–537 MR Zbl

[3] **O Baues**, **W M Goldman**, *Is the deformation space of complete affine structures on the 2–torus smooth?*, from "Geometry and dynamics" (J Eells, E Ghys, M Lyubich, J Palis, J Seade, editors), Contemp. Math. 389, Amer. Math. Soc., Providence, RI (2005) 69–89 MR Zbl

[4] **J-P Benzécri**, *Sur les variétés localement affines et localement projectives*, Bull. Soc. Math. France 88 (1960) 229–332 MR Zbl

[5] **M Boileau**, **B Leeb**, **J Porti**, *Geometrization of 3–dimensional orbifolds*, Ann. of Math. 162 (2005) 195–290 MR Zbl

[6] **C Chabauty**, *Limite d'ensembles et géométrie des nombres*, Bull. Soc. Math. France 78 (1950) 143–151 MR Zbl

[7] **D Cooper**, **J Danciger**, **A Wienhard**, *Limits of geometries*, Trans. Amer. Math. Soc. 370 (2018) 6585–6627 MR Zbl

[8] **D Cooper**, **C D Hodgson**, **S P Kerckhoff**, *Three-dimensional orbifolds and cone-manifolds*, MSJ Memoirs 5, Math. Soc. Japan, Tokyo (2000) MR Zbl

[9] **J Danciger**, *Geometric transitions*: *from hyperbolic to AdS geometry*, PhD thesis, Stanford University (2011) MR Available at `https://www.proquest.com/docview/2452509608`

[10] **J Danciger**, *A geometric transition from hyperbolic to anti-de Sitter geometry*, Geom. Topol. 17 (2013) 3077–3134 MR Zbl

[11] **J Danciger**, *Ideal triangulations and geometric transitions*, J. Topol. 7 (2014) 1118–1154 MR Zbl

[12] **J Danciger**, **F Guéritaud**, **F Kassel**, *Geometry and topology of complete Lorentz spacetimes of constant curvature*, Ann. Sci. Éc. Norm. Supér. 49 (2016) 1–56 MR Zbl

[13] **F Fillastre**, **A Seppi**, *Spherical, hyperbolic, and other projective geometries: convexity, duality, transitions*, from "Eighteen essays in non-Euclidean geometry" (V Alberge, A Papadopoulos, editors), IRMA Lect. Math. Theor. Phys. 29, Eur. Math. Soc. (2019) 321–409 MR Zbl

[14] **W M Goldman**, *Geometric structures on manifolds and varieties of representations*, from "Geometry of group representations" (W M Goldman, A R Magid, editors), Contemp. Math. 74, Amer. Math. Soc., Providence, RI (1988) 169–198 MR Zbl

[15] **W M Goldman**, *Locally homogeneous geometric manifolds*, from "Proceedings of the International Congress of Mathematicians, II" (R Bhatia, A Pal, G Rangarajan, V Srinivas, M Vanninathan, editors), Hindustan, New Delhi (2010) 717–744 MR Zbl

[16] **M Heusener**, **J Porti**, **E Suárez**, *Regenerating singular hyperbolic structures from Sol*, J. Differential Geom. 59 (2001) 439–478 MR Zbl

[17] **C D Hodgson**, *Degeneration and regeneration of geometric structures on three-manifolds*, PhD thesis, Princeton University (1986) MR Available at `https://www.proquest.com/docview/303436581`

[18] **A Leitner**, *A classification of subgroups of* $\mathrm{SL}(4,\mathbb{R})$ *isomorphic to* $\mathbb{R}^3$ *and generalized cusps in projective* 3 *manifolds*, Topology Appl. 206 (2016) 241–254 MR Zbl

[19] **T Nagano**, **K Yagi**, *The affine structures on the real two-torus, I*, Osaka Math. J. 11 (1974) 181–210 MR Zbl

[20] **J Porti**, *Regenerating hyperbolic and spherical cone structures from Euclidean ones*, Topology 37 (1998) 365–392 MR Zbl

[21] **J Porti**, *Regenerating hyperbolic cone structures from Nil*, Geom. Topol. 6 (2002) 815–852 MR Zbl

[22] **J Porti**, *Regenerating hyperbolic cone 3–manifolds from dimension* 2, Ann. Inst. Fourier (Grenoble) 63 (2013) 1971–2015 MR Zbl

[23] **W P Thurston**, *The geometry and topology of three-manifolds*, lecture notes, Princeton University (1979) Available at `http://msri.org/publications/books/gt3m`

Department of Mathematics, Stanford University
Stanford, CA, United States

`trettel@stanford.edu`

`www.stevejtrettel.site`

Received: 10 May 2018 Revised: 5 October 2021

Algebraic & Geometric Topology **23**:4 (2023) 1501–1550
DOI: 10.2140/agt.2023.23.1501
Published: 14 June 2023

The realization problem for noninteger Seifert fibered surgeries

AHMAD ISSA
DUNCAN MCCOY

Conjecturally, the only knots in S^3 with noninteger surgeries producing Seifert fibered spaces are torus knots and cables of torus knots. We make progress on the associated realization problem. Let Y be a small Seifert fibered space arising by p/q–surgery on a knot in S^3, where p/q is positive and a noninteger. Let e denote the weight of the central vertex in the minimal star-shaped plumbing that Y bounds. We show that if $e \leq -2$ or $e \geq 3$, then Y can be obtained by p/q–surgery on a torus knot or a cable of a torus knot.

57M25, 57M27

1 Introduction

One of the simplest operations to produce new 3–manifolds is Dehn surgery on a knot K in S^3. Thus, it is natural to consider how certain 3–manifolds may arise by surgery

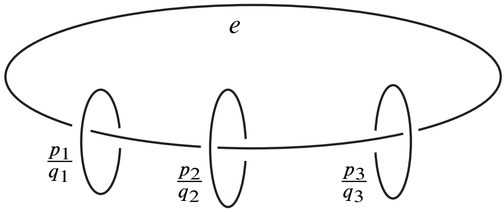


Figure 1: Surgery diagram of the Seifert fibered space $S^2(e; p_1/q_1, p_2/q_2, p_3/q_3)$.

on a knot in S^3. It is, of course, well known that every closed oriented 3–manifold arises by surgery on a *link* in S^3; see Lickorish [18] and Wallace [33]. A natural candidate for studying such questions is the family of Seifert fibred spaces.

Question 1.1 Which Seifert fibered spaces can arise by surgery on a knot in S^3?

As Seifert fibered spaces are not hyperbolic 3–manifolds, this is naturally related to the problem of understanding exceptional surgeries on hyperbolic knots in S^3. One conjecture is the following, which explains why one might consider integer and noninteger Seifert fibered surgeries separately.

Conjecture 1.2 (Gordon [10, Conjecture 4.8]) If $S^3_{p/q}(K)$ is a Seifert fibered space and K is a hyperbolic knot, then $q = 1$.

This has an equivalent formulation which provides a conjectural list of knots in S^3 with noninteger Seifert fibered surgeries; see Proposition 2.2 for a proof of the equivalence.

Conjecture 1.3 If $S^3_{p/q}(K)$ is a Seifert fibered space and $q \geq 2$, then K is a torus knot or a cable of a torus knot.

We consider Question 1.1 for noninteger surgeries and show that for a significant subset of the Seifert fibered spaces the only ones arising by noninteger surgery on a knot in S^3 are the ones predicted by Conjecture 1.3.

Culler, Gordon, Luecke and Shalen's cyclic surgery theorem shows that lens spaces arise by noninteger surgery only on torus knots [4]. Boyer and Zhang have shown that Haken Seifert fibered spaces can arise only by integer surgeries on knots in S^3 [1, Corollary J], a fact that also follows from later work of Gordon and Luecke [11]. Thus it remains to consider noninteger surgeries yielding small Seifert fibered spaces, that is spaces that fiber over S^2 with three exceptional fibers. We use $Y \cong S^2(e; p_1/q_1, p_2/q_2, p_3/q_3)$ to denote the Seifert fibered space obtained according to the surgery diagram in Figure 1.

If Y is a rational homology sphere, then it arises as the boundary of a definite manifold obtained by plumbing sphere bundles according to a star-shaped graph. We define $e(Y) \in \mathbb{Z} \setminus \{0\}$ to be the weight of the central vertex of the unique minimal definite plumbing which Y bounds; see Section 3.1.

Theorem 1.4 *Let Y be a Seifert fibered space over S^2 with three exceptional fibers and $e(Y) \notin \{+1, +2, -1\}$. If there is a knot K in S^3 with $Y \cong S^3_{p/q}(K)$ where $p/q > 0$ and $p/q \in \mathbb{Q} \setminus \mathbb{Z}$, then there is a knot K' which is either a torus knot or a cable of a torus knot with $S^3_{p/q}(K') \cong Y$ and $\Delta_K(t) = \Delta_{K'}(t)$.*

It turns out that the spaces arising in the conclusion of Theorem 1.4 are all L–spaces. Thus, the fact that K and K' have the same Alexander polynomial shows that they have isomorphic knot Floer homology groups; see Ozsváth and Szabó [30]. In order to make full use of Theorem 1.4, one also needs to understand for which surgeries on torus knots or cables of torus knots we have $e(Y) \notin \{+1, +2, -1\}$. Thus, we provide the following result as a companion to Theorem 1.4.

Proposition 1.5 *Let K be a torus knot or a cable of a torus knot. Then for $p/q > 0$ we have that $S^3_{p/q}(K)$ is a Seifert fibered space over S^2 with three exceptional fibers and $e(S^3_{p/q}(K)) \notin \{-1, +1, +2\}$ if and only if*

(i) *K is a torus knot $K = T_{r,s}$ with $r, s > 1$, $p/q > rs - 1$ and $|p - rsq| > 1$, or*

(ii) *K is a cable of a torus knot $K = C_{a,b} \circ T_{r,s}$, where $r, s > 1$, $b/a > rs - 1$ and $p/q = ab \pm 1/q$.*

Since Theorem 1.4 is phrased in terms of positive surgeries, we will reflect Y, if necessary, to assume that it bounds a positive definite plumbing, ie so that $e(Y) \geq 2$. Thus in order to prove Theorem 1.4 we have two possible cases to consider. Either we have $e(Y) = 2$ and $p/q < 0$ or we have that $e(Y) \geq 3$. We deal with these two regimes differently. The main technical content of this paper comes in the analysis of the $e(Y) = 2$ case. The key point is that the definite plumbing bounding a Seifert fibered space is an example of a "sharp" manifold, meaning that, roughly speaking, its intersection form determines the Heegaard Floer d–invariants of its boundary; see Ozsváth and Szabó [29]. This allows us to apply the changemaker lattice surgery obstruction developed by Greene for integer and half-integer surgeries [13; 14] and extended to all noninteger surgeries by Gibbons [7]. This reduces the problem to studying when the intersection form of a star-shaped plumbing can be isomorphic to a changemaker lattice. Almost all previous applications of changemaker lattices have involved studying situations in which changemaker lattices are isomorphic to graph lattices. However, when $e(Y) = 2$ the intersection form of the relevant star-shaped plumbing is not a graph lattice, meaning that new ideas are required to apply the changemaker obstruction. The majority of the technical innovation in this paper comes from circumventing the fact that we are not dealing with a graph lattice. When $e(Y) \geq 3$ the Seifert fibered space is the double branched cover of an alternating Montesinos link. This allows us to apply previous results describing when the double branched cover of an alternating link can arise by noninteger surgery; see McCoy [19]. Although the results of [19] were derived using changemaker lattices, we do not explicitly use lattice theoretic techniques in this part of the proof. We prove the theorem by considering Conway spheres in alternating diagrams of Montesinos links.

The structure of the paper is as follows. We begin in Section 2 by recalling some properties of Seifert fibered surgeries and observing that Conjecture 1.3 is true for surgeries with $q \geq 9$. Sections 3 and 4 contain the necessary background on lattices, with Section 3 discussing the necessary results on the intersection forms of plumbings and Section 4 addressing changemaker lattices. The technical results necessary for the $e(Y) = 2$ case of Theorem 1.4 are developed in Section 5. The $e(Y) \geq 3$ case is studied in Section 6. Finally, in Section 7, we pull together all the necessary results to prove Theorem 1.4 and Proposition 1.5.

Acknowledgements

Issa would like to thank Cameron Gordon for his support, encouragement and helpful conversations. McCoy has been thinking (mostly unsuccessfully) about questions relating to this paper for a number of years. He is grateful to numerous people, most notably Josh Greene and Brendan Owens, for helpful conversations during various incarnations of this project. Both authors would like to thank the anonymous referee for their detailed feedback.

2 Seifert fibered surgeries

In this section we justify the equivalence of Conjecture 1.2 and Conjecture 1.3. We also note that Conjecture 1.2 is true for $q \geq 9$.

Lemma 2.1 *Let K be a knot which is not a torus knot or a cable of a torus knot with a Seifert fibered surgery $S^3_{p/q}(K)$ for some $q \geq 2$. Then there is a hyperbolic knot K' and $q' \geq q$ such that $S^3_{p/q}(K) \cong S^3_{p/q'}(K')$.*

Proof By Thurston's work every knot is either a hyperbolic knot, a satellite knot or a torus knot [31; 32]. Applied to K, this shows that K is a hyperbolic knot or a satellite knot. If K is hyperbolic then we may take $K' = K$ and $q' = q$. Thus suppose that K is a satellite knot. Consider an innermost incompressible torus R in $S^3 \setminus \nu K$. This cuts $S^3 \setminus \nu K$ into two components. One of these is the complement of a knot $K' \subset S^3$ and on the other side of the complement of a knot $C \subseteq S^1 \times D^2$ in a solid torus. The innermost assumption on R implies that K' is either a hyperbolic knot or a torus knot. Since $S^3_{p/q}(K)$ is a small Seifert fibered space [1, Corollary J], it is irreducible and atoroidal. Therefore after performing surgery, the torus R must bound a solid torus. In particular, C must be a knot in $S^1 \times D^2$ with a nontrivial $S^1 \times D^2$ surgery. By the work of Gabai [5, Lemma 2.3], this implies that C is either a torus knot or a 1–bridge braid in the solid torus. However Gabai has also shown that 1–bridge braids admit only integer solid torus surgeries [6, Lemma 3.2]. Thus C must a torus knot in $S^1 \times D^2$. This implies that K is a cable of K'. As we are assuming that K is not a cable of a torus knot, it follows that K' is a hyperbolic knot. Since the torus R bounds a solid torus after performing surgery on C, it follows that $S^3_{p/q}(K) \cong S^3_{p'/q'}(K')$, where p'/q' is the slope on R which bounds a disk after this surgery. By considering how the homology of a solid torus changes under surgery one can see that $p'/q' = p/(qw^2)$, where $w \geq 2$ is the winding number of C [9, Lemma 3.3]. □

This allows us to prove the following two useful results.

Proposition 2.2 *Conjecture 1.2* $\Longleftrightarrow$ *Conjecture 1.3*

Proof The implication Conjecture 1.2 $\Longleftarrow$ Conjecture 1.3 follows from the fact that torus knots and cables of torus knots are not hyperbolic knots. The reverse implication follows from Lemma 2.1, since Conjecture 1.2 asserts that no hyperbolic knot K' satisfying the conclusion of the lemma can exist. □

Proposition 2.3 *If $S^3_{p/q}(K)$ is a Seifert fibered space and $q \geq 9$ then K is a cable of a torus knot or a torus knot.*

Proof Lackenby and Meyerhoff have shown that the distance between exceptional fillings on a hyperbolic knot is eight [17]. Therefore if K' is a hyperbolic knot such that $S^3_{p/q'}(K')$ is a Seifert fibered space, then $q' \leq 8$. Hence the proposition follows from Lemma 2.1. □

3 Seifert fibered spaces and plumbings

We use $S^2(e; p_1/q_1, p_2/q_2, p_3/q_3)$ to denote the space obtained by surgery on the link as in Figure 1, where $e \in \mathbb{Z}$ and for each i the integers p_i and q_i are coprime. This is a Seifert fibered space with three exceptional fibers provided that $|p_i| > 1$ for $i = 1, 2, 3$. By performing Rolfsen twists on the p_i/q_i–framed components, we see that there is an orientation preserving homeomorphism between $S^2(e; p_1/q_1, p_2/q_2, p_3/q_3)$ and $S^2(e'; p'_1/q'_1, p'_2/q'_2, p'_3/q'_3)$ whenever

$$e - \frac{q_1}{p_1} - \frac{q_2}{p_2} - \frac{q_3}{p_3} = e' - \frac{q'_1}{p'_1} - \frac{q'_2}{p'_2} - \frac{q'_3}{p'_3} \tag{3-1}$$

and there is a permutation π of $\{1, 2, 3\}$ such that

$$\frac{q_i}{p_i} \equiv \frac{q'_{\pi(i)}}{p'_{\pi(i)}} \bmod 1 \quad \text{for } i = 1, 2, 3. \tag{3-2}$$

Conversely it follows from the classification of Seifert fibered space (see, for example, the results in [26, Section 5.3]) that conditions (3-1) and (3-2) are, in fact, necessary for there to be an orientation preserving homeomorphism between

$$S^2\left(e; \frac{p_1}{q_1}, \frac{p_2}{q_2}, \frac{p_3}{q_3}\right) \quad \text{and} \quad S^2\left(e'; \frac{p'_1}{q'_1}, \frac{p'_2}{q'_2}, \frac{p'_3}{q'_3}\right).$$

The generalized Euler invariant of $Y \cong S^2(e; p_1/q_1, p_2/q_2, p_3/q_3)$ is defined to be

$$\varepsilon(Y) := e - \frac{q_1}{p_1} - \frac{q_2}{p_2} - \frac{q_3}{p_3}.$$

By the above discussion, one sees that $\varepsilon(Y)$ is a topological invariant. Reversing the orientation on the Seifert fibered space Y yields the Seifert fibered space

$$-Y \cong S^2\left(-e; -\frac{p_1}{q_1}, -\frac{p_2}{q_2}, -\frac{p_3}{q_3}\right).$$

Thus we see that the generalized Euler characteristic satisfies

$$\varepsilon(-Y) = -\varepsilon(Y).$$

Using the surgery description of Y in Figure 1, one finds that the order of its first homology can be calculated as

$$|H_1(Y; \mathbb{Z})| = |(p_1 p_2 p_2)\varepsilon(Y)|.$$

It follows that Y is a rational homology sphere if and only if $\varepsilon(Y) \neq 0$. Thus if Y is a Seifert fibered space rational homology sphere, Y can be oriented so that $\varepsilon(Y) > 0$.

3.1 Minimal definite plumbings

Let Y be a Seifert fibered rational homology sphere with three exceptional fibers oriented so that $\varepsilon(Y) > 0$. The discussion at the start of this section shows that Y has a unique description in the form

$$Y \cong S^2\left(e; \frac{p_1}{q_1}, \frac{p_2}{q_2}, \frac{p_3}{q_3}\right),$$

where $e > 0$ and $p_1/q_1, p_2/q_2, p_3/q_3 > 1$. We define $e(Y)$ to be the value e in this presentation with the convention that $e(-Y) = -e(Y)$. This is the invariant $e(Y)$ appearing in the statement of Theorem 1.4. As we will see, this quantity is precisely the weight of the central vertex in the minimal definite plumbing that Y bounds.

There is a unique continued fraction expansion

$$\frac{p_1}{q_1} = [a_1, \dots, a_k]^- = a_1 - \cfrac{1}{a_2 - \cfrac{1}{\ddots\; a_{k-1} - \cfrac{1}{a_k}}},$$

where $k \geq 1$ and $a_j \geq 2$ for $j \in \{1, \dots, k\}$. Similarly, we write $p_2/q_2 = [b_1, \dots, b_l]^-$ and $p_3/q_3 = [c_1, \dots, c_m]^-$, where $l, m \geq 1$, and $b_j \geq 2$ and $c_j \geq 2$ for all j. Performing a sequence of reverse slam dunks to convert the fractional surgery coefficients to integer coefficients, we see that Y has a surgery description as shown in Figure 2. Since these surgery coefficients are integers, this can also be viewed as a Kirby diagram for a 4–manifold X with $\partial X = Y$. This manifold is diffeomorphic to one obtained by plumbing disk bundles over S^2 according to the star-shaped graph given in Figure 3; see Section 6.1 of [8]. This X is precisely the unique minimal positive definite plumbing that Y bounds [25, Theorem 5.2]. Given the plumbing diagram as in Figure 3, we can define an integer lattice $(\Lambda_\Gamma, Q_\Gamma)$, where Λ_Γ is the free abelian group generated by the vertices of Γ and $Q_\Gamma\colon \Lambda_\Gamma \times \Lambda_\Gamma \to \mathbb{Z}$ is the bilinear pairing with

$$Q_\Gamma(u, v) = \begin{cases} \mathrm{w}(v) & \text{if } u = v, \\ -1 & \text{if vertices } u \text{ and } v \text{ are connected by an edge}, \\ \ \ 0 & \text{otherwise}, \end{cases}$$

where u and v are vertices of Γ and $\mathrm{w}(v)$ denotes the weight of vertex v. The lattice $(\Lambda_\Gamma, Q_\Gamma)$ is naturally isomorphic to the intersection form of X, and hence is positive definite. We write $x \cdot y$ to denote the pairing $Q_X(x, y)$ and $\|x\|^2$ to denote $Q_X(x, x)$.

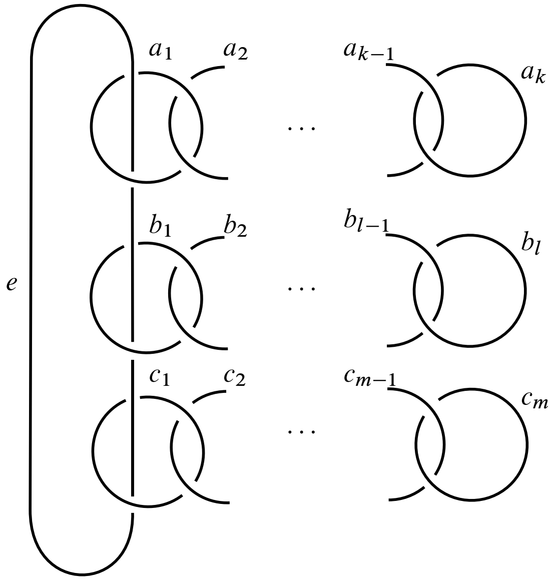


Figure 2: The Kirby diagram for X (also a surgery diagram for $\partial X = Y$).

3.2 Quasialternating plumbings

In order to prove Theorem 1.4 we need to understand the properties of lattices arising as the intersection forms in the case $e = 2$. For topological reasons we need only consider a special subset of such forms. The following was proven by the first author in his classification of quasialternating Montesinos links [15].

Lemma 3.1 *Let $Y = S^2(e; p_1/q_1, p_2/q_2, p_3/q_3)$ with $e \geq 2$ be such that Y is the boundary of the* (*canonical*) *positive definite plumbing* 4*–manifold X. Then the following are equivalent*:

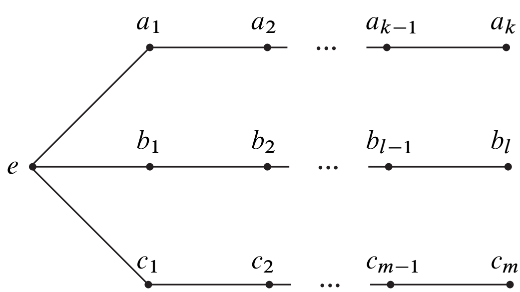


Figure 3: Weighted star-shaped plumbing graph Γ.

(i) *Y bounds a negative definite 4–manifold W with $H_1(W)$ torsion free.*

(ii) *Y is homeomorphic to the double branched cover of a quasialternating Montesinos link.*

(iii) *Either $e \geq 3$, or $e = 2$ and $q_i/p_i + q_j/p_j < 1$ for some $i, j \in \{1, 2, 3\}$ with $i \neq j$.*

(iv) *If A is a matrix representing some embedding $H_2(X) \hookrightarrow \mathbb{Z}^n$ with $n \in \mathbb{Z}_{>0}$ of the intersection lattice of X into a standard positive diagonal lattice with respect to a pair of bases, then A^T is surjective.*

On account of the condition Lemma 3.1(ii):

Definition 3.2 Let Γ be a star-shaped plumbing graph as in Figure 3. We say that Γ is *quasialternating* if $e = 2$ and the continued fractions

$$\frac{p_1}{q_1} = [a_1, \dots, a_k]^-$$

and

$$\frac{p_2}{q_2} = [b_1, \dots, b_l]^-$$

satisfy $q_1/p_1 + q_2/p_2 < 1$. We also call the corresponding lattice Λ_Γ *quasialternating*.

In order to study quasialternating lattices, it will be convenient to define the following quadratic form:

Definition 3.3 Suppose $k > 0$ and $n_1, \dots, n_k \in \mathbb{Z}$. We denote by $Q_{n_1,\dots,n_k}$ the quadratic form given by

$$Q_{n_1,\dots,n_k}(x_1, \dots, x_k) = n_1 x_1^2 - 2x_1x_2 + n_2x_2^2 - \cdots - 2x_{k-1}x_k + n_kx_k^2 \tag{3-3}$$

for all $x_1, \dots, x_k \in \mathbb{Z}$.

We will begin by proving some preparatory inequalities on quadratic forms of this type.

Lemma 3.4 *Let $c_1, \dots, c_m \geq 2$ be integers and $z_1, \dots, z_m$ be integers. We have the following inequalities:*

(i) *If at least one z_i is nonzero, then*

$$Q_{c_1,\dots,c_m}(z_1, \dots, z_m) \geq 2 + \sum_{i=1}^{m} (c_i - 2)|z_i|.$$

(ii) $$Q_{1,c_1,\dots,c_m}(c,z_1,\dots,z_m)\geq |c|+\sum_{i=1}^{m}(c_i-2)|z_i|.$$

(iii) *If* $c\neq 0$ *or* $z_i\neq 0$ *for some* i, *then*
$$Q_{1,c_1,\dots,c_m}(c,z_1,\dots,z_m)+|c|\geq 2+\sum_{i=1}^{m}(c_i-2)|z_i|.$$

Proof We prove (i) first. Since $c_i\geq 2$ for all $1\leq i\leq m$, we can complete the square to obtain
$$Q_{c_1,\dots,c_m}(z_1,\dots,z_m)=z_1^2+(z_1-z_2)^2+\dots+(z_{m-1}-z_m)^2+z_m^2+\sum_{i=1}^{m}z_i^2(c_i-2).$$
If z_i is nonzero for some i, then at least two of the terms
$$z_1^2,(z_1-z_2)^2,\dots,(z_{m-1}-z_m)^2,z_m^2$$
must be nonzero. Since these terms are all integers, this gives the desired inequality when combined with the previous equation.

Now we prove (ii) and (iii). Since $c_i\geq 2$ for all $1\leq i\leq m$, we can complete the square to obtain

$$\begin{aligned}(3\text{-}4)\quad & Q_{1,c_1,\dots,c_m}(c,z_1,\dots,z_m)\\ &=(c-z_1)^2+(z_1-z_2)^2+\dots+(z_{m-1}-z_m)^2+z_m^2+\sum_{i=1}^{m}z_i^2(c_i-2).\end{aligned}$$

However, notice that we have
$$\begin{aligned}(c-z_1)^2&+(z_1-z_2)^2+\dots+(z_{m-1}-z_m)^2+z_m^2\\ &\geq |c-z_1|+|z_1-z_2|+\dots+|z_{m-1}-z_m|+|z_m|\\ &\geq |(c-z_1)+\dots+(z_{m-1}-z_m)+z_m|=|c|.\end{aligned}$$
Combining this with (3-4) proves (ii).

To prove (iii) observe that if at least one of $c,z_1,\dots,z_m$ is nonzero then at least two of the terms
$$|c|,(c-z_1)^2,(z_1-z_2)^2,\dots,(z_{m-1}-z_m)^2,z_m^2$$
must be nonzero. Since each of these terms are integers, this gives the desired inequality when combined with (3-4). □

Lemma 3.5 *Let $a_1, \dots, a_k, b_1, \dots, b_l \geq 2$ be integers and let $p_1/q_1 = [a_1, \dots, a_k]^-$ and $p_2/q_2 = [b_1, \dots, b_l]^-$ where $(p_i, q_i) = 1$ and $p_i > q_i \geq 1$ for $i \in \{1, 2\}$. Suppose that $q_1/p_1 + q_2/p_2 < 1$. Then for any integers $x_1, \dots, x_k, y_1, \dots, y_l, c$ with at least one of c or the x_i or y_i nonzero, we have*

$$Q_{a_k,\dots,a_1,1,b_1,\dots,b_l}(x_k, \dots, x_1, c, y_1, \dots, y_l) + |c| \geq 2 + \sum_{i=1}^{k}(a_i - 2)|x_i| + \sum_{i=1}^{l}(b_i - 2)|y_i|. \tag{3-5}$$

Proof First observe that

$$Q_{a_1,\dots,a_k,1,b_1,\dots,b_l}(0, \dots, 0, c, 0, \dots, 0) + |c| = c^2 + |c|.$$

So if $c \neq 0$ and $x_1 = \dots = x_k = y_1 = \dots = y_l = 0$, then

$$Q_{a_k,\dots,a_1,1,b_1,\dots,b_l}(x_k, \dots, x_1, c, y_1, \dots, y_l) \geq 2,$$

which is the desired inequality. Thus we may assume that at least one of the x_i or y_j terms is nonzero.

Consider the integers $x_1, \dots, x_k, y_1, \dots, y_l, c$. The right hand side of (3-5) is invariant under changing the signs of any subset of these integers. Moreover, the left hand side of (3-5) is minimal when all these integers have the same sign, and is invariant under simultaneously replacing all of the integers by their negatives. Hence, it suffices to consider the case $x_1, \dots, x_k, y_1, \dots, y_l, c \geq 0$.

Now consider

$$\begin{aligned} &a_1 x_1^2 - 2x_1 c + c^2 - 2y_1 c + b_1 y_1^2 + c \\ &\quad = (a_1 - 1)x_1^2 - 2x_1 y_1 + (b_1 - 1)y_1^2 + x_1 + y_1 + \left(x_1 + y_1 - c - \tfrac{1}{2}\right)^2 - \tfrac{1}{4} \\ &\quad \geq (a_1 - 1)x_1^2 - 2x_1 y_1 + (b_1 - 1)y_1^2 + x_1 + y_1, \end{aligned} \tag{3-6}$$

where the inequality follows from the observation that the square of a half integer is always at least a quarter. It follows from (3-6) that

$$Q_{a_k,\dots,a_1,1,b_1,\dots,b_l}(x_l, \dots, x_1, c, y_1, \dots, y_l) + |c| \geq Q_{a_k,\dots,a_1-1,b_1-1,\dots,b_l}(\dots, x_1, y_1, \dots) + |x_1| + |y_1|, \tag{3-7}$$

where we are using the positivity assumption to write $|x_1|$, $|y_1|$ and $|c|$ in place of x_1, y_1 and c.

We will use (3-7) to prove (3-5) by induction.

Note that $q_1/p_1 + q_2/p_2 < 1$ implies that at most one of a_1 and b_1 can equal two.

If $a_1 > 2$ and $b_1 > 2$, then Lemma 3.4(i) applies to show that

$$Q_{a_k,\dots,a_1-1,b_1-1,\dots,b_l}(x_l,\dots,x_1,y_1,\dots,y_l) \\ \geq 2+\sum |x_i|(a_i-2)-x_1+\sum |y_i|(b_i-2)-y_1.$$

Combining this with (3-7) gives the desired inequality.

Thus it suffices to consider the possibility that $a_1 = 2$ or $b_1 = 2$. Without loss of generality we can assume that $a_1 = 2$. If $k = 1$, then Lemma 3.4(iii) combined with (3-7) gives the desired bound.

Thus, it remains to consider the case that $a_1 = 2$ and $k > 1$. Let

$$\frac{p_1'}{q_1'} = [a_2,\dots,a_k]^- \quad \text{and} \quad \frac{p_2'}{q_2'} = [b_1-1,b_2,\dots,b_l]^-.$$

We wish to show that these satisfy $q_1'/p_1' + q_2'/p_2' < 1$. Since $a_1 = 2$, we have that $p_1/q_1 = 2 - q_1'/p_1'$. We also have that $p_2'/q_2' = p_2/q_2 - 1$. The condition that $q_1/p_1 + q_2/p_2 < 1$ implies that $p_1/q_1 > p_2/p_2 - q_2$. Thus we see that

$$\frac{q_1'}{p_1'} + \frac{q_2'}{p_2'} = 2 - \frac{p_1}{q_1} + \frac{q_2}{p_2-q_2} < 2 - \frac{p_2}{p_2-q_2} + \frac{q_2}{p_2-q_2} = 1,$$

as required.

This allows us to prove the lemma inductively, by considering

$$Q_{a_k,\dots,a_2,1,b_1-1,\dots,b_l}(x_k,\dots,x_1,y_1,\dots,y_l),$$

with x_1 taking the role of c. □

With these inequalities in place, we can prove our key result on quasialternating lattices:

Lemma 3.6 *Let Λ be a quasialternating lattice associated to a graph Γ and let $V \subseteq \Lambda$ be the basis elements corresponding to the vertices of Γ. Then for any nonzero $x = \sum_{v\in V} c_v v$, we have*

$$\|x\|^2 \geq 2 + \sum_{v\in V} |c_v|(\|v\|^2 - 2).$$

Proof Suppose that Λ is the lattice corresponding to the star-shaped plumbing in Figure 3 with $e = 2$, and

$$\frac{p_1}{q_1} = [a_1,\dots,a_k]^-, \quad \frac{p_2}{q_2} = [b_1,\dots,b_l]^- \quad \text{and} \quad \frac{p_3}{q_3} = [c_1,\dots,c_m]^-,$$

where $a_i, b_i, c_i \geq 2$ and $q_1/p_1 + q_2/p_2 < 1$. Thus if we take

$$c_v = \begin{cases} x_i & \text{if } v \text{ is the } a_i\text{–weighted vertex,} \\ y_i & \text{if } v \text{ is the } b_i\text{–weighted vertex,} \\ z_i & \text{if } v \text{ is the } c_i\text{–weighted vertex,} \\ c & \text{if } v \text{ is the central vertex,} \end{cases}$$

then it is not hard to verify that $\|x\|^2$ can be calculated as

$$\|x\|^2 = Q_{a_k,\dots,a_1,1,b_1,\dots,b_l}(x_k,\dots,x_1,c,y_1,\dots,y_l) + Q_{1,c_1,\dots,c_m}(c,z_1,\dots,z_m).$$

If $c = 0$, then this simplifies to

$$\|x\|^2 = Q_{a_k,\dots,a_1}(x_k,\dots,x_1) + Q_{b_1,\dots,b_l}(y_1,\dots,y_l) + Q_{c_1,\dots,c_m}(z_1,\dots,z_m).$$

In this case the required inequality follows from Lemma 3.4(i).

Thus it suffices to suppose that $c \neq 0$. In this case, we can apply Lemma 3.4(ii) to the second summand of the first equation for $\|x\|^2$ above. This gives

$$\|x\|^2 \geq Q_{a_k,\dots,a_1,1,b_1,\dots,b_l}(x_k,\dots,x_1,c,y_1,\dots,y_l) + |c| + \sum_{i=1}^{m} |z_i|(c_i - 2).$$

By applying Lemma 3.5, we get the desired inequality. □

Lemma 3.6 has several consequences that will be of use later. To describe these consequences we need the following lattice-theoretic concepts:

Definition 3.7 Let Λ be an integer lattice and let $v \in \Lambda$.

- The vector v is *irreducible* if, for all $x, y \in \Lambda$, if $v = x + y$ and $x \cdot y \geq 0$ then either $x = 0$ or $y = 0$.
- The vector v is *unbreakable* if, for all $x, y \in \Lambda$, if $v = x + y$ and $x \cdot y = -1$ then either $\|x\|^2 = 2$ or $\|y\|^2 = 2$.

Lemma 3.8 *Let Λ be a quasialternating lattice associated to a graph Γ and let $V \subseteq \Lambda$ be the basis elements corresponding to the vertices of Γ. Then*:

(i) *If $x \in \Lambda$ is nonzero, then $\|x\|^2 \geq 2$.*

(ii) *If $x = \sum_{v \in V} c_v v$, then if $c_w \neq 0$ for some $w \in V$, then $\|x\|^2 \geq \|w\|^2$.*

(iii) *Any vertex $v \in V$ is irreducible.*

(iv) *Any vertex $v \in V$ is unbreakable.*

Proof The statements (i) and (ii) follow immediately from Lemma 3.6.

Suppose that a vertex v can be written as $v = x + y$ for $x, y \in \Lambda$. If we write $x = \sum c_w w$ and $y = \sum d_w w$, then since the vertices are a basis for Λ, we see that we must have $c_v \neq 0$ or $d_v \neq 0$. Without loss of generality assume that $c_v \neq 0$. Thus by (ii), $\|x\|^2 \geq \|v\|^2$. However, we also have

$$\|v\|^2 = \|x + y\|^2 = \|x\|^2 + 2(x \cdot y) + \|y\|^2,$$

showing that

$$0 \leq \|y\|^2 \leq -2(x \cdot y).$$

Thus if $x \cdot y \geq 0$, then $\|y\|^2 = 0$ implying that $y = 0$. This shows irreducibility. If $x \cdot y = -1$, then $y \neq 0$ and $\|y\|^2 \leq 2$. By (i) this means $\|y\|^2 = 2$. Thus we have shown unbreakability. □

The following observation will also be useful.

Lemma 3.9 *Suppose Λ is a quasialternating lattice with the vertex basis V. If $x = \sum_{v \in V} c_v v \in \Lambda$ is irreducible, then we have $c_v \geq 0$ for all v or $c_v \leq 0$ for all v.*

Proof Let $P = \{v \in V : c_v > 0\}$ and $N = \{v \in V : c_v < 0\}$ and let $w_+ = \sum_{v \in P} c_v v$ and $w_- = \sum_{v \in N} c_v v$. We have $x = w_+ + w_-$ and $w_+ \cdot w_- \geq 0$. Since x is irreducible this implies that $x = w_+$ or $x = w_-$, proving that the c_v must all have the same sign, as required. □

4 Changemaker lattices

In this section we recall the changemaker theorem and the properties of changemaker lattices. The changemaker theorem was first developed by Greene for integer surgeries in his work on the lens space realization problem [12] and the cabling conjecture [14], and for half-integer surgeries in his work on 3–braid knots with unknotting number one [13]. It was extended to general noninteger slopes by Gibbons [7]. A proof of the changemaker theorem at the level of generality stated here can be found in the second author's thesis [20].

The changemaker theorems are obstructions to manifolds arising by positive surgery and bounding sharp negative definite manifolds. Recall that given a negative-definite manifold X with $\partial X = Y$ equipped with a spin^c–structure $\mathfrak{s}$ which restricts to $\mathfrak{t}$ on Y, there is an upper bound [28]:

$$d(Y, \mathfrak{t}) \geq \tfrac{1}{4}(c_1(\mathfrak{s})^2 + b_2(X)). \tag{4-1}$$

Here $d(Y, \mathfrak{t})$ denotes the d–invariant from Heegaard Floer homology. A sharp manifold is one for which (4-1) is sufficient to determine all d–invariants on the boundary.

Definition 4.1 A negative definite manifold X with boundary Y is sharp if for every $\mathfrak{t} \in \mathrm{Spin}^c(Y)$ there is $\mathfrak{s} \in \mathrm{Spin}^c(X)$ such that $\mathfrak{s}$ restricts to $\mathfrak{t}$ and $\mathfrak{s}$ attains equality in (4-1), that is,

$$d(Y, \mathfrak{t}) = \tfrac{1}{4}(c_1(\mathfrak{s})^2 + b_2(X)).$$

Definition 4.2 We say that a tuple of increasing positive integers $(\sigma_1, \dots, \sigma_t)$ satisfies the *changemaker condition* if, for every

$$1 \le n \le \sigma_1 + \dots + \sigma_t,$$

there is $A \subseteq \{1, \dots, t\}$ such that $n = \sum_{i \in A} \sigma_i$.

The changemaker has an equivalent formulation which will sometimes be useful:

Proposition 4.3 (Brown, [2]) *A tuple $(\sigma_1, \dots, \sigma_t)$ of increasing positive integers satisfies the changemaker condition if and only if*

$$\sigma_1 = 1 \qquad \text{and} \qquad \sigma_i \le \sigma_1 + \dots + \sigma_{i-1} + 1 \quad \text{for } 1 < i \le t.$$

The key definition we will need is that of a changemaker lattice.

Definition 4.4 Let $p/q > 0$ be given by the continued fraction

$$p/q = [a_0, \dots, a_l]^- = a_0 - \cfrac{1}{a_1 - \cfrac{1}{\ddots - \cfrac{1}{a_l}}},$$

where $a_0 \ge 1$, and $a_i \ge 2$ for $i \ge 1$. Suppose further that $\{f_0, \dots, f_s, e_1, \dots, e_t\}$ is an orthonormal basis of $\mathbb{Z}^{t+s+1}$, where $s = \sum_{i=1}^{l}(a_i - 1)$. Let $w_0, \dots, w_l \in \mathbb{Z}^{s+t+1}$ be such that:

(I) w_0 has norm $\|w_0\|^2 = a_0$ and takes the form

$$w_0 = \begin{cases} \sigma_1 e_1 + \dots + \sigma_t e_t & \text{if } l = 0, \\ f_0 + \sigma_1 e_1 + \dots + \sigma_t e_t & \text{if } l > 0, \end{cases}$$

where $(\sigma_1, \dots, \sigma_t)$ is a tuple satisfying the changemaker condition.

(II) For $k \geq 1$,

$$w_k = -f_{\alpha_{k-1}} + f_{\alpha_{k-1}+1} + \cdots + f_{\alpha_k},$$

where $\alpha_0 = 0$ and $\alpha_k = \sum_{i=1}^{k}(a_i - 1)$.

Then we say that the orthogonal complement

$$L = \langle w_0, \ldots, w_l \rangle^\perp \subseteq \mathbb{Z}^{s+t+1}$$

is a *p/q–changemaker lattice*.

Moreover, we say that the σ_i are the *changemaker coefficients* of L and that the σ_i satisfying $\sigma_i > 1$ are the *stable coefficients* of L.

Some remarks on this definition are in order.

Remark 4.5 (1) As the α_i are defined so that $\alpha_i - \alpha_{i-1} = a_i - 1$, the w_i satisfy

$$w_i \cdot w_j = \begin{cases} a_i & \text{if } i = j, \\ -1 & \text{if } |i-j| = 1, \\ 0 & \text{if } |i-j| > 1. \end{cases}$$

(2) By definition, we have $\alpha_l = s$. Thus for every $0 \leq j \leq s$ there is w_k with $w_k \cdot f_j = 1$. As $w_0 \cdot e_i > 0$ for every $1 \leq i \leq t$, this shows that there are no vectors of norm one in a changemaker lattice.

(3) A p/q–changemaker lattice is determined up to isomorphism by its stable coefficients. Given the stable coefficients, the remaining changemaker coefficients are all equal to one and the number of remaining coefficients is determined by the requirement that $a_0 = \|w_0\|^2 = \lceil p/q \rceil$. All other w_i are determined by the continued fraction expansion for p/q.

We are now ready to state the changemaker surgery obstruction.

Theorem 4.6 [20, Theorem 2.1] *Let $K \subseteq S^3$ be such that $S^3_{p/q}(K)$ bounds a sharp manifold X for $p/q > 0$. Then the intersection form Q_X satisfies*

$$-Q_X \cong L \oplus \mathbb{Z}^S,$$

where $S \geq 0$ is an integer and

$$L = \langle w_0, \ldots, w_l \rangle^\perp \subseteq \mathbb{Z}^{s+t+1}$$

is a p/q–changemaker lattice such that, for all $0 \le i \le n/2$,

$$8V_i = \min_{\substack{|c\cdot w_0|=n-2i \\ c\in \mathrm{Char}(\mathbb{Z}^{s+t+1})}} \|c\|^2 - (s+t+1), \tag{4-2}$$

where $n = \lceil p/q \rceil$. □

Here the V_i are a nonincreasing sequence of nonnegative integers that are determined by the knot Floer complex CFK^∞ of K.

Remark 4.7 It is clear from (4-2) that the vector w_0 determines the V_i. It turns out that the sequence of V_i, along with (4-2), is sufficient to determine the stable coefficients of w_0 [21]. In particular, this means that the intersection form Q_X is determined by the knot, the surgery slope p/q and the second Betti number of X.

In the case where K is an L–space knot (a knot with positive L–space surgeries) the V_i can be computed from the Alexander polynomial. For an L–space knot we may write its Alexander polynomial in the form

$$\Delta_K(t) = a_0 + \sum_{i=1}^{g} a_i(t^i + t^{-i}),$$

where $g = g(K)$ is the genus of K and the nonzero values of the a_i alternate in sign and take values $a_i = \pm 1$. We also assume that $\Delta_K(1) = 1$. With these conventions, we define the torsion coefficients of $\Delta_K(t)$ to be

$$t_i(K) = \sum_{j \ge 1} j a_{|i|+j}.$$

For K an L–space knot we have that $V_i = t_i(K)$.

Remark 4.8 The torsion coefficients are sufficient to determine the Alexander polynomial. For $j \ge 1$, we can recover a_j by the relation

$$a_j = t_{j-1}(K) - 2t_j(K) + t_{j+1}(K).$$

Since we are normalizing so that $\Delta_K(1) = 1$, this is also sufficient to recover a_0.

When applied to Seifert fibered surgeries, Theorem 4.6 yields:

Lemma 4.9 *Let $Y = S^2(2; p_1/q_1, p_2/q_2, p_3/q_3)$ be a Seifert fibered space bounding positive-definite plumbed 4–manifold X_Γ such that $Y \cong S^3_{-p/q}(K)$ for some $K \subseteq S^3$ and $p/q > 0$. Then Y is an L–space and $Q_\Gamma \cong L$, where L is the p/q–changemaker lattice determined by the Alexander polynomial of $\Delta_K(t)$.*

Proof Since Y arises by surgery of a negative slope, Y bounds a negative definite manifold W with $H_1(W;\mathbb{Z}) = 0$. Combined with the positive definite plumbing, this shows that Y satisfies condition (i) of Lemma 3.1. Consequently, Y satisfies the other conditions of Lemma 3.1. This shows that Y is the double branched cover of a quasialternating link and, consequently, is an L–space.

Reversing orientations shows that $-Y \cong S^3_{p/q}(\bar{K})$. Ozsváth and Szabó have shown that the negative definite plumbing $-X_\Gamma$ is a sharp 4–manifold [29, Corollary 1.5]. Since the intersection form of $-X_\Gamma$ is isomorphic to $-Q_\Gamma$, Theorem 4.6 applies to show that Q_Γ is isomorphic to $L \oplus \mathbb{Z}^S$ for some $S \geq 0$, where L is the p/q–changemaker lattice whose stable coefficients are determined by the Alexander polynomial of K. However, since Y satisfies the conditions of Lemma 3.1, Lemma 3.8 applies to Q_Γ. This shows in particular that Q_Γ contains no vectors of norm one and hence that $S = 0$, as required. □

4.1 Standard bases

Having stated the changemaker surgery obstruction, we now discuss the properties of changemaker lattices that will be required. We begin first by constructing a basis for a p/q–changemaker lattice; see also [19; 20]. Let

$$L = \langle w_0, \dots, w_l \rangle^{\perp} \subseteq \mathbb{Z}^{s+t+1}$$

be a p/q–changemaker lattice for $p/q = n - r/q$ for $n > 1$ and $1 \leq r < q$. Let

$$w_0 = f_0 + \sigma_1 e_1 + \dots + \sigma_t e_t$$

and $0 = \alpha_0 < \dots < \alpha_l = s$ be as in the definition of L. Consider the set

$$M = \{0, \dots, s\} \setminus \{\alpha_1, \dots, \alpha_{l-1}\}.$$

Write M as

$$M = \{\beta_0, \dots, \beta_m\},$$

where the β_i are ordered to be increasing. Notice that $\beta_0 = 0$ and $\beta_m = \alpha_l = s$. For $0 \leq k < m$ define

$$\mu_k = \begin{cases} f_0 + \dots + f_{\beta_1} & \text{if } k = 0, \\ -f_{\beta_k} + f_{\beta_k+1} + \dots + f_{\beta_{k+1}} & \text{if } k > 0. \end{cases}$$

These are constructed so that $\mu_k \in L$ for $k > 0$. By construction the μ_i pair as follows:

$$\mu_i \cdot \mu_j = \begin{cases} \|\mu_i\|^2 & \text{if } i = j, \\ -1 & \text{if } |i-j| = 1, \\ 0 & \text{if } |i-j| > 1. \end{cases} \tag{4-3}$$

In particular this means, for any $0 \le a \le b \le m$,

$$\|\mu_a + \cdots + \mu_b\|^2 = 2 + \sum_{i=a}^{b} (\|\mu_i\|^2 - 2). \tag{4-4}$$

It will also be useful to note that the μ_i are determined by r/q by the continued fraction identity:

Lemma 4.10 [20, Lemma 4.8] *The μ_i satisfy*

$$[\|\mu_0\|^2, \dots, \|\mu_m\|^2]^- = \frac{q}{q-r}.$$ □

Remark 4.11 Of particular interest will be the cases where $p/q = n - 1/q$ and $p/q = n - (q-1)/q$. In these cases Lemma 4.10 says:

(i) If $p/q = n - 1/q$, then $m = q - 2$ and

$$\|\mu_0\|^2 = \cdots = \|\mu_{q-2}\|^2 = 2.$$

(ii) If $p/q = n - (q-1)/q$, then there is just μ_0 and it satisfies $\|\mu_0\|^2 = q$.

For $1 \le k \le t$, we say that σ_k is *tight* if

$$\sigma_k = 1 + \sigma_1 + \cdots + \sigma_{k-1}.$$

If σ_k is not tight, then Proposition 4.3 shows that there is a subset $A \subseteq \{1, \dots, k-1\}$ such that $\sigma_k = \sum_{i \in A} \sigma_i$. For each k, let A_k denote the maximal such subset with respect to the lexicographical ordering on subsets of $\{1, \dots, k-1\}$. Define ν_k by

$$\nu_k = \begin{cases} -e_k + e_{k-1} + \cdots + e_1 + \mu_0 & \text{if } \sigma_k \text{ is tight}, \\ -e_k + \sum_{i \in A_k} e_i & \text{otherwise}. \end{cases}$$

Note that in any changemaker lattice $\sigma_1 = 1$ is always tight and we have $\nu_1 = -e_1 + \mu_0$. We say that a standard basis element ν_k is *gapless* if it takes the form[1]

$$\nu_k = -e_k + e_{k-1} + \cdots + e_l$$

for some $l < k$.

[1] Such elements were called *just right* by Greene.

Remark 4.12 The lexicographical maximality condition on A_k has the following useful consequences.

(i) For $k > 1$, we always have $\nu_k \cdot e_{k-1} = 1$. When σ_k is tight this is by definition. When σ_k is not tight, Proposition 4.3 shows that we can construct the set A_k by a "greedy algorithm". Under such an algorithm, $k-1$ is the always the first element to be included in A_k.

(ii) If $v \in L$ takes the form
$$v = -e_k + e_{k-1} + \cdots + e_l,$$
then $v = \nu_k$ is necessarily a *gapless* standard basis vector.

We say that
$$S = \{\nu_1, \ldots, \nu_t, \mu_1, \ldots, \mu_m\}$$
is the *standard basis* for L. The standard basis is, in fact, a basis for L.

Lemma 4.13 [20, Proposition 4.9] *The standard basis S is a basis for L.* □

Recall that the notions of irreducibility and unbreakability are given in Definition 3.7.

Lemma 4.14 [20, Lemma 4.13] *Every element $v \in S$ is irreducible.* □

We will also require the following structure result on certain irreducible and unbreakable elements of L. It is an extension of Lemmas 4.16 and 4.17 of [20].

Lemma 4.15 *Let $v \in L$ be irreducible and unbreakable with $v \cdot f_i \neq 0$ for some i.*

(i) *If $v \cdot f_0 = 0$, then v takes the form*
$$\pm v = \mu_a + \cdots + \mu_b,$$
where there is at most one c in the range $a \le c \le b$ with $\|\mu_c\|^2 > 2$.

(ii) *If $v \cdot f_0 \neq 0$, then v takes the form*
$$\pm v = -e_g + e_{k-1} + \cdots + e_1 + \mu_0 + \cdots + \mu_b,$$
where σ_k is tight, $\sigma_g = \sigma_k$ and $\|\mu_i\|^2 = 2$ for $1 \le i \le b$.

Proof Since v is irreducible, it follows from Lemmas 4.16 and 4.17 of [20] that if $v \cdot f_0 = 0$, then v takes the form
$$\pm v = \mu_a + \cdots + \mu_b$$

for $1 \leq a \leq b \leq m$. We claim the unbreakability of v implies that there is at most one $a \leq c \leq b$ with $\|\mu_c\|^2 > 2$. Take c to be minimal such that $\|\mu_c\|^2 > 2$. If $c < b$, then

$$(\mu_a + \cdots + \mu_c) \cdot (\mu_{c+1} + \cdots + \mu_b) = -1.$$

Thus, the unbreakability of v implies that we must have $\|\mu_{c+1} + \cdots + \mu_b\|^2 = 2$, and hence by (4-4) that $\|\mu_{c+1}\|^2 = \cdots = \|\mu_b\|^2 = 2$. Similarly, if $a < c$ then

$$(\mu_a + \cdots + \mu_{c-1}) \cdot (\mu_c + \cdots + \mu_b) = -1,$$

implying that $\|\mu_a\|^2 = \cdots = \|\mu_{c-1}\|^2 = 2$, as required.

Now suppose that $v \cdot f_0 \neq 0$. In this case Lemmas 4.16 and 4.17 of [20] imply that v takes the form

$$\pm v = x_I + x_F,$$

where $x_I \neq 0$ and $x_I \cdot f_i = 0$ for all i and x_F takes the form

$$x_F = \mu_0 + \cdots + \mu_b.$$

Since $\mu_1, \ldots, \mu_b$ are in L, we have $x_I + \mu_0 \in L$. We also have $\|x_I + \mu_0\|^2 > \|\mu_0\|^2 \geq 2$. So by applying the unbreakability condition to $(x_I + \mu_0) \cdot (\mu_1 + \cdots + \mu_b) = -1$ we obtain that

$$\|\mu_1 + \cdots + \mu_b\|^2 = 2.$$

Using (4-4), this implies that

$$\|\mu_1\|^2 = \cdots = \|\mu_b\|^2 = 2,$$

as required.

Now we study the structure of x_I. Let $k \geq 1$ be minimal such that $x_I \cdot e_k \leq 0$. By Proposition 4.3, there is a subset $B \subseteq \{1, \ldots, k-1\}$ such that

$$\sigma_k - 1 = \sum_{i \in B} \sigma_i.$$

Thus we can consider

$$z = -e_k + \sum_{i \in B} e_i + x_F \in L.$$

Note that, by assumption, we have $x_I \cdot e_i \geq 1$ for all $i < k$ and hence for all $i \in B$. Thus we obtain the bound

$$(x_I + x_F - z) \cdot z = -(x_I \cdot e_k + 1) + \sum_{i \in B} (x_I \cdot e_i - 1) \geq -x_I \cdot e_k - 1 \geq -1. \tag{4-5}$$

Thus, by the assumption of irreducibility,

$$(x_I + x_F - z) \cdot z = \begin{cases} 0 & \text{if } z = x_I + x_F, \\ -1 & \text{otherwise.} \end{cases}$$

Suppose first that $z = x_I + x_F$. Since k was chosen to be minimal such that $x_I \cdot e_k \le 0$,

$$x_I + x_F = -e_k + e_{k-1} + \cdots + e_1 + \mu_0 + \cdots + \mu_b,$$

which is in the required form. Thus we can assume that

$$(x_I + x_F - z) \cdot z = -1,$$

which can only occur if $x_I \cdot e_k = 0$. Since $\|z\|^2 > 2$, it follows from the indecomposability condition that $x_I + x_F - z$ has norm two. We have $(x_I + x_F - z) \cdot e_k = -(z \cdot e_k) = 1$. Thus $x_I + x_F - z$ takes the form

$$x_I + x_F - z = e_k - \varepsilon e_g$$

for some $g \neq k$ and some $\varepsilon \in \{\pm 1\}$. The fact that $(x_I + x_F - z) \cdot w_0 = 0$ shows that $\sigma_g = \sigma_k$ and $\varepsilon = 1$. Thus

$$x_I + x_F = z + e_k - e_g$$

for some g with $\sigma_g = \sigma_k$. Since k is minimal with $v \cdot e_k \le 0$, it follows that $g > k$ and

$$x_I + x_F = -e_g + e_{k-1} + \cdots + e_1 + \mu_0 + \cdots + \mu_b,$$

as required. □

Remark 4.16 When rewritten in terms of the orthonormal basis for $\mathbb{Z}^{s+t+1}$ the two types of vector arising in the previous lemma are

$$\mu_a + \cdots + \mu_b = -f_{\beta_a} + f_{\beta_c+1} + \cdots + f_{\beta_{c+1}-1} + f_{\beta_{b+1}}$$

where, if it exists, c is unique in the range $a \le c \le b$ with $\|\mu_c\|^2 > 2$, and

$$-e_g + e_{k-1} + \cdots + e_1 + f_0 + \cdots + f_{\beta_1 - 1} + f_{\beta_{b+1}}.$$

We end with a final useful observation:

Remark 4.17 There is a certain redundancy in the choice of indexing of $f_0, \dots, f_s$ and $e_1, \dots, e_t$. Whenever $\sigma_a = \sigma_b$ for $a \neq b$ (equivalently if $e_a - e_b \in L$), then we can reindex the e_i to exchange e_a and e_b. Similarly given f_a and f_b such that $f_a - f_b \in L \setminus \{0\}$, then we can exchange f_a and f_b. More formally, this is the observation that automorphism of $\mathbb{Z}^{s+t+1}$ exchanging e_a and e_b or f_a and f_b preserves L as subset of $\mathbb{Z}^{s+t+1}$. We will make frequent use of such relabeling in Section 5.

5 Analysis for the $e = 2$ case

Although the formal proof of Theorem 1.4 is stated in Section 7, this section contains the analysis necessary to prove Theorem 1.4 for $e = 2$. The section culminates in Lemma 5.15, which combines with Lemma 4.9 to give the proof.

Let L be a p/q–changemaker lattice

$$L = \langle w_0, \dots, w_l \rangle^{\perp} \subseteq \langle f_0, \dots, f_s, e_1, \dots, e_t \rangle = \mathbb{Z}^{s+t+1}$$

for $q > 1$. Suppose that L is isomorphic to the intersection form of some plumbing Γ (as in Figure 3) with $e = 2$. Let V denote the image of the vertices of Γ in L. In a mild abuse of notation we will simply refer to the elements of V as the vertices of Γ. We seek to understand the structure of V and Γ. The eventual aim is to show that if Y is the Seifert fibered space for which Γ is the canonical plumbing then Y arises by p/q–surgery. In order to do this, we will take L to have standard basis elements

$$\{\nu_1, \dots, \nu_t, \mu_1, \dots, \mu_m\},$$

as defined in Section 4.1.

Key to this section will be the observation that Γ is quasialternating. Consequently the results of Section 3 apply, showing in particular that the vertices are irreducible and unbreakable.

Proposition 5.1 *The plumbing graph Γ is quasialternating.*

Proof Let A be the matrix representing the inclusion $L \to \mathbb{Z}^{s+t+1}$ with respect to the standard basis for L and the orthonormal basis for $\mathbb{Z}^{s+t+1}$. By ordering the basis vectors appropriately A^T takes the form

$$A^T = \begin{pmatrix} \nu_t \cdot f_s & \dots & \nu_t \cdot f_0 & \nu_t \cdot e_1 & \dots & \nu_t \cdot e_t \\ \vdots & & \vdots & \vdots & & \vdots \\ \nu_1 \cdot f_s & \dots & \nu_1 \cdot f_0 & \nu_1 \cdot e_1 & \dots & \nu_1 \cdot e_t \\ \mu_1 \cdot f_s & \dots & \mu_1 \cdot f_0 & \mu_1 \cdot e_1 & \dots & \mu_1 \cdot e_t \\ \vdots & & \vdots & \vdots & & \vdots \\ \mu_m \cdot f_s & \dots & \mu_m \cdot f_0 & \mu_m \cdot e_1 & \dots & \mu_m \cdot e_t \end{pmatrix}.$$

However by definition of the standard basis elements, this matrix is in row echelon form and the first nonzero entry in each row is -1. Consequently A^T is surjective over the integers. This shows that Lemma 3.1(iv) is satisfied. Therefore, Lemma 3.1(iii) applies to show Γ is quasialternating. □

Now we set about understanding the vertices of Γ in L.

Lemma 5.2 *We may assume that $\mu_1, \dots, \mu_m$ are vertices.*

Proof We prove the lemma inductively by establishing that if $\mu_{k+1}, \dots, \mu_m$ are vertices, then we may further assume that μ_k is a vertex.

Since the vertices of Γ span L, there are integers c_v such that $\mu_k = \sum_{v \in \Gamma} c_v v$. For any v with $c_v \neq 0$, Lemma 3.8(ii) shows that $\|v\|^2 \leq \|\mu_k\|^2$. We may write each v as an integer combination of the standard basis elements in a unique way. Thus we see there must be some v with $c_v \neq 0$, for which μ_k appears with nonzero coefficient when v is expressed as an integer combination of standard basis elements. As v is irreducible and unbreakable, Lemma 4.15 combined with the fact that $\|v\|^2 \leq \|\mu_k\|^2$ shows that v takes the form

$$\pm v = \mu_a + \cdots + \mu_b,$$

where $a \leq k \leq b$ and there is at most one c in the range $a \leq c \leq b$ with $\|\mu_c\|^2 \geq 2$. If such a c exists, then we have $k = c$, since $\|v\|^2 = \|\mu_c\|^2 \leq \|\mu_k\|^2$. Thus we have $\|\mu_i\|^2 = 2$ for $a \leq i < k$ and $k < i \leq b$. If $a < k$, then a relabelling of the f_i (the one exchanging the roles of f_{β_a} and f_{β_k}) allows us to assume that $a = k$.

If $k = m$, then we have shown that we can assume $\pm\mu_m$ is a vertex. So, by multiplying all vertices by -1 if necessary, we can assume that μ_m is a vertex. This deals with the base case of the induction.

Thus, suppose that $k < m$. By the previous discussion we can assume there is a vertex v of the form $v = \varepsilon(\mu_k + \cdots + \mu_b)$. One can easily calculate that

$$\mu_i \cdot v = \begin{cases} 0 & \text{if } k < i < b, \\ \varepsilon & \text{if } i = b, \\ -\varepsilon & \text{if } i = b+1, \\ 0 & \text{if } i > b+1. \end{cases} \tag{5-1}$$

Since $\mu_{k+1}, \dots, \mu_m$ form a connected chain of vertices, v can pair nontrivially with at most one of them and this pairing must be -1. Thus it follows from (5-1), that we must have either $b = k$ and $\varepsilon = 1$, or $b = m$ and $\varepsilon = -1$. In the former case we must have $v = \mu_k$ as required. In the latter,

$$v = -(\mu_k + \cdots + \mu_m). \tag{5-2}$$

However in this case we have that

$$\mu_{k+1} = -f_{\beta_{k+1}} + f_{\beta_{k+2}}, \dots, \mu_m = -f_{s-1} + f_s$$

are all of norm two and that

$$v = f_{\beta_k} - f_{\beta_k+1} - \cdots - f_{\beta_{k+1}-1} - f_s.$$

Thus if we relabel the f_i so as to reverse the order of $f_{\beta_{k+1}}, \dots, f_s$, then the set of vertices $\{v, \mu_{k+1}, \dots, \mu_m\}$ becomes $\{-\mu_k, \dots, -\mu_m\}$. Therefore, after multiplying every vertex by -1, we may assume that we have the desired set of vertices.

This verifies the inductive step and completes the proof. □

Lemma 5.3 *Let v be a vertex distinct from $\mu_1, \dots, \mu_m$ with $v \cdot f_i \neq 0$ for some i. Then v takes the form*

$$v = -e_g + e_{k-1} + \cdots + e_1 + \mu_0 \tag{5-3}$$

or

$$v = e_g - e_{k-1} - \cdots - e_1 - \mu_0 - \cdots - \mu_m, \tag{5-4}$$

where $k \leq g$, and this latter case can occur only if

$$\|\mu_1\|^2 = \cdots = \|\mu_m\|^2 = 2.$$

Proof Since every vertex is irreducible and unbreakable, by Lemma 4.15 we see that either v is a linear combination of $\mu_1, \dots, \mu_m$ or it has $v \cdot f_0 \neq 0$. Since the vertices are linearly independent, we must have $v \cdot f_0 \neq 0$. By Lemma 4.15 we may assume that such a vertex takes the form

$$v = \varepsilon(-e_g + e_{k-1} + \cdots + e_1 + \mu_0 + \cdots + \mu_b) \tag{5-5}$$

for some $\varepsilon \in \{\pm 1\}$ and $g \geq k$ with $\sigma_k = \sigma_g$ and σ_k is tight, and $\|\mu_1\|^2 = \cdots = \|\mu_b\|^2 = 2$. Since the μ_i form a linear chain of vertices, we see that v can have nonzero pairing with at most one of them. However, as we have the pairings

$$\mu_i \cdot v = \begin{cases} 0 & \text{if } 0 < i < b, \\ \varepsilon & \text{if } i = b, \\ -\varepsilon & \text{if } i = b+1, \\ 0 & \text{if } i > b+1, \end{cases} \tag{5-6}$$

either $\varepsilon = 1$ and $b = 0$, or $\varepsilon = -1$ and $b = m$. In the $\varepsilon = 1$ and $b = 0$ case, this puts v in the form of (5-3). In the $\varepsilon = -1$ and $b = m$ case, this puts v in the form of (5-4). □

Lemma 5.4 *We may assume that ν_1 is a vertex.*

Proof Expressing ν_1 as a linear combination of vertices, we see that there must be a vertex v with $v \cdot f_0 \neq 0$, and $\|v\|^2 \leq \|\nu_1\|^2 = \|\mu_0\|^2 + 1$ by Lemma 3.8. We see that such a vertex must take either the form

$$v = -e_g + \mu_0, \tag{5-7}$$

coming from (5-3), or the form

$$v = e_g - \mu_0 - \cdots - \mu_m, \tag{5-8}$$

coming from (5-4). In both cases $\sigma_g = \sigma_1 = 1$, and in the latter case $\|\mu_i\|^2 = 2$ for $1 \leq i \leq m$. By relabelling the e_i, we may assume that $g = 1$. Thus there is nothing further to check when v take the form given in (5-7). So suppose that v takes the form given in (5-8). In this case, we apply an argument similar to the one at the end of the proof of Lemma 5.2. We can relabel the f_i so as to reverse the order of $f_{\beta_1}, \dots, f_s$. Under this relabelling the vertices $\mu_1, \dots, \mu_m$ become $-\mu_m, \dots, -\mu_1$ and v becomes $-\nu_1$. Thus by reversing signs on all vertices, we can assume that $\nu_1, \mu_1, \dots, \mu_m$ are all vertices, as required. □

Lemma 5.5 *If $v \notin \{\nu_1, \mu_1, \dots, \mu_m\}$ is a vertex, then either*

(a) $v \cdot e_1 = 0$ *and* $v \cdot f_i = 0$ *for all* $0 \leq i \leq s$, *or*

(b) $v \cdot e_1 = 1$ *and* $v \cdot f_i = 0$ *for all* $0 \leq i \leq s$, *or*

(c) $p/q = n - 1/q$ *and v can be assumed to take the form*

$$v = e_k - e_{k-1} - \cdots - e_1 - \mu_0 - \cdots - \mu_m,$$

where $k > 1$ and σ_k is tight.

Moreover, there is at most one vertex of type (c).

Proof Let $v \neq \nu_1, \mu_1, \dots, \mu_m$ be a vertex with $v \cdot f_i \neq 0$ for some i. By Lemma 5.3, there are two possible forms for v. First assume that v takes the form given in (5-3). In this case, we have $v \cdot \nu_1 \geq \|\mu_0\|^2 - 1 > 0$, which is impossible unless $v = \nu_1$. Thus v must take the form given in (5-4).

If $m = 0$, then

$$v \cdot \nu_1 = -\|\mu_0\|^2 - v \cdot e_1 \in \{-\|\mu_0\|^2 \pm 1, -\|\mu_0\|^2\}.$$

However since v and ν_1 are both vertices, $v \cdot \nu_1 \in \{0, -1\}$. As $\|\mu_0\|^2 \geq 2$, this implies that $v \cdot e_1 = -1$ and $\|\mu_0\|^2 = 2$. This implies that $q = 2$ and $k > 1$ (see Remark 4.11).

If $m > 0$, the argument is similar. Since $\mu_m \cdot v = -1$ and $\nu_1, \mu_1, \dots, \mu_m$ form a linear chain of vertices, we must have $v \cdot \nu_1 = 0$. This implies that

$$v \cdot \nu_1 = -(\|\mu_0\|^2 - 1) - v \cdot e_1 = 0.$$

This shows that $\|\mu_0\|^2 = 2$ and $v \cdot e_1 = -1$. In either case this shows that p/q takes the form $p/q = n - 1/q$ (see Remark 4.11). Since $v \cdot e_1 = -1$, it follows that $k > 1$.

To see that such a v is necessarily unique, suppose that v and w are both vertices of the form given in (5-4). For such vertices we have

$$v \cdot w \geq \|\mu_0 + \dots + \mu_m\|^2 - 1 > 0,$$

which is impossible, unless $v = w$.

Given that such a v is unique and $k > 1$, we see that there is no loss of generality in relabelling the e_i to assume that $g = k$. This shows that v can be taken to be in the form given by (c).

Finally, consider the case that v is a vertex with $v \cdot f_i = 0$ for all i. Since ν_1 is a vertex, we have

$$v \cdot \nu_1 = -v \cdot e_1 \in \{0, -1\}.$$

This shows that v is in the form described by (a) or (b), as required. □

Given a vertex $v \neq \nu_1, \mu_1, \dots, \mu_m$, we refer to it as being of type (a), (b) or (c) if it satisfies conditions (a), (b) or (c) from Lemma 5.5, respectively. This allows us to show that the vertex set satisfies the following trichotomy.

Lemma 5.6 *The vertex set takes one and only one of the following forms*:

(I) *There are no type* (c) *vertices and* ν_1 *is adjacent to a single vertex of type* (b).

(II) $p/q = n - (q-1)/q$ *and* ν_1 *is adjacent to two vertices of type* (b).

(III) $p/q = n - 1/q$ *and there is a unique vertex of type* (c) *and at most one vertex of type* (b).

Proof As we are assuming that the central vertex is of norm two and $\|\nu_1\|^2 > 2$, we see that ν_1 is not the central vertex of Γ. Thus ν_1 pairs with at most two vertices in the graph. Since a vertex of type (b) is always adjacent to ν_1, this shows there are at most two vertices of type (b).

Suppose that the vertex set contains two vertices of type (b). We will show that the vertex set is of type II. Since both vertices of type (b) pair with ν_1, the vertex μ_1 cannot exist and hence p/q takes the form $p/q = n - (q-1)/q$ by Remark 4.11. If the vertex

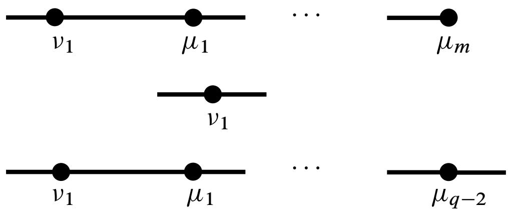


Figure 4: From top to bottom: types I, II and III.

set further contains a vertex of type (c), this would be a third vertex adjacent to ν_1. Thus the vertex set is of type II.

Now suppose that the vertex set contains a vertex of type (c). We will show that the vertex set is of type III. By the argument in the previous paragraph, the vertex set contains at most one vertex of type (b). By Lemma 5.5 the existence of a vertex of type (c) shows that p/q takes the form $p/q = n - 1/q$. Thus the vertex set is of type III.

Finally suppose that the vertex set contains no vertex of type (c) and at most one vertex of type (b). Note that if there is no vertex of type (b) then the graph Γ would have a connected component consisting of the linear chain $\nu_1, \mu_1, \dots, \mu_m$, which is incompatible with our assumptions on Γ. Thus the vertex set contains a unique vertex of type (b) and is hence of type I, as required. □

The local structure of each of these three types is shown in Figure 4. It turns out that a type I vertex set corresponds to surgery on a torus knot. Type II and III vertex sets both correspond to surgery on a cable of a torus knot.

5.1 Type I and II

Now that we understand vertices pairing nontrivially with the f_i, we turn our attention to the remaining vertices. In the case where there are no vertices of type (c), these vertices can be taken to be exactly the standard basis elements.

Lemma 5.7 *If the vertex set of Γ is of type I or II, then we can assume that the vertices are the standard basis elements and are all gapless.*

Proof We prove inductively that we can take the vertices to be standard basis elements. By Lemmas 5.2 and 5.4, we can assume that $\nu_1, \mu_1, \dots, \mu_m$ are vertices. This is the base case.

Now assume that $\mu_1, \dots, \mu_m, \nu_1, \dots, \nu_k$ are all vertices.

Claim *Suppose that v is a vertex which is not one of $\mu_1, \dots, \mu_m, \nu_1, \dots, \nu_k$. Then v has the properties*

(i) $v \cdot f_i = 0$ *for all* i,

(ii) $v \cdot e_i \geq 0$ *for* $1 \leq i \leq k$, *and*

(iii) *if* $v \cdot e_j > 0$ *for some* $j < k$, *then* $v \cdot e_i > 0$ *for all* $j \leq i \leq k$.

Proof By the assumption that there are no type (c) vertices, we have $v \cdot f_i = 0$ for all i. Now suppose that $v \cdot e_i \neq 0$ for some $1 \leq i \leq k$. Let $l \geq 1$ be minimal such that $v \cdot e_l \neq 0$. In this case we have $v \cdot \nu_l = -v \cdot e_l$. As v and ν_l are both vertices then this shows that $v \cdot e_l = 1$. Now let $g > l$ be minimal such that $v \cdot e_g \leq 0$. By Remark 4.12, we have that $\nu_g \cdot e_{g-1} = 1$. Therefore we see that

$$v \cdot \nu_g \geq -v \cdot e_g + v \cdot e_{g-1} > 0.$$

From this we conclude that either $v = \nu_g$ or ν_g is not a vertex. In either case this implies $g > k$. This gives (ii) and (iii). $\square$

Let $v_1, \dots, v_N$, be the vertices which are not already known to be standard basis elements. The preceding claim shows that each v_j can be written as $v_j = v'_j + v^+_j$, where

$$v'_j \cdot e_i = 0 \quad \text{for } i \leq k$$

and

$$v^+_j \cdot e_i \geq 0 \quad \text{for } i \leq k \qquad \text{and} \qquad v^+_j \cdot e_i = 0 \quad \text{for } i > k.$$

Now consider ν_{k+1}. There are integers α_i and β_j such that

$$\nu_{k+1} = \sum_{i=1}^{k} \alpha_i \nu_i + \sum_{j=1}^{N} \beta_j v_j. \tag{5-9}$$

A priori one might expect the μ_i to appear in this sum. However it follows from considering the pairing with the f_i that there is no need to include them. By construction of the standard basis vectors $\nu_{k+1} \cdot f_i$ can be nonzero only if $i \leq \beta_1$. If there were μ_i appearing in the sum (5-9), then we would have $\nu_k \cdot f_i \neq 0$ for some $i > \beta_1$, contradicting this.

Since ν_{k+1} is irreducible, Lemma 3.9 shows that all nonzero α_i and β_j must have the same sign.

Now if we write ν_{k+1} in the form $\nu_{k+1} = -e_{k+1} + \nu^+$, then (5-9) yields

$$\nu^+ = \sum_{i=1}^{k} \alpha_i \nu_i + \sum_{j=1}^{N} \beta_j v_j^+. \tag{5-10}$$

By taking the pairing of (5-10) with w_0 and observing that, by construction, $w_0 \cdot v_j = 0$ for all j, we obtain

$$\sigma_{k+1} = \nu^+ \cdot w_0 = \sum_{j=1}^{N} \beta_j (v_j^+ \cdot w_0). \tag{5-11}$$

Since $\sigma_{k+1} > 0$ and $v_j^+ \cdot w_0 \geq 0$ for all j, this shows that the α_i and β_j must all be nonnegative.

Let $\|x\|_1$ denote the ℓ_1–norm

$$\|x\|_1 = \sum_{i=1}^{t} |x \cdot e_i| + \sum_{j=0}^{s} |x \cdot f_j|.$$

Since the coefficients of ν^+ are equal to 0 or 1, we have $\|\nu^+\|_1 = \|\nu^+\|^2$. However by writing ν^+ as a sum in (5-10) and computing $\|\nu^+\|_1$ we obtain

$$\|\nu^+\|^2 = \sum_{i=1}^{k} \alpha_i (\|\nu_i\|_1 - 2) + \sum_{j=1}^{N} \beta_i \|v_i^+\|_1, \tag{5-12}$$

where the $\|\nu_i\|_1 - 2$ terms come from the fact that $\nu_i \cdot e_i = -1$ and $\nu_i \cdot e_j \geq 0$ for $j \neq i$.

By the inequality in Lemma 3.6 we have the bound

$$\begin{aligned} \|\nu_{k+1}\|^2 &= \|\nu^+\|^2 + 1 \\ &\geq 2 + \sum_{i=1}^{k} \alpha_i (\|\nu_i\|^2 - 2) + \sum_{j=1}^{N} \beta_j (\|v_j\|^2 - 2) \\ &= 2 + \sum_{i=1}^{k} \alpha_i (\|\nu_i\|_1 - 2) + \sum_{j=1}^{N} \beta_j (\|v_j^+\|^2 + \|v_j'\|^2 - 2) \\ &= \|\nu^+\|^2 + 2 + \sum_{j=1}^{N} \beta_j (\|v_j^+\|^2 - \|v_j^+\|_1 + \|v_j'\|^2 - 2), \end{aligned} \tag{5-13}$$

where (5-12) was used to obtain the last line. Comparing the first and last lines in (5-13) shows that

$$\sum_{j=1}^{N} \beta_i (\|v_i^+\|^2 - \|v_i^+\|_1 + \|v_i'\|^2 - 2) \leq -1. \tag{5-14}$$

Since there must be at least one negative summand on the left hand side of (5-14), we can assume that

$$\|v_1^+\|^2 - \|v_1^+\|_1 + \|v_1'\|^2 \le 1 \quad \text{and} \quad \beta_1 \ge 1.$$

Since $\|v_1'\|^2 \ge 1$ and $\|v_1^+\|^2 \ge \|v_1^+\|_1$, we must have $\|v_1'\|^2 = 1$ and $\|v_1^+\|^2 = \|v_1^+\|_1$. However $\|v_1^+\|^2 = \|v_1^+\|_1$ only if $v_1^+ \cdot e_j \in \{0, \pm 1\}$ for all j. By the restrictions on v_1 proven in the claim at the start of the proof, this shows that v_1 takes the form

$$v_1 = -e_g + e_k + \cdots + e_l \tag{5-15}$$

for some $g > k$ and $l \le k$.

Since $g > k$ we have that $\sigma_g \ge \sigma_{k+1}$. On the other hand, the condition $v_1 \cdot w_0 = 0$ implies that $v_1^+ \cdot w_0 = e_g \cdot w_0 = \sigma_g$. Furthermore, computing as in (5-11) and using the fact that $\beta_1 \ge 1$ and that $\beta_j(v_j^+ \cdot w_0) \ge 0$ for each j shows that

$$\sigma_{k+1} = \sum_{i=1}^{N} \beta_i (v_j^+ \cdot w_0) \ge v_1^+ \cdot w_0.$$

Thus we have $\sigma_{k+1} = \sigma_g$. By relabelling we can assume that $v_1 = -e_{k+1} + e_k + \cdots + e_l$. As mentioned in Remark 4.12 it follows that $\nu_{k+1} = v_1$ is a gapless standard basis vector. Thus we have shown we may assume that ν_{k+1} is vertex. This completes the inductive step of the proof. □

This has several useful consequences.

Remark 5.8 Suppose that Γ is a plumbing whose intersection form is isomorphic to a p/q–changemaker lattice L with type I or type II vertex set.

(i) Since the vertices can be taken to be standard basis elements of L, the plumbing graph Γ is completely determined by L.

(ii) L can have no tight standard basis elements except ν_1. Since a type III vertex set implies the existence of a tight standard basis element, this shows that the type of vertex set is intrinsic to the lattice L rather than the plumbing Γ or the choice of vertex set.

(iii) Since there can be no tight standard basis elements we have $\nu_2 = -e_2 + e_1$ as one type (b) vertex. In the type II case the other type (b) vertex must take the form

$$-e_g + e_{g-1} + \cdots + e_1,$$

for some $g > 1$. This shows that Γ takes the form shown in Figure 5. Recall that in the type II case there is no vertex μ_1; see Remark 4.11(ii).

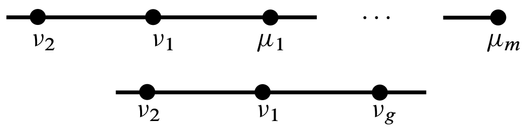


Figure 5: Further structure of Γ in the type I (top) and II (bottom) cases.

We now show that under some circumstances the converse to Remark 5.8(i) holds. This will be useful for recovering the Alexander polynomial from the structure of Γ.

Lemma 5.9 *Suppose that q is a positive integer and Γ is a plumbing graph with intersection form isomorphic to an $(N+1/q)$–changemaker lattice L for some integer $N \geq 0$. If q is larger than the number of vertices of Γ, then L, and hence N, are uniquely determined by Γ.*[2]

Proof Since Γ must have at least four vertices $q > 2$. Thus there can be no vertices of type (c), showing that the vertex set must be of type I or II. Recall from Remark 4.11 that there are no vertices of the form μ_i when p/q takes the form $p/q = n + 1/q$. Thus by Lemma 5.7 we can assume that the vertices are the standard basis elements $\nu_1, \dots, \nu_t$. For $k > 1$, we have

$$\|\nu_k\|^2 \leq k \leq t,$$

where the upper bound involving k comes from observing that the largest possible norm of a nontight standard basis element occurs when $\nu_k = -e_k + e_{k-1} + \cdots + e_1$. However, using Lemma 4.10 we have that $\|\nu_1\|^2 = q + 1$. Therefore, the assumption that $q > t$ implies that ν_1 is the unique vertex of norm $q + 1$ in Γ. Now we can see inductively that the remaining vertices have unique embeddings as gapless standard basis elements. If we have a vertex v whose image is not among $\nu_1, \dots, \nu_k$ but pairs with some ν_l for $l \leq k$, then v must be embedded as $v = -e_g + e_{g-1} + \cdots + e_l$, where $g = l + \|v\|^2 - 1$, in order to ensure that $v \cdot \nu_l = -1$ and v has the correct norm. Thus the choice of ν_1 determines the rest of the embedding and hence the standard basis vectors of L. However, one can easily recover the structure of L from its standard basis elements. $\square$

The following example shows that the requirement that q be sufficiently large is necessary for the conclusion of Lemma 5.9 to hold.

[2]The value of N can also be determined by comparing the discriminant of both lattices.

Example 5.10 The two $\frac{133}{2}$–changemaker lattices

$$\langle f_1 - f_0,\ f_0 + e_1 + e_2 + e_3 + 2e_4 + 3e_5 + 5e_6 + 5e_7\rangle^{\perp}$$

and

$$\langle f_1 - f_0,\ f_0 + e_1 + e_2 + 2e_3 + 2e_4 + 2e_5 + 4e_6 + 6e_7\rangle^{\perp}$$

are both isomorphic to the same plumbing lattice. This can be seen by writing down the standard bases in each case. This example arises from the fact that $\frac{133}{2}$–surgery on $T_{5,13}$ and the $(2,33)$–cable of $T_{3,5}$ both yield the Seifert fibered space $S^2\left(2; \frac{13}{5}, \frac{5}{3}, \frac{3}{1}\right)$.

5.2 The marked vertex

Now let Δ be a star-shaped or linear plumbing whose intersection form is isomorphic to an $\left(n-\frac{1}{2}\right)$–changemaker lattice L' by an isomorphism which carries the vertices of Δ to gapless standard basis elements of L'. We define the *marked vertex* of Δ to be the vertex of Δ which corresponds to $\nu_1 = -e_1 + f_0 + f_1$. Note that this definition depends a priori on the lattice L' and the choice of isomorphism. In practice, we will always have a fixed lattice L' and a choice of isomorphism in mind, so it will be convenient to think of the marked vertex as being a property of Δ. Although we will be primarily interested in the case where Δ is a star-shaped plumbing with $e = 2$, we extend the definition to include the degenerate case that Δ is a linear plumbing as these will arise in the course of some ensuing proofs.

Example 5.11 Consider the $\frac{17}{2}$–changemaker lattice

$$L' = \langle f_1 - f_0,\ f_0 + e_1 + e_2 + e_3 + e_4 + 2e_5\rangle^{\perp}.$$

The standard basis elements for this lattice are $\nu_1 = -e_1 + f_0 + f_1$, $\nu_2 = -e_2 + e_1$, $\nu_3 = -e_3 + e_2$, $\nu_4 = -e_4 + e_3$ and $\nu_5 = -e_5 + e_4 + e_3$. These are gapless and form the set of vertices for a plumbing Δ shown in Figure 6.

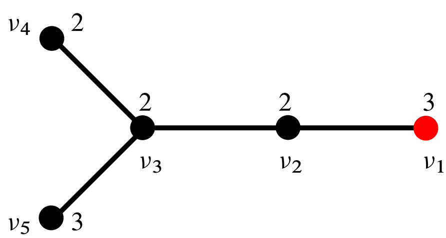


Figure 6: The plumbing Δ corresponding to Example 5.11 with the marked vertex on the right indicated in red.

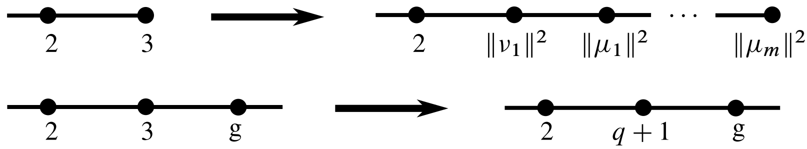


Figure 7: Obtaining Γ from Δ in the type I (top) or II (bottom) case. In both cases the marked vertices are the vertices of weight three in the plumbings on the left hand side.

For each changemaker lattice isomorphic to the intersection form of a plumbing graph Γ with $e=2$, we will produce a plumbing graph Δ whose intersection form is isomorphic to a half-integer changemaker lattice with vertices mapping to gapless standard basis elements such that Γ is obtained by modifying Δ near its marked vertex. We will then use this Δ to construct a knot in S^3 which surgers to give the Seifert fibered space corresponding to Γ.

First we show how to obtain an appropriate Δ. In the type I and II cases this is an easy consequence of Lemma 5.7. Recall that the stable coefficients of a changemaker lattice are defined in Definition 4.4.

Lemma 5.12 *Let L be a p/q–changemaker lattice, where $p/q = n - r/q$ with $1 \le r < q$. Suppose that L is isomorphic to the intersection form of a plumbing Γ with $e=2$ and the vertex set is of type I or II. Then the $\left(n-\frac{1}{2}\right)$–changemaker lattice L' with the same stable coefficients as L is isomorphic to the intersection form of a plumbing Δ, where the vertex set is of type I or II. Moreover Γ is obtained by replacing the marked vertex of Δ by a chain of vertices of weights $\|\nu_1\|^2, \|\mu_1\|^2, \dots, \|\mu_m\|^2$; see Figure 7.*

Proof Let $\nu_1, \dots, \nu_t, \mu_1, \dots, \mu_m$ be the standard basis elements of L. By Lemma 5.7 we can assume that these are the vertices of Γ and by the type I or II assumption none of $\nu_2, \dots, \nu_t$ are tight. Thus the standard basis for L' is

$$-e_1 + f_0 + f_1, \nu_2, \dots, \nu_t.$$

These standard basis elements pair exactly like the vertices of the plumbing graph Δ obtained from Γ by deleting the vertices $\mu_1, \dots, \mu_m$ and changing the weight of ν_1 to three. $\square$

The type III case is a little more subtle:

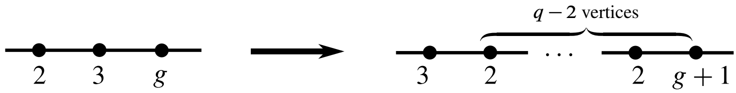


Figure 8: Obtaining Γ from Δ in the type III case. The marked vertex is the vertex of weight three in the plumbing on the left hand side.

Lemma 5.13 *Let L be a p/q–changemaker lattice, where $p/q = n - 1/q$ and $q > 1$. Suppose that L is isomorphic to the intersection form of a plumbing Γ with $e = 2$ and the vertex set is of type III. Then the $\left(n+\frac{1}{2}\right)$–changemaker lattice L' with the same stable coefficients as L is isomorphic to the intersection form of a plumbing Δ, where the vertex set is of type II. Moreover Γ is obtained by increasing the weight of the two vertices adjacent to the marked vertex of Δ by one and converting the marked vertex to a chain of $q-2$ vertices of weight two; see Figure 8.*

Proof It will be convenient to write L' as

$$L' = \langle f_0 + e_0 + \sigma_1 e_1 + \cdots + \sigma_t e_t, f_1 - f_0 \rangle^{\perp} \subseteq \langle f_0, f_1, e_0, \dots, e_t \rangle = \mathbb{Z}^{t+3}.$$

This differs from the notation in Section 4 only by a shift in the indices of the e_i. We will show that L' is isomorphic to the intersection form of the relevant plumbing.

Let $\mu_1, \dots, \mu_m, v_1, \dots, v_t$ be the vertices of Γ, where we assume that $v_1 = \nu_1$ and v_2 is the unique type (c) vertex. By Lemma 5.5 we may assume that v_2 takes the form $v_2 = -(\nu_k + \mu_1 + \cdots + \mu_m)$, where $k > 1$ and ν_k is tight. We modify these to obtain a collection of vectors $v_0', \dots, v_t' \in L'$ as follows. Take $v_0' = -e_0 + f_0 + f_1$, $v_1' = -e_1 + e_0$, $v_2' = e_k - e_{k-1} - \cdots - e_0$ and $v_k' = v_k$ for $k > 2$. By construction we have that each of the v_i' is in L'.

Claim *The vectors $v_0', \dots, v_t'$ span L'.*

Proof Consider the standard basis $\nu_1, \dots, \nu_t$ for L. Since the standard basis elements for L and the vertices of Γ both form bases for L, there are integers α_{ik}, β_{jk} such that

$$\nu_k = \sum_{i=1}^{t} \alpha_{ik} v_i + \sum_{j=1}^{m} \beta_{jk} \mu_j.$$

Consider instead the vectors $\nu_1', \dots, \nu_t'$ in L' defined by

$$\nu_k' = \sum_{i=1}^{t} \alpha_{ik} v_i'.$$

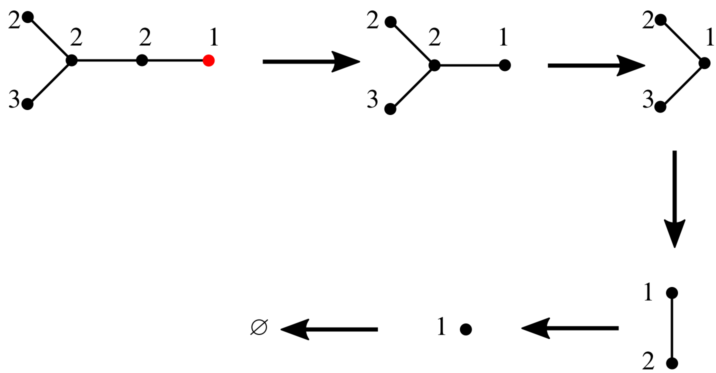


Figure 9: The sequence of blowdowns from Δ' to the empty plumbing when the marked vertex of the plumbing Δ of Example 5.11 is changed to one.

By construction we have, for all $j \geq 1$, that $v_j \cdot e_i = v'_j \cdot e_i$ for $i \geq 1$ and $v_j \cdot f_0 = v'_j \cdot e_0$. Thus we see that $\nu'_k = \nu_k$ unless ν_k is tight, in which case $\nu'_k = -e_k + e_{k-1} + \cdots + e_0$. In either case we see that, up to reindexing the e_i to agree with the notation in Section 4, the vectors $v'_0, \nu'_1, \ldots, \nu'_t$ are precisely the standard basis vectors for L'. Since they are a linear combination of the v'_i, this proves that the v'_i span L'. □

Let Δ be the plumbing graph obtained by replacing the linear chain in Γ given by $v_1, \mu_1, \ldots, \mu_m, v_2$ by the linear chain of vectors of norm 2, 3 and $\|v_2\|^2 - 1$. By construction, the v'_i almost pair as the vertices of Δ: the only exception being that $v'_2 \cdot v'_0 = 1$. However as Δ is a tree, we can choose signs $\varepsilon_i = \pm 1$ so that $\varepsilon_0 = \varepsilon_1 = 1$, $\varepsilon_2 = -1$ and $\varepsilon_0 v'_0, \ldots, \varepsilon_r v'_t$ pair as the vertices of Δ. Thus as the v'_i span L' we see that the intersection form of Δ is isomorphic to L'. By construction the vertex set given by $\varepsilon_0 v'_0, \ldots, \varepsilon_t v'_t$ is of type II. □

Finally, we observe that changing the weight on a marked vertex to one results in a plumbing representing S^3. Figure 9 illustrates how the plumbing from Example 5.11 blows down to the empty plumbing when the weight on the marked vertex is changed to one.

Lemma 5.14 *Let Δ be a star-shaped plumbing or a linear plumbing whose intersection form is isomorphic to a half-integer changemaker lattice L by an isomorphism mapping vertices to gapless standard basis elements. Let Δ' be the plumbing obtained from Δ by changing the weight of the marked vertex to one. Then Δ' can be reduced to the empty plumbing by a sequence of blowdowns on weight one vertices.*

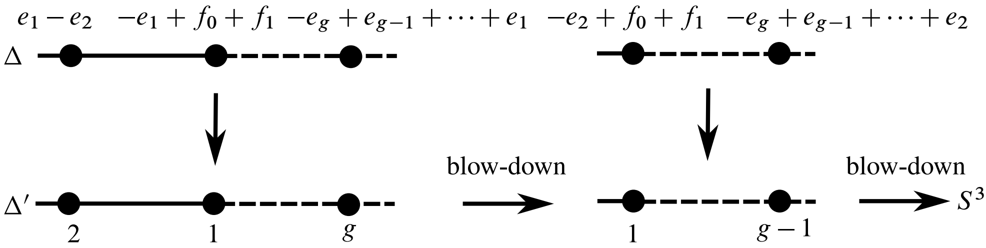


Figure 10: Showing inductively that Δ' blows down.

In particular, the 4*–manifold* X *obtained by plumbing disk-bundles according to* Δ' *has boundary* $\partial X \cong S^3$ *and the corresponding surgery diagram for* S^3 *can be reduced to the empty diagram by performing a sequence of Rolfsen twists on* 1*–framed unknots.*

Proof We will prove this inductively on the number of vertices in Δ. Suppose that Δ is a tree whose intersection form is isomorphic to a half-integer changemaker lattice for which each vertex is a gapless standard basis element of L. When L has rank one Δ consists of just a single vertex, ν_1. The lemma is clearly true in this case.

So now suppose that L has rank $t > 1$ and the vertices of Δ are gapless standard basis elements $\nu_1, \dots, \nu_t$. With the exception of ν_1, these basis elements are not tight since they must have pairing $\nu_1 \cdot \nu_k \in \{0, -1\}$. Thus we must have $\sigma_2 = 1$ and $\nu_2 = -e_2 + e_1$. Note that any other vertex pairing with ν_1 must take the form $\nu_g = -e_g + e_{g-1} + \cdots + e_1$ for some $g > 2$. If it exists then this ν_g is unique. For if we had $\nu_k = -e_k + e_{k-1} + \cdots + e_1$ for some $k > g$, then

$$\nu_k \cdot \nu_g = g - 1 > 0,$$

which is impossible for distinct vertices.

Thus if we obtain Δ' by changing the weight of the marked vertex ν_1 to have weight one, we may perform a blow-down on this weight one vertex in Δ'. This produces a new plumbing $\widetilde{\Delta}'$ with one fewer vertex. Since blowing down a weight one vertex decreases the weight of its neighbors by one, $\widetilde{\Delta}'$ contains a vertex of weight one. Let $\widetilde{\Delta}$ be the plumbing obtained by changing the weight of this vertex to three. These operations are illustrated in Figure 10.

The intersection form of $\widetilde{\Delta}$ embeds into the diagonal lattice that is generated by $e_2, \dots, e_t, f_0, f_1$ by taking vertices $\nu_2', \dots, \nu_t'$, where $\nu_2' = -e_2 + f_0 + f_1$, if there is $\nu_g = -e_g + e_{g-1} + \cdots + e_1$ then $\nu_g' = -e_g + e_{g-1} + \cdots + e_2$, and $\nu_k' = \nu_k$ for all other k. However, these $\nu_2', \dots, \nu_t'$ are precisely the standard basis elements for some half-integer changemaker lattice

$$L' = \langle w_0', f_1 - f_0 \rangle^{\perp} \subseteq \langle f_0, f_1, e_2, \dots, e_t \rangle$$

of rank $t-1$, where $w_0' = f_0 + \sigma_2' e_2 + \cdots + \sigma_t' e_t$ is defined by choosing the σ_i' inductively so that $\sigma_2' = 1$ and σ_k' is chosen to ensure that $\nu_k' \cdot w_0' = 0$. Moreover, these standard basis elements for L' are gapless by construction. Thus we have an isomorphism from the intersection form of $\widetilde{\Delta}$ to a half-integer changemaker lattice which maps vertices to gapless standard basis elements. Moreover, the vertex corresponding to ν_2' is the marked vertex of $\widetilde{\Delta}$. Thus $\widetilde{\Delta}'$ is obtained by changing the marked vertex in $\widetilde{\Delta}$. Since $\widetilde{\Delta}$ has $t-1$ vertices, we can assume inductively that $\widetilde{\Delta}'$ can be blown down to the empty diagram. Since $\widetilde{\Delta}'$ is obtained from Δ' by a blow-down, it follows that Δ' can also be blown down to the empty plumbing, as required.

The statement about Rolfsen twists follows, since a blow-down on the plumbing graph is achieved by a Rolfsen twist in the corresponding surgery diagram. □

5.3 From lattices to surgeries

Now we show how to pass from changemaker lattices to knots with Seifert fibered space surgeries.

Lemma 5.15 *Let Γ be a plumbing graph with $e = 2$ whose intersection form is isomorphic to a p/q–changemaker lattice L, where $p/q \in \mathbb{Q} \setminus \mathbb{Z}$. If Y is the corresponding Seifert fibered space then there is a knot K' which is either a torus knot or a cable of a torus knot such that $S^3_{-p/q}(K') \cong Y$ and the Alexander polynomial of K' is determined by the stable coefficients of L.*[3]

Proof First consider the following construction. Let Δ be a plumbing isomorphic to an $\left(n-\frac{1}{2}\right)$–changemaker lattice L' with the same stable coefficients as L and with vertices of type I or II. Note here that n is the integer $n = \lceil p/q \rceil$. By Lemma 5.7, we can assume that the vertices of Δ in L' are gapless standard basis vectors and Δ has a marked vertex as defined at the start of Section 5.2. Let Δ' be the plumbing obtained by changing the weight of the marked vertex in Δ to one and let D be the surgery diagram corresponding to Δ'. By Lemma 5.14, D is a surgery diagram for S^3. Thus

[3]That is to say that the torsion coefficients of $\Delta_{K'}(t)$ can be computed from L by (4-2). As in Remark 4.8, this allows us to calculate $\Delta_{K'}(t)$ from L (see also Lemma 4.9).

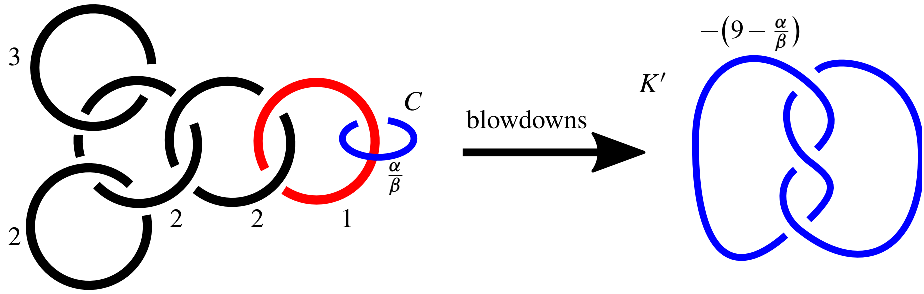


Figure 11: The construction of Lemma 5.15 applied to the plumbing from Example 5.11. After performing the necessary blowdowns, the curve C becomes the trefoil and the α/β surgery coefficient becomes $\alpha/\beta - 9$.

if we let C be the meridian of the unique 1–framed unknot in D, then C describes a knot $K' \subseteq S^3$. Note that even though C is unknotted in the diagram D, the knot K' will be nontrivial in general (see, for example, Figure 11).

Let Y' be the 3–manifold obtained by performing α/β–surgery on C for some $\alpha/\beta \in \mathbb{Q}$. By Lemma 5.14, we may perform a sequence of Rolfsen twists on 1–framed unknots to obtain a surgery description of Y' involving only the component given by C (ie we obtain the surgery description for Y in terms of the knot K'). Since each such Rolfsen twist decreases the framing on C by a nonnegative integer, we see that $Y' \cong S^3_{-(N-\alpha/\beta)}(K')$ for some integer $N > 0$ which is independent of α/β.

Now consider the special case where $\alpha/\beta = -1/d$ for $d \geq 2$. In this case we may perform a slam dunk on the component C to obtain a framing of $1+d$ on the component with which C is linked. Observe that this is the surgery diagram corresponding to the

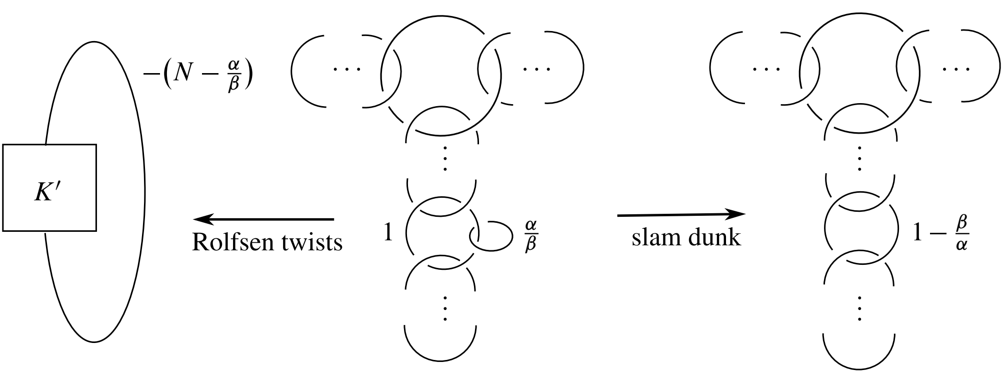


Figure 12: The knot K'.

plumbing graph Δ_d obtained by changing the weight of the marked vertex of Δ to $d+1$. If X_d is the plumbed 4–manifold corresponding to Δ_d, then we have that

$$S^3_{-(N+1/d)}(K') \cong \partial X_d.$$

It follows from Lemma 4.9 that the intersection form of Δ_d is isomorphic to an $(N+1/d)$–changemaker lattice whose stable coefficients compute the Alexander polynomial $\Delta_{K'}(t)$. However, the intersection form of Δ_d is isomorphic to the $(n-1+1/d)$–changemaker lattice with the same stable coefficients as L'. This isomorphism can be seen by observing that the standard basis elements of this $(n-1+1/d)$–changemaker lattice form a set of vertices for the plumbing Δ_d (see Lemma 5.12). Since d can be taken to be arbitrarily large, it follows from Lemma 5.9 that $N=n-1$ and the Alexander polynomial of K' is computed from the stable coefficients of L'. Moreover, as all these surgeries are Seifert fibered spaces, Proposition 2.3 implies that K' is either a torus knot or a cable of a torus knot. With this construction in hand we prove the lemma.

Type I or type II Suppose that L is of type I or II. Write $p/q = n - r/q$, where $1 \le r < q$. The standard basis elements $\nu_1, \mu_1, \dots, \mu_m$ of L form a chain of vertices in Γ. Take L' to be the $\left(n-\frac{1}{2}\right)$–changemaker lattice with the same stable coefficients as L. By Lemma 5.12, L' is isomorphic to the intersection form of the plumbing Δ obtained by deleting $\mu_1, \dots, \mu_m$ and changing the weight on ν_1 to be three. Let K' be the knot constructed from L' as in the first part of this proof. We have shown that the Alexander polynomial of K' is determined by the stable coefficients of L and that K' is either a torus knot or a cable of a torus knot. It remains to check that $S^3_{-p/q}(K') \cong Y$. We obtain a surgery diagram for $S^3_{-p/q}(K')$ by taking the diagram D and performing $(r/q-1)$–surgery on the meridian of C. Performing a slam dunk allows us to absorb C into the 1–framed component and replace the framing on this component by

$$1 + \frac{q}{q-r} = 1 - \frac{1}{\frac{r}{q} - 1}.$$

In the type II case, we have $r/q = (q-1)/q$. Thus after performing this slam dunk we obtain a $(1+q)$–framed component, giving us the surgery diagram corresponding to the plumbing Γ (see Figure 14). This shows that $S^3_{-p/q}(K')$ is the required Seifert fibered space.

In the type I case, we perform a sequence of reverse slam dunks to obtain an integer surgery diagram. Using Lemma 4.10 and $\|\nu_1\|^2 = 1 + \|\mu_0\|^2$, we see that

$$1 + \frac{q}{q-r} = [\|\nu_1\|^2, \|\mu_1\|^2, \dots, \|\mu_m\|^2]^-.$$

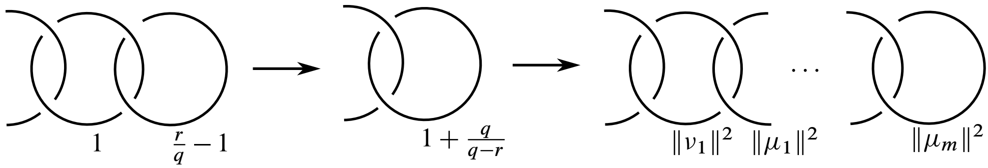


Figure 13: Surgery calculus in the type I case.

Thus if we perform a sequence of reverse slam dunks to convert this to a surgery diagram with integer coefficients, then this gives a chain of unknots with surgery coefficients $\|\nu_1\|^2, \|\mu_1\|^2, \dots, \|\mu_m\|^2$. This is illustrated in Figure 13. However, this surgery diagram is precisely the surgery diagram for Y corresponding to Γ, so we have shown that $S^3_{-p/q}(K')$ is the required Seifert fibered space.

Type III When the vertices of Γ are of type III and $p/q = n - 1/q$, take L' to be the $\bigl(n+\frac{1}{2}\bigr)$–changemaker lattice with the same stable coefficients as L. By Lemma 5.13 this is isomorphic to the intersection form of a plumbing Δ with type II vertices.

Let K' be the knot constructed from L' as in the first part of the proof. Such a knot is either a torus knot or a cable of a torus knot and has the required Alexander polynomial. Thus it remains only to check that it has the desired surgery. We obtain a surgery diagram for $S^3_{-p/q}(K')$ by performing $1/q$–surgery on the curve C. By performing a slam dunk, this can be absorbed to a give a $(1-q)$–framed unknot. This results in a chain of unknotted components with framings 2, $1-q$ and g, for some g. By performing a sequence of $q-2$ blow-ups introducing 1–framed components, we can increase the $1-q$ framing to -1. Then can we blow the -1–components down to obtain a chain of unknots with every framing at least two. The result of these operations is to replace the chain with weights $2, 1-q, g$, by a chain with weights

$$3, \underbrace{2, \dots, 2}_{q-2}, g+1.$$

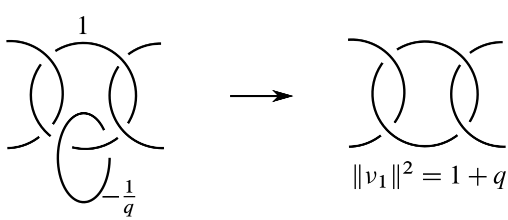


Figure 14: Surgery calculus in the type II case, where $r = q-1$.

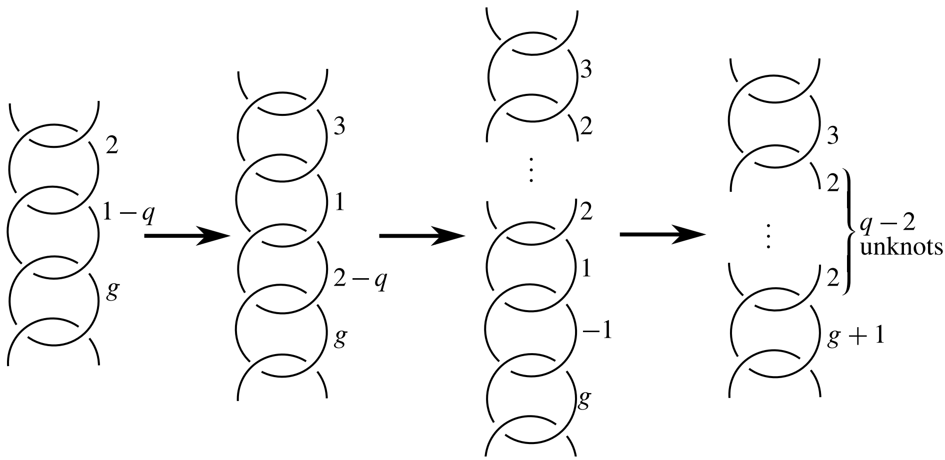


Figure 15: Surgery calculus in the type III case.

This is shown in Figure 15. However, this diagram is precisely the surgery diagram for Y corresponding to Γ. Thus we have shown that $S^3_{-p/q}(K')$ is the required Seifert fibered space. □

Remark 5.16 Some observations on the preceding lemma are in order.

(i) Although we used Proposition 2.3 to deduce that the knot K' is a torus knot or a cable of a torus knot, one can also deduce this fact directly by studying how the curve C sits inside the surgery diagram for S^3.

(ii) One can check that the knot K' constructed in the previous lemma is a torus knot in the type I case and a cable of a torus knot in the type II and III cases.

6 Analysis for the $e \geq 3$ case

In this section, we develop the methods to prove Theorem 1.4 for $e \geq 3$. In this case the surgered Seifert fibered space is the double branched cover of an alternating Montesinos link. This allows us to apply results of [19; 20] which characterize when the double branched cover of an alternating link can arise by noninteger surgery. Before we state these results we will set out some conventions.

A *tangle* $T = (B^3, A)$ will always be a properly embedded 1–manifold A in B^3 where $\partial B^3 \cap A$ consists of four points. Thus the double branched cover of a tangle T will always be a 3–manifold with torus boundary. When considering isotopies between tangles, we will allow isotopies that move ∂B^3. In particular, we will allow isotopies that exchange boundary points of A. If two tangles T and T' are isotopic, then their double branched covers are homeomorphic. For the purposes of this paper, one may take a *rational tangle* to simply mean a tangle whose double branched cover is a solid torus. The notion of slope for rational tangles will not be used.

A *Conway sphere* for a knot K is an embedded sphere in S^3 intersecting the knot transversely in four points. A Conway sphere is said to be *visible* in a diagram if it intersects the plane of the diagram in a connected simple closed curve and intersects the diagram transversely in four points. Note that a Conway sphere always separates a diagram into two tangles.

The following is an amalgamation of Theorems 7.1 and 7.12 of [20].

Theorem 6.1 *Let L be an alternating knot or link such that $S^3_{p/q}(K) \cong \Sigma(L)$ for some knot $K \subseteq S^3$ and $p/q \in \mathbb{Q} \setminus \mathbb{Z}$. Then L has a reduced alternating diagram D with a visible Conway sphere C which separates D into two tangles such that*

(i) *one tangle is a rational tangle containing at least one crossing which can be replaced with a single crossing to obtain an almost-alternating diagram of the unknot, and*

(ii) *the double branched cover of the other tangle is homeomorphic to the complement of a knot $K' \subseteq S^3$ with $\Delta_K(t) = \Delta_{K'}(t)$ and $S^3_{p/q}(K') \cong S^3_{p/q}(K) \cong \Sigma(L)$.*

Recall that an almost-alternating diagram is one that can be transformed into an alternating diagram by changing a single crossing. Although Theorem 6.1 only guarantees the existence of a single diagram for L with a nice Conway sphere, we can easily obtain a similar condition on any alternating diagram of L. This uses the fact that any two reduced alternating diagrams of the same alternating link are related by flypes and planar isotopy [22]. See Figure 17 for an example of a flype.

Proposition 6.2 *Let L be an alternating knot or link such that $S^3_{p/q}(K) \cong \Sigma(L)$ for some knot $K \subseteq S^3$ and $p/q \in \mathbb{Q} \setminus \mathbb{Z}$. Then for any reduced alternating diagram D of L there is a visible Conway sphere C separating D into two tangles such that*

(i) *one tangle is a single crossing,*

(ii) *the double branched cover of the other tangle is homeomorphic to the complement of a knot $K' \subseteq S^3$ with $\Delta_K(t) = \Delta_{K'}(t)$ and $S^3_{p/q}(K') \cong S^3_{p/q}(K) \cong \Sigma(L)$.*

Proof First we will show that there is some reduced alternating diagram for L with the required property. To do this take the diagram D of L along with the Conway

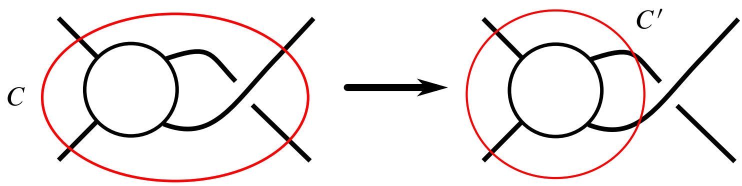


Figure 16: Shrinking C to obtain C'.

sphere C guaranteed by Theorem 6.1. The rational tangle side of C contains at least one crossing. We will show that if C contains more than one crossing, then it can be "shrunk" until it contains a single crossing. It follows from the results of [16, Section 4] that in any alternating diagram of a rational tangle at least one pair of arcs emerging from the boundary sphere must meet in a crossing.

Thus we can assume that C appears as in Figure 16. Take C' to be the Conway sphere obtained by shrinking C to omit this crossing. Notice that the tangles on the outside of C and C' are isotopic by an isotopy swapping the two right-most endpoints to eliminate a crossing. Thus we see that the branched cover of the exterior of C' is still the knot complement $S^3 \setminus \nu K'$. Continuing this way we can reduce C until it contains a single crossing, thus giving a Conway sphere in D with the required properties.

Thus suppose that we have a diagram D with a Conway sphere C with the desired properties. Now let D' be any other reduced alternating diagram for L. This can be obtained from D by a sequence of planar isotopies and flypes. It is clear that planar isotopies preserve the required property, so we only need to check that the existence of C is preserved under flypes. Consider a flype as depicted in Figure 17. When C is contained in one of the tangles marked F or B, then it is clear that the image of C under the flype will again be a Conway sphere with the required properties. Thus we need only consider the case that C encloses the crossing destroyed by the flype. In this case we take C' to be a Conway sphere in D' containing only the crossing created by

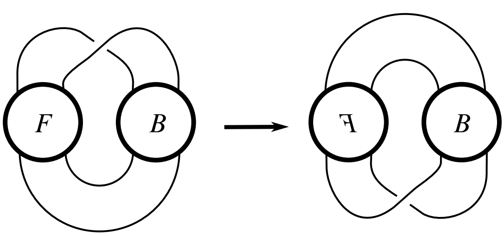


Figure 17: A flype.

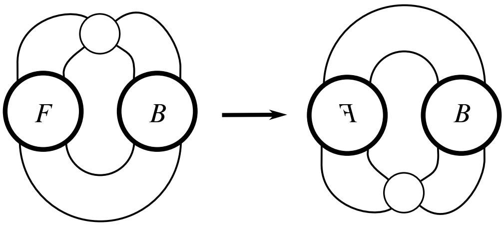


Figure 18: The choice of C and C' after flyping.

the flype; see Figure 18. Consider the tangles on the outside of C and C'. It is not hard to see that these tangles are related by a sequence of isotopies and mutations. Since isotopies and mutations do preserve the homeomorphism type of the double branched cover, C' has the required properties. □

Combining Propositions 6.2 and 2.3 allows us to prove Theorem 1.4 for $e \geq 3$:

Lemma 6.3 *Let $Y = S^2(e; p_1/q_1, p_2/q_2, p_3/q_3)$ be a Seifert fibered space with $e \geq 3$ such that $S^3_{p/q}(K) \cong Y$ for some $K \subseteq S^3$ and $p/q \in \mathbb{Q} \setminus \mathbb{Z}$. Then there is a knot $K' \subseteq S^3$ which is either a torus knot or a cable of a torus knot with $S^3_{p/q}(K') \cong Y$ and $\Delta_K(t) = \Delta_{K'}(t)$.*

Proof It follows from the classification of Montesinos knots in terms of their double branched covers (see, for example, [3, Chapter 12]) that such a Y is the double branched cover of an alternating Montesinos link L with three arms. Such a link has a diagram of the form D shown in Figure 19, where the rectangular boxes are twist regions, each containing some number of crossings.

By Proposition 6.2, there is a Conway sphere C containing on one side a single crossing c and on the other a tangle such that the double branched cover of its exterior is homeomorphic to the complement of a knot K' in S^3 such that $S^3_{p/q}(K') \cong \Sigma(L) \cong Y$ and with $\Delta_K(t) = \Delta_{K'}(t)$. Thus we need only to check that K' is a torus knot or a cable of a torus knot.

The crossing c lies in some twist region R of D. For any $n > 1$, let D_n be the alternating diagram obtained by replacing c with a twist region containing n crossings, where we insert the crossings so that the twist region R is extended in length. Since we started with a diagram of the form shown in Figure 19 and extended the length of a twist region, we see that D_n is still in the form given in Figure 19. Thus D_n is a Montesinos knot or link and its double branched cover $\Sigma(D_n)$ is a Seifert fibered

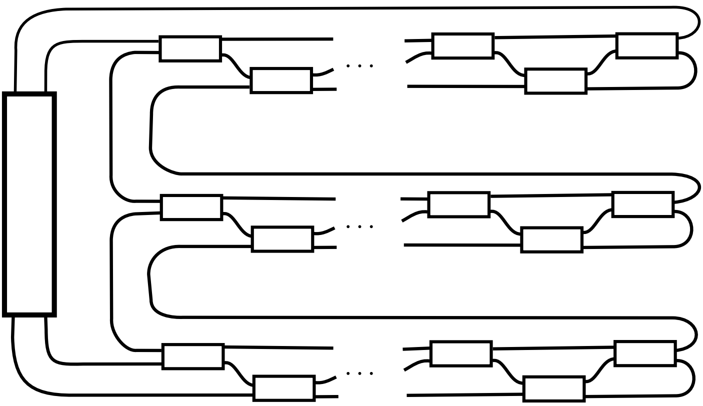


Figure 19: A diagram for a three armed Montesinos link. The boxes represent twist regions.

space [3, Chapter 12]. Since we obtained D_n by replacing the crossing c with a rational tangle, the Montesinos trick shows that there is a rational number $p_n/q_n \in \mathbb{Q}$ such that $S^3_{p_n/q_n}(K') \cong \Sigma(D_n)$ [23]. Since the crossing numbers of the D_n are monotonically increasing, we see that the D_n are diagrams for distinct knots or links. As there are only finitely many nonsplit alternating knots or links with a given determinant, we see that $\det D_n$, and hence $|p_n|$, tends to infinity. By [21, Theorem 1.1] any such p_n/q_n satisfies $|p_n/q_n| \leq 4g(K') + 3$. So for n sufficiently large we have $q_n \geq 9$. Thus Proposition 2.3 applies to show that K' is either a torus knot or a cable of a torus knot, as required. □

7 Proofs of Theorem 1.4 and Proposition 1.5

Theorem 1.4 *Let Y be a Seifert fibered space over S^2 with three exceptional fibers and $e(Y) \notin \{+1, +2, -1\}$. If there is a knot K in S^3 with $Y \cong S^3_{p/q}(K)$ where $p/q > 0$ and $p/q \in \mathbb{Q} \setminus \mathbb{Z}$, then there is a knot K' which is either a torus knot or a cable of a torus knot with $S^3_{p/q}(K') \cong Y$ and $\Delta_K(t) = \Delta_{K'}(t)$.*

Proof We have two cases to consider: either $e(Y) = -2$ or $|e(Y)| \geq 3$. First suppose that $e(Y) = -2$. Since

$$e(-Y) = 2 \quad \text{and} \quad -Y \cong S^3_{-p/q}(\overline{K}),$$

where $\overline{K}$ is knot obtained by reflecting K, Lemma 4.9 shows that the intersection form of the canonical plumbing bounding $-Y$ is isomorphic to a p/q–changemaker lattice whose stable coefficients compute the Alexander polynomial of $\overline{K}$. Lemma 5.15 then shows that the existence of this changemaker lattice allows us to construct a knot $\overline{K'}$ which is a torus knot or a cable of a torus knot with $\Delta_{\overline{K'}}(t) = \Delta_{\overline{K}}(t)$ and $S^3_{-p/q}(\overline{K}) \cong S^3_{-p/q}(\overline{K'})$. Reflecting yields a knot K' with the desired properties. After reflecting suitably, the case that $|e(Y)| \geq 3$ is precisely given by Lemma 6.3. □

We now turn to the proof of Proposition 1.5. First, note that the Seifert invariants of surgeries on a torus knot can easily be calculated directly; see, for example, [24; 27, Lemma 4.4].

Proposition 7.1 *For coprime $r, s > 1$, let $Y \cong S^3_{p/q}(T_{r,s})$. Then:*

(i) *Y is reducible if $p/q = rs$.*

(ii) *Y is a lens space if $p/q = rs \pm 1/q$.*

(iii) *Otherwise Y is the small Seifert fibered space with three exceptional fibers*

$$Y \cong S^2\left(1; \frac{r}{s'}, \frac{s}{r'}, \frac{p}{q} - rs\right),$$

where integers s' and r' satisfy $1 \leq s' < r$, $1 \leq r' < s$ and $s'/r + r'/s = 1 + 1/rs$.

The corresponding result for negative torus knots can be obtained by changing orientations, since $S^3_{p/q}(T_{r,s}) \cong -S^3_{-p/q}(T_{-r,s})$. Next we calculate $e(Y)$ for these surgeries.

Lemma 7.2 *Let $Y \cong S^3_{p/q}(T_{r,s})$ be a small Seifert fibered space, where $r, s > 1$. Then $e(Y)$ satisfies*

(i) *$e(Y) = -1$ if $p/q < 0$,*

(ii) *$e(Y) = 2$ if $0 < p/q < rs - 1$,*

(iii) *$e(Y) \geq 3$ if $rs - 1 < p/q < rs$, and*

(iv) *$e(Y) \leq -2$ if $p/q > rs$.*

Proof By Proposition 7.1, we have that

$$Y \cong S^2\left(1; \frac{r}{s'}, \frac{s}{r'}, \frac{p}{q} - rs\right). \tag{7-1}$$

This shows that

$$\varepsilon(S^3_{p/q}(T_{r,s})) = \frac{1}{rs}\left(\frac{\frac{p}{q}}{rs - \frac{p}{q}}\right),$$

which implies that $\varepsilon(Y) > 0$ if $0 < p/q < rs$ and $\varepsilon(Y) < 0$ if $p/q < 0$ or $p/q > rs$.

First suppose that $0 < p/q < rs$, so that $\varepsilon(Y) > 0$. We apply Rolfsen twists to (7-1) to show that Y takes the form

$$Y \cong S^2\left(1+n; \frac{r}{s'}, \frac{s}{r'}, \frac{p-rsq}{q+n(p-rsq)}\right),$$

where n is such that $(p-rsq)/(q+n(p-rsq)) > 1$. One can check that the necessary value of n is $n = \lceil q/(rsq-p)\rceil$. Thus we have that $e(Y) = 2$ if $0 < p/q < rs-1$ and $e(Y) \geq 3$ if $rs-1 < p/q < rs$, as required.

By Rolfsen twisting twice, we see that Y can be written in the form

$$Y \cong S^3_{p/q}(T_{r,s}) \cong S^2\left(-1; -\frac{r}{r-s'}, -\frac{s}{s-r'}, \frac{p}{q} - rs\right). \tag{7-2}$$

If $p/q < 0$, then $\varepsilon(Y) < 0$ and we have that $p/q - rs < -1$. Thus the description in (7-2) shows that $e(Y) = -1$ in this case. If $p/q > rs$, then $\varepsilon(Y) < 0$, but $p/q - rs > 0$. Thus by Rolfsen twisting we see that Y takes the form

$$Y \cong S^2\left(-1-n; -\frac{r}{r-s'}, -\frac{s}{s-r'}, \frac{p-qrs}{q-n(p-qrs)}\right),$$

where $n \geq 1$ is chosen to ensure that $(p-qrs)/(q-n(p-qrs)) < -1$. This shows that $e(Y) \leq -2$ in this case. □

This allows us to determine the surgeries arising in the conclusion of Theorem 1.4.

Proposition 1.5 *Let K be a torus knot or a cable of a torus knot. Then for $p/q > 0$ we have that $S^3_{p/q}(K)$ is a Seifert fibered space over S^2 with three exceptional fibers and $e(S^3_{p/q}(K)) \notin \{-1, +1, +2\}$ if and only if*

(i) *K is a torus knot $K = T_{r,s}$ with $r, s > 1$, $p/q > rs - 1$ and $|p - rsq| > 1$, or*

(ii) *K is a cable of a torus knot $K = C_{a,b} \circ T_{r,s}$, where $r, s > 1$, $b/a > rs - 1$ and $p/q = ab \pm 1/q$.*

Proof If K is a torus knot, then this is a consequence of Lemma 7.2 and Proposition 7.1. The result is deduced for cables of torus knots by using the fact that Seifert fibered surgeries on $C_{a,b} \circ T_{r,s}$ take the form

$$S^3_{ab\pm 1/q}(C_{a,b} \circ T_{r,s}) \cong S^3_{(qab\pm 1)/(qa^2)}(T_{r,s}),$$

where a is the winding number of the pattern torus knot; see [9, Lemma 3.3]. The condition that $b/a > rs-1$ is a consequence of the fact that $(qab \pm 1)/(qa^2) > rs - 1$ if and only if $b/a > rs - 1$. □

References

[1] **S Boyer**, **X Zhang**, *Exceptional surgery on knots*, Bull. Amer. Math. Soc. 31 (1994) 197–203 MR Zbl

[2] **J L Brown, Jr**, *Note on complete sequences of integers*, Amer. Math. Monthly 68 (1961) 557–560 MR Zbl

[3] **G Burde**, **H Zieschang**, *Knots*, 2nd edition, De Gruyter Studies in Math. 5, de Gruyter, Berlin (2003) MR Zbl

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

[5] **D Gabai**, *Surgery on knots in solid tori*, Topology 28 (1989) 1–6 MR Zbl

[6] **D Gabai**, *1–Bridge braids in solid tori*, Topology Appl. 37 (1990) 221–235 MR Zbl

[7] **J Gibbons**, *Deficiency symmetries of surgeries in S^3*, Int. Math. Res. Not. 2015 (2015) 12126–12151 MR Zbl

[8] **R E Gompf**, **A I Stipsicz**, *4–Manifolds and Kirby calculus*, Graduate Studies in Math. 20, Amer. Math. Soc., Providence, RI (1999) MR Zbl

[9] **C M Gordon**, *Dehn surgery and satellite knots*, Trans. Amer. Math. Soc. 275 (1983) 687–708 MR Zbl

[10] **C M Gordon**, *Dehn filling: a survey*, from "Knot theory" (V F R Jones, J Kania-Bartoszyńska, J H Przytycki, P Traczyk, V G Turaev, editors), Banach Center 42, Polish Acad. Sci. Inst. Math., Warsaw (1998) 129–144 MR Zbl

[11] **C M Gordon**, **J Luecke**, *Non-integral toroidal Dehn surgeries*, Comm. Anal. Geom. 12 (2004) 417–485 MR Zbl

[12] **J E Greene**, *The lens space realization problem*, Ann. of Math. 177 (2013) 449–511 MR Zbl

[13] **J E Greene**, *Donaldson's theorem, Heegaard Floer homology, and knots with unknotting number one*, Adv. Math. 255 (2014) 672–705 MR Zbl

[14] **J E Greene**, *L-space surgeries, genus bounds, and the cabling conjecture*, J. Differential Geom. 100 (2015) 491–506 MR Zbl

[15] **A Issa**, *The classification of quasi-alternating Montesinos links*, Proc. Amer. Math. Soc. 146 (2018) 4047–4057 MR Zbl

[16] **L H Kauffman**, **S Lambropoulou**, *On the classification of rational tangles*, Adv. in Appl. Math. 33 (2004) 199–237 MR Zbl

[17] **M Lackenby**, **R Meyerhoff**, *The maximal number of exceptional Dehn surgeries*, Invent. Math. 191 (2013) 341–382 MR Zbl

[18] **W B R Lickorish**, *A representation of orientable combinatorial 3–manifolds*, Ann. of Math. 76 (1962) 531–540 MR Zbl

[19] **D McCoy**, *Non-integer surgery and branched double covers of alternating knots*, J. Lond. Math. Soc. 92 (2015) 311–337 MR Zbl

[20] **D McCoy**, *Alternating surgeries*, PhD thesis, University of Glasgow (2016) Available at `https://www.maths.gla.ac.uk/~bowens/theses/2016McCoyPhD.pdf`

[21] **D McCoy**, *Bounds on alternating surgery slopes*, Algebr. Geom. Topol. 17 (2017) 2603–2634 MR Zbl

[22] **W Menasco**, **M Thistlethwaite**, *The classification of alternating links*, Ann. of Math. 138 (1993) 113–171 MR Zbl

[23] **J M Montesinos**, *Surgery on links and double branched covers of S^3*, from "Knots, groups, and 3–manifolds: papers dedicated to the memory of R H Fox" (L P Neuwirth, editor), Ann. of Math. Studies 84, Princeton Univ. Press (1975) 227–259 MR Zbl

[24] **L Moser**, *Elementary surgery along a torus knot*, Pacific J. Math. 38 (1971) 737–745 MR Zbl

[25] **W D Neumann**, **F Raymond**, *Seifert manifolds, plumbing, μ–invariant and orientation reversing maps*, from "Algebraic and geometric topology" (L P Neuwirth, editor), Lecture Notes in Math. 664, Springer (1978) 163–196 MR Zbl

[26] **P Orlik**, *Seifert manifolds*, Lecture Notes in Math. 291, Springer (1972) MR Zbl

[27] **B Owens**, **S Strle**, *Dehn surgeries and negative-definite four-manifolds*, Selecta Math. 18 (2012) 839–854 MR Zbl

[28] **P Ozsváth**, **Z Szabó**, *Absolutely graded Floer homologies and intersection forms for four-manifolds with boundary*, Adv. Math. 173 (2003) 179–261 MR Zbl

[29] **P Ozsváth**, **Z Szabó**, *On the Floer homology of plumbed three-manifolds*, Geom. Topol. 7 (2003) 185–224 MR Zbl

[30] **P Ozsváth**, **Z Szabó**, *On knot Floer homology and lens space surgeries*, Topology 44 (2005) 1281–1300 MR Zbl

[31] **R Riley**, *An elliptical path from parabolic representations to hyperbolic structures*, from "Topology of low-dimensional manifolds" (R A Fenn, editor), Lecture Notes in Math. 722, Springer (1979) 99–133 MR Zbl

[32] **W P Thurston**, *Three-dimensional manifolds, Kleinian groups and hyperbolic geometry*, Bull. Amer. Math. Soc. 6 (1982) 357–381 MR Zbl

[33] **A H Wallace**, *Modifications and cobounding manifolds*, Canadian J. Math. 12 (1960) 503–528 MR Zbl

Department of Mathematics, University of British Columbia
Vancouver, BC, Canada

Département de mathématiques, Université du Québec à Montréal
Montreal, QC, Canada

aissa@math.ubc.ca, mc_coy.duncan@uqam.ca

Received: 22 October 2018 Revised: 14 June 2021

Algebraic & Geometric Topology **23**:4 (2023) 1551–1582
DOI: 10.2140/agt.2023.23.1551
Published: 14 June 2023

Bialgebraic approach to rack cohomology

SIMON COVEZ
MARCO ANDRÉS FARINATI
VICTORIA LEBED
DOMINIQUE MANCHON

We interpret the complexes defining rack cohomology in terms of a certain dg bialgebra. This yields elementary algebraic proofs of old and new structural results for this cohomology theory. For instance, we exhibit two explicit homotopies controlling structure defects on the cochain level: one for the commutativity defect of the cup product, and the other for the "Zinbielity" defect of the dendriform structure. We also show that, for a quandle, the cup product on rack cohomology *restricts* to, and the Zinbiel product *descends* to quandle cohomology. Finally, for rack cohomology with suitable coefficients, we complete the cup product with a compatible coproduct.

16T10, 20N02, 55N35, 57M27

1 Introduction

A *quandle* X is an algebraic structure whose axioms describe what survives from a group when one only looks at the conjugacy operation. Quandles have been intensively studied since the 1982 work of D Joyce [14] and S Matveev [19], who showed how to extract powerful knot invariants from them. But the history of quandles goes back much further. *Racks*, which slightly generalize quandles, can be traced back to an unpublished 1959 correspondence between J Conway and G Wraith; keis, or involutive quandles, appeared in 1943 in an article of M Takasaki [29], and other related structures were mentioned as early as the end of the 19th century. A thorough account, focusing on topological aspects through the rack space, can be found in R Fenn, C Rourke and B Sanderson [13]. Other viewpoints and various applications to algebra, topology, and set theory are treated, for instance, in Andruskiewitsch and Graña [1], Dehornoy [8], Elhamdadi and Nelson [9], Kinyon [15] and Przytycki [21].

Rack cohomology $H_{\mathrm{R}}(X)$ was defined by Fenn, Rourke and Sanderson in 1993 [11]. It strengthened and extended the applications of racks. The *cup product* $\smile$ on $H_{\mathrm{R}}(X)$ was first described in topological terms by F Clauwens [5]. Later S Covez proposed a cubical interpretation, which allowed him to refine the cup product to a *dendriform structure* using shuffle combinatorics, and further to a *Zinbiel*[1] *product* $\underset{\leftarrow}{\smile}$ using acyclic models [6; 7]. This yields in particular the commutativity of $\smile$.

Rack cohomology is a particular case of the cohomology of set-theoretic solutions to the *Yang–Baxter equation*, as constructed by J S Carter, M Elhamdadi and M Saito [2]. This very general cohomology theory still belongs to the cubical context. It thus carries a commutative cup product, explicitly described by M Farinati and J García Galofre [10]. V Lebed [17] gave it two new interpretations: in terms of M Rosso's quantum shuffles [26], and via graphical calculus. She gave a graphical version of an explicit homotopy on the cochain level, which controls the commutativity defect of $\smile$.

Our purpose is to study a *differential graded bialgebra* $B(X)$ that is attached to any rack X, and governs the algebraic structure of its cohomology. This construction was first unveiled in [10]. We show that $B(X)$ is graded cocommutative up to an explicit homotopy, which implies the commutativity of the cup product $\smile$ on $H_{\mathrm{R}}(X)$. This yields a purely algebraic version of the diagrammatic construction from [17]. On

[1]The word *Zinbiel* is the mirror image of the word *Leibniz*, and the corresponding structures are Koszul-dual. The definitions will be recalled in Section 7.

a quotient $\bar{B}(X)$ of $B(X)$, we refine the coproduct to a dg codendriform structure. Even better, this codendriform structure is co-Zinbiel up to another explicit homotopy, which has not been described before. As a result, the associative commutative cup product on $H_{\mathrm{R}}(X)$ stems from a Zinbiel product $\underset{\leftarrow}{\smile}$, which coincides with the one from [7]. Both the rack cohomology $H_{\mathrm{R}}(X)$ and our dg bialgebra $B(X)$ can be enriched with coefficients. For finite X and suitably chosen coefficients, we complete $\smile$ with a compatible coassociative coproduct. Here again our bialgebraic interpretation considerably simplifies all verifications.

The rack cohomology of a quandle received particular attention, starting from the work of Carter, Jelsovsky, Kamada, Langford and Saito [3] and R A Litherland and S Nelson [18]. It is known to split into two parts, called *quandle* and *degenerate*: $H_{\mathrm{R}} = H_{\mathrm{Q}} \oplus H_{\mathrm{D}}$. The degenerate part H_{D} is far from being empty, since it contains (a shifted copy of) H_{Q} as a direct factor: $H_{\mathrm{D}} = H_{\mathrm{Q}}[1] \oplus H_{\mathrm{L}}$. However, as an abelian group, it does not carry any new information. As shown by J Przytycki and K Putyra [23], it is completely determined by H_{Q}. We recover these cohomology decompositions at the bialgebraic level, and show that the cup product on rack cohomology restricts to H_{Q}. This result is new, to our knowledge. Rather unexpectedly, our proof heavily uses the Zinbiel product $\underset{\leftarrow}{\smile}$ refining $\smile$, even though $\underset{\leftarrow}{\smile}$ does not *restrict* to H_{Q}. What we show is that $\underset{\leftarrow}{\smile}$ *induces* a Zinbiel product on H_{Q}; for this we need to regard H_{Q} as a quotient rather than a subspace of H_{R}. Our results suggest the questions:

(i) Does H_{D} allow one to reconstruct H_{Q} as a Zinbiel, or at least associative, algebra?

(ii) In the opposite direction, does H_{Q} determine the whole rack cohomology H_{R} as a Zinbiel, or at least associative, algebra?

Acknowledgements Farinati is a research member of Conicet, partially supported by PIP 112-200801-00900, PICT 2006 00836, UBACyT X051 and MathAmSud 10-math-01 OPECSHA. Lebed was partially supported by a Hamilton Research Fellowship (Hamilton Mathematics Institute, Trinity College Dublin).

2 Cohomology of shelves and racks

We start with recalling the classical complexes defining rack homology and cohomology. They will be given a bialgebraic interpretation in Section 3. The consequences of this interpretation will be explored in the remainder of the paper.

A *shelf* is a set X together with a binary operation $\triangleleft\colon X\times X\to X$ given by $(x,y)\mapsto x\triangleleft y$ (sometimes denoted by $x\triangleleft y=x^y$) satisfying the self-distributivity axiom

$$(x\triangleleft y)\triangleleft z=(x\triangleleft z)\triangleleft(y\triangleleft z) \tag{1}$$

for all $x,y,z\in X$. In exponential notation, it reads $(x^y)^z=(x^z)^{(y^z)}$. A shelf is called a *rack* if the maps $-\triangleleft y\colon X\to X$ are bijective for all $y\in X$, a *spindle* if $x\triangleleft x=x$ for all $x\in X$, and a *quandle* if it is both a rack and a spindle. The fundamental family of examples of quandles is given by groups X with $x\triangleleft y=y^{-1}xy$.

Define $C_n(X)=\mathbb{Z}X^n$ to be the free abelian group with basis X^n, and put

$$C^n(X)=\mathbb{Z}^{X^n}\cong\operatorname{Hom}(C_n(X),\mathbb{Z}).$$

Define the differential $\partial\colon C_\bullet(X)\to C_{\bullet-1}(X)$ as the linearization of

$$\partial(x_1\cdots x_n)=\sum_{i=1}^{n}(-1)^{i-1}(x_1\cdots\hat{x}_i\cdots x_n-x_1^{x_i}\cdots x_{i-1}^{x_i}x_{i+1}\cdots x_n), \tag{2}$$

where $\hat{x}_i$ means that x_i was omitted. Here and below we denote by $x_1\cdots x_n$ the element $(x_1,\dots,x_n)$ of X^n. For cohomology, the differential will be $\partial^*\colon C^\bullet(X)\to C^{\bullet+1}(X)$. These maps are of square zero (by direct computation or see Remark 11 later) and define, respectively, the *rack*[2] homology $H^{\mathrm{R}}(X)$ and cohomology $H_{\mathrm{R}}(X)$ of the shelf X.

In knot theory, a quandle Q can be used to color arcs of knot diagrams; a coloring rule involving the operation $\triangleleft$ is imposed at each crossing. The three quandle axioms are precisely what is needed for the number of Q–colorings of a diagram to depend on the underlying knot only. These Q–coloring counting invariants can be enhanced by Boltzmann-type weights, computed using a 2–cocycle of Q. Similarly, n–cocycles of Q yield invariants of $(n-1)$–dimensional surfaces knotted in $\mathbb{R}^{n+1}$; see [4; 24]. Now, together with arcs, one can color diagram regions. The colors can be taken from a Q–set, and the weights are given by cocycles with coefficients, which we will describe next.

Given a shelf X, an X*–set* is a set S together with a map $\blacktriangleleft\colon S\times X\to S$ satisfying

$$(x\blacktriangleleft y)\blacktriangleleft z=(x\blacktriangleleft z)\blacktriangleleft(y\triangleleft z) \tag{3}$$

for all $x\in S$ and $y,z\in X$. The basic examples are

(i) X itself, with $\triangleleft$ as the action map $\blacktriangleleft$;

[2] The terminology is inconsistent here. Indeed rack (co)homology was originally defined for racks, and only later was it realized that it works and is interesting, more generally, for shelves. The same goes for the quandle (co)homology of spindles, considered in Section 8.

(ii) any set with the trivial action $x \blacktriangleleft y = x$;

(iii) the *structure monoid* $M(X)$ of X (denoted simply by M if X is understood), which is a quadratic monoid

$$M(X) = \langle X : yx^y = xy \text{ for all } x, y \in X \rangle,$$

and where the action map $\blacktriangleleft$ is concatenation in M.

An X–set can also be seen as a set with an action of the monoid $M(X)$.

More generally, an *X–module* is an abelian group R together with a map $\blacktriangleleft : R \times X \to R$ (often written exponentially) which is linear in R, and obeys relation (3) for all $x \in R$ and $y, z \in X$. In other words, it is an $M(X)$–module. The basic examples are the linearization $\mathbb{Z}S$ of an X–set S, or any abelian group with the trivial action.

Take a shelf X and an X–module R. Take the free R–module $C_n(X,R) = RX^n$ with basis X^n, and put $C^n(X,R) = \mathrm{Hom}(C_n(X,R), \mathbb{Z})$. The differential ∂ on C_n is the linearization of

$$\partial(rx_1 \cdots x_n) = \sum_{i=1}^{n} (-1)^{i-1}(rx_1 \cdots \hat{x}_i \cdots x_n - r^{x_i} x_1^{x_i} \cdots x_{i-1}^{x_i} x_{i+1} \cdots x_n), \tag{4}$$

and the differential on C^n is the induced one. Again, these maps are of square zero and define, respectively, the *rack homology* $H^{\mathrm{R}}(X,R)$ *and cohomology* $H_{\mathrm{R}}(X,R)$ *of the shelf* X *with coefficients in* R. If R is the linearization of an X–set S, we use the notation $C_{\mathrm{R}}(X,S)$, $H_{\mathrm{R}}(X,S)$ etc. Choosing S to be the empty set, one recovers the previous definitions. Another interesting coefficient choice is the *structure algebra* $A(X)$ of X (often denoted simply by A), which is the monoid algebra of $M(X)$:

$$A(X) = \mathbb{Z}M(X) \cong \mathbb{Z}\langle X \rangle / \langle yx^y - xy : x, y \in X \rangle.$$

Declaring every $x \in X$ group-like, one gets an associative bialgebra structure on A. This coefficient choice is universal in the following sense. Any X–module R is a right $M(X)$–module, hence a right $A(X)$–module. Then one has an obvious isomorphism of chain complexes

$$C_n(X,R) \cong R \otimes_{A(X)} C_n(X, M(X)),$$

where A acts on the first factor of $C_n(X,M) \cong A \otimes_{\mathbb{Z}} \mathbb{Z}X^n$ by multiplication on the left, and the differential acts on the second factor of $R \otimes_A C_n(X,M)$.

3 A dg bialgebra associated to a shelf

The algebraic objects introduced in this section are aimed to yield a simple and explicit description of a differential graded algebra structure on the complex $(C^\bullet(X), \partial^*)$ above, which is commutative, and in fact even Zinbiel, up to explicit homotopies.

Fix a shelf X. All the (bi)algebra structures below will be over $\mathbb{Z}$, and will be (co)unital. Also, the tensor product $A \otimes B$ of two graded algebras will always be endowed with the product algebra structure involving the Koszul sign

$$(a_1 \otimes b_1)(a_2 \otimes b_2) = (-1)^{|b_1||a_2|} a_1 a_2 \otimes b_1 b_2,$$

where $b_1 \in B$ and $a_2 \in A$ are homogeneous of degree $|b_1|$ and $|a_2|$, respectively. The Koszul sign also appears when $A^* \otimes B^*$ acts on $A \otimes B$. Similarly, by a (co)derivation on a graded (co)algebra we will always mean a super(co)derivation, and by commutativity we will mean supercommutativity.

Define $B(X)$ (also denoted by B) as the algebra freely generated by two copies of X with the relations

$$B(X) = \mathbb{Z}\langle x, e_y : x, y \in X \rangle / \langle y x^y - x y, y e_{x^y} - e_x y : x, y \in X \rangle.$$

The interest of this construction lies in the rich structure it carries:

Theorem 1 *For any shelf X, $B(X)$ is a differential graded bialgebra and a differential graded $A(X)$–bimodule, where*:

- *The grading is given by declaring $|e_x| = 1$ and $|x| = 0$ for all $x \in X$.*
- *The differential d is the unique derivation of degree -1 determined by*

$$d(e_x) = 1 - x \quad \text{and} \quad d(x) = 0 \quad \text{for all } x \in X.$$

- *The comultiplication $\Delta\colon B \to B \otimes B$ and the counit $\varepsilon\colon B \to k$ are defined on the generators by*

$$\begin{matrix} \Delta(e_x) = e_x \otimes x + 1 \otimes e_x, & \varepsilon(e_x) = 0, \\ \Delta(x) = x \otimes x, & \varepsilon(x) = 1 \end{matrix} \qquad \text{for all } x \in X,$$

  and extended multiplicatively.
- *The A–actions $\lambda\colon A \otimes B \to B$ and $\rho\colon B \otimes A \to B$ are defined by*

$$x \cdot b \cdot y = x b y \quad \text{for all } x, y \in X \text{ and } b \in B.$$

By differential graded bialgebra we mean that the differential is both a derivation with respect to multiplication, and a coderivation with respect to comultiplication.

Notice that B is neither commutative nor cocommutative in general.

As usual, from Theorem 1 one deduces dg algebra and dg $A(X)$–bimodule structures on the graded dual $B^*(X)$ of $B(X)$.

Proof Since the relations are homogeneous, B is a graded algebra. In order to see that d is well defined, one must check that the relations $yx^y \sim xy$ and $ye_{x^y} \sim e_x y$ are compatible with d. The first relation is easier:

$$d(yx^y - xy) = d(y)x^y + yd(x^y) - d(x)y - xd(y) = 0 + 0 - 0 - 0 = 0.$$

For the second relation, one has

$$\begin{aligned} d(ye_{x^y} - e_x y) &= yd(e_{x^y}) - d(e_x)y = y(1-x^y) - (1-x)y = y - yx^y - y + xy \\ &= xy - yx^y. \end{aligned}$$

So the ideal of relations defining B is stable by d.

Since d is a derivation and d^2 vanishes on generators, we have $d^2 = 0$, hence a structure of differential graded algebra on B.

Next, we need to check that Δ is well defined. The first relation is easy since all $x \in X$ are group-like in B:

$$\begin{aligned} \Delta(xy - yx^y) &= (x \otimes x)(y \otimes y) - (y \otimes y)(x^y \otimes x^y) = xy \otimes xy - yx^y \otimes yx^y \\ &= xy \otimes (xy - yx^y) + (xy - yx^y) \otimes yx^y. \end{aligned}$$

For the second relation, we check

$$\begin{aligned} \Delta(ye_{x^y} - e_x y) &= (y \otimes y)(e_{x^y} \otimes x^y + 1 \otimes e_{x^y}) - (e_x \otimes x + 1 \otimes e_x)(y \otimes y) \\ &= ye_{x^y} \otimes yx^y + y \otimes ye_{x^y} - e_x y \otimes xy - y \otimes e_x y \\ &= (ye_{x^y} - e_x y) \otimes yx^y + e_x y \otimes (yx^y - xy) + y \otimes (ye_{x^y} - e_x y). \end{aligned}$$

So the ideal defining the relations is also a coideal.

Clearly, Δ respects the grading.

Let us now check that d is a coderivation. It is enough to see this on generators:

$$\begin{aligned} \Delta(d(x)) &= \Delta(0) = 0 = d(x) \otimes x + x \otimes d(x) = (d \otimes 1 + 1 \otimes d)\Delta(x), \\ \Delta(d(e_x)) &= \Delta(1-x) = 1 \otimes 1 - x \otimes x. \end{aligned}$$

This coincides with

$$\begin{aligned} (d \otimes 1 + 1 \otimes d)\Delta(e_x) &= (d \otimes 1 + 1 \otimes d)(e_x \otimes x + 1 \otimes e_x) \\ &= (1-x) \otimes x + 1 \otimes (1-x) = 1 \otimes x - x \otimes x + 1 \otimes 1 - 1 \otimes x \\ &= 1 \otimes 1 - x \otimes x. \end{aligned}$$

The map ε is also well defined, since

$$\varepsilon(xy - yx^y) = \varepsilon(ye_{x^y} - e_x y) = 0.$$

An easy verification on the generators shows that it is indeed a counit.

Finally, the formula $x \cdot b \cdot y = xby$ obviously defines commuting degree-preserving A–actions on B. By the definition of the differential d, one has $d(xby) = xd(b)y$, thus d respects this bimodule structure. □

Example 2 Let us compute $\Delta(e_x e_y)$. By definition, $\Delta(e_x e_y) = \Delta(e_x)\Delta(e_y)$ in $B \otimes B$, and this is equal to

$$\begin{aligned}(e_x \otimes x + 1 \otimes e_x)(e_y \otimes y + 1 \otimes e_y) &= e_x e_y \otimes xy + e_x \otimes x e_y - e_y \otimes e_x y + 1 \otimes e_x e_y \\ &= e_x e_y \otimes xy + e_x \otimes x e_y - e_y \otimes y e_{x^y} + 1 \otimes e_x e_y.\end{aligned}$$

Note the Koszul sign appearing in the product $(1 \otimes e_x)(e_y \otimes y) = -e_y \otimes e_x y$.

The structure on $B(X)$ survives in homology:

Proposition 3 *For any shelf X the homology $H(X)$ of $B(X)$ inherits a graded algebra structure. Moreover, the $A(X)$–actions on $B(X)$ induce trivial actions on $H(X)$:*

$$x \cdot h \cdot y = h \quad \text{for all } x, y \in X \text{ and } h \in H.$$

Dually, the cohomology $H^*(X)$ of $B(X)^*$ inherits a graded algebra structure and trivial $A(X)$–actions.

Here and below, by $B(X)^*$ we mean the graded dual of $B(X)$.

Proof The only nonclassical statement here is the triviality of the induced actions. Take $x \in X$ and $b \in B$. By the definition of d, one has $d(e_x b) = d(e_x)b - e_x d(b)$, hence

$$d(e_x b) = (1 - x)b - e_x d(b). \tag{5}$$

If b is a cycle, this shows that $x \cdot b = b$ modulo a boundary. Hence the induced left A–action on H is trivial. The cases of the right action and the actions in cohomology are analogous. □

The proposition implies the following remarkable property of H: if $b \in B$ is a representative of some homology class in H, and if one lets an $x \in X$ act upon all the letters from X occurring in b (where the action is $y \mapsto y^x$), then one obtains another representative of the same homology class.

One can also define a version of the bialgebra $B(X)$ with coefficients in any unital commutative ring k:

$$B(X,k) = k\langle x, e_y : x, y \in X\rangle / \langle yx^y - xy,\, ye_{x^y} - e_x y : x, y \in X\rangle.$$

In particular, all the tensor products should be taken over k. Theorem 1 and its proof extend verbatim to this setting. For suitable coefficients k, one can say even more:

Proposition 4 *For any shelf X and any field k, the homology $H(X,k)$ of $B(X,k)$ inherits a graded bialgebra structure. If, moreover, X is finite, then the cohomology $H^*(X,k)$ of $B(X,k)^*$ also inherits a graded bialgebra structure.*

This results from the following general observation; it is surely known to specialists, however the authors were unable to find it in the literature:

Lemma 5 *Let k be a field.*

(i) *If $\left(C = \bigoplus C_i, d, \Delta\right)$ is a k–linear dg coassociative coalgebra, then Δ induces a coproduct on the homology H of (C,d).*

(ii) *If $\left(C = \bigoplus C_i, d, \cdot\right)$ is a k–linear dg algebra of finite dimension in each degree, then $\cdot$ induces a coproduct on the cohomology H^* of (C^*, d^*).*

Proof (i) The relation $\Delta d = (d \otimes \mathrm{Id} + \mathrm{Id} \otimes d)\Delta$ implies that Δ survives in the quotient $C/\operatorname{Im}(d)$. To restrict it further to $H = \operatorname{Ker}(d)/\operatorname{Im}(d)$, we shall check that

$$\Delta(\operatorname{Ker}(d)) \subseteq \operatorname{Ker}(d) \otimes \operatorname{Ker}(d) + \operatorname{Im}(d) \otimes C + C \otimes \operatorname{Im}(d).$$

Since k is a field, the space $K := \operatorname{Ker}(d)$ has a complement L in C, on which d is injective. Putting $I := \operatorname{Im}(d)$, one has

$$\begin{aligned} &(d \otimes \mathrm{Id})(L \otimes L) = I \otimes L, && (\mathrm{Id} \otimes d)(L \otimes L) = L \otimes I,\\ &(d \otimes \mathrm{Id} + \mathrm{Id} \otimes d)(K \otimes L) = K \otimes I, && (d \otimes \mathrm{Id} + \mathrm{Id} \otimes d)(L \otimes K) = I \otimes K,\\ &(d \otimes \mathrm{Id} + \mathrm{Id} \otimes d)(K \otimes K) = 0. \end{aligned}$$

Moreover, in the first two lines all the maps are bijective. Now, from

$$(d \otimes \mathrm{Id} + \mathrm{Id} \otimes d)\Delta(K) = \Delta d(K) = 0$$

and from the disjointness of L and K (and hence I), one sees that $\Delta(K)$ cannot have components in $L \otimes L$, and its components from $K \otimes L$ (resp. $L \otimes K$) necessarily lie in $I \otimes L$ (resp. $L \otimes I$).

(ii) Due to the finite dimension in each degree, the product on (C,d) induces a coproduct on (C^*, d^*), to which we apply the first statement. □

Remark 6 The dg bialgebra $B(X)$ admits a variation $B'(X)$, where one adds the inverses x^{-1} of the generators $x \in X$, with $|x^{-1}|=0$, $d(x^{-1})=0$, $\Delta(x^{-1})=x^{-1}\otimes x^{-1}$ and $\varepsilon(x^{-1})=1$. One obtains a dg Hopf algebra, with the antipode defined on the generators by

$$s(x)=x^{-1}, \quad s(x^{-1})=x \quad \text{and} \quad s(e_x)=-e_x x^{-1},$$

and extended superantimultiplicatively, in the sense of $s(ab)=(-1)^{d(a)d(b)}s(b)s(a)$ for homogeneous a and b. Indeed, one easily verifies that this map

- is well defined, that is, compatible with the defining relations of $B'(X)$;
- is of degree 0, that is, $|s(b)|=|b|$ for all homogeneous b;
- yields the inverse of Id in the convolution algebra, that is, if $\Delta(b)=\sum_i a_i \otimes c_i$ for a given $b \in B'(X)$, one has
$$\sum_i s(a_i)c_i=\sum_i a_i s(c_i)=\varepsilon(b);$$
- commutes with the differential d, that is, $ds=sd$.

For the square of the antipode, one computes $s^2(e_x)=xe_xx^{-1}$. In a spindle it equals e_x, yielding $s^2=\mathrm{Id}$. In general s need not be of finite order; for the rack $X=\mathbb{Z}$ with $x^y=x+1$, one has $s^2\cdot(e_x)=e_{x-1}$. In a rack, one simplifies $s^2(e_x)=e_{x\tilde{\triangleleft}x}$, where the operation $\tilde{\triangleleft}$ is defined by $(x\triangleleft y)\mathbin{\tilde{\triangleleft}} y=x$ for all $x,y\in X$. The map $x\mapsto x\mathbin{\tilde{\triangleleft}} x$ plays an important role in the study of racks; see for instance [28]. Finally, in the computation

$$\begin{aligned} s(e_{x_1}e_{x_2}\cdots e_{x_{n-1}}e_{x_n}) &= (-1)^{\binom{n}{2}}(-e_{x_n}x_n^{-1})(-e_{x_{n-1}}x_{n-1}^{-1})\cdots(-e_{x_2}x_2^{-1})(-e_{x_1}x_1^{-1}) \\ &= (-1)^{\frac{1}{2}n(n+1)}e_{x_n}e_{x_{n-1}^{x_n}}\cdots e_{x_2^{x_3\cdots x_n}}e_{x_1^{x_2\cdots x_n}}x_n^{-1}\cdots x_1^{-1} \end{aligned}$$

one recognizes the *remarkable map*

$$(x_1,x_2,\ldots,x_n)\mapsto(x_1^{x_2\cdots x_n},x_2^{x_3\cdots x_n},\ldots,x_n)$$

of Przytycki [20].

4 The bialgebra encodes the cohomology

We will now show that the dg bialgebra $B(X)$ knows everything about the homology $(C_\bullet,\partial)$ and the cohomology $(C^\bullet,\partial^*)$ of our shelf X, and about its variations $(C_\bullet^M,\partial)$ and $(C_M^\bullet,\partial^*)$ with coefficients in the structure monoid $M(X)$.

First, we need to modify B slightly:

Lemma 7 *The following data define a dg coalgebra and a right dg A–module*:

$$(\mathbb{Z} \otimes_A B, \mathrm{Id}_{\mathbb{Z}} \otimes d, \mathrm{Id}_{\mathbb{Z}} \otimes \Delta, \mathrm{Id}_{\mathbb{Z}} \otimes \varepsilon, \mathrm{Id}_{\mathbb{Z}} \otimes \rho).$$

Here the grading is the one induced from B, and the A–action on $\mathbb{Z}$ is the trivial one: $\lambda \cdot x = \lambda$ *for all* $x \in X$ *and* $\lambda \in \mathbb{Z}$.

The dg coalgebra from the lemma will be denoted by $\overline{B} = \overline{B}(X)$. It has obvious variants $\overline{B}(X, k)$ with coefficients in any unital commutative ring k.

Proof On the level of abelian groups, one has

$$\overline{B} \simeq B/\langle xb - b : x \in X, b \in B\rangle. \tag{6}$$

The grading survives in this quotient since $|x| = 0$, and the degree -1 differential survives since $d(x) = 0$ and $|x| = 0$ imply $d(xb) = xd(b) \sim d(b)$. Further, we have $\Delta(xb) = (x \otimes x)\Delta(b) \sim \Delta(b)$, so Δ induces a coproduct on $\overline{B}$ compatible with the grading and the differential. For the counit, we have $\varepsilon(xb) = \varepsilon(x)\varepsilon(b) = \varepsilon(b)$. Finally, the right A–action also descends to $\overline{B}$, as $(xb) \cdot y = xby = x(by) \sim by = b \cdot y$. □

Remark 8 Observe that we lose the product in the quotient $\overline{B}$. Indeed, for all $x, y \in X$ we have $y \sim 1$, but $e_x \cdot y = e_x y = y e_{x^y} \sim e_{x^y} \nsim e_x = e_x \cdot 1$.

Lemma 9 *As a left A–module*, B *can be presented as*

$$B \cong A \otimes \mathbb{Z}\langle X\rangle.$$

Proof Consider the map

$$A \otimes \mathbb{Z}\langle X\rangle \to B \quad \text{given by } x_1 \cdots x_k \otimes y_1 \cdots y_n \mapsto x_1 \cdots x_k e_{y_1} \cdots e_{y_n}.$$

It is well defined since the relations in A hold true for the corresponding generators of B.

Going in the opposite direction is trickier. A monomial b in B is a product of generators of the form x and e_y. Let $a(b)$ be what remains in b when all generators of the form e_y are erased. Further, start with a new copy of b and erase all generators of the form x one by one, starting from the left. When erasing a generator x, replace all generators of the form e_y to its left by e_{y^x}. After that replace all the e_y by y. This yields a monomial $t(b) \in \mathbb{Z}\langle X\rangle$. Analyzing the defining relations of B, and using the self-distributivity axiom (1) for X, one sees that we obtain a well-defined map

$$B \to A \otimes \mathbb{Z}\langle X\rangle \quad \text{given by } b \mapsto a(b) \otimes t(b).$$

Both maps are clearly A–equivariant, and are mutually inverse. □

From this follows:

Proposition 10 *One has the following isomorphisms of complexes*:

$$(C_\bullet^M,\partial)\cong B,\quad (C_M^\bullet,\partial^*)\cong B^*,\quad (C_\bullet,\partial)\cong \bar{B},\quad (C^\bullet,\partial^*)\cong \bar{B}^*,$$
$$(C_\bullet(X,kM(X)),\partial)\cong B(X,k),\quad (C^\bullet(X,kM(X)),\partial^*)\cong B^*(X,k),$$
$$(C_\bullet(X,k),\partial)\cong \bar{B}(X,k),\qquad (C^\bullet(X,k),\partial^*)\cong \bar{B}^*(X,k).$$

Here B^* denotes the graded dual of B with the induced differential, and similarly for $\bar{B}^*$. In the last isomorphisms, the ring k is considered as a trivial X–module.

Proof The preceding lemma yields isomorphisms of abelian groups

$$B\cong C_\bullet^M \quad\text{and}\quad \bar{B}\cong \mathbb{Z}\langle X\rangle = C_\bullet,$$

and their k–versions.

To compute the differential induced on this by d, we use that d is a derivation:

$$\begin{aligned}
d(e_{x_1}\cdots e_{x_n})
&=\sum_{i=1}^n(-1)^{i-1}e_{x_1}\cdots e_{x_{i-1}}d(e_{x_i})e_{x_{i+1}}\cdots e_{x_n}\\
&=\sum_{i=1}^n(-1)^{i-1}e_{x_1}\cdots e_{x_{i-1}}(1-x_i)e_{x_{i+1}}\cdots e_{x_n}\\
&=\sum_{i=1}^n(-1)^{i-1}e_{x_1}\cdots e_{x_{i-1}}e_{x_{i+1}}\cdots e_{x_n}-\sum_{i=1}^n(-1)^{i-1}e_{x_1}\cdots e_{x_{i-1}}x_ie_{x_{i+1}}\cdots e_{x_n}.
\end{aligned}$$

Using the relation $e_xy=ye_{x^y}$, one gets

$$e_{x_1}\cdots e_{x_{i-1}}x_ie_{x_{i+1}}\cdots e_{x_n}=x_ie_{x_1^{x_i}}\cdots e_{x_{i-1}^{x_i}}e_{x_{i+1}}\cdots e_{x_n},$$

which in $C_\bullet^M$ corresponds to $x_ix_1^{x_i}\cdots x_{i-1}^{x_i}x_{i+1}\cdots x_n$. We thus recover the differential (4). In the quotient $\bar{B}$, the last computation simplifies as $e_{x_1^{x_i}}\cdots e_{x_{i-1}^{x_i}}e_{x_{i+1}}\cdots e_{x_n}$, and we recover the differential (2). □

Remark 11 This proposition provides a very simple proof that $\partial^2=0$ in $C_\bullet(X,M(X))$ and its versions.

From the proof of Lemma 9 one deduces the useful fact:

Lemma 12 *The map $A\to B$ given by $x\mapsto x$ is an injective algebra morphism.*

In what follows we will often identify A with its isomorphic image in B.

The isomorphisms in Proposition 10 allow one to transport the structure from B and $\bar{B}$ to rack (co)homology:

Theorem 13 *Take a shelf X and a field k. Then*

(i) *the chain complex $(C_\bullet(X), \partial)$ carries a coassociative coproduct,*

(ii) *the cochain complex $(C^\bullet(X), \partial^*)$ carries an associative product,*

(iii) *the chain complex $(C_\bullet(X, M(X)), \partial)$ carries a bialgebra structure,*

(iv) *the cochain complex $(C^\bullet(X, M(X)), \partial^*)$ carries an associative product, enriched to a bialgebra structure when X is finite.*

This induces

(i) *associative products on $H^*(X)$, $H^*(X, M(X))$, $H^*(X,k)$, and $H(X, M(X))$,*

(ii) *a coassociative coproduct on $H(X,k)$,*

(iii) *a bialgebra structure on $H(X, kM(X))$,*

(iv) *an associative product on $H^*(X, kM(X))$, which is completed to a bialgebra structure for finite X.*

The product in cohomology is called the *cup product*, and is denoted by $\smile$.

Example 14 Take $f, g \in C^2(X)$. To compute $f \smile g$, one needs to compute the summands in $\Delta(e_x e_y e_z e_t)$ with two tensors of type e_u in each factor. Using the computation from Example 2

$$\begin{aligned}\Delta(e_x e_y e_z e_t) &= \Delta(e_x e_y)\Delta(e_z e_t)\\ &= (e_x e_y \otimes xy + 1 \otimes e_x e_y + e_x \otimes x e_y - e_y \otimes y e_{x^y})\\ &\qquad\qquad \cdot (e_z e_t \otimes zt + 1 \otimes e_z e_t + e_z \otimes z e_t - e_t \otimes t e_{z^t})\\ &= e_x e_y \otimes xy e_z e_t + e_z e_t \otimes e_x e_y zt - e_x e_z \otimes x e_y z e_t + e_x e_t \otimes x e_y t e_{z^t}\\ &\qquad\qquad + e_y e_z \otimes y e_{x^y} z e_t - e_y e_t \otimes y e_{x^y} t e_{z^t} + \cdots,\end{aligned}$$

where the dots hide terms on which f and g vanish. Pushing the e_u to the right and the elements of X to the left, we get

$$\begin{aligned}&e_x e_y \otimes xy e_z e_t + e_z e_t \otimes zx e_{x^{zt}} e_{y^{zt}} - e_x e_z \otimes xz e_{y^z} e_t\\ &\qquad + e_x e_t \otimes xt e_{y^t} e_{z^t} + e_y e_z \otimes yz e_{x^{yz}} e_t - e_y e_t \otimes yt e_{x^{yt}} e_{z^t} + \cdots,\end{aligned}$$

so finally $(f \smile g)(e_x e_y e_z e_t)$ is equal to

$$f(e_x e_y)g(e_z e_t) + f(e_z e_t)g(e_{x^{zt}} e_{y^{zt}}) - f(e_x e_z)g(e_{y^z} e_t) \\ + f(e_x e_t)g(e_{y^t} e_{z^t}) + f(e_y e_z)g(e_{x^{yz}} e_t) - f(e_y e_t)g(e_{x^{yt}} e_{z^t}).$$

This formula is to be compared with (23) of [5]. A full explanation of this agreement is given in the next section.

The last piece of structure to be extracted from Proposition 10 is the A–action:

Proposition 15 *For a shelf X, the complex $(C^\bullet(X), \partial^*)$ is a left $A(X)$–module, with*

$$(x \cdot f)(x_1 \cdots x_n) := (x_1^x \cdots x_n^x),$$

where $f \in C^n(X)$ and $x, x_1, \dots, x_n \in X$. The induced $A(X)$–action in cohomology is trivial.

Proof This directly follows from Propositions 3 and 10. □

This property of rack cohomology was first noticed by Przytycki and Putyra [22]. In our bialgebraic interpretation it becomes particularly natural.

5 An explicit expression for the cup product in cohomology

To give an explicit formula for the cup product in rack cohomology, we need to compute $\Delta(e_{x_1} \cdots e_{x_n})$ for any $x_1, \dots, x_n$ in the rack X, generalizing Example 2. For this we will introduce some notation. First, for any $n \geq 1$ and for any $i \in \{1, \dots, n\}$ we define two maps $\delta_i^0, \delta_i^1 : X^n \to X^{n-1}$ by

$$\delta_i^0(x_1, \dots, x_n) = (x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n),$$
$$\delta_i^1(x_1, \dots, x_n) = (x_1 \lhd x_i, \dots, x_{i-1} \lhd x_i, x_{i+1}, \dots, x_n).$$

The above identification of B with $A \otimes \mathbb{Z}\langle X \rangle$ given by $ae_{x_1} \cdots e_{x_n} \leftrightarrow a \otimes x_1 \cdots x_n$ allows one to transport δ_i^0 and δ_i^1 to A–linear endomorphisms of B:

$$\delta_i^0(ae_{x_1} \cdots e_{x_n}) = \begin{cases} ae_{x_1} \cdots e_{x_{i-1}} e_{x_{i+1}} \cdots e_{x_n} & \text{if } i \leq n, \\ 0 & \text{if } i > n, \end{cases}$$
$$\delta_i^1(ae_{x_1} \cdots e_{x_n}) = \begin{cases} ax_i e_{x_1 \lhd x_i} \dots e_{x_{i-1} \lhd x_i} e_{x_{i+1}} \cdots e_{x_n} & \text{if } i \leq n, \\ 0 & \text{if } i > n. \end{cases}$$

A straightforward computation using self-distributivity yields

$$\delta_i^\varepsilon \delta_j^\eta = \delta_{j-1}^\eta \delta_i^\varepsilon \tag{7}$$

for any $i<j$ and any $\varepsilon,\eta\in\{0,1\}$. Identities (7) are the defining axioms for $\square$–sets [27]; see also [5; 12]. Now, the boundary (2) can be rewritten as

$$\partial=\sum_{i\geq 1}(-1)^{i-1}(\delta_i^0-\delta_i^1). \tag{8}$$

For any finite subset S of $\mathbb{N}$ and for $\varepsilon\in\{0,1\}$, we denote by δ_S^ε the composition, in increasing order, of the maps δ_a^ε for $a\in S$.

Proposition 16 *Given a rack X, the coproduct in $B(X)$ can be computed by the formula*

$$\Delta(ae_{x_1}\cdots e_{x_n})=\sum_{S\subset\{1,\dots,n\}}\epsilon(S)a\delta_S^0(e_{x_1}\cdots e_{x_n})\otimes a\delta_{S^c}^1(e_{x_1}\cdots e_{x_n}) \tag{9}$$

for all $a\in\langle X\rangle$ and $x_i\in X$. Here $S^c=\{1,\dots,n\}\setminus S$, and $\epsilon(S)$ is the signature of the unshuffle permutation of $\{1,\dots,n\}$ which puts S^c on the left and S on the right.

We use the canonical form $ae_{x_1}\cdots e_{x_n}$ of a monomial in $B(X)$.

Proof Since $\Delta(a)=a\otimes a$, we can omit this part of our monomial. Let us then proceed by induction on n, the case $n=1$ being immediate:

$$\begin{aligned}
&\Delta(e_{x_1}\cdots e_{x_n})\\
&\quad=\Delta(e_{x_1}\cdots e_{x_{n-1}})\Delta(e_{x_n})\\
&\quad=\Bigg(\sum_{B\subset\{1,\dots,n-1\}}\epsilon(B)\delta_B^0(e_{x_1}\cdots e_{x_{n-1}})\otimes\delta_{B^c}^1(e_{x_1}\cdots e_{x_{n-1}})\Bigg)(e_{x_n}\otimes x_n+1\otimes e_{x_n})\\
&\quad=\sum_{B\subset\{1,\dots,n-1\}}(-1)^{|B|}\epsilon(B)\delta_B^0(e_{x_1}\cdots e_{x_{n-1}})e_{x_n}\otimes\delta_{B^c}^1(e_{x_1}\cdots e_{x_{n-1}})x_n\\
&\qquad\qquad+\sum_{B\subset\{1,\dots,n-1\}}\epsilon(B)\delta_B^0(e_{x_1}\cdots e_{x_{n-1}})\otimes\delta_{B^c}^1(e_{x_1}\cdots e_{x_{n-1}})e_{x_n}\\
&\quad=\sum_{\substack{S\subset\{1,\dots,n\}\\ n\notin S}}\epsilon(S)\delta_S^0(e_{x_1}\cdots e_{x_n})\otimes\delta_{S^c}^1(e_{x_1}\cdots e_{x_n})\\
&\qquad\qquad+\sum_{\substack{S\subset\{1,\dots,n\}\\ n\in S}}\epsilon(S)\delta_S^0(e_{x_1}\cdots e_{x_n})\otimes\delta_{S^c}^1(e_{x_1}\cdots e_{x_n})\\
&\quad=\sum_{S\subset\{1,\dots,n\}}\epsilon(S)\delta_S^0(e_{x_1}\cdots e_{x_n})\otimes\delta_{S^c}^1(e_{x_1}\cdots e_{x_n}).
\end{aligned}$$

$\square$

Corollary 17 *The cup product in rack cohomology induced from the coproduct in B coincides with the cup product given by Clauwens in* [5, (32)].

Proof This is immediate by comparing (9) with [5, (32)]. The overall sign $(-1)^{km}$ in [5] is the Koszul sign. □

6 The cup product is commutative

In the preceding section we established that our cup product on rack cohomology coincides with Clauwens's product. The latter comes from the cohomology of a topological space, and is thus commutative (where, as usual, we mean supercommutative). We will now give a direct algebraic proof based on an explicit homotopy argument. This homotopy is a specialization of the graphically defined map, constructed for solutions to the Yang–Baxter equation by Lebed [17].

Let us start with a low-degree illustration:

Example 18 Take $f, g \in C^1(X)$, identified with A–linear maps from B to $\mathbb{Z}$ (also denoted by f and g) determined by the values $f(e_x) := f(x)$ and $g(e_x) := g(x)$ for $x \in X$, and vanishing in degrees other than 1. Then the cup product $f \smile g \in C^2(X)$ is defined by

$$\begin{aligned}(f \smile g)(e_x e_y) &= (f \otimes g)\Delta(e_x e_y)\\ &= (f \otimes g)(e_x e_y \otimes xy + 1 \otimes e_x e_y + e_x \otimes x e_y - e_y \otimes y e_{x^y})\end{aligned}$$

(see Example 2). Since f and g vanish on elements of degree 0 and 2, and are left A–linear (where x and y act on $\mathbb{Z}$ trivially), the only remaining terms are

$$-f(e_x)g(x e_y) + f(e_y)g(y e_{x^y}) = -f(e_x)g(e_y) + f(e_y)g(e_{x^y}).$$

Note the Koszul sign $(-1)^{|g||e_x|} = -1$, and similarly in the second term. So the product is in general not commutative. On the other hand, the cocycle condition $\partial^* g = 0$ means precisely $g(e_x) = g(e_{x^y})$ for all x and y, so the cup product restricted to 1–cocycles is commutative.

Now, take $f, g \in C^1(X, M(X))$, identified with maps $B \to \mathbb{Z}$ vanishing in degrees other than 1. Then, for monomials $a \in A$ and $x, y \in X$, one computes

$$(f \smile g)(a e_x e_y) = -f(a e_x)g(a x e_y) + f(a e_y)g(a e_x y)$$

and

$$\begin{aligned}&(f \smile g + g \smile f)(a e_x e_y) \\ &\quad = -f(a e_x)g(a x e_y) + f(a e_y)g(a e_x y) - g(a e_x)f(a x e_y) + g(a e_y)f(a e_x y).\end{aligned} \tag{10}$$

Suppose that f and g are 1–cocycles. This yields the relation

$$\begin{aligned}0 &= (-d^* f)(ae_x e_y) = f(d(ae_x e_y)) = f(d(a)e_x e_y) + f(ad(e_x)e_y) - f(ae_x d(e_y)) \\ &= f(a(1-x)e_y) - f(ae_x(1-y)) = f(ae_y) - f(axe_y) - f(ae_x) + f(ae_x y),\end{aligned}$$

and similarly for g. Note the Koszul sign in $-d^* f = fd$. Define a map $h\colon B \to \mathbb{Z}$ by

$$h(ae_x) = f(ae_x)g(ae_x)$$

for all monomials $a \in A$ and all $x \in X$. Then (10) becomes

$$\begin{aligned}&-f(ae_x)g(axe_y) + f(ae_y)(g(ae_x) - g(ae_y) + g(axe_y)) \\ &\qquad\qquad -g(ae_x)(f(ae_x y) - f(ae_x) + f(ae_y)) + g(ae_y)f(ae_x y) \\ &= (f(ae_y) - f(ae_x))g(axe_y) - h(ae_y) + h(ae_x) + (g(ae_y) - g(ae_x))f(ae_x y) \\ &= (f(axe_y) - f(ae_x y))g(axe_y) - h(ae_y) + h(ae_x) + (g(axe_y) - g(ae_x y))f(ae_x y) \\ &= h(axe_y) - h(ae_y) + h(ae_x) - h(ae_x y) = (d^* h)(ae_x e_y),\end{aligned}$$

yielding the relation $f \smile g + g \smile f = d^* h$, and hence the supercommutativity of the cup product of degree 1 cohomology classes.

Lemma 19 *Let $h\colon B \to B \otimes B$ be the degree 1 linear map defined on monomials in B written in the canonical form as follows: $h(a) = 0$, and*

$$h(ae_{x_1} \cdots e_{x_n}) := \sum_{i=1}^{n} (-1)^{i-1} (a \otimes a)(\tau\Delta)(e_{x_1} \cdots e_{x_{i-1}})(e_{x_i} \otimes e_{x_i})\Delta(e_{x_{i+1}} \cdots e_{x_n}),$$

where $\tau\colon B \otimes B \to B \otimes B$ is the signed flip $\tau(a \otimes b) = (-1)^{d(a)d(b)} b \otimes a$ for homogeneous a and b. Then for any homogeneous $b_1, b_2 \in B$ we have

$$h(b_1 b_2) = h(b_1)\Delta(b_2) + (-1)^{|b_1|}(\tau\Delta)(b_1)h(b_2). \tag{11}$$

Also, h induces a map $\bar{B} \to \bar{B} \otimes \bar{B}$.

The induced map will still be denoted by h.

Proof Using the fact that both Δ and $\tau\Delta$ are algebra morphisms, one rewrites the definition of h as

$$\begin{aligned}&h(ae_{x_1} \cdots e_{x_n}) \\ &\qquad = \sum_{i=1}^{n} (-1)^{i-1} (a \otimes a)(\tau\Delta)(e_{x_1}) \cdots (\tau\Delta)(e_{x_{i-1}})(e_{x_i} \otimes e_{x_i})\Delta(e_{x_{i+1}}) \cdots \Delta(e_{x_n}).\end{aligned}$$

This immediately yields (11) on $b_1 = a_1 e_{x_1} \cdots e_{x_p}$ (ie any monomial in B) and $b_2 = e_{x_{p+1}} \cdots e_{x_{p+q}}$. To check (11) on general monomials b_1 and $b_2' = a_2 b_2$, with

$a_2 \in \langle X \rangle$ and b_2 a product of the e_x, one observes that the maps h, Δ, and $\tau\Delta$ are X–equivariant both on the left and on the right, which gives

$$\begin{aligned} h(b_1(a_2b_2)) = h((b_1a_2)b_2) &= h(b_1a_2)\Delta(b_2) + (-1)^{|b_1a_2|}(\tau\Delta)(b_1a_2)h(b_2) \\ &= h(b_1)(a_2 \otimes a_2)\Delta(b_2) + (-1)^{|b_1|}(\tau\Delta)(b_1)(a_2 \otimes a_2)h(b_2) \\ &= h(b_1)\Delta(a_2b_2) + (-1)^{|b_1|}(\tau\Delta)(b_1)h(a_2b_2). \end{aligned}$$

Finally, h survives in the quotient $\bar{B}$ since its X–equivariance reads

$$h(ae_{x_1}\cdots e_{x_n}) = (a \otimes a)h(e_{x_1}\cdots e_{x_n}) \sim h(e_{x_1}\cdots e_{x_n}),$$

with the usual notation. □

For example, an easy computation gives

$$h(e_xe_y) = (xe_y + e_x) \otimes e_xe_y - e_xe_y \otimes (e_xy + e_y),$$

which in $\bar{B}$ becomes

$$h(e_xe_y) = (e_y + e_x) \otimes e_xe_y - e_xe_y \otimes (e_{x^y} + e_y).$$

Proposition 20 *The map h is a homotopy between Δ and $\tau\Delta$:*

$$(d \otimes \mathrm{Id}_B + \mathrm{Id}_B \otimes d)h + hd = \Delta - \tau\Delta. \tag{12}$$

Of course, h remains a homotopy in $\bar{B}$ as well.

Proof We will use the shorthand notation $dh := (d \otimes \mathrm{Id}_B + \mathrm{Id}_B \otimes d)h$.

If x is a degree zero generator of B then $(dh+hd)(x)=0$ and $\Delta(x)=x\otimes x=(\tau\Delta)(x)$, hence (12) holds. Now, for a degree one generator e_x we have

$$\begin{aligned} (dh+hd)(e_x) &= d(e_x \otimes e_x) + 0 = (1-x) \otimes e_x - e_x \otimes (1-x) \\ &= -x \otimes e_x + e_x \otimes x + 1 \otimes e_x - e_x \otimes 1 = (\Delta - \tau\Delta)(e_x). \end{aligned}$$

The proof can then be carried out by induction on the degree, using (11):

$$\begin{aligned} &(dh+hd)(b_1b_2) \\ &\quad = d(h(b_1)\Delta(b_2) + (-1)^{|b_1|}(\tau\Delta)(b_1)h(b_2)) + h(db_1 \cdot b_2 + (-1)^{|b_1|}b_1 \cdot db_2) \\ &\quad = dh(b_1)\Delta(b_2) + (-1)^{|b_1|+1}h(b_1)d\Delta(b_2) + (-1)^{|b_1|}d(\tau\Delta)(b_1)h(b_2) \\ &\qquad\qquad + (\tau\Delta)(b_1)dh(b_2) + hd(b_1)\Delta(b_2) + (-1)^{|b_1|+1}(\tau\Delta)(db_1)h(b_2) \\ &\qquad\qquad\qquad + (-1)^{|b_1|}h(b_1)\Delta(db_2) + (\tau\Delta)(b_1)hd(b_2) \\ &\quad = (\Delta - \tau\Delta)(b_1)\Delta(b_2) + (\tau\Delta)(b_1)(\Delta - \tau\Delta)(b_2) \\ &\quad = \Delta(b_1)\Delta(b_2) - (\tau\Delta)(b_1)(\tau\Delta)(b_2) = (\Delta - \tau\Delta)(b_1b_2). \end{aligned}$$

□

Theorem 21 *For a shelf X, the map h induces a homotopy between the cup product $\smile$ and its opposite version $\smile^{op} := \smile\tau$ on $C^\bullet(X)$ and $C^\bullet(X, M(X))$.*

Using a standard argument we obtain an elementary algebraic proof of the commutativity of the cup product on the rack cohomologies $H_{\mathrm{R}}(X)$ and $H_{\mathrm{R}}(X, M(X))$. The same result holds for the more general cohomologies $H_{\mathrm{R}}(X, k)$ and $H_{\mathrm{R}}(X, kM(X))$.

Proof The cup product of two cochains f and g is given by the convolution product

$$f \smile g = \mu(f \otimes g)\Delta,$$

where the coproduct Δ is taken in $\bar{B}$, and μ is the multiplication in $\mathbb{Z}$. Hence for any homogeneous $x \in \bar{B}$ of degree $|f| + |g|$ we have

$$\begin{aligned}(f \smile g - (-1)^{|f||g|} g \smile f)(x) &= \sum_{(x)} (-1)^{|f||g|} f(x_1)g(x_2) - g(x_1)f(x_2) \\ &= \sum_{(x)} (-1)^{|x_1||x_2|} f(x_1)g(x_2) - f(x_2)g(x_1) \\ &= \mu(f \otimes g)(\Delta - \tau\Delta)(x) \\ &= \mu(f \otimes g)(hd + dh)(x).\end{aligned}$$

We use Sweedler's notation for $\Delta(x)$. Hence $H\colon \mathrm{Hom}_A(B, \mathbb{Z})^{\otimes 2} \to \mathrm{Hom}_A(B, \mathbb{Z})$ defined by

$$H(f \otimes g) := \mu(f \otimes g)h$$

is a homotopy between $\smile$ and $\smile^{op}$. The proof for the cohomology with coefficients in M is similar. □

7 Rack cohomology is a Zinbiel algebra

To better understand the coproduct Δ on $\bar{B}(X)$, we now refine it to an (almost) *dg codendriform structure*. That is, in positive degree it decomposes as $\Delta = \overleftarrow{\Delta} + \overrightarrow{\Delta}$, the two parts $\overleftarrow{\Delta}$ and $\overrightarrow{\Delta}$ being compatible. Moreover, we establish the relation $\overrightarrow{\Delta} = \tau\overleftarrow{\Delta}$ (where τ is as usual the signed flip), up to an explicit homotopy $\bar{h}$. This homotopy is inspired by the homotopy h from Section 6, and is, to our knowledge, new. We thus recover the *Zinbiel product* on rack cohomology, first described by Covez in [7].

Coalgebras need not be unital in this section. General definitions are given over a unital commutative ring k; in particular, all the tensor products are taken over k here.

Definition 22 A graded coalgebra $\bigl(C = \bigoplus_{i\geqslant 0} C_i, \Delta\bigr)$ is called *+–codendriform* if there exist two maps of degree 0 on its positive-degree part $C^+ = \bigoplus_{i\geqslant 1} C_i$, denoted by $\overleftarrow{\Delta}\colon C^+ \to C^+ \otimes C$ and $\overrightarrow{\Delta}\colon C^+ \to C \otimes C^+$, satisfying

$$(\overleftarrow{\Delta} \otimes \mathrm{Id})\overleftarrow{\Delta} = (\mathrm{Id} \otimes \Delta)\overleftarrow{\Delta}, \tag{13}$$

$$(\mathrm{Id} \otimes \overrightarrow{\Delta})\overrightarrow{\Delta} = (\Delta \otimes \mathrm{Id})\overrightarrow{\Delta}, \tag{14}$$

$$(\mathrm{Id} \otimes \overleftarrow{\Delta})\overrightarrow{\Delta} = (\overrightarrow{\Delta} \otimes \mathrm{Id})\overleftarrow{\Delta}, \tag{15}$$

and where Δ decomposes as $\overleftarrow{\Delta} + \overrightarrow{\Delta}$ on C^+ and Δ is coassociative on C_0. It is called *+–co-Zinbiel* if moreover $\overrightarrow{\Delta} = \tau\overleftarrow{\Delta}$, where τ is the signed flip. A *dg +–codendriform* or *+–co-Zinbiel coalgebra* carries in addition a differential d satisfying

$$\overleftarrow{\Delta} d = (d \otimes \mathrm{Id})\overleftarrow{\Delta} + (\mathrm{Id} \otimes d)\overleftarrow{\Delta} \quad \text{on } \bigoplus_{i\geqslant 2} C_i, \tag{16}$$

$$\overrightarrow{\Delta} d = (d \otimes \mathrm{Id})\overrightarrow{\Delta} + (\mathrm{Id} \otimes d)\overrightarrow{\Delta} \quad \text{on } \bigoplus_{i\geqslant 2} C_i, \tag{17}$$

$$\Delta d = (d \otimes \mathrm{Id})\overleftarrow{\Delta} + (\mathrm{Id} \otimes d)\overrightarrow{\Delta} \quad \text{on } C_1. \tag{18}$$

Dually, one defines (dg) *+–dendriform* and *+–Zinbiel* algebras.

In the case when the 0–degree part C_0 is empty, one recovers the familiar *(co)dendriform* and *(co)Zinbiel* structures. One can play with this idea further, and extend a positively graded codendriform coalgebra $(C^+, \overleftarrow{\Delta}^+, \overrightarrow{\Delta}^+)$ by a unit: $C := C^+ \oplus k1$, with $\Delta(1) = \overleftarrow{\Delta}(1) = \overrightarrow{\Delta}(1) = 1 \otimes 1$, and $\overleftarrow{\Delta}(c) = \overleftarrow{\Delta}^+(c) + c \otimes 1$ and $\overrightarrow{\Delta}(c) = \overrightarrow{\Delta}^+(c) + 1 \otimes c$ for all $c \in C$. One can also go in the opposite direction:

Lemma 23 *Let $(C, \Delta, \overleftarrow{\Delta}, \overrightarrow{\Delta})$ be a +–codendriform coalgebra. Denote by $\varepsilon\colon C \to C^+$ and $\iota\colon C^+ \to C$ the obvious projection and inclusion, where $C^+ := \bigoplus_{i>0} C_i$. Put $\Delta^+ := (\varepsilon \otimes \varepsilon)\Delta\iota$, $\overleftarrow{\Delta}^+ := (\varepsilon \otimes \varepsilon)\overleftarrow{\Delta}\iota$ and $\overrightarrow{\Delta}^+ := (\varepsilon \otimes \varepsilon)\overrightarrow{\Delta}\iota$. Then $(C^+, \overleftarrow{\Delta}^+, \overrightarrow{\Delta}^+)$ is a codendriform coalgebra.*

The proof is straightforward. These observations explain our choice of the name. In the literature there exist alternative approaches to such "almost codendriform" structures.

Finally, one easily checks that a +–codendriform structure refines a coassociative one:

Lemma 24 *In a (dg) +–codendriform coalgebra, the coproduct Δ is necessarily coassociative. It is also compatible with the differential: writing $\Delta(b) = \sum_i a_i \otimes c_i$ for a given $b \in C$, with all the a_i homogeneous, one has*

$$\Delta(d(b)) = \sum_i d(a_i) \otimes c_i + \sum_i (-1)^{|a_i|} a_i \otimes d(c_i).$$

Let us now return to shelves and their associated dg bialgebras.

Proposition 25 *Let X be a shelf. Define two maps $\overleftarrow{\Delta}\colon B(X)^+ \to B(X)^+ \otimes B(X)$ and $\overrightarrow{\Delta}\colon B(X)^+ \to B(X) \otimes B(X)^+$ by*

$$\overleftarrow{\Delta}(ae_{x_1}\cdots e_{x_n}) = (ae_{x_1} \otimes ax_1)\Delta(e_{x_2}\cdots e_{x_n}),$$
$$\overrightarrow{\Delta}(ae_{x_1}\cdots e_{x_n}) = (a \otimes ae_{x_1})\Delta(e_{x_2}\cdots e_{x_n}),$$

where as usual we use the canonical form of monomials in $B(X)$, and extend this definition by linearity. These maps and the coproduct Δ yield a $+$–codendriform structure on $B(X)$.

Proof Put $\Delta^2 = (\Delta \otimes \mathrm{Id})\Delta = (\mathrm{Id} \otimes \Delta)\Delta$. Then both sides of (13) act on a canonical monomial as follows:

$$ae_{x_1}\cdots e_{x_n} \mapsto (ae_{x_1} \otimes ax_1 \otimes ax_1)\Delta^2(e_{x_2}\cdots e_{x_n}).$$

Similarly, both sides of (14) and (15) act by

$$ae_{x_1}\cdots e_{x_n} \mapsto (a \otimes a \otimes ae_{x_1})\Delta^2(e_{x_2}\cdots e_{x_n})$$

and

$$ae_{x_1}\cdots e_{x_n} \mapsto (a \otimes ae_{x_1} \otimes ax_1)\Delta^2(e_{x_2}\cdots e_{x_n}),$$

respectively. Thus our maps satisfy relations (13)–(15). Finally, in positive degree their sum clearly yields Δ, and in degree 0 the coproduct Δ is coassociative. □

The maps above are not compatible with the differential in general, since

$$\overleftarrow{\Delta}d(e_xe_y) = e_y \otimes y - xe_y \otimes xy - e_x \otimes x + e_xy \otimes xy,$$
$$(d \otimes \mathrm{Id})\overleftarrow{\Delta} + (\mathrm{Id} \otimes d)\overleftarrow{\Delta}(e_xe_y) = e_y \otimes xy - xe_y \otimes xy - e_x \otimes x + e_xy \otimes xy + 1 \otimes xe_y - x \otimes xe_y.$$

As usual, the solution is to work in the quotient $\overline{B}(X)$. Indeed, $\overleftarrow{\Delta}$ and $\overrightarrow{\Delta}$ descend to maps $\overline{B}(X)^+ \to \overline{B}(X)^+ \otimes \overline{B}(X)$ and $\overline{B}(X)^+ \to \overline{B}(X) \otimes \overline{B}(X)^+$, still denoted by $\overleftarrow{\Delta}$ and $\overrightarrow{\Delta}$, and one has:

Proposition 26 *The induced maps $\overleftarrow{\Delta}$ and $\overrightarrow{\Delta}$ make $\overline{B}(X)$ a dg $+$–codendriform coalgebra.*

Proof Recall the interpretation (6) of $\overline{B}$ as the quotient of B by $xb \sim b$ for all $x \in X$ and $b \in B$. This yields that the maps $\overleftarrow{\Delta}$ and $\overrightarrow{\Delta}$ are symmetric in $\overline{B}$:

$$\overleftarrow{\Delta}(e_{x_1}\cdots e_{x_n}) = (e_{x_1} \otimes 1)\Delta(e_{x_2}\cdots e_{x_n}), \quad \overrightarrow{\Delta}(e_{x_1}\cdots e_{x_n}) = (1 \otimes e_{x_1})\Delta(e_{x_2}\cdots e_{x_n}).$$

Also, it turns (5) into

$$d(e_{x_1}\cdots e_{x_n}) = -e_{x_1} d(e_{x_2}\cdots e_{x_n}). \tag{19}$$

We can now establish relation (16):

$$\begin{aligned}\overleftarrow{\Delta} d(e_{x_1}\cdots e_{x_n}) = \overleftarrow{\Delta}(-e_{x_1} d(e_{x_2}\cdots e_{x_n})) &= -(e_{x_1}\otimes 1)\Delta d(e_{x_2}\cdots e_{x_n})\\ &= -(e_{x_1}\otimes 1)(d\otimes \mathrm{Id})\Delta(e_{x_2}\cdots e_{x_n}) - (e_{x_1}\otimes 1)(\mathrm{Id}\otimes d)\Delta(e_{x_2}\cdots e_{x_n})\\ &= (d\otimes \mathrm{Id})(e_{x_1}\otimes 1)\Delta(e_{x_2}\cdots e_{x_n}) + (\mathrm{Id}\otimes d)(e_{x_1}\otimes 1)\Delta(e_{x_2}\cdots e_{x_n})\\ &= (d\otimes \mathrm{Id}+\mathrm{Id}\otimes d)\overleftarrow{\Delta}(e_{x_1}\cdots e_{x_n}).\end{aligned}$$

Relation (17) is proved similarly. Finally, relation (18) follows from

$$\Delta d = (d\otimes \mathrm{Id} + \mathrm{Id}\otimes d)\Delta$$

in degree 1. □

Proposition 27 *Define the map* $\bar{h}\colon B(X)\to B(X)\otimes B(X)$ *by* $\bar{h}(a)=0$ *and*

$$\bar{h}(ae_{x_1}\cdots e_{x_n}) = -(ax_1\otimes ae_{x_1})h(e_{x_2}\cdots e_{x_n}).$$

It induces a map $\bar{B}\to\bar{B}\otimes\bar{B}$*, still denoted by* $\bar{h}$*, which is a homotopy between* $\overrightarrow{\Delta}$ *and* $\tau\overleftarrow{\Delta}$*.*

Proof The map $\bar{h}$ clearly descends to $\bar{B}$. For this induced map, one has

$$\bar{h}(e_{x_1}\cdots e_{x_n}) = -(1\otimes e_{x_1})h(e_{x_2}\cdots e_{x_n}).$$

It remains to check the relation

$$(d\otimes \mathrm{Id}_{\bar{B}} + \mathrm{Id}_{\bar{B}}\otimes d)\bar{h} + \bar{h}d = \tau\overleftarrow{\Delta} - \overrightarrow{\Delta}\colon \bar{B}^+\to\bar{B}\otimes\bar{B}.$$

Using (19), one computes

$$\begin{aligned}(d\otimes \mathrm{Id})\bar{h}(e_{x_1}\cdots e_{x_n}) &= -(d\otimes \mathrm{Id})(1\otimes e_{x_1})h(e_{x_2}\cdots e_{x_n})\\ &= (1\otimes e_{x_1})(d\otimes \mathrm{Id})h(e_{x_2}\cdots e_{x_n}),\\ (\mathrm{Id}\otimes d)\bar{h}(e_{x_1}\cdots e_{x_n}) &= -(\mathrm{Id}\otimes d)(1\otimes e_{x_1})h(e_{x_2}\cdots e_{x_n})\\ &= (1\otimes e_{x_1})(\mathrm{Id}\otimes d)h(e_{x_2}\cdots e_{x_n}),\\ \bar{h}d(e_{x_1}\cdots e_{x_n}) = -\bar{h}(e_{x_1}d(e_{x_2}\cdots e_{x_n})) &= (1\otimes e_{x_1})hd(e_{x_2}\cdots e_{x_n}).\end{aligned}$$

The sum yields

$$
\begin{aligned}
(1\otimes e_{x_1})((d\otimes \mathrm{Id}+\mathrm{Id}\otimes d)h+hd)(e_{x_2}\cdots e_{x_n})&\\
&=(1\otimes e_{x_1})(\Delta-\tau\Delta)(e_{x_2}\cdots e_{x_n})\\
&=(1\otimes e_{x_1})\Delta(e_{x_2}\cdots e_{x_n})-\tau((e_{x_1}\otimes 1)\Delta(e_{x_2}\cdots e_{x_n}))\\
&=\overrightarrow{\Delta}(e_{x_1}e_{x_2}\cdots e_{x_n})-\tau\overleftarrow{\Delta}(e_{x_1}e_{x_2}\cdots e_{x_n}),
\end{aligned}
$$

as desired. □

As usual, using Lemma 9 one deduces from Proposition 25 a $+$–dendriform structure on the complex defining rack cohomology, and from Proposition 27 a $+$–Zinbiel product on the rack cohomology. Lemma 23 then yields dendriform and Zinbiel structures in positive degree:

Theorem 28 *For a shelf X, the complex $\left(\bigoplus_{n\geqslant 1} C^n(X),\partial^*\right)$ admits a dendriform structure, which is Zinbiel up to a homotopy induced by $\bar{h}$. The rack cohomology of X thus receives a strictly Zinbiel product.*

Remark 29 The dendriform structures above are not surprising. In [16; 17], rack cohomology is interpreted in terms of quantum shuffle algebras, which are key examples of dendriform structures. The shuffle interpretation generalizes to the cohomology of solutions to the Yang–Baxter equation, where dendriform structures reappear as well. The Zinbiel structure in cohomology is on the contrary remarkable, and does not extend to the YBE setting. Shuffles also suggest that, for $B(X)^+$, the codendriform structure and the associative product are compatible, in the sense of [25]. However, this does not seem to yield Zinbiel-coassociative structures on rack cohomologies. If we choose to work without coefficients (ie in $\bar{B}(X)^+$), the dendriform structure is compatible with the differential but the coproduct is lost. If we take coefficients $kM(X)$ (ie we work in $B(X)^+$), where k is a field and X is finite, the coproduct is preserved but the dendriform structure does not survive in cohomology.

8 Quandle cohomology inside rack cohomology

If X is a spindle (eg a quandle), then the complex $C_\bullet(X,k)$ has a *degenerate subcomplex*

$$C_\bullet^{\mathrm{D}}(X,k)=\langle x^2 : x\in X\rangle.$$

In other words, it is the linear envelope of all monomials with repeating neighbors. Carter et al [3] defined the *quandle (co)homology* of X via the complexes

$$C_\bullet^{\mathrm{Q}}(X,k) := C_\bullet(X,k)/C_\bullet^{\mathrm{D}}(X,k) \quad \text{and} \quad C_{\mathrm{Q}}^\bullet(X,k) := \mathrm{Hom}(C_\bullet^{\mathrm{Q}}(X,k),k).$$

Litherland and Nelson [18] showed that in this case the complex $C_\bullet(X,k)$ splits:

$$C = C^{\mathrm{N}} \oplus C^{\mathrm{D}}.$$

The quandle (co)homology is then the (co)homology of the complement C^{N}. We will now show that this decomposition is already visible at the level of the dg bialgebra $B(X)$. Moreover, in the bialgebraic setting it will be particularly easy to prove that

- the Zinbiel product on rack cohomology induces one on quandle cohomology but does not restrict to quandle cohomology,
- the associative cup product on rack cohomology restricts to quandle cohomology.

Proposition 30 *Let X be a spindle. In $B(X)$, consider the ideal*

$$B^{\mathrm{D}}(X) := \langle e_x^2 : x \in X \rangle,$$

and the left sub-$A(X)$–module $B^{\mathrm{N}}(X)$ generated by the elements 1 and

$$(e_{x_1} - e_{x_2})(e_{x_2} - e_{x_3}) \cdots (e_{x_{n-1}} - e_{x_n}) e_{x_n} \quad \text{where } n \geqslant 1 \text{ and all } x_i \in X. \tag{20}$$

Then B decomposes as a dg A–bimodule:

$$B(X) = B^{\mathrm{N}}(X) \oplus B^{\mathrm{D}}(X). \tag{21}$$

Moreover B^{D} is a coideal and B^{N} is a left coideal and a left codendriform coideal of B.

Proof The expression (20) vanishes when $x_i = x_{i+1}$ for some i. Moreover, one has

$$e_{x_1} \cdots e_{x_n} = (e_{x_1} - e_{x_2})(e_{x_2} - e_{x_3}) \cdots (e_{x_{n-1}} - e_{x_n}) e_{x_n} + \text{terms from } B^{\mathrm{D}}.$$

This implies the decomposition (21) of abelian groups.

The subspaces B^{N} and B^{D} are homogeneous, and for any $y \in X$ one has

$$(e_{x_1} - e_{x_2}) \cdots (e_{x_{n-1}} - e_{x_n}) e_{x_n} y = y (e_{x_1^y} - e_{x_2^y}) \cdots (e_{x_{n-1}^y} - e_{x_n^y}) e_{x_n^y}.$$

So, B^{N} and B^{D} are graded sub-A–bimodules of B.

Let us now check that B^{D} is a differential coideal. Indeed, using the property $x^x = x$ of a spindle, one computes

(22) $d(e_x^2) = d(e_x)e_x - e_x d(e_x) = (1-x)e_x - e_x(1-x) = e_x - xe_x - e_x + xe_{x^x} = 0,$

(23) $\Delta(e_x^2) = e_x^2 \otimes x^2 + 1 \otimes e_x^2 + e_x \otimes xe_x - e_x \otimes xe_{x^x} = e_x^2 \otimes x^2 + 1 \otimes e_x^2.$

To check that B^{N} is a subcomplex of B, we need its alternative description:

Lemma 31 *$B^{\mathrm{N}}(X)$ is the left sub-$A(X)$–module generated by the elements 1 and*

(24) $(e_{x_1} - e_{y_1})(e_{x_2} - e_{y_2}) \cdots (e_{x_{n-1}} - e_{y_{n-1}})e_{x_n}$ *where $n \geqslant 1$ and all $x_i, y_i \in X$.*

Proof It is sufficient to represent an element of the form (24) as a linear combination of elements of the form (20). This can be done inductively using the following observation:

$$\begin{aligned}
&(e_x - e_y)(e_{x_1} - e_{x_2})(e_{x_2} - e_{x_3}) \cdots (e_{x_{n-1}} - e_{x_n})e_{x_n} \\
&\quad = (e_x - e_{x_1})(e_{x_1} - e_{x_2})(e_{x_2} - e_{x_3}) \cdots (e_{x_{n-1}} - e_{x_n})e_{x_n} \\
&\qquad\qquad - (e_y - e_{x_1})(e_{x_1} - e_{x_2})(e_{x_2} - e_{x_3}) \cdots (e_{x_{n-1}} - e_{x_n})e_{x_n}. \qquad \square
\end{aligned}$$

Now, for $a \in A$ and $x_1, \dots, x_n \in X$, we have

$$\begin{aligned}
&d(a(e_{x_1} - e_{x_2})(e_{x_2} - e_{x_3}) \cdots (e_{x_{n-1}} - e_{x_n})e_{x_n}) \\
&\quad = ad(e_{x_1} - e_{x_2})(e_{x_2} - e_{x_3}) \cdots (e_{x_{n-1}} - e_{x_n})e_{x_n} \\
&\qquad\qquad - a(e_{x_1} - e_{x_2})d((e_{x_2} - e_{x_3}) \cdots (e_{x_{n-1}} - e_{x_n})e_{x_n}) \\
&\quad = a(x_2 - x_1)(e_{x_2} - e_{x_3}) \cdots (e_{x_{n-1}} - e_{x_n})e_{x_n} \\
&\qquad\qquad - a(e_{x_1} - e_{x_2})d((e_{x_2} - e_{x_3}) \cdots (e_{x_{n-1}} - e_{x_n})e_{x_n}).
\end{aligned}$$

An inductive argument using the lemma shows that this lies in B^{N}.

It remains to prove that Δ, $\overleftarrow{\Delta}$ and $\overrightarrow{\Delta}$ send B^{N} to $B \otimes B^{\mathrm{N}}$. In degree 0 everything is clear. In higher degree, from

$$\Delta(e_x - e_y) = e_x \otimes x - e_y \otimes y + 1 \otimes (e_x - e_y)$$

one sees that any of Δ, $\overleftarrow{\Delta}$ and $\overrightarrow{\Delta}$ sends an expression of the form (24) to a linear combination of tensor products, where on the right one has a product of terms of the form z, $e_x - e_y$ and possibly an e_u at the end. By Lemma 31, all these right parts lie in B^{N}. $\square$

The proposition describes all the structure inherited from B by B^{D} and B^{N}. Indeed,

- B^{D} is not a subcoalgebra, as follows from (23),

- B^{D} is not a coideal in the dendriform sense, since

$$\overleftarrow{\Delta}(e_x^2) = e_x^2 \otimes x^2 + e_x \otimes x e_x, \tag{25}$$

- B^{N} is not a subalgebra of B, since $e_x \in B^{\mathrm{N}}$ for any $x \in X$, whereas $e_x^2 \in B^{\mathrm{D}}$,
- B^{N} is not a subcoalgebra either, since one has

$$\Delta((e_x - e_y)e_y) = e_y^2 \otimes (x-y)y + \text{terms from } B^{\mathrm{N}} \otimes B^{\mathrm{N}}, \tag{26}$$

  and $e_y^2 \otimes (x-y)y$ is a nonzero term from $B^{\mathrm{D}} \otimes B^{\mathrm{N}}$ in general.

In particular, there is no natural way to define a codendriform structure on B^{N}. Passing to the quotient $\overline{B}$ does not solve this problem: $\overline{B}^{\mathrm{D}}$ is still not a codendriform coideal because of (25), and $\overline{B}^{\mathrm{N}}$ is not a subcoalgebra of $\overline{B}$. Indeed, even if (26) implies $\Delta((e_x - e_y)e_y) \in \overline{B}^{\mathrm{N}} \otimes \overline{B}^{\mathrm{N}}$, things go wrong in degree 3, since

$$\Delta((e_x - e_y)(e_y - e_z)e_z) = e_z^2 \otimes (e_{X^Y} - e_{X^z} - e_Y + e_{Y^z}) + \text{terms from } \overline{B}^{\mathrm{N}} \otimes \overline{B}^{\mathrm{N}},$$

where $X = x^z$ and $Y = y^z$. One gets a term from $\overline{B}^{\mathrm{D}} \otimes \overline{B}^{\mathrm{N}}$ which does not vanish in general. However, since $e_{X^Y} = e_X = e_{X^z}$ and $e_{Y^z} = e_Y$ modulo the boundary, this term disappears in homology. More generally:

Proposition 32 *Let X be a spindle. The homology $\overline{H}(X)$ of $\overline{B}(X)$ decomposes as a graded abelian group*:

$$\overline{H}(X) = \overline{H}^{\mathrm{N}}(X) \oplus \overline{H}^{\mathrm{D}}(X). \tag{27}$$

If k is a field, then one obtains a decomposition

$$\overline{H}(X,k) = \overline{H}^{\mathrm{N}}(X,k) \oplus \overline{H}^{\mathrm{D}}(X,k), \tag{28}$$

with $\overline{H}^{\mathrm{D}}$ a coassociative coideal and $\overline{H}^{\mathrm{N}}$ a co-Zinbiel (and hence coassociative) coalgebra.

Dually, the cohomology $\overline{H}^\bullet(X)$ of $\overline{B}(X)^*$ decomposes as

$$\overline{H}^\bullet(X) = \overline{H}^\bullet_{\mathrm{N}}(X) \oplus \overline{H}^\bullet_{\mathrm{D}}(X), \tag{29}$$

with $\overline{H}^\bullet_{\mathrm{D}}$ a Zinbiel (and hence associative) ideal, and $\overline{H}^*_{\mathrm{N}}$ an associative subalgebra of $\overline{H}^\bullet$. The same holds for $\overline{H}^\bullet(X,k)$.

Proof Proposition 30 yields the desired decompositions, and, together with Propositions 4 and 26, shows that $\overline{H}^{\mathrm{D}}$ is a coideal and $\overline{H}^{\mathrm{N}}$ a left codendriform coideal. In particular,

$$\overleftarrow{\Delta}((\overline{H}^{\mathrm{N}})^+) \subseteq (\overline{H}^{\mathrm{N}})^+ \otimes \overline{H}^{\mathrm{N}} \oplus (\overline{H}^{\mathrm{D}})^+ \otimes \overline{H}^{\mathrm{N}}$$

and

$$\overrightarrow{\Delta}((\overline{H}^{\mathrm{N}})^+) \subseteq \overline{H}^{\mathrm{N}} \otimes (\overline{H}^{\mathrm{N}})^+ \oplus \overline{H}^{\mathrm{D}} \otimes (\overline{H}^{\mathrm{N}})^+.$$

But Proposition 27 yields the relation $\overrightarrow{\Delta} = \tau\overleftarrow{\Delta}$ in homology, hence the terms in $(\overline{H}^{\mathrm{D}})^+ \otimes \overline{H}^{\mathrm{N}}$ and $\overline{H}^{\mathrm{D}} \otimes (\overline{H}^{\mathrm{N}})^+$ above must be trivial. This shows that $\overline{H}^{\mathrm{N}}$ is in fact a co-Zinbiel coalgebra.

The proof for the cohomology $\overline{H}^\bullet$ is analogous. □

Again, this proposition describes all the structure inherited by $\overline{H}^{\mathrm{D}}(X,k)$: it is neither a subcoalgebra, nor a codendriform coideal. Indeed, computations (23) and (25) still yield counterexamples, since e_x^2 and e_x represent nontrivial classes in $\overline{H}^{\mathrm{D}}$ and $\overline{H}^{\mathrm{N}}$, respectively.

Now, in order to understand what our proposition means for quandle cohomology, we need to recall Lemma 9 and observe that the construction of B^{D} precisely repeats that of the degenerate complex. This yields:

Lemma 33 *For any spindle X, one has isomorphisms of complexes*

$$(C_\bullet^{\mathrm{Q}}, \partial) \cong \overline{B}^{\mathrm{N}} \quad \text{and} \quad (C^\bullet_{\mathrm{Q}}, \partial^*) \cong \overline{B}^*_{\mathrm{N}}.$$

Proposition 32 then translates as follows:

Theorem 34 *The rack cohomology of a spindle X decomposes into quandle and degenerate parts, so one has the isomorphism*

$$H_{\mathrm{R}}(X) \simeq H_{\mathrm{Q}}(X) \oplus H_{\mathrm{D}}(X)$$

of graded abelian groups. Moreover,

- *H_{Q} is an associative subalgebra of H_{R} and H_{D} is an associative ideal,*
- *H_{D} is a Zinbiel ideal, hence H_{Q} carries an induced Zinbiel product.*

The situation is rather subtle here. The Zinbiel product on rack cohomology does not restrict to the quandle cohomology; to get a Zinbiel product on H_{Q}, we need to consider it as a quotient of H_{R}. However, the associative product induced by the Zinbiel product does restrict to H_{Q}.

9 Quandle cohomology vs rack cohomology

The rack cohomology of spindles and quandles shares a lot with the Hochschild cohomology of monoids and groups. This analogy suggests that the degenerate subcomplex C^{D} can be ignored, and that the rack cohomology H_{R} and the quandle cohomology H_{Q} carry the same information about a spindle. Litherland and Nelson [18] showed this is not as straightforward as that: the degenerate part is highly nontrivial, and in particular contains the entire quandle part:

$$C^{\mathrm{D}}_{\bullet} \simeq C^{\mathrm{Q}}_{\bullet-1} \oplus C^{\mathrm{L}}_{\bullet} \quad \text{for } \bullet \geqslant 2.$$

Here $C^{\mathrm{L}}_{\bullet}(X) := \mathbb{Z}X \otimes C^{\mathrm{D}}_{\bullet-1}(X)$ is the *late degenerate subcomplex*, which is the linear envelope of all monomials with repetition at some place other than the beginning. This refines the rack cohomology splitting from Theorem 34:

$$H^{\bullet}_{\mathrm{R}} \simeq H^{\bullet}_{\mathrm{Q}} \oplus H^{\bullet-1}_{\mathrm{Q}} \oplus H^{\bullet}_{\mathrm{L}} \quad \text{for } \bullet \geqslant 2. \tag{30}$$

We will now recover this decomposition in our bialgebraic setting. However, our methods are not sufficient for coupling this decomposition with the algebraic structure on H_{R}:

Question 35 Do the cup product and the Zinbiel product on the rack cohomology of a spindle respect the decomposition (30) in any sense? In particular, can the quandle cohomology regarded as a *Zinbiel algebra* be reconstructed from the degenerate cohomology?

Now, even though H_{D} is big, it is degenerate in a certain sense. Indeed, Przytycki and Putyra [23] showed the quandle cohomology H_{Q} of a spindle to completely determine its rack cohomology H_{R}, and hence H_{D}, on the level of abelian groups. In light of the preceding section, the following question becomes particularly interesting:

Question 36 Can Zinbiel and associative structures on the rack cohomology of a spindle be recovered from the corresponding structures on its quandle cohomology?

Let us now return to our dg bialgebra $B(X)$:

Proposition 37 *Let X be a spindle. Put*

$$B^{\mathrm{L}}(X) := B^{+}(X) \otimes_{A(X)} B^{\mathrm{D}}(X),$$

where B^+ is the positive-degree part of B. One has the following isomorphism of graded $A(X)$–bimodules:

$$B^{\mathrm{D}}_\bullet(X) \simeq B^{\mathrm{L}}_\bullet(X) \oplus B^{\mathrm{Q}}_{\bullet-1}(X) \quad \textit{for } \bullet \geqslant 2. \tag{31}$$

This results immediately from the following technical lemma:

Lemma 38 *Let X be a spindle. Define a map $s\colon B^+(X) \to B^{\mathrm{D}}(X)$ as follows: take any element from B^+ written using the generators of the form x and e_y, and in each of its monomials replace the first letter of the form e_y by $e_y e_y$. Then*

- *s is a well-defined injective A–bilinear map of degree* 1,
- *one has the decomposition of graded A–bimodules*

$$B^{\mathrm{D}} = B^{\mathrm{L}} \oplus s((B^{\mathrm{N}})^+).$$

The map s yields the first degeneracy s_1 for the cubical structure underlying quandle cohomology, hence the notation.

Proof To show that s is well defined, one needs to check that it is compatible with the relation $e_x y = y e_{x^y}$ in B, that is, we should have $e_x e_x y = y e_{x^y} e_{x^y}$. This is indeed true:

$$e_x e_x y = e_x y e_{x^y} = y e_{x^y} e_{x^y}.$$

This map is A–bilinear and of degree 1 by construction. Injectivity becomes clear if one writes all the monomials in B in the canonical form $x_1 \cdots x_k e_{y_1} \cdots e_{y_n}$.

Further, the map $b \otimes b' \mapsto bb'$ identifies B^{L} with an A–invariant subspace of B^{D}, which is again clear using canonical forms.

Now, $s(B^+)$ and B^{L} are graded A–subbimodules of B^{D}, with $s(B^+) + B^{\mathrm{L}} = B^{\mathrm{D}}$ and $s(B^+) \cap B^{\mathrm{L}} = s(B^{\mathrm{D}})$. As usual, this is clear using canonical forms. Since $s(B^+) = s(B^{\mathrm{D}} \oplus (B^{\mathrm{N}})^+) = s(B^{\mathrm{D}}) \oplus s((B^{\mathrm{N}})^+)$, we obtain the desired decomposition $B^{\mathrm{D}} = B^{\mathrm{L}} \oplus s((B^{\mathrm{N}})^+)$. □

As usual, decomposition (31) implies

$$\bar{B}^{\mathrm{D}}_\bullet(X) \simeq \bar{B}^{\mathrm{L}}_\bullet(X) \oplus \bar{B}^{\mathrm{Q}}_{\bullet-1}(X) \quad \text{for } \bullet \geqslant 2, \tag{32}$$

with obvious notation. And as usual this decomposition respects more structure than (31):

Proposition 39 *Let X be a spindle. Then (32) is an isomorphism of differential graded $A(X)$–bimodules.*

Proof Let us check that the differential preserves $\bar{B}^{\mathrm{L}}$. An element of $\bar{B}^{\mathrm{L}}$ is a linear combination of (the $\sim$–equivalence classes of) terms of the form $e_x b$, with $b \in B^{\mathrm{D}}$. Since $d(e_x b) = -e_x d(b)$ in $\bar{B}$ by relation (6), and since d preserves B^{D}, we have $d(e_x b) \in \bar{B}^{\mathrm{L}}$.

Now, the map s from Lemma 38 induces a map $s\colon \bar{B}^+ \to \bar{B}^{\mathrm{D}}$. We need to check that the differential preserves $s((\bar{B}^{\mathrm{N}}_\bullet)^+) \simeq \bar{B}^{\mathrm{Q}}_{\bullet-1}(X)$. Adapting the arguments from the proof of Proposition 30, we see that an element of $s((\bar{B}^{\mathrm{N}})^+)$ is a linear combination of the classes of terms of the form $(e_x^2 - e_y^2)b$, with $b \in B^{\mathrm{N}}$. Then (22) yields

$$d((e_x^2 - e_y^2)b) = d(e_x^2 - e_y^2)b + (e_x^2 - e_y^2)d(b) = (e_x^2 - e_y^2)d(b).$$

Since d preserves B^{N}, we conclude $d((e_x^2 - e_y^2)b) \in s((B^{\mathrm{N}})^+)$. □

References

[1] **N Andruskiewitsch**, **M Graña**, *From racks to pointed Hopf algebras*, Adv. Math. 178 (2003) 177–243 MR Zbl

[2] **J S Carter**, **M Elhamdadi**, **M Saito**, *Homology theory for the set-theoretic Yang–Baxter equation and knot invariants from generalizations of quandles*, Fund. Math. 184 (2004) 31–54 MR Zbl

[3] **J S Carter**, **D Jelsovsky**, **S Kamada**, **L Langford**, **M Saito**, *Quandle cohomology and state-sum invariants of knotted curves and surfaces*, Trans. Amer. Math. Soc. 355 (2003) 3947–3989 MR Zbl

[4] **S Carter**, **S Kamada**, **M Saito**, *Surfaces in 4–space*, Encyclopaedia of Mathematical Sciences 142, Springer (2004) MR Zbl

[5] **F Clauwens**, *The algebra of rack and quandle cohomology*, J. Knot Theory Ramifications 20 (2011) 1487–1535 MR Zbl

[6] **S Covez**, *On the conjectural Leibniz cohomology for groups*, J. K–Theory 10 (2012) 519–563 MR Zbl

[7] **S Covez**, *Rack homology and conjectural Leibniz homology*, preprint (2014) arXiv 1402.1625

[8] **P Dehornoy**, *Braids and self-distributivity*, Progr. Math. 192, Birkhäuser, Basel (2000) MR Zbl

[9] **M Elhamdadi**, **S Nelson**, *Quandles—an introduction to the algebra of knots*, Student Mathematical Library 74, Amer. Math. Soc., Providence, RI (2015) MR Zbl

[10] **M A Farinati**, **J García Galofre**, *A differential bialgebra associated to a set theoretical solution of the Yang–Baxter equation*, J. Pure Appl. Algebra 220 (2016) 3454–3475 MR Zbl

[11] **R Fenn**, **C Rourke**, **B Sanderson**, *An introduction to species and the rack space*, from "Topics in knot theory" (M E Bozhüyük, editor), NATO Adv. Sci. Inst. Ser. C: Math. Phys. Sci. 399, Kluwer, Dordrecht (1993) 33–55 MR Zbl

[12] **R Fenn**, **C Rourke**, **B Sanderson**, *Trunks and classifying spaces*, Appl. Categ. Structures 3 (1995) 321–356 MR Zbl

[13] **R Fenn**, **C Rourke**, **B Sanderson**, *The rack space*, Trans. Amer. Math. Soc. 359 (2007) 701–740 MR Zbl

[14] **D Joyce**, *A classifying invariant of knots, the knot quandle*, J. Pure Appl. Algebra 23 (1982) 37–65 MR Zbl

[15] **M K Kinyon**, *Leibniz algebras, Lie racks, and digroups*, J. Lie Theory 17 (2007) 99–114 MR Zbl

[16] **V Lebed**, *Homologies of algebraic structures via braidings and quantum shuffles*, J. Algebra 391 (2013) 152–192 MR Zbl

[17] **V Lebed**, *Cohomology of idempotent braidings with applications to factorizable monoids*, Internat. J. Algebra Comput. 27 (2017) 421–454 MR Zbl

[18] **R A Litherland**, **S Nelson**, *The Betti numbers of some finite racks*, J. Pure Appl. Algebra 178 (2003) 187–202 MR Zbl

[19] **S V Matveev**, *Distributive groupoids in knot theory*, Mat. Sb. 119(161) (1982) 78–88 MR Zbl In Russian; translated in Math. USSR-Sb 47 (1984) 73–83

[20] **J H Przytycki**, *Distributivity versus associativity in the homology theory of algebraic structures*, Demonstratio Math. 44 (2011) 823–869 MR Zbl

[21] **J H Przytycki**, *Knots and distributive homology: from arc colorings to Yang–Baxter homology*, from "New ideas in low dimensional topology" (L H Kauffman, V O Manturov, editors), Ser. Knots Everything 56, World Sci., Hackensack, NJ (2015) 413–488 MR Zbl

[22] **J H Przytycki**, **K K Putyra**, *Homology of distributive lattices*, J. Homotopy Relat. Struct. 8 (2013) 35–65 MR Zbl

[23] **J H Przytycki**, **K K Putyra**, *The degenerate distributive complex is degenerate*, Eur. J. Math. 2 (2016) 993–1012 MR Zbl

[24] **J H Przytycki**, **W Rosicki**, *Cocycle invariants of codimension 2 embeddings of manifolds*, from "Knots in Poland, III" (J H Przytycki, P Traczyk, editors), Banach Center Publ. 103, Polish Acad. Sci. Inst. Math., Warsaw (2014) 251–289 MR Zbl

[25] **M Ronco**, *Primitive elements in a free dendriform algebra*, from "New trends in Hopf algebra theory" (N Andruskiewitsch, W R Ferrer Santos, H-J Schneider, editors), Contemp. Math. 267, Amer. Math. Soc., Providence, RI (2000) 245–263 MR Zbl

[26] **M Rosso**, *Groupes quantiques et algèbres de battage quantiques*, C. R. Acad. Sci. Paris Sér. I Math. 320 (1995) 145–148 MR Zbl

[27] **J-P Serre**, *Homologie singulière des espaces fibrés: applications*, Ann. of Math. 54 (1951) 425–505 MR Zbl

[28] **M Szymik**, *Permutations, power operations, and the center of the category of racks*, Comm. Algebra 46 (2018) 230–240 MR Zbl

[29] **M Takasaki**, *Abstraction of symmetric transformations*, Tôhoku Math. J. 49 (1943) 145–207 MR Zbl In Japanese

Lycée Sainte-Pulchérie
Istanbul, Turkey

Departamento de Matemática FCEyN, Universidad de Buenos Aires
Buenos Aires, Argentina

Normandie Univ, UNICAEN, CNRS, LMNO
Caen, France

Laboratoire de Mathématiques Blaise Pascal, Université Clermont-Auvergne
Clermont-Ferrand, France

simon.covez@gmail.com, mfarinat@dm.uba.ar, lebed@unicaen.fr, dominique.manchon@uca.fr

Received: 7 May 2019 Revised: 18 November 2021

Algebraic & Geometric Topology **23**:4 (2023) 1583–1600
DOI: 10.2140/agt.2023.23.1583
Published: 14 June 2023

Rigidity at infinity for the Borel function of the tetrahedral reflection lattice

ALESSIO SAVINI

If Γ is the fundamental group of a complete finite volume hyperbolic 3–manifold, Guilloux conjectured that the Borel function on the $\mathrm{PSL}(n,\mathbb{C})$–character variety of Γ should be rigid at infinity, that is it should stay bounded away from its maximum at ideal points.

We prove Guilloux's conjecture in the particular case of the reflection group associated to a regular ideal tetrahedron of $\mathbb{H}^3$.

57T10; 53C35, 57M27

1 Introduction

Let Γ be the fundamental group of a finite volume complete hyperbolic 3–manifold M. In the attempt to explore the rigidity properties of Γ, many mathematicians studied the space of representations of Γ inside a semisimple Lie group G. For instance, when $G = \mathrm{PSL}(n,\mathbb{C})$, Bucher, Burger and Iozzi [7] introduced the *Borel function* on the character variety $X(\Gamma, \mathrm{PSL}(n,\mathbb{C}))$ using bounded cohomology techniques. The Borel function is continuous with respect to the topology of pointwise convergence and its absolute value is bounded by the volume of M multiplied by a suitable constant depending on n. Additionally, the maximum is attained only by the conjugacy class of the representation $\pi_n \circ i$ (or by its complex conjugate), where $i : \Gamma \to \mathrm{PSL}(2,\mathbb{C})$ is the standard lattice embedding and $\pi_n : \mathrm{PSL}(2,\mathbb{C}) \to \mathrm{PSL}(n,\mathbb{C})$ is the irreducible representation. When $n = 2$ the Borel function boils down to the volume function introduced for instance by Dunfield [11] or Francaviglia [12] and its rigid behavior can be translated in terms of the Mostow rigidity theorem [20].

Beyond their intrinsic interest, the previous results have several important consequences for the birationality properties of the character variety $X(\Gamma, \mathrm{PSL}(n,\mathbb{C}))$. For example, both Dunfield [11] and Klaff and Tillmann [18] used the properties of the volume function to prove that the component of the variety $X(\Gamma, \mathrm{PSL}(2,\mathbb{C}))$ containing the holonomy of M is birational to its image through the peripheral holonomy map, which is obtained by restricting any representation to the fundamental groups of the cusps. A similar result has been obtained by Guilloux [17] for the geometric component of the $\mathrm{PSL}(n,\mathbb{C})$–character variety, but the author needed to conjecture that outside of an analytic neighborhood of the class of the representation $\pi_n \circ i$ the Borel function is bounded away from its maximum value.

In this paper we focus our attention on the reflection group associated to a regular ideal tetrahedron and we prove a weak version of [17, Conjecture 1] for every $n \geq 2$.

Theorem 1.1 *Let Γ be the reflection group associated to the regular ideal tetrahedron $(0,1,e^{\pi i/3},\infty)$ and let $\Gamma_0 < \mathrm{PSL}(2,\mathbb{C})$ be a torsion-free finite index subgroup of Γ. Let $\rho_k \colon \Gamma_0 \to \mathrm{PSL}(n,\mathbb{C})$ be a sequence of representations and assume that each ρ_k admits an equivariant measurable map $\varphi_k \colon \mathbb{P}^1(\mathbb{C}) \to \mathscr{F}(n,\mathbb{C})$. Suppose that $\lim_{k\to\infty} \beta_n(\rho_k) = \binom{n+1}{3}\operatorname{Vol}(\Gamma_0\backslash\mathbb{H}^3)$. Then there must exist a sequence $(g_k)_{k\in\mathbb{N}}$ of elements in $\mathrm{PSL}(n,\mathbb{C})$ such that for every $\gamma \in \Gamma_0$,*

$$\lim_{k\to\infty} g_k \rho_k(\gamma) g_k^{-1} = (\pi_n \circ i)(\gamma),$$

where $i \colon \Gamma_0 \to \mathrm{PSL}(2,\mathbb{C})$ is the standard lattice embedding and

$$\pi_n \colon \mathrm{PSL}(2,\mathbb{C}) \to \mathrm{PSL}(n,\mathbb{C})$$

is the irreducible representation.

This phenomenon, called *rigidity at infinity*, was proved by the author and Francaviglia [14, Theorem 1.1] for $n = 2$ and any nonuniform lattice of $\mathrm{PSL}(2,\mathbb{C})$ — notice that the same phenomenon holds for all rank-one representations of any rank-one lattice [21]. However, since in that case our proof exploited the existence of natural maps for nonelementary representations — see for instance Besson, Courtois and Gallot [2; 3; 4] and Francaviglia [13] — we could not use the same argument here.

For our purposes, the existence of a boundary map φ_k is crucial. Indeed, the possibility to express the Borel invariant $\beta_n(\rho_k)$ as the integral over a fundamental domain for $\Gamma_0\backslash\mathrm{PSL}(2,\mathbb{C})$ of the pullback of the Borel cocycle along the boundary map φ_k together with the maximality hypothesis allows us to prove the existence of a suitable sequence $(g_k)_{k\in\mathbb{N}}$ of elements in $\mathrm{PSL}(n,\mathbb{C})$ such that the sequence $(g_k\varphi_k(\gamma\underline{\xi}))_{k\in\mathbb{N}}$ is bounded, where $\underline{\xi} = (0,1,e^{\pi i/3},\infty)$ and γ is any element of Γ_0. The boundedness of the previous sequence implies the boundedness of $(g_k\rho_k g_k^{-1}(\gamma))_{k\in\mathbb{N}}$ for every $\gamma \in \Gamma_0$ and hence we reach our conclusion.

Organization

Section 2 is dedicated to preliminary definitions. We start with the notion of bounded cohomology for a locally compact group, then we recall the definition of the Borel cocycle and the Borel class. We finally introduce the Borel invariant for a representation $\rho\colon \Gamma \to \mathrm{PSL}(n,\mathbb{C})$ and we recall its rigidity property. Section 3 is devoted to the proof of the main theorem.

Acknowledgements

I would like to thank Alessandra Iozzi for having proposed this nice problem to me. I am also grateful to Marc Burger and Stefano Francaviglia for the enlightening discussions and the help they gave me. I thank Michelle Bucher and Antonin Guilloux for the interest they showed in this problem. I finally thank the referees for their suggestions to improve the quality of the paper.

2 Preliminary definitions

2.1 Bounded cohomology of semisimple Lie groups

Given a locally compact group G there exist several ways to introduce the notion of continuous bounded cohomology of G. The standard one relies on the complex of continuous bounded functions on tuples of G. Since we will deal only with semisimple Lie groups and their lattices, we are going to follow a different approach. Indeed, in this case, one can introduce the continuous bounded cohomology of G via the complex of essentially bounded measurable functions on the Furstenberg boundary. This definition is equivalent to the standard one thanks to the work by Burger and Monod [9, Corollary 1.5.3]. More generally, one can use any strong resolution of $\mathbb{R}$ via relatively injective G–modules to compute the continuous bounded cohomology of G. For a more detailed exposition about these notions, we refer the reader to Monod's book [19].

Let G be a semisimple Lie group of noncompact type and let $B(G)$ be its Furstenberg boundary. The latter can be identified with G/P, where P is a minimal parabolic subgroup of G. For instance, when $G = \mathrm{PSL}(2,\mathbb{C})$, its Furstenberg boundary is $B(G) = \mathbb{P}^1(\mathbb{C})$. Recall that $B(G)$ admits a canonical quasi-invariant measure obtained by the Haar measurable structure on the group G.

We define the space of *bounded measurable functions* on the Furstenberg boundary as

$$\mathcal{B}^\infty(B(G)^{n+1},\mathbb{R}) := \{f : B(G)^{n+1} \to \mathbb{R} \mid f \text{ is measurable}\}.$$

By introducing the usual equivalence relation $f \sim g$, where f and g are equivalent if and only if they coincide up to a measure zero subset, we can define the space of *essentially bounded measurable functions* as

$$L^\infty(B(G)^{n+1}, \mathbb{R}) := \mathcal{B}^\infty(B(G)^{n+1}, \mathbb{R})/\sim.$$

From now on, with an abuse of notation, we are going to write only f when we refer to its equivalence class $[f]_\sim$.

The space $L^\infty(B(G)^{n+1}, \mathbb{R})$ admits a natural G–module structure given by

$$(gf)(\xi_0, \dots, \xi_n) := f(g^{-1}\xi_0, \dots, g^{-1}\xi_n)$$

for every element $g \in G$, every function $f \in L^\infty(B(G)^{n+1}, \mathbb{R})$ and almost every $\xi_0, \dots, \xi_n \in B(G)$. Together with the *standard homogeneous coboundary operator*

$$\delta^n : L^\infty(B(G)^{n+1}, \mathbb{R}) \to L^\infty(B(G)^{n+2}, \mathbb{R}),$$

$$\delta^n f(\xi_0, \dots, \xi_{n+1}) = \sum_{i=0}^{n+1} (-1)^i f(\xi_0, \dots, \xi_{i-1}, \xi_{i+1}, \dots \xi_{n+1}),$$

we obtain a cochain complex $(L^\infty(B(G)^{\bullet+1}, \mathbb{R}), \delta^\bullet)$.

If we define the space of *G–invariant functions* as

$$L^\infty(B(G)^{n+1}, \mathbb{R})^G := \{f \in L^\infty(B(G)^{n+1}, \mathbb{R}) \mid gf = f \text{ for all } g \in G\},$$

we can restrict the coboundary operators to that collection of spaces getting a subcomplex $(L(B(G)^{\bullet+1}, \mathbb{R})^G; \delta^\bullet_|)$.

Definition 2.1 The *continuous bounded cohomology* of G is the cohomology of the subcomplex $(L^\infty(B(G)^{\bullet+1}, \mathbb{R})^G; \delta^\bullet_|)$ and it is denoted by $H^\bullet_{\mathrm{cb}}(G, \mathbb{R})$. In a similar way, if $\Gamma < G$ is a lattice, its bounded cohomology groups are given by the cohomology of the subcomplex $(L^\infty(B(G)^{\bullet+1}, \mathbb{R})^\Gamma; \delta^\bullet_|)$ and they are denoted by $H^\bullet_b(\Gamma, \mathbb{R})$.

Notice that in the case of a lattice we omitted the subscript c, since the topology inherited by Γ from G is the discrete one and the continuity issue becomes trivial. For both the group G and its lattices, from now on, we are going to omit the real coefficients when we refer to the continuous bounded cohomology groups.

Remarkably, one can consider the complex of bounded measurable functions

$$(\mathcal{B}^\infty(B(G)^{\bullet+1}, \mathbb{R}), \delta^\bullet)$$

to gain precious information about the continuous bounded cohomology of G. Here $\delta^\bullet$ still denotes the standard coboundary operator.

Proposition 2.2 [8, Proposition 2.1] *If we add to the complex* $(\mathcal{B}^\infty(B(G)^{\bullet+1},\mathbb{R}),\delta^\bullet)$ *the inclusion of coefficient* $\mathbb{R}\hookrightarrow\mathcal{B}^\infty(B(G),\mathbb{R})$, *we get back a strong resolution of* $\mathbb{R}$. *Hence there exists a canonical map*

$$\mathfrak{c}^\bullet\colon H^\bullet(\mathcal{B}^\infty(B(G)^{\bullet+1},\mathbb{R})^G)\to H^\bullet_{\mathrm{cb}}(G).$$

We conclude the section by observing that both Definition 2.1 and Proposition 2.2 are still valid if we consider the subcomplex of alternating cochains. Recall that an essentially bounded function or a bounded measurable function $f\colon B(G)^{n+1}\to\mathbb{R}$ is *alternating* if for every permutation $\sigma\in S_{n+1}$, it holds that

$$f(x_{\sigma(0)},\dots,x_{\sigma(n)})=\operatorname{sgn}(\sigma)f(x_0,\dots,x_n),$$

where sgn is the sign of the permutation.

2.2 The Borel cocycle

A *complete flag* F of $\mathbb{C}^n$ is a sequence of nested subspaces

$$F^0\subset F^1\subset\dots F^{n-1}\subset F^n$$

where $\dim_{\mathbb{C}} F^i=i$ for $i=1,\dots,n$. Let $\mathscr{F}(n,\mathbb{C})$ be the space parametrizing all the possible complete flags of $\mathbb{C}^n$. This is a complex variety which can be thought of as a homogeneous space obtained as the quotient of $\mathrm{PSL}(n,\mathbb{C})$ by any of its Borel subgroups. In this way $\mathscr{F}(n,\mathbb{C})$ is the realization of the Furstenberg boundary associated to $\mathrm{PSL}(n,\mathbb{C})$.

An *affine flag* (F,v) of $\mathbb{C}^n$ is a complete flag F together with a decoration

$$v=(v^1,\dots,v^n)\in(\mathbb{C}^n)^n$$

such that

$$F^i=\mathbb{C}v^i+F^{i-1}$$

for $i=1,\dots n$. For any 4–tuple of affine flags $\boldsymbol{F}=((F_0,v_0),\dots,(F_3,v_3))$ of $\mathbb{C}^n$ and given a multi-index $\boldsymbol{J}\in\{0,\dots,n-1\}^4$, we set

$$\mathcal{Q}(\boldsymbol{F},\boldsymbol{J}):=\left[\frac{\langle F_0^{j_0+1},\dots,F_3^{j_3+1}\rangle}{\langle F_0^{j_0},\dots,F_3^{j_3}\rangle};(v_0^{j_0+1},\dots,v_3^{j_3+1})\right].$$

In the notation above we denoted by $[V,(x_0,\dots,x_k)]$ the equivalence class of a complex m–dimensional vector space V together with a $(k+1)$–tuple of spanning vectors $(x_0,\dots,x_k)\in V^{k+1}$ modulo the diagonal action of $\mathrm{GL}(m,\mathbb{C})$.

Since the hyperbolic volume function $\mathrm{Vol}\colon \mathbb{P}^1(\mathbb{C})^4 \to \mathbb{R}$ can be thought of as defined on $(\mathbb{C}^2 \setminus \{0\})^4$, we can actually extend it on $(\mathbb{C}^2)^4$. Using such an extension, we define the cocycle B_n as

$$B_n((F_0, v_0), \dots, (F_3, v_3)) := \sum_{\boldsymbol{J} \in \{0,\dots,n-1\}^4} \mathrm{Vol}\, \mathcal{Q}(\boldsymbol{F}, \boldsymbol{J}), \tag{1}$$

where we consider the volume function exactly when the dimension of the vector space appearing in $\mathcal{Q}(\boldsymbol{F}, \boldsymbol{J})$ is equal to 2, and we set the volume equal to zero otherwise.

In the particular case of generic flags (see Definition 2.7), the definition of the Borel cocycle is given by Goncharov [15]. Its extension to the whole space of 4–tuples of flags is due to Bucher, Burger and Iozzi, who proved the following.

Proposition 2.3 [7, Corollary 13, Theorem 14] *The function B_n does not depend on the decoration used to compute it and hence it descends naturally to a function*

$$B_n\colon \mathscr{F}(n, \mathbb{C})^4 \to \mathbb{R}$$

on 4–tuples of flags which is defined everywhere. Moreover, that function is a measurable $\mathrm{PSL}(n, \mathbb{C})$*–invariant alternating cocycle whose absolute value is bounded by* $\binom{n+1}{3} \nu_3$*, where ν_3 is the volume of a positively oriented regular ideal tetrahedron in* $\mathbb{H}^3$.

As a consequence of Proposition 2.2, the function B_n naturally determines a bounded cohomology class in $H^3_{\mathrm{cb}}(\mathrm{PSL}(n, \mathbb{C}))$, which we are going to denote by $\beta_b(n)$.

Definition 2.4 The cocycle B_n is called a *Borel cocycle* and the class $\beta_b(n)$ is called a *bounded Borel class*.

Bucher, Burger and Iozzi [7, Theorem 2] proved that the cohomology group

$$H^3_{\mathrm{cb}}(\mathrm{PSL}(n, \mathbb{C}))$$

is a one-dimensional real vector space generated by the bounded Borel class. This generalizes a previous result by Bloch [5] for $\mathrm{PSL}(2, \mathbb{C})$.

We are going now to recall the main rigidity property of the Borel cocycle. Denote by $\mathcal{V}_n\colon \mathbb{P}^1(\mathbb{C}) \to \mathscr{F}(n, \mathbb{C})$ the Veronese map. Recall that, if $\mathcal{V}_n^i(\xi)$ is the i–dimensional space of the flag $\mathcal{V}_n(\xi)$ and ξ has homogeneous coordinates $[x : y]$, then we define $\mathcal{V}_n^{n-i}(\xi)$ as the $(n-i)$–dimensional subspace with basis

$$\left(0, \dots, 0, x^i, \binom{i}{1} x^{i-1} y, \dots, \binom{i}{j} x^{i-j} y^j, \dots, \binom{i}{i-1} x y^{i-1}, y^i, 0, \dots, 0\right)^T$$

where the first are k zeros and the last are $n-i-k-1$ zeros, for $k = 0, \dots, n-1-i$.

Definition 2.5 Let $(F_0,\dots,F_3)\in\mathscr{F}(n,\mathbb{C})^4$ be a 4–tuple of flags. We say that the 4–tuple is *maximal* if
$$|B_n(F_0,\dots,F_3)|=\binom{n+1}{3}\nu_3.$$

Maximal flags can be described in terms of the Veronese embedding. More precisely:

Theorem 2.6 [7, Theorem 19, Corollary 20] *Let (F_0,F_1,F_2,F_3) be a maximal 4–tuple of flags in $\mathscr{F}(n,\mathbb{C})$. Then there must exist a unique element $g\in\mathrm{PSL}(n,\mathbb{C})$ such that*
$$g(F_0,F_1,F_2,F_3)=(\mathcal{V}_n(0),\mathcal{V}_n(1),\mathcal{V}_n(\pm e^{\frac{i\pi}{3}}),\mathcal{V}_n(\infty))$$
where the sign $\pm$ reflects the sign of $B_n(F_0,\dots,F_3)$. Additionally, if (F_0,F_1,F_2,F_3) and (F_0,F_1,F_2,F_3') are both maximal with the same sign, then $F_3=F_3'$.

Now we discuss the continuity property of the Borel cocycle. The latter is measurable and not continuous since for instance one can consider a maximal 4–tuple of flags (F_0,F_1,F_2,F_3) and apply the sequence $(\pi_n(g)^k)_{k\in\mathbb{N}}$ to it, where $g\in\mathrm{PSL}(2,\mathbb{C})$ is loxodromic and $\pi_n\colon\mathrm{PSL}(2,\mathbb{C})\to\mathrm{PSL}(n,\mathbb{C})$ is the irreducible representation. In this way we get a sequence of maximal 4–tuples which degenerates at the limit and for that sequence the Borel cocycle is not continuous.

Nevertheless one can say something relevant about continuity when a 4–tuple of flags (F_0,F_1,F_2,F_3) satisfies a particular condition called *general position*.

Definition 2.7 Let $(F_0,F_1,F_2,F_3)\in\mathscr{F}(n,\mathbb{C})^4$ be a 4–tuple of flags. We say that the flags are in *general position* if
$$\dim_{\mathbb{C}}\langle F_0^{j_0},\dots F_3^{j_3}\rangle=j_0+\dots+j_3,$$
whenever $j_0+\dots+j_3\le n$.

For a 4–tuple of flags in general position and a multi-index $\boldsymbol{J}$ such that
$$j_0+\cdots+j_3=n-2,$$
the projection of the 4–tuple $(v_0^{j_0+1},\dots,v_3^{j_3+1})$ to the 2–dimensional vector space appearing in $\mathcal{Q}(\boldsymbol{F},\boldsymbol{J})$ gives us back a 4–tuple of distinct points on a projective line. Since such a 4–tuple varies continuously and the volume function Vol is continuous on 4–tuples of distinct points in $\mathbb{P}^1(\mathbb{C})$, we get that the Borel cocycle is continuous on $\mathrm{PSL}(n,\mathbb{C})$–orbits of 4–tuples of flags in general position.

The Borel cocycle can be used to understand when 4 flags are in general position.

Lemma 2.8 *Let $(F_0, F_1, F_2, F_3) \in \mathscr{F}(n, \mathbb{C})^4$ be a 4–tuple of flags. If*

$$\left| B_n(F_0, F_1, F_2, F_3) - \binom{n+1}{3} \nu_3 \right| < \varepsilon$$

for some $\varepsilon > 0$ sufficiently small, then the flags are in general position.

Proof We are going to denote by $C_k(n)$ the number of all the possible partitions of n by k integers.

Our proof will follow the line of [7, Lemma 15]. We will argue by induction on n. Suppose $n = 2$. The flags boil down to lines in $\mathbb{C}^2$ and those lines are in general position only if they are distinct. Since the Borel invariant is equal to zero when evaluated at two lines that coincide, the claim follows.

Assume now that the statement is true for $n-1$. Given a flag $F \in \mathscr{F}(n, \mathbb{C})$ we are going to denote by $\bar{F} \in \mathscr{F}(\mathbb{C}^n / \langle F_0^1 \rangle)$ the complete flag of the quotient $\mathbb{C}^n / \langle F_0^1 \rangle$ obtained by projecting F. Take the minimal value j such that $F_0^1 \subset F_1^j$.

We define the sets

$$\begin{aligned}
\mathcal{J}_1 &:= \{ \boldsymbol{J} \in \{0, \dots, n-1\}^4 \mid j_0 = j_1 = 0,\ 0 \le j_2, j_3 \le n-2 \}, \\
\mathcal{J}_2 &:= \{ \boldsymbol{J} \in \{0, \dots, n-1\}^4 \mid j_0 = 0,\ 0 < j_1 \le n-2,\ 0 \le j_2, j_3 \le n-2 \}, \\
\mathcal{J}_3 &:= \{ \boldsymbol{J} \in \{0, \dots, n-1\}^4 \mid 0 < j_0 \le n-2,\ 0 \le j_1, j_2, j_3 \le n-2 \}.
\end{aligned}$$

By following the same computation of Bucher, Burger and Iozzi [7, Equation 8, Lemma 17], we have

$$\begin{aligned}
(2) \quad \varepsilon &> \binom{n+1}{3} \nu_3 - B_n(F_0, \dots, F_3) \\
&= C_4(n-2)\nu_3 - \sum_{\boldsymbol{J} \in \{0,\dots,n-1\}^4} \operatorname{Vol} \mathcal{Q}(\boldsymbol{F}, \boldsymbol{J}) \\
&= \Big(C_4(n-3)\nu_3 - \sum_{\mathcal{J}_3} \operatorname{Vol} \mathcal{Q}(\boldsymbol{F}, \boldsymbol{J}) \Big) + \Big(C_3(n-3)\nu_3 - \sum_{\mathcal{J}_2} \operatorname{Vol} \mathcal{Q}(\boldsymbol{F}, \boldsymbol{J}) \Big) \\
&\qquad + \Big(C_2(n-2)\nu_3 - \sum_{\mathcal{J}_1} \operatorname{Vol} \mathcal{Q}(\boldsymbol{F}, \boldsymbol{J}) \Big),
\end{aligned}$$

where we used the fact that $C_4(n-2) = \binom{n+1}{3}$ and the recursive relation

$$C_k(n) = C_{k-1}(n) + C_k(n-1).$$

Notice that in the last line of the equation we removed the vanishing terms whose multi-index $\boldsymbol{J}$ does not lie in any $\mathcal{J}_i$ for $i = 1, 2, 3$.

It follows that if the Borel invariant is ε–near to its maximal value, then the sums over the sets $\mathcal{J}_1$, $\mathcal{J}_2$ and $\mathcal{J}_3$ are ε–near to their maximal values. By the symmetry in the roles played by the indices appearing in $\boldsymbol{J}$, we must have

$$\sum_{\substack{j_0=j_2=0 \\ 0\le j_1,j_3\le n-2}} \operatorname{Vol} \mathcal{Q}(\boldsymbol{F}, \boldsymbol{J}) > C_2(n-2)\nu_3 - \varepsilon.$$

Using the particular choice of j and following the same argument of [7, Lemma 15], we get that

$$(j-1)\nu_3 \ge C_2(n-2)\nu_3 - \varepsilon = (n-1)\nu_3 - \varepsilon,$$

and since ε is sufficiently small and j is an integer, j must be equal to n. This implies that $F_0^1 \subset F_1^n \setminus F_1^{n-1}$. A similar condition holds also for F_2 and F_3. In this way we get that

$$\overline{F_k^i} = \overline{F}_k^i,$$

for $k = 1, 2, 3$ and $0 \le i \le n-1$, whereas

$$\overline{F_0^i} = \overline{F}_0^{i-1},$$

for $i \ge 1$.

Consider now $0 \le j_0, j_1, j_2, j_3 \le n$ such that $j_0 + j_1 + j_2 + j_3 \le n$. The case $j_0 = \ldots = j_3 = 0$ is trivial, so we will assume $j_0 \ge 1$. By (2) we know that the sum over $\mathcal{J}_3$ is ε–near to its maximal value $C_4(n-3)\nu_3$. Thanks to [7, Equation 9], we can write

$$B_{n-1}(\overline{F}_0, \ldots, \overline{F}_3) = \sum_{\mathcal{J}_3} \operatorname{Vol} \mathcal{Q}(\boldsymbol{F}, \boldsymbol{J}) \ge C_4(n-3)\nu_3 - \varepsilon.$$

Hence $\overline{F}_0, \ldots, \overline{F}_3$ are in general position by the inductive hypothesis. In this way we get

$$\begin{aligned}\dim_{\mathbb{C}}\langle F_0^{j_0}, \ldots, F_3^{j_3}\rangle &= \dim_{\mathbb{C}}\langle \overline{F_0^{j_0}}, \ldots, \overline{F_3^{j_3}}\rangle + 1 \\ &= \dim_{\mathbb{C}}\langle \overline{F}_0^{j_0-1}, \overline{F}_1^{j_1}, \ldots, \overline{F}_3^{j_3}\rangle + 1 \\ &= (j_0-1) + j_1 + j_2 + j_3 + 1 \\ &= j_0 + j_1 + j_2 + j_3,\end{aligned}$$

and this finishes the proof of the lemma. □

The previous result is crucial in the proof of the following:

Lemma 2.9 *Let $(F_0^{(k)},\dots,F_3^{(k)})_{k\in\mathbb{N}}$ be a sequence of 4–tuples of flags such that*

$$\lim_{k\to\infty} B_n(F_0^{(k)},\dots,F_3^{(k)}) = \binom{n+1}{3}\nu_3.$$

Given a positively oriented regular ideal tetrahedron $\underline{\xi}=(\xi_0,\dots,\xi_3)$, there exists a sequence $(g_k)_{k\in\mathbb{N}}$ of elements $g_k\in \mathrm{PSL}(n,\mathbb{C})$ such that

$$\lim_{k\to\infty} g_k F_i^{(k)} = \mathcal{V}_n(\xi_i)$$

for $i=0,\dots,3$.

Proof By hypothesis we know that for k large enough,

$$\left| B_n(F_0^{(k)},\dots,F_3^{(k)}) - \binom{n+1}{3}\nu_3 \right| < \varepsilon$$

for $\varepsilon>0$ fixed. By Lemma 2.8, up to discarding the first terms of the sequence, we can suppose that $F_0^{(k)}, F_1^{(k)}, F_2^{(k)}, F_3^{(k)}$ are in general position. If F_0 and F_1 are flags and L is a line, using the transitivity of $\mathrm{PSL}(n,\mathbb{C})$ on triples (F_0,F_1,L) in general position [7, Lemma 23], we can find a unique element $g_k\in\mathrm{PSL}(n,\mathbb{C})$ such that

$$g_k F_0^{(k)} = \mathcal{V}_n(\xi_0), \quad g_k F_1^{(k)} = \mathcal{V}_n(\xi_1), \quad g_k (F_2^{(k)})^1 = \mathcal{V}_n^1(\xi_2).$$

On the subset of 4–tuples of flags (F_0,F_1,F_2,F_3) in general position such that $F_0=\mathcal{V}_n(\xi_0)$, $F_1=\mathcal{V}_n(\xi_1)$ and $F_2^1=\mathcal{V}_n^1(\xi_2)$ the Borel cocycle is continuous (since we fixed a set of representatives in the $\mathrm{PSL}(n,\mathbb{C})$–orbits) and thus we argue that

$$\lim_{k\to\infty} g_k (F_3^{(k)})^1 = \mathcal{V}_n^1(\xi_3).$$

Imitating the inductive argument in the proof of [7, Theorem 19] one can show that the same holds for the other subspaces of the flags $F_2^{(k)}$ and $F_3^{(k)}$. □

2.3 The Borel invariant for representations into $\mathrm{PSL}(n,\mathbb{C})$

Let Γ be a nonuniform lattice of $\mathrm{PSL}(2,\mathbb{C})$ without torsion and let $\rho\colon\Gamma\to\mathrm{PSL}(n,\mathbb{C})$ be a representation. Define $M:=\Gamma\backslash\mathbb{H}^3$. It is well known that we can decompose the manifold M as $M=N\cup\bigcup_{i=1}^h C_i$, where N is a compact core of M and for every $i=1,\dots,h$ the component C_i is a cuspidal neighborhood diffeomorphic to $T_i\times(0,\infty)$, where T_i is a torus. Since the fundamental group of the boundary ∂N is abelian, the maps $i_b^*\colon H_b^k(M,M\setminus N)\to H_b^k(M)$ induced at the level of bounded cohomology groups are isometric isomorphisms for $k\geq 2$; see [6]. Moreover, it holds that $H_b^k(M,M\setminus N)\cong H_b^k(N,\partial N)$ by homotopy invariance of bounded cohomology.

If we denote by $c\colon H_b^k(N,\partial N)\to H^k(N,\partial N)$ the comparison map, we can consider the composition

$$H_b^3(\mathrm{PSL}(n,\mathbb{C}))\xrightarrow{\rho_b^*} H_b^3(\Gamma)\cong H_b^3(M)\xrightarrow{(i_b^*)^{-1}} H_b^3(N,\partial N)\xrightarrow{c} H^3(N,\partial N),$$

where the isomorphism that appears in this composition holds by Gromov's mapping theorem [16].

Definition 2.10 The *Borel invariant* associated to a representation $\rho\colon\Gamma\to\mathrm{PSL}(n,\mathbb{C})$ is given by

$$\beta_n(\rho):=\langle (c\circ(i_b^*)^{-1}\circ(\rho_b^*))\beta_b(n),[N,\partial N]\rangle,$$

where the bracket $\langle\cdot,\cdot\rangle$ indicates the Kronecker pairing and $[N,\partial N]\in H_3(N,\partial N)$ is a fixed fundamental class.

The definition of the Borel invariant $\beta_n(\rho)$ is due to Bucher, Burger and Iozzi [7]. One can check that $\beta_n(\rho)$ does not depend on the choice of the compact core N and it can be suitably extended also to lattices with torsion. We want to remark that there exist other different approaches to the Borel invariant, for instance the one given by Dimofte, Gabella and Goncharov [10]. However, since they are all equivalent, we will consider [7] as our main reference.

The Borel invariant $\beta_n(\rho)$ remains unchanged on the $\mathrm{PSL}(n,\mathbb{C})$–conjugacy class of a representation ρ; hence it naturally defines a function on the character variety $X(\Gamma,\mathrm{PSL}(n,\mathbb{C}))$ which is continuous with respect to the topology of the pointwise convergence (this is a consequence of Proposition 2.12, for instance). This function, called the Borel function, satisfies a strong rigidity property.

Theorem 2.11 [7, Theorem 1] *Given any representation $\rho\colon\Gamma\to\mathrm{PSL}(n,\mathbb{C})$,*

$$|\beta_n(\rho)|\le\binom{n+1}{3}\mathrm{Vol}(M)$$

and equality holds if and only if ρ is conjugate to $\pi_n\circ i$ or its complex conjugate $\overline{\pi_n\circ i}$, where $i:\Gamma\to\mathrm{PSL}(2,\mathbb{C})$ is the standard lattice embedding and

$$\pi_n:\mathrm{PSL}(2,\mathbb{C})\to\mathrm{PSL}(n,\mathbb{C})$$

is the irreducible representation.

We want to conclude this section by expressing the Borel invariant in terms of boundary maps between Furstenberg boundaries. We first recall the definition of the transfer map $\mathrm{trans}_\Gamma\colon H^3_b(\Gamma) \to H^3_{\mathrm{cb}}(\mathrm{PSL}(2,\mathbb{C}))$. We can define the map

$$\mathrm{trans}_\Gamma\colon L^\infty(\mathbb{P}^1(\mathbb{C})^{n+1},\mathbb{R})^\Gamma \to L^\infty(\mathbb{P}^1(\mathbb{C})^{n+1},\mathbb{R})^{\mathrm{PSL}(2,\mathbb{C})},$$
$$\mathrm{trans}_\Gamma(c)(x_0,\dots,x_n) := \int_{\Gamma\backslash\mathrm{PSL}(2,\mathbb{C})} c(\bar{g}x_0,\dots,\bar{g}x_n)d\mu(\bar{g}),$$

where $\bar{g}$ stands for the equivalence class of g in $\Gamma\backslash\mathrm{PSL}(2,\mathbb{C})$ and μ is any invariant probability measure on $\Gamma\backslash\mathrm{PSL}(2,\mathbb{C})$. Since trans_Γ is a cochain map, we get a well-defined map

$$\mathrm{trans}_\Gamma\colon H^\bullet_b(\Gamma) \to H^\bullet_{\mathrm{cb}}(\mathrm{PSL}(2,\mathbb{C})).$$

Given a representation $\rho\colon \Gamma \to \mathrm{PSL}(n,\mathbb{C})$ we can consider the composition

$$H^3_{\mathrm{cb}}(\mathrm{PSL}(n,\mathbb{C})) \xrightarrow{\rho^*_b} H^3_b(\Gamma) \xrightarrow{\mathrm{trans}_\Gamma} H^3_{\mathrm{cb}}(\mathrm{PSL}(2,\mathbb{C})).$$

We have the following:

Proposition 2.12 [7, Propositions 26 and 28] *Considering the composition of the map ρ^*_b with the transfer map* trans_Γ,

$$(\mathrm{trans}_\Gamma \circ \rho^*_b)(\beta_b(n)) = \frac{\beta_n(\rho)}{\mathrm{Vol}(M)}\beta_b(2).$$

Given a measurable ρ–equivariant map $\varphi\colon \mathbb{P}^1(\mathbb{C}) \to \mathscr{F}(n,\mathbb{C})$, we can rewrite the above equation in terms of cochains as

$$\int_{\Gamma\backslash\mathrm{PSL}(2,\mathbb{C})} B_n(\varphi(g\xi_0),\dots\varphi(g\xi_3))\,d\mu(g) = \frac{\beta_n(\rho)}{\mathrm{Vol}(M)}\,\mathrm{Vol}(\xi_0,\dots,\xi_3) \tag{3}$$

for every $(\xi_0,\dots,\xi_3) \in \mathbb{P}^1(\mathbb{C})^4$.

3 Proof of the main theorem

In this section we prove our main theorem. The proof will follow the strategy adopted by Bucher, Burger and Iozzi for proving [7, Theorem 29].

Let Γ be the reflection group associated to the regular ideal tetrahedron of vertices $(0,1,e^{\pi i/3},\infty) \in \mathbb{P}^1(\mathbb{C})^4$ and let $\Gamma_0 < \mathrm{PSL}(2,\mathbb{C})$ be a torsion-free subgroup of Γ of finite index. From now until the end of the paper, with an abuse of notation, we are going to denote by g both a general element in $\mathrm{PSL}(2,\mathbb{C})$ and its equivalence class in $\Gamma_0\backslash\mathrm{PSL}(2,\mathbb{C})$.

Lemma 3.1 *Let Λ be a torsion-free lattice of $\mathrm{PSL}(2,\mathbb{C})$. Suppose $\rho_k : \Lambda \to \mathrm{PSL}(n,\mathbb{C})$ is a sequence of representations which satisfy $\lim_{k\to\infty} \beta_n(\rho_k) = \binom{n+1}{3}\operatorname{Vol}(\Lambda\backslash\mathbb{H}^3)$. Assume there exists a measurable map $\varphi_k : \mathbb{P}^1(\mathbb{C}) \to \mathscr{F}(n,\mathbb{C})$ which is ρ_k–equivariant. Then, up to passing to a subsequence, for almost every $g \in \operatorname{Isom}(\mathbb{H}^3)$,*

$$\lim_{k\to\infty} B_n(\varphi_k(g\xi_0),\dots,\varphi_k(g\xi_3)) = \binom{n+1}{3}\operatorname{Vol}(g\xi_0,\dots,g\xi_3),$$

where $(\xi_0,\dots,\xi_3) \in \mathbb{P}^1(\mathbb{C})^4$ are the vertices of a regular ideal tetrahedron.

Proof Let $(\xi_0,\dots,\xi_3) \in \mathbb{P}^1(\mathbb{C})^4$ be the vertices of a regular ideal tetrahedron. Without loss of generality we can assume that $\operatorname{Vol}(\xi_0,\dots,\xi_3) = \nu_3$. By Proposition 2.12 we know that (3) holds everywhere and hence we can write

$$\int_{\Lambda\backslash\mathrm{PSL}(2,\mathbb{C})} B_n(\varphi_k(g\xi_0),\dots,\varphi_k(g\xi_3))\, d\mu_{\Lambda\backslash G}(g) = \frac{\beta_n(\rho_k)}{\operatorname{Vol}(\Lambda\backslash\mathbb{H}^3)}\nu_3 \tag{4}$$

for every $k \in \mathbb{N}$, where $\mu_{\Lambda\backslash G}$ is the measure induced by the Haar measure and renormalized to be a probability measure. Since by hypothesis

$$\lim_{k\to\infty} \beta_n(\rho_k) = \binom{n+1}{3}\operatorname{Vol}(\Lambda\backslash\mathbb{H}^3),$$

by taking the limit on both sides of (4) we get

$$\lim_{k\to\infty}\int_{\Lambda\backslash\mathrm{PSL}(2,\mathbb{C})} B_n(\varphi_k(g\xi_0),\dots,\varphi_k(g\xi_3))d\mu_{\Lambda\backslash G}(g) = \binom{n+1}{3}\nu_3. \tag{5}$$

Since by Proposition 2.3 the Borel cocycle satisfies $|B(F_0,\dots,F_3)| \le \binom{n+1}{3}\nu_3$,

$$\binom{n+1}{3}\nu_3 - B_n(F_0,\dots,F_3) = \left|\binom{n+1}{3}\nu_3 - B_n(F_0,\dots,F_3)\right|$$

for every $(F_0,\dots,F_3) \in \mathscr{F}(n,\mathbb{C})^4$. If we denote by

$$\varphi_k^4 : \Lambda\backslash\mathrm{PSL}(2,\mathbb{C}) \to \mathscr{F}(n,\mathbb{C})^4, \quad \varphi_k^4(g) := (\varphi_k(g\xi_0),\dots,\varphi_k(g\xi_3)),$$

then (5) implies

$$\lim_{k\to\infty}\left\| B_n\circ\varphi_k^4 - \binom{n+1}{3}\nu_3\right\|_{L^1(\Lambda\backslash\mathrm{PSL}(2,\mathbb{C}),\mu_{\Lambda\backslash G})} = 0.$$

Since L^1–convergence implies the convergence almost everywhere of a suitable subsequence [1, Section 7], we can extract a subsequence $(\varphi_{k_\ell})_{\ell\in\mathbb{N}}$ such that

$$\lim_{\ell\to\infty} B_n(\varphi_{k_\ell}(g\xi_0),\dots,\varphi_{k_\ell}(g\xi_3)) = \binom{n+1}{3}\nu_3$$

for $\mu_{\Lambda\backslash G}$–almost every $g \in \Lambda\backslash\mathrm{PSL}(2,\mathbb{C})$. By the equivariance of the maps φ_{k_ℓ}, the equality above holds for μ_G–almost every $g \in \mathrm{PSL}(2,\mathbb{C})$.

If σ is a reflection along any face of $(\xi_0,\dots,\xi_3)$, the same argument can be adapted to a tetrahedron $(\sigma\xi_0,\dots,\sigma\xi_3)$ with negative maximal volume $\mathrm{Vol}(\sigma\xi_0,\dots,\sigma\xi_3)=-\nu_3$. Hence the statement follows. □

We can apply the previous theorem for a sequence of representations $\rho_k:\Gamma_0\to\mathrm{PSL}(n,\mathbb{C})$ with boundary maps $\varphi_k:\mathbb{P}^1(\mathbb{C})\to\mathscr{F}(n,\mathbb{C})$ such that

$$\lim_{k\to\infty}\beta_n(\rho_k)=\binom{n+1}{3}\mathrm{Vol}(\Gamma_0\backslash\mathbb{H}^3).$$

With an abuse of notation we are going to denote by $(\varphi_k)_{k\in\mathbb{N}}$ the subsequence that we get from Lemma 3.1.

Our goal now is to show that, up to translating each boundary map φ_k by an element $g_k\in\mathrm{PSL}(n,\mathbb{C})$, the sequence $g_k\varphi_k$ tends to the Veronese embedding on the vertices of the tiling of $\mathbb{H}^3$ by an ideal regular simplex. Denote by $\mathcal{T}_{\mathrm{reg}}\subset\mathbb{P}^1(\mathbb{C})^4$ the subset of 4–tuples which are the vertices of regular ideal tetrahedra. For every element $\underline{\xi}=(\xi_0,\dots,\xi_3)$ we denote by $\Gamma_{\underline{\xi}}$ the subgroup of $\mathrm{Isom}(\mathbb{H}^3)$ generated by the reflections along the faces of $\underline{\xi}$.

We start with the following:

Lemma 3.2 *Let $\underline{\xi}\in\mathcal{T}^\infty$ be a regular tetrahedron. Consider a sequence of measurable maps $\varphi_k:\mathbb{P}^1(\mathbb{C})\to\mathscr{F}(n,\mathbb{C})$. Define*

$$\mathcal{T}^\infty:=\left\{\underline{\xi}\in\mathcal{T}_{\mathrm{reg}}\;\middle|\;\lim_{k\to\infty}B_n(\varphi_k(\underline{\xi}))=\binom{n+1}{3}\mathrm{Vol}(\underline{\xi})\right\}\tag{6}$$

where $\varphi_k(\underline{\xi}):=(\varphi_k(\xi_0),\dots,\varphi_k(\xi_3))$ for every regular tetrahedron $\underline{\xi}=(\xi_0,\dots,\xi_3)\in\mathcal{T}_{\mathrm{reg}}$. Suppose that for every $\gamma\in\Gamma_{\underline{\xi}}$ we have that $\gamma\underline{\xi}\in\mathcal{T}^\infty$. Then there exists a sequence $(g_k)_{k\in\mathbb{N}}$, where each g_k is an element of $\mathrm{PSL}(n,\mathbb{C})$, such that

$$\lim_{k\to\infty}g_k\varphi_k(\alpha)=\mathcal{V}_n(\alpha)$$

for every $\alpha\in\bigcup_{i=0}^3\Gamma_{\underline{\xi}}\xi_i$.

Proof Since by hypothesis the tetrahedron $\underline{\xi}$ is an element of $\mathcal{T}^\infty$, by Lemma 2.9 we can find a sequence $(g_k)_{k\in\mathbb{N}}$ of elements in $\mathrm{PSL}(n,\mathbb{C})$ such that

$$\lim_{k\to\infty}g_k\varphi_k(\xi_i)=\mathcal{V}_n(\xi_i),$$

for $i=0,\dots,3$.

We want now to verify that the sequence $(g_k)_{k\in\mathbb{N}}$ is the one we were looking for. In order to do this we need to verify that

$$\lim_{k\to\infty} g_k\varphi_k(\gamma\xi_i) = \mathcal{V}_n(\gamma\xi_i)$$

for $i=0,\dots,3$ and for every $\gamma\in\Gamma_{\underline{\xi}}$. If γ is an arbitrary element of $\Gamma_{\underline{\xi}}$ we can write it as $\gamma = r_N\cdot r_{N-1}\cdots r_1$, where each r_i is a reflection along a face of the tetrahedron $r_{i-1}\cdots r_1\underline{\xi}$. We are going to prove the statement by induction on N. If $N=0$ there is nothing to prove. Assume the statement holds for $\gamma' = r_{N-1}\cdots r_1$. Denote by $\underline{\eta} = \gamma'\underline{\xi}$. We know that for the vertices of $\underline{\eta}$ we have

$$\lim_{k\to\infty} g_k\varphi_k(\eta_i) = \mathcal{V}_n(\eta_i)$$

for $i=0,\dots,3$. We want to prove that

$$\lim_{k\to\infty} g_k\varphi_k(r_N\eta_i) = \mathcal{V}_n(r_N\eta_i)$$

for $i=0,\dots,3$. Assume r_N is the reflection along the face of $\underline{\eta}$ whose vertices are η_1, η_2 and η_3. In particular we have that $r_N\eta_i=\eta_i$ for $i=1,2,3$, so for these vertices the statement holds. We are left to prove that

$$\lim_{k\to\infty} g_k\varphi_k(r_N\eta_0) = \mathcal{V}_n(r_N\eta_0).$$

The sequence $(g_k\varphi_k(r_N\eta_0))_{k\in\mathbb{N}}$ is a sequence of points in $\mathscr{F}(n,\mathbb{C})$, which is compact. Hence we can extract a subsequence which converges to a point $\alpha_0\in\mathscr{F}(n,\mathbb{C})$. By Lemma 2.8 we know that the 4–tuple $g_k\varphi_k(\underline{\eta})$ is eventually in general position. By the continuity of the Borel cocycle on the set of 4–tuples in general position we get

$$\lim_{k\to\infty} B_n\big(g_k\varphi_k(r_N\eta_0), g_k\varphi_k(\eta_1),\dots,g_k\varphi_k(\eta_3)\big) = B_n(\alpha_0,\mathcal{V}_n(\eta_1),\dots,\mathcal{V}_n(\eta_3)).$$

At the same time, by hypothesis it follows that

$$\lim_{k\to\infty} B_n(g_k\varphi_k(r_N\underline{\eta})) = \binom{n+1}{3}\operatorname{Vol}(r_N\underline{\eta}) = -\binom{n+1}{3}\operatorname{Vol}(\underline{\eta}).$$

On the other hand,

$$B_n(\mathcal{V}_n(r_N\underline{\eta})) = \binom{n+1}{3}\operatorname{Vol}(r_N\underline{\eta}) = -\binom{n+1}{3}\operatorname{Vol}(\underline{\eta}).$$

Hence, by a simple comparison argument, we get

$$B_n(\mathcal{V}_n(r_N\eta_0),\mathcal{V}_n(\eta_1),\dots,\mathcal{V}_n(\eta_3)) = B_n(\alpha_0,\mathcal{V}_n(\eta_1),\dots,\mathcal{V}_n(\eta_3)) = \pm\binom{n+1}{3}\nu_3.$$

As a consequence we must have $\alpha_0 = \mathcal{V}_n(r_N \eta_0)$, but this is equivalent to saying that the sequence $(g_k \varphi_k(r_N \eta_0))_{k\in\mathbb{N}}$ satisfies

$$\lim_{k\to\infty} g_k \varphi_k(r_N \eta_0) = \mathcal{V}_n(r_N \eta_0)$$

for any convergent subsequence of $(g_k \varphi_k(r_N \eta_0))_{k\in\mathbb{N}}$. The statement follows. □

We are now ready to prove the main theorem.

Proof of Theorem 1.1 Define the set

$$\mathcal{T}_\Gamma^\infty := \{\underline{\xi} \in \mathcal{T}^\infty \mid \gamma\underline{\xi} \in \mathcal{T}^\infty \text{ for all } \gamma \in \Gamma_{\underline{\xi}}\}.$$

We claim that this set is a set of full measure in $\mathcal{T}_{\mathrm{reg}}$. By Lemma 3.1, we already know that $\mathcal{T}^\infty$ defined by (6) is a set of full measure. For any $\underline{\eta} \in \mathcal{T}_{\mathrm{reg}}$ we define the evaluation map

$$\mathrm{ev}_{\underline{\eta}} \colon \mathrm{Isom}(\mathbb{H}^3) \to \mathcal{T}_{\mathrm{reg}}, \quad \mathrm{ev}_{\underline{\eta}}(g) := g\underline{\eta}.$$

Set $G^\infty := \mathrm{ev}_{\underline{\eta}}^{-1}(\mathcal{T}^\infty)$ and $G_\Gamma^\infty := \mathrm{ev}_{\underline{\eta}}^{-1}(\mathcal{T}_\Gamma^\infty)$. Let $\underline{\xi} = g\underline{\eta}$. Then $\underline{\xi} \in \mathcal{T}_\Gamma^\infty$ if and only if for any $\gamma \in \Gamma_{\underline{\xi}}$ we have that $\gamma\underline{\xi} = \gamma g\underline{\eta} \in \mathcal{T}^\infty$. Since $\Gamma_{\underline{\xi}} = \Gamma_{g\underline{\eta}} = g\Gamma_{\underline{\eta}} g^{-1}$, any element $\gamma \in \Gamma_{\underline{\xi}}$ can be written as $\gamma = g\gamma_0 g^{-1}$, where $\gamma_0 \in \Gamma_{\underline{\eta}}$. Thus, by a simple substitution, we get that $\underline{\xi} \in \mathcal{T}_\Gamma^\infty$ if and only if for every $\gamma_0 \in \Gamma_{\underline{\eta}}$ we have that $g\gamma_0\underline{\eta} \in \mathcal{T}^\infty$. This argument implies that we can write

$$G_\Gamma^\infty = \bigcap_{\gamma_0 \in \Gamma_{\underline{\eta}}} G^\infty \gamma_0^{-1}.$$

All the sets $G^\infty \gamma_0^{-1}$ are sets of full measure, since they are right-translates of the set of full measure G^∞ by the element γ_0^{-1}. Being a countable intersection of full measure sets, G_Γ^∞ also has full measure. Hence also $\mathcal{T}_\Gamma^\infty$ has full measure, as claimed.

Since all regular ideal tetrahedra are in a unique $\mathrm{Isom}(\mathbb{H}^3)$–orbit, up to conjugating each representation ρ_k, we can assume that $\xi = (0, 1, e^{\frac{\pi i}{3}}, \infty) \in \mathcal{T}_\Gamma^\infty$. With this assumption we have that $\Gamma_{\underline{\xi}} = \Gamma$, the reflection lattice we started with. By applying Lemma 3.2, there must exist a sequence $(g_k)_{k\in\mathbb{N}}$ of elements $g_k \in \mathrm{PSL}(n,\mathbb{C})$ such that

$$\lim_{k\to\infty} g_k \varphi_k(\gamma\underline{\xi}) = \mathcal{V}_n(\gamma\underline{\xi}) = \pi_n(\gamma)\mathcal{V}_n(\underline{\xi})$$

for every $\gamma \in \Gamma$ and hence for every $\gamma \in \Gamma_0$, where $\pi_n \colon \Gamma_0 \to \mathrm{PSL}(n,\mathbb{C})$ is the irreducible representation and $\mathcal{V}_n \colon \mathbb{P}^1(\mathbb{C}) \to \mathscr{F}(n,\mathbb{C})$ is the Veronese embedding. For every $k \in \mathbb{N}$ we define $\tilde{\varphi}_k := g_k \varphi_k$ and $\tilde{\rho}_k := g_k \rho_k g_k^{-1}$. We get that

$$\lim_{k\to\infty} \tilde{\rho}_k(\gamma)\tilde{\varphi}_k(\underline{\xi}) = \lim_{k\to\infty} \tilde{\varphi}_k(\gamma\underline{\xi}) = \mathcal{V}_n(\gamma\underline{\xi}) = \pi_n(\gamma)\mathcal{V}_n(\underline{\xi})$$

for every $\gamma \in \Gamma_0$. In particular notice that both sequences $(\varphi_k(\underline{\xi}))_{k\in\mathbb{N}}$ and $(\varphi_k(\gamma\underline{\xi}))_{k\in\mathbb{N}}$ converge. The element γ acts as $\pi_n(\gamma)$ at the limit, so the sequence $(\tilde{\rho}_k(\gamma))_{k\in\mathbb{N}}$ cannot diverge and it remains bounded in $\mathrm{PSL}(n,\mathbb{C})$. Hence the sequence of representations $(\tilde{\rho}_k)_{k\in\mathbb{N}}$ has to be bounded in the character variety $X(\Gamma_0, \mathrm{PSL}(n,\mathbb{C}))$ and there must exists a subsequence of $(\tilde{\rho}_k)_{k\in\mathbb{N}}$ converging to a suitable representation ρ_∞.

By the continuity of the Borel function on the character variety $X(\Gamma_0, \mathrm{PSL}(n,\mathbb{C}))$ with respect to the pointwise topology, it follows that

$$\beta_n(\rho_\infty) = \lim_{k\to\infty} \beta_n(\tilde{\rho}_k) = \lim_{k\to\infty} \beta_n(\rho_k) = \binom{n+1}{3} \mathrm{Vol}(\Gamma_0\backslash\mathbb{H}^3).$$

By [7, Theorem 1] the representation ρ_∞ must be conjugate to the representation $(\pi_n \circ i)$, where $i : \Gamma_0 \to \mathrm{PSL}(2,\mathbb{C})$ is the standard lattice embedding and

$$\pi_n : \mathrm{PSL}(2,\mathbb{C}) \to \mathrm{PSL}(n,\mathbb{C})$$

is the irreducible representation. Since the argument above holds for every convergent subsequence of $(\tilde{\rho}_k)_{k\in\mathbb{N}}$, the theorem follows. □

We conclude by noticing that in the proof we exploited crucially the combinatorial structure of the reflection group Γ. For this reason it seems unlikely the proof will adapt to more general lattices.

References

[1] **R G Bartle**, *The elements of integration and Lebesgue measure*, Wiley, New York (1995) MR Zbl

[2] **G Besson**, **G Courtois**, **S Gallot**, *Entropies et rigidités des espaces localement symétriques de courbure strictement négative*, Geom. Funct. Anal. 5 (1995) 731–799 MR Zbl

[3] **G Besson**, **G Courtois**, **S Gallot**, *Minimal entropy and Mostow's rigidity theorems*, Ergodic Theory Dynam. Systems 16 (1996) 623–649 MR Zbl

[4] **G Besson**, **G Courtois**, **S Gallot**, *A real Schwarz lemma and some applications*, Rend. Mat. Appl. 18 (1998) 381–410 MR Zbl

[5] **S J Bloch**, *Higher regulators, algebraic K–theory, and zeta functions of elliptic curves*, CRM Monograph Series 11, Amer. Math. Soc., Providence, RI (2000) MR Zbl

[6] **M Bucher**, **M Burger**, **R Frigerio**, **A Iozzi**, **C Pagliantini**, **M B Pozzetti**, *Isometric embeddings in bounded cohomology*, J. Topol. Anal. 6 (2014) 1–25 MR Zbl

[7] **M Bucher**, **M Burger**, **A Iozzi**, *The bounded Borel class and 3–manifold groups*, Duke Math. J. 167 (2018) 3129–3169 MR Zbl

[8] **M Burger**, **A Iozzi**, *Boundary maps in bounded cohomology*, Geom. Funct. Anal. 12 (2002) 281–292 MR Zbl Appendix to [9]

[9] **M Burger**, **N Monod**, *Continuous bounded cohomology and applications to rigidity theory*, Geom. Funct. Anal. 12 (2002) 219–280 MR Zbl

[10] **T Dimofte**, **M Gabella**, **A B Goncharov**, *K–decompositions and 3d gauge theories*, J. High Energy Phys. (2016) art. id. 151 MR Zbl

[11] **N M Dunfield**, *Cyclic surgery, degrees of maps of character curves, and volume rigidity for hyperbolic manifolds*, Invent. Math. 136 (1999) 623–657 MR Zbl

[12] **S Francaviglia**, *Hyperbolic volume of representations of fundamental groups of cusped 3–manifolds*, Int. Math. Res. Not. 2004 (2004) 425–459 MR Zbl

[13] **S Francaviglia**, *Constructing equivariant maps for representations*, Ann. Inst. Fourier (Grenoble) 59 (2009) 393–428 MR Zbl

[14] **S Francaviglia**, **A Savini**, *Volume rigidity at ideal points of the character variety of hyperbolic* 3*–manifolds*, Ann. Sc. Norm. Super. Pisa Cl. Sci. 20 (2020) 1325–1344 MR Zbl

[15] **A B Goncharov**, *Explicit construction of characteristic classes*, from "I M Gelfand Seminar" (S Gelfand, S Gindikin, editors), Adv. Soviet Math. 16, Amer. Math. Soc., Providence, RI (1993) 169–210 MR Zbl

[16] **M Gromov**, *Volume and bounded cohomology*, Inst. Hautes Études Sci. Publ. Math. 56 (1982) 5–99 MR Zbl

[17] **A Guilloux**, *Volume of representations and birationality of peripheral holonomy*, Exp. Math. 27 (2018) 472–477 MR Zbl

[18] **B Klaff**, **S Tillmann**, *A birationality result for character varieties*, Math. Res. Lett. 23 (2016) 1099–1110 MR Zbl

[19] **N Monod**, *Continuous bounded cohomology of locally compact groups*, Lecture Notes in Math. 1758, Springer (2001) MR Zbl

[20] **G D Mostow**, *Quasi-conformal mappings in n–space and the rigidity of hyperbolic space forms*, Inst. Hautes Études Sci. Publ. Math. 34 (1968) 53–104 MR Zbl

[21] **A Savini**, *Rigidity at infinity for lattices in rank-one Lie groups*, J. Topol. Anal. 12 (2020) 113–130 MR Zbl

Section de Mathématiques, University of Geneva
Geneva, Switzerland

alessio.savini@unige.ch

Received: 11 June 2019 Revised: 29 May 2021

Algebraic & Geometric Topology 23:4 (2023) 1601–1639
DOI: 10.2140/agt.2023.23.1601
Published: 14 June 2023

A construction of pseudo-Anosov homeomorphisms using positive twists

YVON VERBERNE

We introduce a construction of pseudo-Anosov homeomorphisms on n–times punctured spheres and surfaces of higher genus using only positive half-twists and Dehn twists. These constructions produce explicit examples of pseudo-Anosov maps with various number-theoretic properties associated to the stretch factors. For instance, we produce examples where the trace field is not totally real and the Galois conjugates of the stretch factor are on the unit circle. It follows that for these examples, no power of these maps can arise from either Thurston's or Penner's constructions.

15A18, 37E30, 57M07

1 Introduction

Pseudo-Anosov homeomorphisms play an important role in the study of homeomorphisms of orientable finite-type surfaces. The Nielsen–Thurston classification theorem states that, up to isotopy, every homeomorphism of a surface is either periodic, reducible, or pseudo-Anosov; see Thurston [14]. Reducible homeomorphisms can be reduced into pieces which are either finite-order or pseudo-Anosov. Therefore, understanding periodic and pseudo-Anosov homeomorphisms helps us understand arbitrary homeomorphisms. In this paper, we focus our attention on pseudo-Anosov homeomorphisms.

Despite the importance of pseudo-Anosov homeomorphisms, constructing explicit examples is difficult. Thurston [14] and Penner [10] each produced simple constructions which produce an infinite number of pseudo-Anosov maps on surfaces. In both constructions, the starting point is a pair of two filling multicurves which are used to produce pseudo-Anosov homeomorphisms.

We introduce a new construction of pseudo-Anosov homeomorphisms on punctured spheres. To prove that this construction produces maps which are different from the maps produced by the constructions of Penner and Thurston, we analyze the number-theoretic properties associated to these homeomorphisms. We also explain how to lift the construction to surfaces of higher genus using a branched cover.

1.1 The stretch factor of a pseudo-Anosov homeomorphism and number-theoretic properties

Let $S = S_{g,n}$ be a surface of genus g with n points removed from its interior. When convenient, we treat the punctures as marked points. A homeomorphism f of a finite type surface S is *pseudo-Anosov* if there is a representative homeomorphism ϕ, a real number $\lambda > 1$ and a pair of transverse measured foliations $\mathcal{F}^u$ and $\mathcal{F}^s$ such that $\phi(\mathcal{F}^u) = \lambda\mathcal{F}^u$ and $\phi(\mathcal{F}^s) = \lambda^{-1}\mathcal{F}^s$. The number λ by which ϕ stretches and contracts its foliations is called the *stretch factor* of f, and $\mathcal{F}^u$ and $\mathcal{F}^s$ are called the *unstable foliation* and *stable foliation*, respectively.

A considerable amount of research has probed the properties of the stretch factors of pseudo-Anosov maps. For example, Thurston [14] proved that the stretch factor of a pseudo-Anosov homeomorphism on a surface $S_{g,0}$ is an algebraic integer whose degree is bounded above by $6g-6$, where g is the genus of the surface.

Hubert and Lanneau [5] proved that for any pseudo-Anosov homeomorphism arising from Thurston's construction, the trace field $\mathbb{Q}(\lambda+\lambda^{-1})$ is always totally real. Shin and Strenner [13] proved that the Galois conjugates of the stretch factor for pseudo-Anosov homeomorphisms arising from Penner's construction are never on the unit circle in $\mathbb{C}$. This result from Shin and Strenner proved that it is not the case that every pseudo-Anosov homeomorphism has a power that arises from Penner's construction. This leads one to ask whether every pseudo-Anosov homeomorphism has a power which arises from either Penner's or Thurston's construction. To this end, one could ask whether it is possible for a pseudo-Anosov homeomorphism to have a trace field which is not totally real and have Galois conjugates of its stretch factor on the unit circle. The results of this paper allow us to construct examples giving a positive answer to this question.

Theorem 1.1 *Let S be either $S_{0,n}$ for $n \geq 8$, or $S_{g,k}$ for $g \geq 3$ and $k \geq 0$. Then there exists a pseudo-Anosov homeomorphism ϕ_S on S with stretch factor λ_{ϕ_S} such that*

(i) *the trace field $\mathbb{Q}(\lambda_{\phi_S}+\lambda_{\phi_S}^{-1})$ is not totally real, and*

(ii) *there exist Galois conjugates of λ_{ϕ_S} on the unit circle.*

In particular, no power of ϕ_S arises from either Penner's or Thurston's constructions.

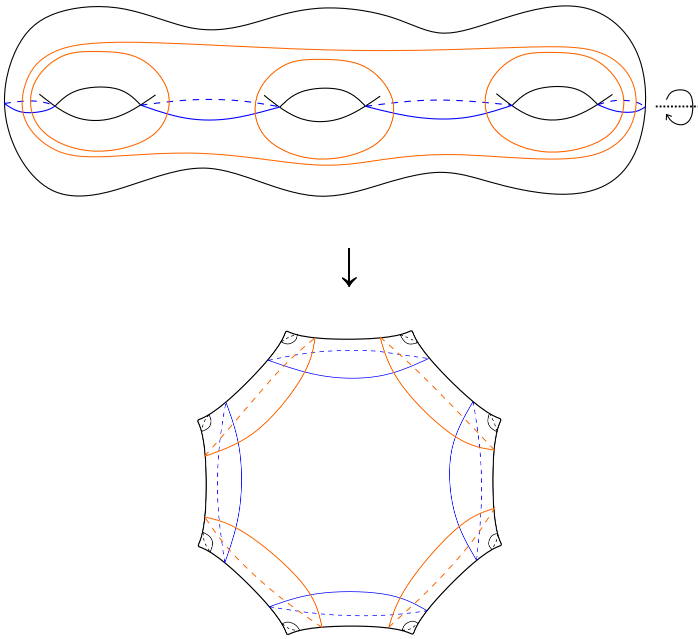


Figure 1: Two-fold branched covering map from $S_{3,0}$ to $S_{0,8}$ induced by the hyperelliptic involution.

In order to prove Theorem 1.1, we begin by constructing a family of pseudo-Anosov homeomorphisms on $S_{0,n}$ for $n \geq 6$. We then show that in this family there exists pseudo-Anosov homeomorphisms where the trace field is not totally real and where there exist Galois conjugates of its stretch factor on the unit circle. If $S_{0,n}$ is a sphere with n marked points, we can find branched covers $S_{g,m} \to S_{0,n}$. See Figure 1 for an illustration. By lifting pseudo-Anosov homeomorphisms from $S_{0,n}$ to $S_{g,m}$ through these branched covers, we obtain the other examples in Theorem 1.1. Since the lifted maps have the same stable and unstable foliations as ϕ, and the stretch factor is a power of λ, we can promote the examples from Theorem 1.4 to examples on surfaces with positive genus.

1.2 Constructing pseudo-Anosov maps on punctured spheres

Each map ϕ_S from Theorem 1.1 is a member of a new, large family of pseudo-Anosov maps on punctured spheres that we introduce. We'll see below that the stretch factors for the maps in this family exhibit various number theoretic properties, similar to the properties in Theorem 1.1.

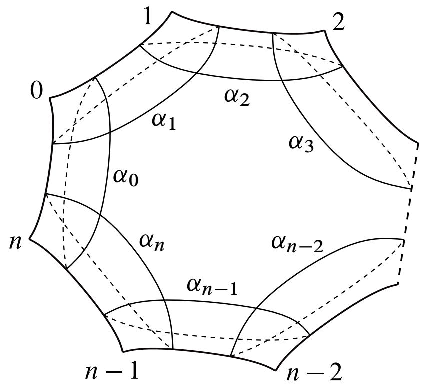


Figure 2: Labeling of the punctures on the n-times punctured sphere.

For the n–times punctured sphere $S_{0,n}$, label the punctures $0, 1, \ldots, n-1$ and fix the curves $\alpha_0, \alpha_1, \ldots, \alpha_{n-1}$ as shown in Figure 2 such that α_i cuts off the punctures i and $i-1 \pmod n$, and $i(\alpha_j, \alpha_{j+1}) = 2$. For each $i \in \{0, \ldots, n\}$, let D_i denote the right half-twist about the curve α_i. The square of a half twist is one example of a Dehn twist.

The basis of our construction of pseudo-Anosov homeomorphisms will consist of products of the D_i's. If $\rho\colon \mathbb{Z}/n \to \mathbb{Z}/n$ is the map $j \mapsto j+1 \bmod n$ which cyclically permutes the set $\{0, \ldots, n\}$, then we say a partition $\mu = \{\mu_0, \ldots, \mu_k\}$ of $\{0, \ldots, n-1\}$ is *evenly spaced* if $k < n$ and $\rho(\mu_i) = \mu_{i+1}$, where the indices are taken mod k. In the case of $n = 6$,

$$\mu = \{\{0,3\}, \{1,4\}, \{2,5\}\} \quad \text{and} \quad \bar{\mu} = \{\{0,2,4\}, \{1,3,5\}\}$$

are both evenly spaced partitions of $\{0, 1, \ldots, 5\}$. If $\mu = \{\mu_0, \ldots, \mu_k\}$ is a partition of $\{0, \ldots, n-1\}$, then for each $j \in \{0, 1, \ldots, k\}$, let D_{μ_j} denote the product of half twists $\prod_{i \in \mu_j} D_i$. Note, the order of this product is irrelevant as for each $i, k \in \mu_j$, D_i commutes with D_k since the curves α_i and α_k are disjoint.

Our first construction of pseudo-Anosov homeomorphisms are products $D_{\mu_0}^{q_0} \cdots D_{\mu_k}^{q_k}$ where the set $\{\mu_0, \ldots, \mu_k\}$ is an evenly spaced partition of $\{0, \ldots, n-1\}$ and each $q_j \geq 2$.

Theorem 1.2 *Let $n \geq 6$, let $q_j \geq 2$ for each $j \in \{0, 1, \ldots, k\}$, and let $\{\mu_0, \ldots, \mu_k\}$ be an evenly spaced partition of $\{0, \ldots, n-1\}$. Then*

$$\phi = \prod_{j=0}^{k} D_{\mu_j}^{q_j}$$

is a pseudo-Anosov homeomorphism of $S_{0,n}$.

Our second construction of pseudo-Anosov homeomorphisms, which will be used to produce the maps for Theorem 1.1, is an augmentation of the construction from Theorem 1.2. We say a partition $\{\mu_0, \dots, \mu_k\}$ of $\{0, 1 \dots, n-1\}$ *reduces to an evenly spaced partition* if there exists $k' \in \{0, 1, \dots, k\}$ and $n' \in \{0, 1, \dots, n-1\}$ such that $\{\mu_0, \mu_1, \dots, \mu_{k'}\}$ is an evenly spaced partition of $\{0, 1, \dots, n'-1\}$ and μ_j contains a single element for all $j > k'$.

Theorem 1.3 *Let $n \geq 7$, let $q_j \geq 2$ for each $j \in \{0, 1, \dots, k\}$, and let*

$$\{\mu_0, \dots, \mu_{k'}, \dots, \mu_k\}$$

be a partition of $\{0, 1, \dots, n-1\}$ that reduces to an evenly spaced partition. Then

$$\phi = \prod_{j=0}^{k} D_{\mu_j}^{q_j}$$

is a pseudo-Anosov homeomorphism of $S_{0,n}$.

To prove that the homeomorphisms in Theorems 1.2 and 1.3 are pseudo-Anosov, we explicitly construct the train tracks for each of the maps. For each pseudo-Anosov homeomorphism, we show that the associated train track has a specific structural form. Additionally, we show that the train track matrix associated to each map is Perron–Frobenius. Lemma 2.4 states that having both the specific form of the train track and the Perron–Frobenius train track matrix concurrently will imply that the homeomorphisms we constructed are pseudo-Anosov.

The stretch factors for the pseudo-Anosov maps arising from Theorems 1.2 and 1.3 exhibit a wide variety of number-theoretic properties. In addition to the map ϕ_S from Theorem 1.1, we produce examples of pseudo-Anosov maps with any desired combination of having totally real trace field, or not, and having Galois conjugates of the stretch factor on the unit circle, or not.

Theorem 1.4 *For any of the following four statements, there exists a pseudo-Anosov homeomorphism whose stretch factor λ satisfies the statement*:

(1) $\mathbb{Q}(\lambda + \lambda^{-1})$ *is totally real and there exists no Galois conjugates of λ on the unit circle.*

(2) $\mathbb{Q}(\lambda + \lambda^{-1})$ *is not totally real and there exist no Galois conjugates of λ on the unit circle.*

(3) $\mathbb{Q}(\lambda + \lambda^{-1})$ *is totally real and there exist Galois conjugates of* λ *on the unit circle.*

(4) $\mathbb{Q}(\lambda + \lambda^{-1})$ *is not totally real and there exist Galois conjugates of* λ *on the unit circle.*

These homeomorphisms are constructed on the surfaces $S_{0,6}$, $S_{0,7}$, $S_{0,8}$, *and* $S_{0,8}$, *respectively.*

We derive Theorem 1.1 as a corollary of Theorem 1.4. As discussed above, statement (4) of Theorem 1.4 proves Theorem 1.1 in the case of the 8–times punctured sphere. By using a branched cover to lift the homeomorphism from Theorem 1.4(4) to other surfaces, we complete the proof of Theorem 1.1.

Outline

In Section 2, we begin by proving an adaptation of the nesting lemma, which was first introduced by Masur and Minsky [8], an important lemma which allows us to determine whether a map is pseudo-Anosov. In Section 3, we will prove Theorems 1.2 and 1.3 which detail the main constructions of pseudo-Anosov homeomorphisms presented in this paper. Section 4 provides additional modifications which can be made to the main constructions in order to produce additional pseudo-Anosov homeomorphisms, and Section 5 details the various number-theoretic properties associated to the stretch factors which will allow us to show that this construction differs from the previous constructions. Lastly, Section 6 proves that the construction produces pseudo-Anosov mapping classes on surfaces of higher genus through a branched cover.

Acknowledgements

I would like to thank Dan Margalit for suggesting I generalize the pseudo-Anosov map from [12], as well as many helpful conversations. I would like to thank Balázs Strenner for suggesting I analyze the number-theoretic properties associated to the stretch factors of the maps produced, and Joan Birman for suggesting I apply the construction to surfaces of higher genus. I would also like to thank Thad Janisse, Chris Leininger, Dan Margalit, Kasra Rafi, Joe Repka, and Balázs Strenner for helpful conversations. Finally, I would like to thank Jacob Russell and the referee for comments on a previous version of the paper.

2 The nesting lemma

In this section, we begin by covering the required background information regarding train tracks and the complex of curves. Afterwards, we prove the nesting lemma. The nesting lemma is inspired by the work of Masur and Minsky in which they show that the diameter of the cure complex is infinite. The nesting lemma in this paper allows us to determine whether a map is pseudo-Anosov by analyzing the train track associated to the map.

2.1 Train tracks

In this section, we recall some of the basic definitions for train tracks. For a thorough treatment of the topic, the author recommends *Combinatorics of train tracks* by Penner and Harer [11].

A *train track* $\tau \subset S$ is an embedded 1–complex whose vertices are called *switches* and edges are called *branches*. Branches are smooth parametrized paths, and at each switch of τ, there is a well-defined tangent space to the branches coming into the switch. The tangent vector at the switch pointing toward the edge can have two possible directions which divides the ends of edges at the switch into two sets. The end of a branch of τ which is incident on a switch is called *incoming* if the one-sided tangent vector of the branch agrees with the direction at the switch and *outgoing* otherwise. Neither the set of incoming nor the set of outgoing branches are permitted to be empty. In this paper, whether a switch is incoming or outgoing is not part of the data in the train track, ie a train track is unoriented.

The valence of each switch in τ is at least 3, except for possibly one bivalent switch in a closed curve component. Finally, we require that every complementary component of $S \setminus \tau$ has a negative generalized Euler characteristic, in particular, for a complementary component $R \in S \setminus \tau$,

$$\chi(R) - \tfrac{1}{2}V(R) < 0$$

where $\chi(R)$ is the usual Euler characteristic and $V(R)$ is the number of outward pointing cusps on $\partial(R)$.

A *train route* is a nondegenerate smooth path in τ. A train route traverses a switch only by passing from an incoming to an outgoing edge (or vice-versa). We call a train track τ *large* if every component of $S \setminus \tau$ is a polygon or a once-punctured polygon, and we call τ *generic* if all switches are trivalent.

If σ is a train track which is a subset of τ, we write $\sigma < \tau$ and say σ is a *subtrack* of τ. In this case we may also say that τ is an *extension* of σ. If there is a homotopy of S such that every train route on σ is taken to a train route on τ we say σ is *carried* on τ and write $\sigma \prec \tau$.

Let $\mathcal{B}$ denote the set of branches of τ. A nonnegative, real-valued function $\mu\colon \mathcal{B} \to \mathbb{R}_+$ is called a *transverse measure* on τ if for each switch, μ satisfies the switch condition: for any switch, the sums of μ over incoming and outgoing branches are equal.

A train track is *recurrent* if there is a transverse measure which is positive on every branch, or equivalently, if each branch is contained in a closed train route.

Let α be a simple closed curve which intersects τ. We say α intersects τ *efficiently* if $\alpha \cup \tau$ has no bigon complementary regions. A track τ is *transversely recurrent* if every branch of τ is crossed by some simple curve α intersecting τ transversely and efficiently. We call a track *birecurrent* if it is both recurrent and transversely recurrent.

Any positive scaling of a transverse measure is also a transverse measure. Therefore, the set of all transverse measures when viewed as a subset of $\mathbb{R}^{\mathcal{B}}$ is a cone over a compact polyhedron in projective space. For a recurrent train track τ, let $P(\tau)$ denote the *polyhedron of measures* supported on τ. By $\mathrm{int}(P(\tau))$ we denote the set of weights on τ which are positive on every branch. We say that σ *fills* τ if $\sigma \prec \tau$ and $\mathrm{int}(P(\sigma)) \subseteq \mathrm{int}(P(\tau))$. Similarly, a curve α *fills* τ if $\alpha \prec \tau$ and α traverses every branch of τ.

One way to obtain a transverse measure on a train track τ is as follows: Fix a reference hyperbolic metric on S. A *geodesic lamination* in S is a closed set foliated by geodesics. A geodesic lamination is *measured* if it supports a measure on arcs transverse to its leaves, which is invariant under isotopies preserving the leaves. The space of all compactly supported measured geodesic laminations on S, with suitable topology, is known as $\mathcal{ML}(S)$, and changing the reference metric on S will yield spaces which are equivalent. A geodesic lamination λ is *carried* on τ if there is a homotopy of S taking λ to a set of train routes. In such a case, λ induces a transverse measure on τ, which in turn uniquely determines λ.

In this paper, we will blur the distinction between $P(\tau)$ as a subset of $\mathcal{ML}(S)$, and as a subset of the space $\mathbb{R}_+^{\mathcal{B}}$ of nonnegative functions on the branch set $\mathcal{B}$ of τ.

Let σ be a large track. A *diagonal extension* of σ is a track κ such that $\sigma < \kappa$ and every branch of $\kappa \setminus \sigma$ is a *diagonal* of σ, ie the endpoints of each edge in $\kappa \setminus \sigma$ terminate in the corner of a complementary region of σ. Let $E(\sigma)$ denote the set of all recurrent diagonal extensions of σ. Note that it is a finite set, and let $PE(\sigma)$ denote $\bigcup_{\kappa\in E(\sigma)} P(\kappa)$. Let $\mathrm{int}(PE(\sigma))$ denote the set of measures $\mu \in PE(\sigma)$ which are positive on every branch of σ.

2.2 The linear algebra of train tracks

For each pseudo-Anosov homeomorphism ϕ, there exists a train track τ which is invariant under the action of ϕ. Under this action, ϕ changes the weights of the branches of τ in a linear way. Thus, the action of the pseudo-Anosov homeomorphism is able to be completely described by the *train track matrix*. In fact, for each pseudo-Anosov mapping class the transition matrix M is *Perron-Frobenius* and the positive eigenvector determines an invariant measure.

Definition 2.1 (Perron–Frobenius matrix) A *Perron–Frobenius matrix* is a matrix with entries $a_{i,j} \geq 0$ such that some power of the matrix has strictly positive entries.

It is also known that the eigenvalue of this eigenvector will correspond to the stretch factor of the pseudo-Anosov map [9].

Theorem 2.2 *Given a pseudo-Anosov mapping class f, there exists a train track τ invariant under the action of f such that the matrix M which determines the action on the transverse measures is Perron–Frobenius. The positive eigenvector determines an invariant measure corresponding to the invariant foliation and the eigenvalue is the stretch factor.*

2.3 Complex of curves

The complex of curves, defined by Harvey [3], is a combinatorial object which encodes the intersection patterns of simple closed curves in S_g.

Definition 2.3 (complex of curves, $C(S_{g,n})$) The *complex of curves* is an abstract simplicial complex associated to a surface S. Its 1–skeleton is given by the following data:

- **Vertices** There is one vertex of $C(S)$ for each isotopy class of essential simple closed curves in S.
- **Edges** There is an edge between any two vertices of $C(S)$ corresponding to isotopy classes a and b with $i(a,b) = 0$, ie a and b are disjoint.

We note that $C(S)$ is a *flag complex*, which means that $k+1$ vertices span a k–simplex of $C(S)$ if and only if they are pairwise connected by edges. We will only make use of the 1–skeleton of the complex of curves, and we will denote the 1–skeleton by $C(S)$. By specifying that each edge has length 1, we turn $C(S)$ into a metric space. We let $d_{C(S)}$ denote the distance function obtained by taking shortest paths.

2.4 The nesting lemma

We are now in a position where we can state and prove the nesting lemma. This lemma is one of the key steps to proving that the homeomorphisms constructed in this paper are pseudo-Anosov.

Lemma 2.4 (the nesting lemma) *Let τ be a large, generic, birecurrent train track. Let $\phi\colon S \to S$ be a map such that τ is carried by ϕ. If the matrix associated to τ is a Perron–Frobenius matrix, then ϕ is a pseudo-Anosov map.*

Before we prove this lemma, we will state some lemmas from Masur and Minsky [8]. Let μ be a measured lamination on a surface S. We define $N(\tau)$ to be the union of $E(\sigma)$ over all large, recurrent subtracks $\sigma < \tau$, and we define $PN(\tau) = \bigcup_{\kappa \in N(\tau)} P(\kappa)$. The following lemma provides a sufficient condition for when μ is contained in $\operatorname{int}(PE(\sigma))$.

Lemma 2.5 [8, Lemma 4.1] *There exists $\delta > 0$ (depending only on S) for which the following holds. Let $\sigma < \tau$ where σ is a large track. If $\mu \in P(\tau)$ and, for every branch b of $\tau \setminus \sigma$ and b' of σ, $\mu(b) < \delta\mu(b')$, then σ is recurrent and $\mu \in \operatorname{int}(PE(\sigma))$.*

Let σ and τ be two large recurrent tracks such that $\sigma \prec \tau$. In this case we say that the train tracks σ and τ are *nested*. The following two lemmas tell us that when we have train tracks which are nested, their diagonal extensions are also nested in a suitable sense. Additionally, these lemmas tell us that the way in which the diagonal branches cover each other is controlled.

Lemma 2.6 [8, Lemma 4.2] *Let σ and τ be large recurrent tracks, and suppose $\sigma \prec \tau$. If σ fills τ, then $PE(\sigma) \subseteq PE(\tau)$. Even if σ does not fill τ, we have $PN(\sigma) \subseteq PN(\tau)$.*

Lemma 2.7 [8, Lemma 4.3] *Let $\sigma \prec \tau$ where σ is a large recurrent track, and let $\sigma' \in E(\sigma)$ and $\tau' \in E(\tau)$ be such that $\sigma' \prec \tau'$. Then any branch b of $\tau' \setminus \tau$ is traversed by branches of σ' with degree at most m_0, a number which depends only on S.*

The final lemma we require from Masur and Minsky is one which gives us a relation between nesting train tracks and their distance in the complex of curves.

Lemma 2.8 [8, Lemma 4.4] *Let α and β be simple, nonperipheral closed curves in a surface S. If σ is a large birecurrent train track and $\alpha \in \mathrm{int}(PE(\sigma))$ (ie α is carried on the maximal train track σ so that it is carried on every branch), then*

$$d_{\mathcal{C}(S)}(\alpha, \beta) \le 1 \implies \beta \in PE(\sigma).$$

In other words,

$$\mathcal{N}_1(\mathrm{int}(PE(\sigma))) \subset PE(\sigma),$$

where $\mathcal{N}_1$ denotes a radius 1 neighborhood in $\mathcal{C}(S)$.

Notice that since there exist simple closed curves on the surface S which are also in $\mathrm{int}(PE(\sigma))$, it means that we can consider a radius 1 neighborhood of $\mathrm{int}(PE(\sigma))$ in $C(S)$.

Before we prove Lemma 2.4, we give a brief outline. Given the train track τ, we consider a measure μ which is positive on each branch. Since $\phi(\tau)$ fills τ, it will follow that there exists some $k \in \mathbb{N}$ such that $\phi^k(PE(\tau)) \subset \mathrm{int}(PE(\tau))$. From this, it follows that $\phi^{jk}(\tau)$ fills $\phi^{(j-1)k}(\tau)$, which will imply that $PE(\phi^{jk}(\tau)) \subset \mathrm{int}(PE(\phi^{(j-1)k}(\tau)))$ for any j. At this point, we will suppose for the sake of contradiction that ϕ is not a pseudo-Anosov map. This would imply that there exists a curve α on the surface S disjoint from τ such that there exists an $m \in \mathbb{N}$ such that $\phi^m(\alpha) = \alpha$. Using Lemma 2.8 and the fact that $PE(\phi^{jk}(\tau)) \subset \mathrm{int}(PE(\phi^{(j-1)k}(\tau)))$ for any j, we will find that $d_{C(S)}(\alpha, \phi^{jk}(\alpha)) \to \infty$, which contradicts that there exists some m such that $\phi^m(\alpha) = \alpha$. Therefore ϕ is a pseudo-Anosov map.

Proof of Lemma 2.4 Let μ be a measure that is positive on every branch of $\phi(\tau)$, ie $\mu \in \mathrm{int}(P(\phi(\tau)))$. Since $\phi(\tau)$ preserves τ, μ is a measure that is positive on any branch of τ, and we have $\mathrm{int}(P(\phi(\tau))) \subseteq \mathrm{int}(P(\tau))$. This implies that $\phi(\tau)$ fills τ.

The next paragraph follows the argument found in [8, Theorem 4.6] which shows that if $\phi(\tau)$ fills τ, then there exists some $k \in \mathbb{N}$ such that $\phi^k(PE(\tau)) \subset \mathrm{int}(PE(\tau))$.

Suppose that $\tau' \in E(\tau)$ is a diagonal extension of τ. By Lemma 2.6, $\phi(\tau')$ is carried by some $\tilde{\tau} \in E(\tau)$. There exists a constant $c_0 = c_0(S)$ such that for some $c \le c_0$ the power $\phi' = \phi^c$ takes τ' to a train track carried by τ', since the number of train tracks in $E(\tau_0)$ is bounded in terms of the topology of S. Let $\mathcal{B}$ represent the branch set of τ', and $\mathcal{B}_\tau \subset \mathcal{B}$ represent the branch set of τ. In the coordinates of $\mathbb{R}^{\mathcal{B}}$ we may represent ϕ' as an integer matrix M, with a submatrix M_τ which gives us the restriction to $\mathbb{R}^{\mathcal{B}_\tau}$. Penner shows in [10] that M_τ^n has all positive entries, where n is the dimension $|\mathcal{B}_\tau|$.

In fact, Penner shows that $|M_\tau^n(x_\tau)| \geq 2|x_\tau|$ for any vector x_τ which represents a measure on τ. Indeed, M_τ has a unique eigenvector in the positive cone of $\mathbb{R}^{\mathcal{B}_\tau}$, which corresponds to $[(\tau,\mu)]$. On the other hand, for a diagonal branch $b \in \mathcal{B}\setminus\mathcal{B}_\tau$, Lemma 2.7 shows that $|M^i(x)| \leq m_\tau|x|$ for all $x \in \mathbb{R}^{\mathcal{B}}$ and all powers $i > 0$. Since τ is generic, we have that any transverse measure x on τ' must put a positive measure on a branch of $\mathcal{B}_\tau$. This implies that given $\delta > 0$ there exists k_1, depending only on δ and S, such that for some $k \leq k_1$ we have $\max_{b\in\mathcal{B}\setminus\mathcal{B}_\tau} \phi^k(x)(b) \leq \delta \min_{b\in\mathcal{B}_\tau} h^k(x)(b)$, for any $x \in P(\tau')$. We apply this to each $\tau' \in E(\tau)$, and by applying Lemma 2.5, we see that, for an appropriate choice of δ,

$$\phi^k(PE(\tau)) \subset \operatorname{int}(PE(\tau)).$$

Using the above argument, we find that $\phi^{jk}(\tau)$ fills $\phi^{(j-1)k}(\tau)$, from which it follows that

$$PE(\phi^{jk}(\tau)) \subset \operatorname{int}(PE(\phi^{(j-1)k}(\tau))) \tag{1}$$

for any j.

By way of contradiction, suppose that ϕ is not a pseudo-Anosov map. Then there exists a curve α on the surface S disjoint from τ such that $\phi^m(\alpha) = \alpha$ for some $m \in \mathbb{N}$.

Since there are elements of $PE(\tau)$ which are not in $\phi^k(PE(\tau))$, it is possible to find $\gamma \in C(S)$ such that $\gamma \notin PE(\tau)$ and $\phi^k(\gamma) \in PE(\tau)$. Then $\phi^{2k}(\gamma) \in \operatorname{int}(PE(\tau))$, which implies that $d_{C(S)}(\gamma, \phi^{2k}(\gamma)) \geq 1$ by Lemma 2.8.

Since $\phi^{jk}(\gamma) \in PE(\phi^{(j-1)k}(\tau))$ for $j \geq 1$, we use (1) to find

$$\begin{aligned}
&\phi^{3k}(\gamma) \in PE(\phi^{2k}(\tau)) \subset \operatorname{int}(PE(\phi^k(\tau))),\\
&\phi^{3k}(\gamma) \in PE(\phi^k(\tau)) \subset \operatorname{int}(PE(\tau)),\\
&\phi^{3k}(\gamma) \in PE(\tau).
\end{aligned}$$

By Lemma 2.8, we have that for any k,

$$\mathcal{N}_1(\operatorname{int}(PE(\phi^k(\tau)))) \subset PE(\phi^k(\tau)).$$

Therefore, we find that

$$\begin{aligned}
\phi^{3k}(\gamma) \in PE(\phi^{2k}(\tau)) \subset \mathcal{N}_1(\operatorname{int}(PE(\phi^k(\tau)))) \subset PE(\phi^k(\tau))&\\
\subset \mathcal{N}_1(\operatorname{int}(PE(\tau))) \subset PE(\tau).&
\end{aligned}$$

Therefore, since $\gamma \notin PE(\tau)$, we have that $d_{C(S)}(\gamma, \phi^{3k}(\gamma)) \geq 2$.

We continue inductively to show that $d_{C(S)}(\gamma, \phi^{jk}(\gamma)) \geq j-1$, which implies that as $j \to \infty$, $d_{C(S)}(\gamma, \phi^{jk}(\gamma)) \to \infty$. Since

$$\begin{aligned} d_{C(S)}(\alpha, \phi^{jk}(\alpha)) &\geq d_{C(S)}(\gamma, \phi^{jk}(\gamma)) - d_{C(S)}(\alpha, \gamma) - d_{C(S)}(h^{jk}(\alpha), h^{jk}(\gamma)) \\ &= d_{C(S)}(\gamma, \phi^{jk}(\gamma)) - 2d_{C(S)}(\alpha, \gamma), \end{aligned}$$

we have $d_{C(S)}(\alpha, \phi^{jk}(\alpha)) \to \infty$, which contradicts that there exists some m such that $\phi^m(\alpha) = \alpha$. Therefore ϕ is a pseudo-Anosov map. □

3 Main construction on n–times punctured spheres

In this section, we will prove Theorem 1.2, that the homeomorphisms induced by evenly spaced partitions are pseudo-Anosov, and Theorem 1.3, that the homeomorphisms induced by partitions which reduce to an evenly spaced partition are pseudo-Anosov.

We begin with the proof of Theorem 1.2.

Theorem 1.2 *Let $n \geq 6$, let $q_j \geq 2$ for each $j \in \{0, 1, \ldots, k\}$, and let $\{\mu_0, \ldots, \mu_k\}$ be an evenly spaced partition of $\{0, \ldots, n-1\}$. Then*

$$\phi = \prod_{j=0}^{k} D_{\mu_j}^{q_j}$$

is a pseudo-Anosov homeomorphism of $S_{0,n}$.

Proof Consider the surface $S_{0,n}$ for some fixed $n \in \mathbb{N}$. Fix $k > 1$, $k \in \mathbb{N}$, and fix a partition $\mu = \{\mu_1, \ldots, \mu_k\}$ of the n punctures of $S_{0,n}$ such that $\rho(\mu_{i-1}) = \rho(\mu_{(i \bmod k)})$. We will prove that

$$\phi = \prod_{i=1}^{k} D_{\mu_i}^{q_i} = D_{\mu_k}^{q_k} \cdots D_{\mu_2}^{q_2} D_{\mu_1}^{q_1}$$

is a pseudo-Anosov mapping class.

We first construct the train track τ so that $\phi(\tau)$ is carried by τ. Consider the partition $\mu = \{\mu_1, \ldots, \mu_k\}$. Construct a k–valent pretrack by having a branch loop around each puncture in the set μ_k, with each of these branches meeting in the center where they are smoothly connected by a k–gon. See Figures 3, left, and 4, left, for examples which correspond to pretracks from Example A.1 of the appendix. For the remaining labeled punctures in μ_i, loop a branch around each puncture and have this branch turn left

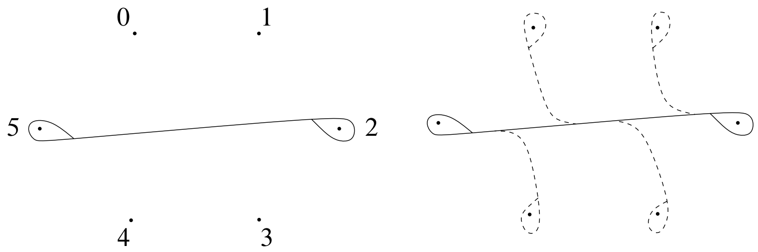


Figure 3: Constructing the train track for the map ϕ_μ from Example A.1.

towards the k–valent pretrack meeting the branch of the pretrack whose label is next in the ordering. See Figures 3, right, and 4, right, for examples which correspond to the train tracks from Example A.1. We notice that the train track has rotational symmetry of order k.

For each k, $D_{\mu_i}^{q_i}$ acts locally the same, by which we mean the following: Each half-twist in $D_{\mu_i}^{q_i}$ involves a branch located around puncture b on the k–valent pretrack and the branch located around puncture b' which is directly next to puncture b in the clockwise direction. As we consider a right half-twist to be positive, we notice that the branch at puncture b will begin to turn into the branch at puncture b', see Figures 11 and 13 for examples. Therefore, after the twist, the branch around puncture b' is now on the k–valent pretrack, and the branch around puncture b is directly next to the branch at puncture b' in the counterclockwise direction. Branches which are neither on the k–valent pretrack nor directly clockwise to the k–valent pretrack are unaffected by $D_{\mu_i}^{q_i}$. Thus, after each application of $D_{\mu_i}^{q_i}$, we arrive at a train track which is able to be obtained by a rotation of $2\pi/n$ of our starting train track. Since τ has a rotational symmetry of order k, we notice that $\phi(\tau)$ is carried by τ.

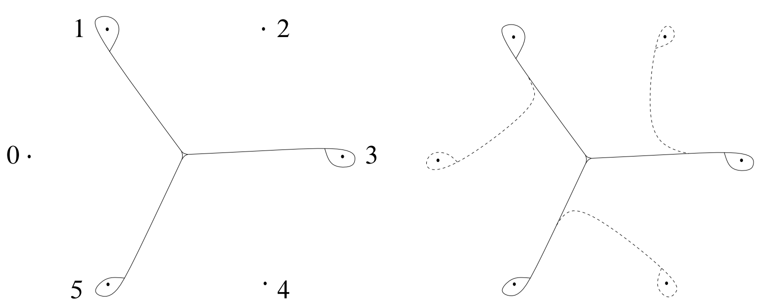


Figure 4: Constructing the train track for the map $\phi_{\bar{\mu}}$ from Example A.1.

Let M_τ denote the matrix representing the induced action of the space of weights on τ. To prove that M_τ is Perron–Frobenius, fix an initial weight on each branch. For each application of $D_{\mu_i}^{q_i}$, the labels on the k–valent pretrack and directly next to the k–valent pretrack in the clockwise direction will become a linear combination of the labels associated to these two branches. In particular, let w be the weight of a branch on the k–valent pretrack, and let w' be the weight of the branch directly next to this branch in the clockwise direction. After applying l half-twists, we see that the weight of branch w is $lw' + (l-1)w$ and the weight of branch w' is $(l+1)w' + lw$. Since τ rotates clockwise by $2\pi/n$ after each application of $D_{\mu_i}^{q_i}$, we know that after k applications of ϕ, the weight of each branch will be a linear combination of the initial weights of each branch and the constants of this linear combination are strictly positive integers. Equivalently, this implies that each entry in M_τ^k is a strictly positive integer value. This implies that the matrix M_τ is Perron–Frobenius.

To finish the proof, we can see by inspection that each of the train tracks which were constructed above are large, generic, and birecurrent. Therefore, we apply Lemma 2.4 which completes the proof that ϕ is a pseudo-Anosov mapping class. □

Using a similar argument as the proof of Theorem 1.2, we now provide a proof for Theorem 1.3.

Theorem 1.3 *Let $n \geq 7$, let $q_j \geq 2$ for each $j \in \{0, 1, \dots, k\}$, and let*

$$\{\mu_0, \dots, \mu_{k'}, \dots, \mu_k\}$$

be a partition of $\{0, 1, \dots, n-1\}$ that reduces to an evenly spaced partition. Then

$$\phi = \prod_{j=0}^{k} D_{\mu_j}^{q_j}$$

is a pseudo-Anosov homeomorphism of $S_{0,n}$.

Proof Fix some value of $n \in \mathbb{N}$, some $k > 1$, $k \in \mathbb{N}$, and a partition $\mu = \{\mu_1, \dots, \mu_k\}$ of the n punctures of $S_{0,n}$ such that $\rho(\mu_{i-1}) = \rho(\mu_{(i \bmod k)})$. We perform the modification outlined in the statement of the theorem to obtain a partition $\mu' = \{\mu'_1, \dots, \mu'_k, \mu'_{k+1}\}$, on the $(n+1)$–times punctured sphere $S_{0,n+1}$, which defines the map

$$\phi' = \prod_{i=1}^{k+1} D_{\mu_i}^{q'_i} = D_{\mu'_{k+1}}^{q'_{k+1}} D_{\mu'_k}^{q'_k} \cdots D_{\mu'_2}^{q'_2} D_{\mu'_1}^{q'_1}.$$

We prove that ϕ' is a pseudo-Anosov mapping class.

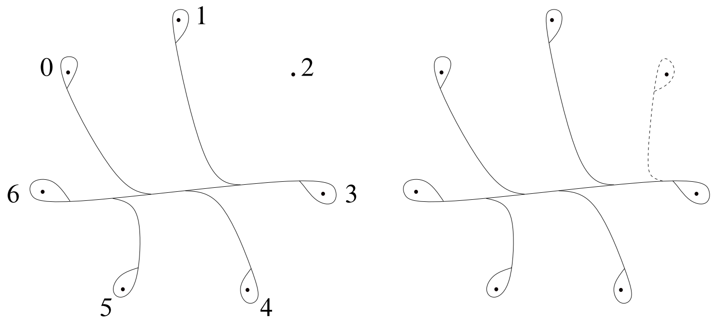


Figure 5: Constructing the train track for the map ψ.

We first construct the train track τ' so that $\phi'(\tau')$ is carried by τ'. We begin by considering the train track τ associated to the map $\phi = \prod_{i=1}^{k} D_{\mu_i}^{q_i}$ defined by the partition μ. We then add in a new puncture onto the sphere between punctures $k-1$ and k, and relabel the punctures. See Figures 5, left, and 6, left, for examples which correspond to the pretracks from Example A.2. Add a branch from puncture $k+1$ so that it turns tangentially into the k–valent pretrack meeting the same branch on the pretrack as the branches associated to punctures $1, \dots, k-1$. See Figures 5, right, and 6, right, for examples which correspond to the train tracks from Example A.2. We denote this modified train track by τ'.

To show that $\phi'(\tau')$ is carried by τ', we notice that by the same reasoning in the proof of Theorem 1.2 that for each $1 \le i < k+1$, the application of $D_{\mu_i'}^{q_i'}$ will result in a train track which is obtained through a rotation by $2\pi/n$ of our starting train track. After the first k applications of $D_{\mu_i'}^{q_i'}$, we have a resulting train track which is obtained by a

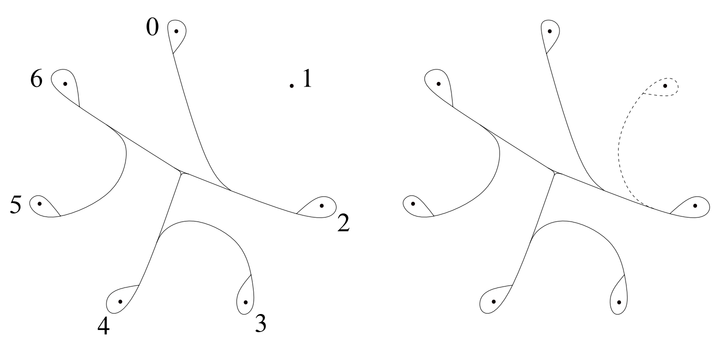


Figure 6: Constructing the train track for the map $\phi_{\bar{\mu}'}$.

rotation by $2\pi k/n$ of our starting train track, which is not quite τ'. By applying the final twist $D_{\mu'_{k+1}}^{q'_{k+1}}$, we find $\phi'(\tau') = \tau'$ and thus $\phi'(\tau')$ is carried by τ'. See Figures 15 and 17 for examples.

By the same reasoning as in the proof of Theorem 1.2, the matrix representing the induced action on the space of weights on τ' will be Perron–Frobenius. To finish the proof, we note that each of the train tracks that we have constructed are large, generic, and birecurrent. Therefore, we can apply Lemma 2.4 which completes the proof that the map is a pseudo-Anosov mapping class. □

4 Modifications of construction

By considering the constructions described in Section 3, we notice that there are additional modifications one can make to the construction to obtain more pseudo-Anosov mapping classes.

Recall that in Theorem 1.3, we added extra branches to the same branch on the $|\mu_k|$–valent pretrack. The first modification in this section is obtained by allowing additional branches to be added to any of the branches of the $|\mu_k|$–valent pretrack. To make this precise, we say a partition $\{\mu_0, \dots, \mu_k\}$ of $\{0, 1, \dots, n-1\}$ 2*–reduces to an evenly spaced partition* if there exists an $n' \in \{0, 1, \dots, n-1\}$ such that $\{\mu'_0, \mu'_1, \dots, \mu'_k\}$ is the partition of $\{0, 1, \dots, n'-1\} \cup \{(n'+1)-1, (n'+2)-1, \dots (n-1)-1\}$ defined by

$$\mu'_j = \{i \mid i \in \mu_j \text{ and } i < n'\} \cup \{i-1 \mid i \in \mu_k \text{ and } i > n'\}, \tag{2}$$

and $\{\mu'_0, \mu'_1, \dots, \mu'_k\}$ reduces to an evenly spaced partition. Similarly, we can iteratively define a partition which k*–reduces to an evenly spaced partition*. Theorem 1.3 holds for any partition which k-reduces to an evenly spaced partition, and the proof follows by considering a modification of the train track analogous to the modification found in the proof of Theorem 1.3.

To obtain a second modification, consider a map ϕ from any of the possible maps found in Theorem 1.3 or any of the modifications outlined above. It is possible to find an additional pseudo-Anosov map which has the same train track as ϕ. Since the train track rotates by $2\pi/n$ for the first k applications of $D_{\mu'_i}^{q'_i}$, we can define a map that will continue to rotate the train track by $2\pi/n$ in place of doing the final twist(s) $D_{\mu'_{k+1}}^{q'_{k+1}}$.

For example, consider the first map from Example A.2. After applying $D^2_{\mu'_3} D^2_{\mu'_2} D^2_{\mu'_1}$, where $\mu' = \{\{1,5\},\{2,6\},\{3,7\}\} = \{\mu'_1, \mu'_2, \mu'_3\}$, the train track has rotated by $6\pi/7$. We can apply the rotations associated to punctures 4 and 1 next, then around puncture 5 and 2, around punctures 6 and 3, and finally around punctures 7 and 4, which will have rotated our train track by a full rotation. In other words, you will obtain a new "partition" containing the sets $\mu''_i = \{i, i + \lceil 7/3 \rceil\}$ for $1 \le i \le 7$. More precisely, we obtain the following additional construction:

Theorem 4.1 *Consider the surface $S_{0,n}$. Consider one of the maps from Theorem 1.2, in particular, consider a partition of the n punctures into $1 < k < n$ sets $\{\mu_i\}_{i=1}^k$ such that the partition is evenly spaced. Apply any number of applications of Theorem 1.3 to obtain a new partition $\mu' = \{\mu'_1, \dots, \mu'_k, \mu'_{k+1}, \dots, \mu'_{k+l}\}$, where $|\mu_{k+j}| < |\mu_1|$ for all $1 \le j \le l$ which defines a map on the p–sphere, where $p = \sum_{i=1}^{k+l} |\mu_i|$. Consider the train track τ' associated to this map. Define the partition μ'' to be the partition containing the sets $\mu''_i = \{i, i + \lceil p/k \rceil, \dots, i + (|\mu_1| - 1)\lceil p/k \rceil\}$, where $1 \le i \le p$. Then μ'' defines the pseudo-Anosov mapping class*

$$\phi' = \prod_{i=1}^{p} D^{q''_i}_{\mu''_i} = D^{q''_p}_{\mu''_p} \cdots D^{q''_2}_{\mu''_2} D^{q''_1}_{\mu''_1}$$

on $S_{0,p}$, where $q''_j = \{q''_{j_1}, \dots, q''_{j_l}\}$ is the set of powers associated to each μ''_i.

Remark 4.2 The proofs for the modifications found in this section follow the same format as the proofs for the theorems found in Section 3.

5 Number-theoretic properties

In this section, we prove Theorem 1.4 which describes the number-theoretic properties associated to the maps arising from Theorems 1.2 and 1.3. To prove Theorem 1.4, we provide explicit examples of pseudo-Anosov mapping classes resulting from the constructions outlined in Theorems 1.2 and 1.3 which have the specific number-theoretic properties we are looking for. An important consequence of Theorem 1.4 is that the construction outlined in this paper differs from the constructions of both Penner and Thurston.

5.1 Number-theoretic properties

We begin by introducing two properties of pseudo-Anosov homeomorphisms: the placement of the Galois conjugates of the stretch factor, and the trace field.

5.1.1 Galois theory We recall that if K is a field containing the subfield F, then K is said to be an *extension field* (or simply an *extension*) of F. We denote this by K/F. Let $\mathrm{Aut}(K/F)$ be the collection of automorphisms of K which fix F. If K/F is a field extension, then K is said to be *Galois* over F and K/F is a *Galois extension* if $|\mathrm{Aut}(K/F)| = [K : F]$. In a Galois extension K/F, the other roots of the minimal polynomial over F of any element $\alpha \in K$ are precisely the distinct conjugates of α under the Galois group K/F. Therefore, the *Galois conjugates* are precisely the other roots of the minimal polynomial over F of an element $\alpha \in K$.

Let ϕ be a pseudo-Anosov homeomorphism with stretch factor λ. Let $L/\mathbb{Q}$ be the Galois extension where L is the splitting field of the minimal polynomial of λ. Whether there exist Galois conjugates of the stretch factor on the unit circle is a number-theoretic property of ϕ.

5.1.2 Trace fields The *trace field* of a linear group is the field generated by the traces of its elements. In particular the trace field of a group $\Gamma \subset SL_2(\mathbb{R})$ is the subfield of $\mathbb{R}$,

$$\{\mathrm{tr}(A) \mid A \in \Gamma\}.$$

Work by Hubbard and Masur shows that for each pseudo-Anosov homeomorphism, one can obtain a corresponding flat structure [4]. An *affine diffeomorphism* is a diffeomorphism which preserves the flat structure, and they form a group which we call the *affine diffeomorphism group*. See [15] for a discussion of flat surfaces and the affine diffeomorphism group. Kenyon and Smillie [6] proved that if the affine diffeomorphism group of a surface contains an orientation preserving pseudo-Anosov element f with largest eigenvalue λ, then the trace field is $\mathbb{Q}(\lambda + \lambda^{-1})$.

5.2 Proof of Theorem 1.4

We now provide a proof of Theorem 1.4 by providing explicit examples of maps satisfying each of the number-theoretic properties.

Theorem 1.4 *For any of the following four statements, there exists a pseudo-Anosov homeomorphism whose stretch factor λ satisfies the statement*:

(1) $\mathbb{Q}(\lambda + \lambda^{-1})$ *is totally real and there exists no Galois conjugates of λ on the unit circle.*

(2) $\mathbb{Q}(\lambda + \lambda^{-1})$ *is not totally real and there exist no Galois conjugates of λ on the unit circle.*

(3) $\mathbb{Q}(\lambda+\lambda^{-1})$ *is totally real and there exist Galois conjugates of* λ *on the unit circle.*

(4) $\mathbb{Q}(\lambda+\lambda^{-1})$ *is not totally real and there exist Galois conjugates of* λ *on the unit circle.*

These homeomorphisms are constructed on the surfaces $S_{0,6}$, $S_{0,7}$, $S_{0,8}$, *and* $S_{0,8}$, *respectively.*

Proof First, we provide an example where the Galois conjugates are never on the unit circle, and that the field $\mathbb{Q}(\lambda+1/\lambda)$ is totally real. This proves case (1).

Case (1) Consider the pseudo-Anosov map $\phi_{2,2,2}$ on $S_{0,6}$ which is induced by the partition $\mu_{2,2,2}=\{\{0,3\},\{1,4\},\{2,5\}\}$ as studied in Example A.1,

$$\phi_{2,2,2}=D_5^2D_2^2D_4^2D_1^2D_3^2D_0^2.$$

See Example A.1 for details. We find that the induced action on the space of weights on $\tau_{2,2,2}$ is given by the matrix

$$M=\begin{pmatrix} 3 & 2 & 0 & 0 & 0 & 2\\ 6 & 3 & 2 & 4 & 0 & 4\\ 12 & 6 & 3 & 6 & 0 & 8\\ 0 & 0 & 2 & 3 & 2 & 0\\ 4 & 0 & 4 & 6 & 3 & 2\\ 6 & 0 & 8 & 12 & 6 & 3 \end{pmatrix}$$

which has the characteristic polynomial

$$p_{2,2,2}=(x-1)^2(x+1)^2(x^2-18x+1).$$

The polynomial $p_{2,2,2}$ is the characteristic polynomial associated to the action of this map on the train track $\tau_{2,2,2}$. We notice that the leading eigenvalue $\lambda_{2,2,2}$ is a root of the factor

$$p_{\phi_{2,2,2},\lambda_{2,2,2}}(x)=x^2-18x+1,$$

which is an irreducible polynomial with real roots. Since $\lambda_{2,2,2}\in\mathbb{R}$ is not on the unit circle, as $\lambda_{2,2,2}>1$, $\lambda_{2,2,2}^{-1}$ is also not on the unit circle. Therefore, the Galois conjugates of the stretch factor are not on the unit circle.

To show that $\mathbb{Q}(\lambda_{2,2,2}+\lambda_{2,2,2}^{-1})$ is totally real, we notice that we can write

$$\frac{p_{2,2,2}}{x}=\left(x+\frac{1}{x}\right)-18=q\left(x+\frac{1}{x}\right).$$

By considering the roots of $q(y) = y - 18$, we notice that the only root is $y = 18$ which implies that the field $\mathbb{Q}(\lambda_{2,2,2} + \lambda_{2,2,2}^{-1})$ is totally real. This completes the proof of case (1).

Next, we will provide an example where the Galois conjugates are never on the unit circle, and that the field $\mathbb{Q}(\lambda + \lambda^{-1})$ is not totally real. This will prove case (2).

Case (2) We consider the map $\phi_{3,3,1} = D_2^2 D_6^2 D_4^2 D_1^2 D_5^2 D_3^2 D_0^2$ on $S_{0,7}$ induced by the partition $\mu_{3,3,1} = \{\{0,3,5\},\{1,4,6\},\{2\}\}$ from Example A.2. We find that the induced action on the space of weights on $\tau_{3,3,1}$ is given by the matrix

$$M = \begin{pmatrix} 3 & 2 & 0 & 0 & 0 & 0 & 2 \\ 6 & 3 & 2 & 4 & 0 & 0 & 4 \\ 12 & 6 & 3 & 6 & 0 & 0 & 8 \\ 0 & 0 & 2 & 3 & 2 & 4 & 0 \\ 0 & 0 & 4 & 6 & 3 & 6 & 0 \\ 4 & 0 & 0 & 0 & 2 & 3 & 2 \\ 6 & 0 & 0 & 0 & 4 & 6 & 3 \end{pmatrix}$$

which has the characteristic polynomial

$$p_{3,3,1} = (x+1)(x^3 - 15x^2 + 7x - 1)(x^3 - 7x^2 + 15x - 1).$$

The polynomial $p_{3,3,1}$ is the characteristic polynomial associated to the action of this map on the train track We notice that the leading eigenvalue, $\lambda_{3,3,1}$, is a root of the polynomial

$$p_{\phi_{3,3,1},\lambda_{3,3,1}}(x) = x^3 - 15x^2 + 7x - 1.$$

The roots of this polynomial are

$$5 + \tfrac{1}{3}\sqrt[3]{2916 - 12\sqrt{19}} + \left(\tfrac{2}{3}\right)^{2/3}\sqrt[3]{243 + \sqrt{93}},$$

$$5 + \tfrac{1}{3}(-1 + i\sqrt{3})\sqrt[3]{2916 - 12\sqrt{19}} - \frac{(1 + i\sqrt{3})\left(\sqrt[3]{\tfrac{1}{2}(243 + \sqrt{93})}\right)}{3^{2/3}},$$

$$5 - \tfrac{1}{3}(-1 + i\sqrt{3})\sqrt[3]{2916 - 12\sqrt{19}} + \frac{(1 + i\sqrt{3})\left(\sqrt[3]{\tfrac{1}{2}(243 + \sqrt{93})}\right)}{3^{2/3}}.$$

By the rational root theorem,

$$p_{\phi_{\bar{\mu}'},\lambda}(x) = x^3 - 15x^2 + 7x - 1$$

is irreducible over $\mathbb{Q}$. None of the roots of $p_{\phi_{3,3,1},\lambda_{3,3,1}}(x)$ are on the unit circle, so we have that there are no Galois conjugates of the stretch factor on the unit circle. We now consider the polynomial

$$\left(\frac{1}{x^3}\right)(x^3-15x^2+7x-1)(x^3-7x^2+15x-1)=x^3-22x^2+127x-276.$$

We rewrite this polynomial as

$$q\left(x+\frac{1}{x}\right)=\left(x+\frac{1}{x}\right)^3-22\left(x+\frac{1}{x}\right)^2+124\left(x+\frac{1}{x}\right)-232.$$

We calculate that the roots of the polynomial $q(y)$ are

$$\begin{gathered}\tfrac{1}{3}\big(22+\sqrt[3]{1801-9\sqrt{26554}}+\sqrt[3]{1801+9\sqrt{26554}}\big),\\ \tfrac{1}{6}\big(44+i(\sqrt{3}+i)\sqrt[3]{1801-9\sqrt{26554}}+(-1-i\sqrt{3})\sqrt[3]{1801+9\sqrt{26554}}\big),\\ \tfrac{1}{6}\big(44+(-1-i\sqrt{3})\sqrt[3]{1801-9\sqrt{26554}}+i(\sqrt{3}+i)\sqrt[3]{1801+9\sqrt{26554}}\big).\end{gathered}$$

By unique factorization, $q(y)$ is irreducible. Additionally, since two of the roots are imaginary, the field $\mathbb{Q}(\lambda_{3,3,1}+\lambda_{3,3,1}^{-1})$ is not totally real. This completes the proof of case (2).

We now provide an example where there are Galois conjugates of the stretch factor on the unit circle, and that the field $\mathbb{Q}(\lambda+\lambda^{-1})$ is totally real. This will prove case (3).

Case (3) We now consider the pseudo-Anosov map

$$\phi_{2,2,2,1,1}=D_4^2D_3^2D_7^2D_2^2D_6^2D_1^2D_5^2D_0^2$$

on $S_{0,8}$, which is induced by the partition $\mu_{2,2,2,1,1}=\{\{0,5\},\{1,6\},\{2,7\},\{3\},\{4\}\}$. This is first map from Example A.1 with the modification from Theorem 1.3 applied twice so that there are two partitions with one element each. The train track $\tau_{2,2,2,1,1}$, where $\phi_{2,2,2,1,1}(\tau_{2,2,2,1,1})$ is carried by $\phi_{2,2,2,1,1}$, is depicted in Figure 7.

The matrix associated to this map is

$$M=\begin{pmatrix}3&2&0&0&0&0&0&2\\6&3&2&4&0&0&0&4\\12&6&3&6&0&0&0&8\\0&0&2&3&2&0&0&0\\0&0&4&6&3&2&0&0\\0&0&8&12&6&3&2&0\\4&0&16&24&12&6&3&2\\6&0&32&48&24&12&6&3\end{pmatrix}$$

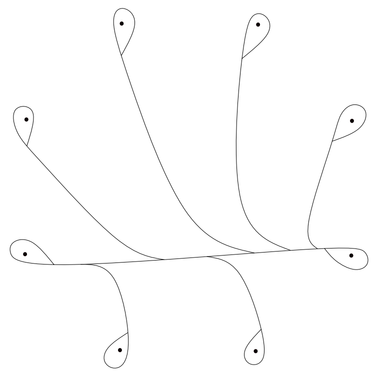


Figure 7: The train track associated to $\phi_{2,2,2,1,1}$.

which has the characteristic polynomial

$$p_{2,2,2,1,1}(x) = (x+1)^4(x^4 - 28x^3 + 6x^2 - 28x + 1).$$

The polynomial $p_{2,2,2,1,1}(x)$ is the characteristic polynomial associated to the action of this map on the train track $\tau_{2,2,2,1,1}$. Our leading eigenvalue $\lambda_{2,2,2,1,1}$ is a root of

$$p_{\lambda_{2,2,2,1,1},\phi_{2,2,2,1,1}}(x) = x^4 - 28x^3 + 6x^2 - 28x + 1.$$

The roots of this polynomial are

$$\lambda^{-1} = 7 + 4\sqrt{3} - 2\sqrt{24 + 14\sqrt{3}}, \quad \lambda = 7 + 4\sqrt{3} + 2\sqrt{24 + 14\sqrt{3}},$$
$$x_1 = 7 - 4\sqrt{3} - 2i\sqrt{14\sqrt{3} - 24}, \quad x_2 = 7 - 4\sqrt{3} + 2i\sqrt{14\sqrt{3} - 24}.$$

To prove that $p_{\lambda_{2,2,2,1,1},\phi_{2,2,2,1,1}}(x)$ is irreducible we use the following fact, a proof of which is found in [1].

Fact 5.1 *If $f(x) \in \mathbb{Z}[x]$ is primitive of degree $d \geq 1$ and there are at least $2d+1$ different integers a such that $|f(a)|$ is 1 or a prime number, then $f(x)$ is irreducible in $\mathbb{Q}[x]$.*

Thus, it suffices find 9 values of x such that $p_{\lambda,\psi}(x)$ is prime. Indeed, the following tuples $(x, p_{\lambda_{2,2,2,1,1},\phi_{2,2,2,1,1}}(x))$ contain 9 x–values such that $p_{\lambda_{2,2,2,1,1},\phi_{2,2,2,1,1}}(x)$ is prime:

$$(-24, 722977), \quad (-16, 182209), \quad (-12, 70321), \quad (0, 1), \quad (2, -239),$$
$$(6, -4703), \quad (8, -10079), \quad (24, -52511), \quad (38, 556321).$$

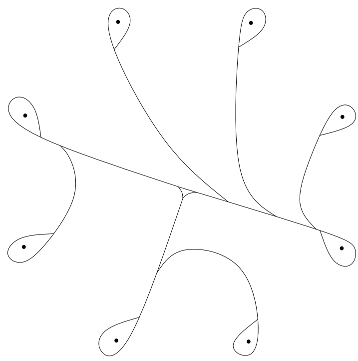


Figure 8: The train track associated to $\phi_{3,3,1,1}$.

Therefore, $p_{\lambda_{2,2,2,1,1},\phi_{2,2,2,1,1}}(x)$ is irreducible over $\mathbb{Q}$. Notice that $|x_1| = 1$ and $|x_2| = 1$, which we can verify by direct computation or by applying [7, Theorem 1]. This implies that there are Galois conjugates of the stretch factor on the unit circle. We now show that the field $\mathbb{Q}(\lambda_{2,2,2,1,1} + \lambda_{2,2,2,1,1}^{-1})$ is totally real by writing

$$\frac{p_{\lambda,\psi}}{x} = \left(x + \frac{1}{x}\right)^2 - 28\left(x + \frac{1}{x}\right) + 4 = q\left(x + \frac{1}{x}\right).$$

We notice that the roots of $q(y) = y^2 - 28y + 4$ are

$$14 - 8\sqrt{3} \quad \text{and} \quad 14 + 8\sqrt{3},$$

which implies that $q(y)$ is irreducible by unique factorization. Additionally, since both roots are real we find that the field $\mathbb{Q}(\lambda_{2,2,2,1,1} + \lambda_{2,2,2,1,1}^{-1})$ is totally real.

Lastly, we provide two examples where there are Galois conjugates of the stretch factor on the unit circle, and where the field $\mathbb{Q}(\lambda + \lambda^{-1})$ is not totally real. This will prove case (4).

Case (4) We begin with an example on $S_{0,8}$ where there are Galois conjugates of the stretch factor on the unit circle, and where the field $\mathbb{Q}(\lambda + 1/\lambda)$ is not totally real. We consider the second map from Example A.1 and apply the modification from Theorem 1.3 twice to obtain the partition $\mu_{3,3,1,1} = \{\{0,4,6\},\{1,5,7\},\{2\},\{3\}\}$. This induces the map

$$\phi_{3,3,1,1} = D_3^2 D_2^2 D_7^2 D_5^2 D_1^2 D_6^2 D_4^2 D_0^2.$$

The train track $\tau_{3,3,1,1}$, where $\phi_{3,3,1,1}(\tau_{3,3,1,1})$ is carried by $\phi_{3,3,1,1}$, is depicted in Figure 8.

The matrix associated to this map is

$$M = \begin{pmatrix} 3 & 2 & 0 & 0 & 0 & 0 & 0 & 2 \\ 6 & 3 & 2 & 0 & 0 & 0 & 0 & 4 \\ 12 & 6 & 3 & 2 & 4 & 0 & 0 & 8 \\ 24 & 12 & 6 & 3 & 6 & 0 & 0 & 16 \\ 0 & 0 & 0 & 2 & 3 & 2 & 4 & 0 \\ 0 & 0 & 0 & 4 & 6 & 3 & 6 & 0 \\ 4 & 0 & 0 & 0 & 0 & 2 & 3 & 2 \\ 6 & 0 & 0 & 0 & 0 & 4 & 6 & 3 \end{pmatrix}$$

which has the characteristic polynomial

$$p_{\phi_{3,3,1,1}}(x) = x^8 - 24x^7 + 156x^6 - 424x^5 - 186x^4 - 424x^3 + 156x^2 - 24x + 1.$$

The polynomial $p_{\phi_{3,3,1,1}}(x)$ is the characteristic polynomial associated to the action of this map on the train track $\tau_{3,3,1,1}$. To show that $p_{\phi_{3,3,1,1}}(x)$ is irreducible, it suffices find 17 values of x such that $p_{\phi_{3,3,1,1}}(x)$ is prime by Fact 5.1. Indeed, the following tuples $(x, p_{\phi_{3,3,1,1}}(x))$ contain 17 x–values such that $p_{\phi_{3,3,1,1}}(x)$ is prime:

$$(-160, 496582824202141441), \quad (-102, 14653782370731169),$$
$$(-90, 5537981240501761), \quad (-76, 1495649690458849),$$
$$(-52, 81377571089569), \quad (-46, 32071763417569), \quad (-40, 11167704826561),$$
$$(-22, 134573887009), \quad (0, 1), \quad (8, -7522751), \quad (26, 59124433057),$$
$$(72, 502376857985089), \quad (86, 2218259932983937),$$
$$(90, 3237148147105441), \quad (120, 34853759407811521),$$
$$(158, 331770565001360449), \quad (164, 449704327465370209).$$

Therefore, we have that $p_{\phi_{3,3,1,1}}(x)$ is irreducible, and the leading eigenvalue $\lambda_{3,3,1,1}$ is a root of $\phi_{3,3,1,1}$. By applying [7, Theorem 1] we find that there are roots of this polynomial which are on the unit circle, and thus, there exist Galois conjugates of $\lambda_{3,3,1,1}$ on the unit circle. We now write

$$\frac{p_{\phi_{3,3,1,1}}}{x^4} = \left(x + \frac{1}{x}\right)^4 - 24\left(x + \frac{1}{x}\right)^3 + 152\left(x + \frac{1}{x}\right)^2 - 352\left(x + \frac{1}{x}\right) - 496.$$

We prove that $q(y) = y^4 - 24y^3 + 152y^2 - 352y - 496$ is an irreducible polynomial as follows. By Gauss's lemma, a primitive polynomial is irreducible over the integers if and only if it is irreducible over the rational numbers. Since $q(y)$ is primitive, it suffices to show that $q(y)$ is irreducible over the integers. The rational root theorem gives us that $q(y)$ has no roots, so if it is reducible then $q(y) = (y^2 + ay + b)(y^2 + cy + d)$. Therefore, suppose that $q(y) = (y^2 + ay + b)(y^2 + cy + d)$. Expanding gives rise to the system of equations

$$a + c = -24, \quad ac + b + d = 152, \quad ad + bc = -352, \quad bd = -496. \tag{3}$$

Substituting $a = -24 - c$ and $b = -496/d$ into the second and third equations give

$$-24c - c^2 - \frac{496}{d} + d = 152, \quad (-24 - c)d - \frac{496c}{d} = -352. \tag{4}$$

We solve for c in the second equation to find

$$c = \frac{24d^2 - 352d}{-d^2 - 496}.$$

Substituting this into the first equation gives

$$-24\left(\frac{24d^2 - 352d}{-d^2 - 496}\right)d - \left(\frac{24d^2 - 352d}{-d^2 - 496}\right)^2 d + d^2 - 152d - 496 = 0,$$

which has no integer roots. Therefore, there is no d satisfying the conditions we require, so $q(y)$ is irreducible. Finally, by using the formulas for the roots of a quartic equation, we see that $q(y)$ has two imaginary roots. Therefore the field $\mathbb{Q}(\lambda_{3,3,1,1} + \lambda_{3,3,1,1}^{-1})$ is not totally real.

We end with an example on $S_{0,10}$ where there are Galois conjugates of the stretch factor on the unit circle, and where the field $\mathbb{Q}(\lambda + 1/\lambda)$ is not totally real. For this example, we will apply the modification from Theorem 1.3 to the map associated to the partition

$$\mu_{3,3,3} = \{\{1, 4, 7\}, \{2, 5, 8\}, \{3, 6, 9\}\}$$

to obtain the partition

$$\mu_{3,3,3,1} = \{\{1, 5, 8\}, \{2, 6, 9\}, \{3, 7, 10\}, \{4\}\}.$$

This induces the map

$$\phi_{3,3,3,1} = D_4 D_{10} D_7 D_3 D_9 D_6 D_2 D_8 D_5 D_1.$$

The train track $\tau_{3,3,3,1}$, where $\phi_{3,3,3,1}(\tau_{3,3,3,1})$ is carried by $\phi_{3,3,3,1}$, is depicted in Figure 9.

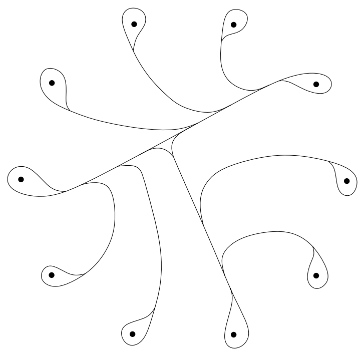


Figure 9: The train track associated to $\phi_{3,3,3,1}$.

The matrix associated to this map is

$$M = \begin{pmatrix} 3 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 \\ 6 & 3 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 4 \\ 12 & 6 & 3 & 2 & 4 & 0 & 0 & 0 & 0 & 8 \\ 24 & 12 & 6 & 3 & 6 & 0 & 0 & 0 & 0 & 16 \\ 0 & 0 & 0 & 2 & 3 & 2 & 4 & 0 & 0 & 0 \\ 0 & 0 & 0 & 4 & 6 & 3 & 2 & 4 & 0 & 0 \\ 0 & 0 & 0 & 8 & 12 & 6 & 3 & 6 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 2 & 3 & 2 & 0 \\ 4 & 0 & 0 & 0 & 0 & 0 & 4 & 6 & 3 & 2 \\ 6 & 0 & 0 & 0 & 0 & 0 & 8 & 12 & 6 & 3 \end{pmatrix}$$

which has the characteristic polynomial

$$p_{\psi_{3,3,3,1}}(x) = x^{10}-30x^9+285x^8-1864x^7-30x^6+204x^5-30x^4-1864x^3+285x^2-30x+1.$$

The polynomial $p_{\phi_{3,3,3,1}}(x)$ is the characteristic polynomial associated to the action of this map on the train track $\tau_{3,3,3,1}$. To show that $p_{\phi_{3,3,3,1}}(x)$ is irreducible, it suffices to find 21 values of x such that $p_{\phi_{3,3,3,1}}(x)$ is prime by Fact 5.1. Indeed, the following tuples $(x, p_{\phi_{3,3,3,1}}(x))$ contain 21 x–values such that $p_{\phi_{3,3,3,1}}(x)$ is prime:

$$(-224, 3624925501673023779835 53), \quad (-208, 17446898676424026808 2529),$$
$$(-168, 2129596989101254012 1329), \quad (-160, 13184110259956391044801),$$
$$(-76, 931111550641774 5721), \quad (-72, 5528387844285055921),$$
$$(-54, 351062674729634953), \quad (-52, 245103960106710121),$$
$$(-12, 405908347321), \quad (-10, 87091192801), \quad (0, 1), \quad (2, -189671),$$

$$(6,-285219647),\quad (24,6967292292721),\quad (74,3161394113461923721),$$
$$(186,41960187610521563501353),\quad (204,107295840626496890031721),$$
$$(216,191678753902872238701553),\quad (234,431604542240603942600521),$$
$$(258,1160267613906359066071321),\quad (264,1464201236305006987479121).$$

Therefore, we have that $p_{\phi_{3,3,3,1}}(x)$ is irreducible; thus the leading eigenvalue $\lambda_{3,3,3,1}$ is a root of $p_{\phi_{3,3,3,1}}(x)$. Applying [7, Theorem 1] we find that there are roots of this polynomial which are on the unit circle. We now write

$$\frac{p_{\psi_{3,3,3,1}}}{x^5}=\Big(x+\frac{1}{x}\Big)^5-30\Big(x+\frac{1}{x}\Big)^4+280\Big(x+\frac{1}{x}\Big)^3+1744\Big(x+\frac{1}{x}\Big)^2-880\Big(x+\frac{1}{x}\Big)-3307.$$

We rewrite the above as $q(y)=y^5-30y^4+280y^3+1744y^2-880y-3307$. To see that $q(y)$ is irreducible over $\mathbb{Q}$, it suffices to show that there are 11 values of y such that $q(y)$ is prime. Indeed, the following tuples $(y,q(y))$ contain 11 y–values such that $q(y)$ is prime:

$$(-12,-1096363),\quad (-10,-500107),\quad (-6,-42379),$$
$$(-4,1493),\quad (-2.2677),\quad (0,-3307),\quad (2,3701),$$
$$(4,32341),\quad (14,479861),\quad (16,658453),\quad (22,1928821).$$

Finally, we notice that the discriminant of the polynomial $q(y)$ is calculated to be

$$-10301707504334020544219.$$

As the discriminant is negative, we know that there must exist nonreal roots; therefore the field $\mathbb{Q}(\lambda_{3,3,3,1}+\lambda_{3,3,3,1}^{-1})$ is not totally real.

As we have provided explicit examples of pseudo-Anosov homeomorphisms for each case of Theorem 1.4, we have completed the proof. □

6 Construction on surfaces of higher genus

In this section, we prove Theorem 1.1. To prove this theorem, we begin by taking the map ϕ on $S_{0,8}$ found in Theorem 1.4(4) and show that by further puncturing $S_{0,8}$, we obtain pseudo-Anosov homeomorphisms which differ from the Penner and Thurston constructions on surfaces $S_{0,n}$ for $n\geq 8$. Furthermore, we lift the constructed pseudo-Anosov mapping classes on $S_{0,2g+2}$ to pseudo-Anosov mapping classes on surfaces of genus $g>0$ through a *branched cover* by treating the marked points as punctures; see Figure 1 for an example. A branched cover $S_{g,0}\to S_{0,2g+2}$ is a true covering map in the complement of a finite set of points of $S_{0,2g+2}$. These points are called the *branch points*. We now provide a proof for Theorem 1.1.

Theorem 1.1 *Let S be either $S_{0,n}$ for $n \geq 8$, or $S_{g,k}$ for $g \geq 3$ and $k \geq 0$. Then there exists a pseudo-Anosov homeomorphism ϕ_S on S with stretch factor λ_{ϕ_S} such that*

(i) *the trace field $\mathbb{Q}(\lambda_{\phi_S} + \lambda_{\phi_S}^{-1})$ is not totally real, and*

(ii) *there exist Galois conjugates of λ_{ϕ_S} on the unit circle.*

In particular, no power of ϕ_S arises from either Penner's or Thurston's constructions.

Proof In the proof of Theorem 1.4(4), we have already shown this theorem to be true for the surfaces $S_{0,8}$ and $S_{0,10}$. To show that there exists a pseudo-Anosov homeomorphism on punctured spheres $S_{0,n}$ where $n \geq 9$, we consider the map from Theorem 1.4(4) on $S_{0,8}$.

Let ϕ be the pseudo-Anosov homeomorphism of $S_{0,8}$ found in Theorem 1.4(4). For any pseudo-Anosov homeomorphism of a surface S, the set of periodic points is dense in S [2]. Since dense sets in a surface contain an infinite number of elements, this implies that there must be at least $n-8$ periodic points. There exists a power of the pseudo-Anosov map, ϕ^k, which fixes the $n-8$ periodic points. If we delete these $n-8$ fixed points, ϕ^k restricts to a pseudo-Anosov map on an n–times punctured sphere. Additionally, the foliations from the map ϕ on $S_{0,8}$ are the same foliations as for the map ϕ^k on $S_{0,n}$, but now the stretch factor is λ^k. Since the algebraic properties are invariant under powers of the pseudo-Anosov homeomorphism, we see that we have proven the claim for surfaces $S_{0,n}$ for $n \geq 9$.

From the above, we have the desired pseudo-Anosov mapping classes for all spheres with at least 8 punctures. The hyperelliptic involution of S_g induces a branched double cover $S_g \to S_{0,2g+2}$, where $S_{0,2g+2}$ is a sphere with $2g+2$ marked points. The pseudo-Anosov maps we constructed above for $S_{0,2g+2}$ lift to pseudo-Anosov maps on S_g with the same stretch factor. Therefore, we have that the same number-theoretic properties hold for the pseudo-Anosov maps on surfaces $S_{g,0}$, where $g \geq 3$. Since we have pseudo-Anosov maps on a closed surface of genus g, we once again have that periodic points are dense. Using a similar argument as above, this implies that we may puncture the surface of genus g any number of times, to obtain the desired pseudo-Anosov mapping classes on surfaces $S_{g,k}$ for $g \geq 3$ and $k \geq 0$. □

Appendix Introductory examples

In this section, we present two detailed examples on how to apply Theorems 1.2 and 1.3. Consider the six-times punctured sphere. We will begin by constructing two pseudo-Anosov maps on the six-times punctured sphere using Theorem 1.2.

Example A.1 Consider the six-times punctured sphere and label the punctures of the sphere as introduced in Figure 2. Up to spherical symmetry, there are two unique partitions of the six punctures so that the labels of the punctures are evenly spaced, namely

$$\mu = \{\{0,3\},\{1,4\},\{2,5\}\} = \{\mu_1,\mu_2,\mu_3\}$$

and

$$\bar{\mu} = \{\{0,2,4\},\{1,3,5\}\} = \{\bar{\mu}_1,\bar{\mu}_2\}.$$

Recall that we define the half-twist associated to puncture j as the half-twist around the curve separating punctures j and $j-1$. Therefore, these partitions can define the two maps,

$$\phi = D_5^2 D_2^2 D_4^2 D_1^2 D_3^2 D_0^2 = D_{\mu_3}^2 D_{\mu_2}^2 D_{\mu_1}^2,$$

and

$$\phi_{\bar{\mu}} = D_5^2 D_3^2 D_1^2 D_4^2 D_2^2 D_0^2 = D_{\bar{\mu}_2}^2 D_{\bar{\mu}_1}^2,$$

respectively. We prove that both maps are pseudo-Anosov.

We first prove that $\phi_\mu = D_{\mu_3}^2 D_{\mu_2}^2 D_{\mu_1}^2$ is a pseudo-Anosov map on $S_{0,6}$. To prove that ϕ_μ is a pseudo-Anosov map, we find a train track τ_μ on $S_{0,6}$ so that $\phi_\mu(\tau_\mu)$ is carried by τ_μ and show that the matrix presentation of ϕ_μ in the coordinates given by τ_μ is a Perron–Frobenius matrix.

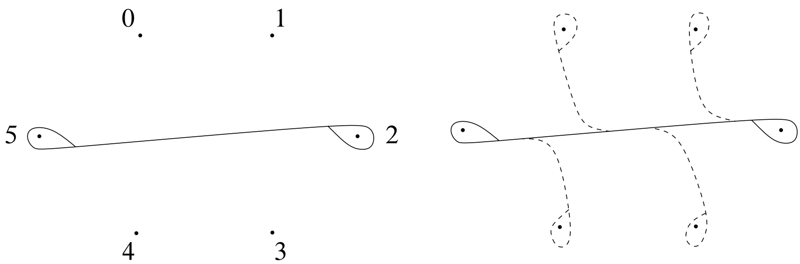


Figure 10: Constructing the train track for the map ϕ_μ.

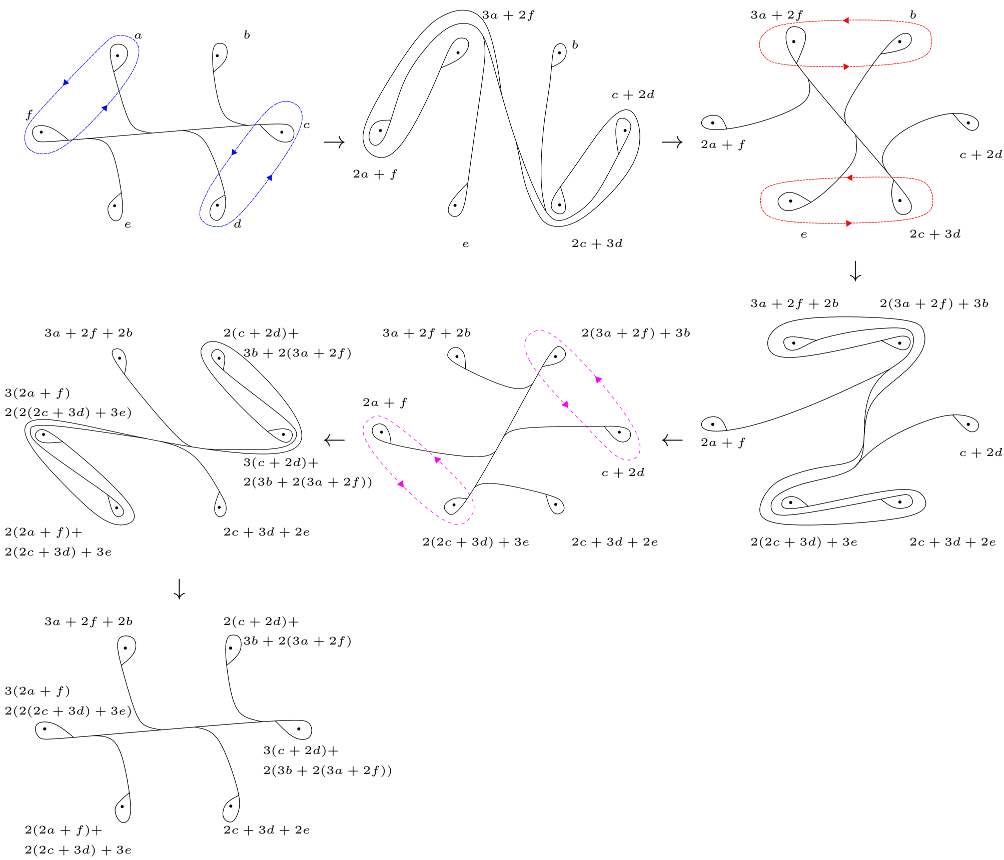


Figure 11: The train track $\phi_\mu(\tau_\mu)$ is carried by τ_μ.

First, we describe how we construct the train track for the map ϕ_μ based on the partition μ. Notice that μ has three subsets containing two punctures. As the punctures in each partition are such that $|i-j| \geq 2 \bmod 6$, the twists associated to the punctures in each subset are disjoint. Since there are two twists in each subset and the partition is evenly spaced, the train track has rotational symmetry of order two. Therefore, we construct a two-valent pretrack around the punctures labeled 2 and 5, pictured in Figure 10, left. Since there are three subsets, there are two branches turning tangentially into each of the two nodes on the two-valent pretrack, where these branches will turn left towards the pretrack, pictured in Figure 10, right.

The series of images in Figure 11 depict the train track τ_μ and its images under successive applications of the Dehn twists associated to ϕ_μ. These images prove that $\phi(\tau)$ is carried by τ_μ, and for every application of $D^2_{\mu_i}$, the train track τ_μ rotates

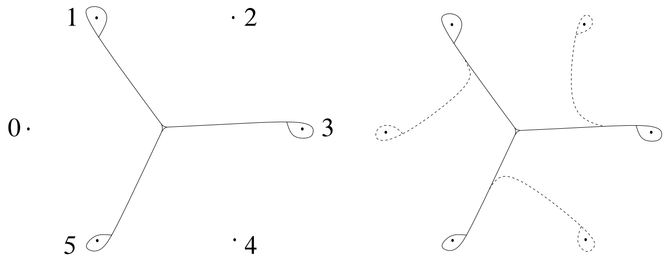


Figure 12: Constructing the train track for the map $\phi_{\bar{\mu}}$.

clockwise by $2\pi/6$. By keeping track of the weights on τ_μ, we calculate that the induced action on the space of weights on τ_μ is given by the matrix

$$A = \begin{pmatrix} 3 & 2 & 0 & 0 & 0 & 2 \\ 6 & 3 & 2 & 4 & 0 & 4 \\ 12 & 6 & 3 & 6 & 0 & 8 \\ 0 & 0 & 2 & 3 & 2 & 0 \\ 4 & 0 & 4 & 6 & 3 & 2 \\ 6 & 0 & 8 & 12 & 6 & 3 \end{pmatrix}.$$

Note that the space of admissible weights on τ_μ is the subset of $\mathbb{R}^6$ given by positive real numbers a, b, c, d, e and f such that $a+b+f = c+d+e$. The linear map described above preserves this subset. The square of the matrix A is strictly positive, which implies that the matrix is Perron–Frobenius. In fact, the top eigenvalue is $9+4\sqrt{5}$, which is associated to a unique irrational measured lamination F carried by τ_μ that is fixed by ϕ_μ. Lastly, since the train track τ_μ is large, generic, and birecurrent, we can apply Lemma 2.4 which finishes the proof that this map is pseudo-Anosov.

Notice that we can perform each of the half twists to any power and still have the exact same train track constructed above. However, the labels associated to the branches will subsequently increase or decrease in value according to how many twists are applied to each curve. Since all the twists are positive, we still have that all values in the resulting matrix will be positive and will be Perron–Frobenius. An application of Lemma 2.4 will give our desired result.

We will now show that $\phi_{\bar{\mu}} = D^2_{\bar{\mu}_2} D^2_{\bar{\mu}_1}$ is a pseudo-Anosov on $S_{0,6}$, which follows a similar argument as above.

We again analyze the partition $\bar{\mu}$ as it determines the construction of our train track $\tau_{\bar{\mu}}$. Notice that $\bar{\mu}$ has two subsets containing three twists each. Since there are three

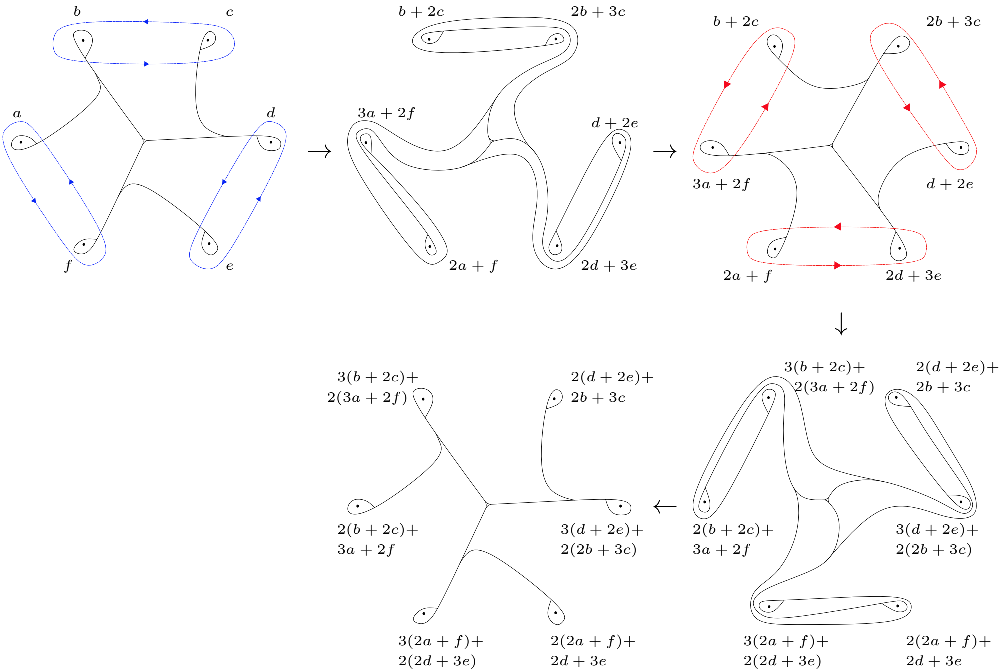


Figure 13: The train track $\phi_{\bar{\mu}}(\tau_{\bar{\mu}})$ is carried by $\tau_{\bar{\mu}}$.

punctures in each subset and the partition is evenly spaced, the train track has rotational symmetry of order three. Therefore, we will construct a three-valent pretrack around the punctures labeled 1, 3 and 5, pictured in Figure 12, left. Since there are two subsets, there is one branch turning tangentially towards each of the three nodes on the three-valent pretrack, where these branches will be turning left towards the pretrack, pictured in Figure 12, right.

The series of images in Figure 13 depict the train track $\tau_{\bar{\mu}}$ and its images under successive applications of the Dehn twists associated to $\phi_{\bar{\mu}}$. These images prove that $\phi_{\bar{\mu}}(\tau_{\bar{\mu}})$ is indeed carried by $\tau_{\bar{\mu}}$. We again notice that for every application of $D^2_{\bar{\mu}_i}$, the train track $\tau_{\bar{\mu}}$ rotates clockwise by $2\pi/6$. By keeping track of the weights on $\tau_{\bar{\mu}}$, we calculate that the induced action on the space of weights on τ is given by the matrix

$$B = \begin{pmatrix} 3 & 2 & 4 & 0 & 0 & 2 \\ 6 & 3 & 6 & 0 & 0 & 4 \\ 0 & 2 & 3 & 2 & 4 & 0 \\ 0 & 4 & 6 & 3 & 6 & 0 \\ 4 & 0 & 0 & 2 & 3 & 2 \\ 6 & 0 & 0 & 4 & 6 & 3 \end{pmatrix}.$$

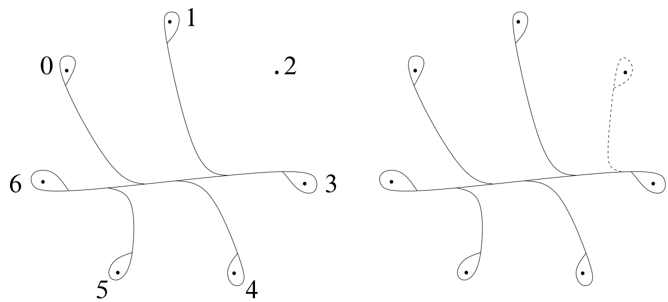


Figure 14: Constructing the train track for the map ψ.

The space of admissible weights on $\tau_{\bar{\mu}}$ is the subset of $\mathbb{R}^6$ given by positive real numbers a, b, c, d, e and f such that $b-a$, $d-c$ and $f-e$ are all positive and satisfy the triangle inequalities. The linear map described above preserves this subset. The square of the matrix B is strictly positive, which implies that the matrix is Perron–Frobenius. The top eigenvalue is $7+4\sqrt{3}$, which is associated to a unique irrational measured lamination F carried by $\tau_{\bar{\mu}}$ that is fixed by $\phi_{\bar{\mu}}$. As the train track is large, generic, and birecurrent, we may apply Lemma 2.4 to finish the proof that this map is pseudo-Anosov.

We now modify the pseudo-Anosov maps from Example A.1 to find two pseudo-Anosov maps on the seven-times punctured sphere. To achieve this, we apply Theorem 1.3 once to each of the maps found in Example A.1. For each of these maps, we note that we can apply the modification more than once to obtain additional pseudo-Anosov maps defined on spheres with more punctures.

Example A.2 We consider the seven-times punctured sphere with the labeling as introduced in Theorem 1.2. After applying Theorem 1.3 to the two partitions found in Example A.1, we obtain two partitions

$$\mu' = \{\{0,4\},\{1,5\},\{2,6\},\{3\}\} = \{\mu'_1, \mu'_2, \mu'_3, \mu'_4\},$$

and

$$\bar{\mu}' = \{\{0,3,5\},\{1,4,6\},\{2\}\} = \{\bar{\mu}'_1, \bar{\mu}'_2, \bar{\mu}'_3\}.$$

We begin by proving that the $\phi_{\mu'} = D^2_{\mu'_4} D^2_{\mu'_3} D^2_{\mu'_2} D^2_{\mu'_1}$ induced by the partition μ' is pseudo-Anosov. First, we describe how to construct the train track associated to $\phi_{\mu'}$, denoted by $\tau_{\mu'}$, from the train track τ_μ associated to the map ϕ_μ from the previous example. Consider the train track τ and place an extra puncture between the punctures

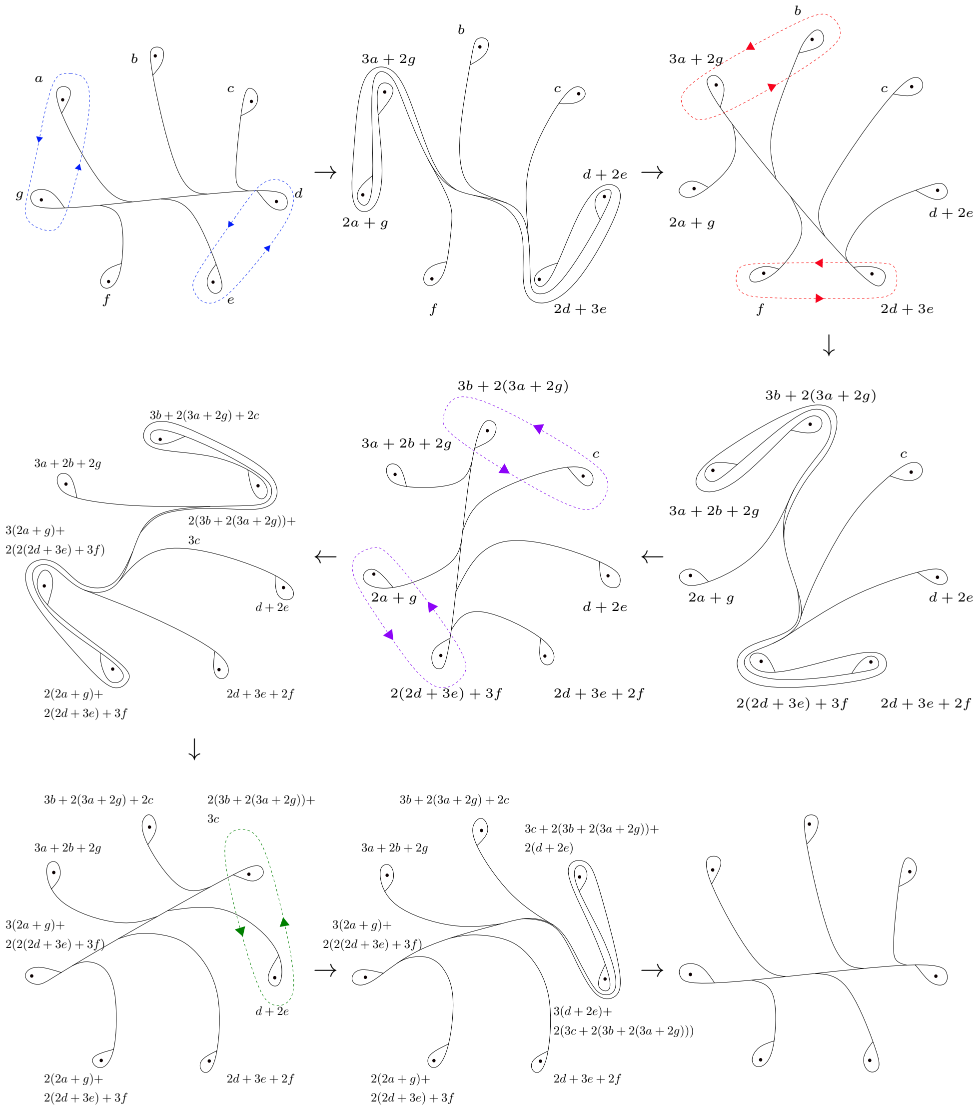


Figure 15: The train track $\phi_{\mu'}(\tau_{\mu'})$ is carried by $\tau_{\mu'}$.

labeled 1 and 2 in the previous example. Relabel the punctures so that the labeling is as in Theorem 1.2; see Figure 14, left. Therefore, we have a train track without a branch around the puncture labeled 2, but the rest of the train track is as in Example A.1 (up to relabeling). We construct a branch around the puncture labeled 2 which will turn tangentially towards the two valent pretrack, turning left towards the puncture labeled 3; see Figure 14, right.

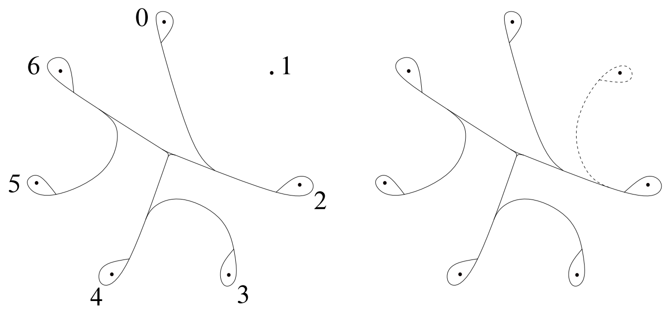


Figure 16: Constructing the train track for the map $\phi_{\bar{\mu}'}$.

The series of images in Figure 15 depict the train track $\tau_{\mu'}$ and its images under successive applications of the Dehn twists associated to $\phi_{\mu'}$. These images prove that $\phi_{\mu'}(\tau_{\mu'})$ is carried by $\tau_{\mu'}$. By keeping track of the weights on $\tau_{\mu'}$, we calculate that the induced action on the space of weights on $\tau_{\mu'}$ is given by the matrix

$$C = \begin{pmatrix} 3 & 2 & 0 & 0 & 0 & 0 & 2 \\ 6 & 3 & 2 & 0 & 0 & 0 & 4 \\ 12 & 6 & 3 & 2 & 0 & 0 & 8 \\ 24 & 12 & 6 & 3 & 6 & 0 & 16 \\ 0 & 0 & 0 & 2 & 3 & 2 & 0 \\ 4 & 0 & 0 & 4 & 6 & 3 & 2 \\ 6 & 0 & 0 & 8 & 12 & 6 & 3 \end{pmatrix}.$$

The space of admissible weights on τ' is the subset of $\mathbb{R}^7$ given by the positive real numbers a, b, c, d, e, f and g such that $a+b+d+f=c+e+g$. The linear map described above preserves this subset. The square of the matrix C is strictly positive, which implies that the matrix is Perron–Frobenius. Additionally, the top eigenvalue is approximately 22.08646, which is associated to a unique irrational measured lamination F carried by $\tau_{\mu'}$ which is fixed by $\phi_{\mu'}$. As the train track is large, generic, and birecurrent, we may apply Lemma 2.4 to finish the proof that this map is pseudo-Anosov.

We now show that the map $\phi_{\bar{\mu}'} = D^2_{\bar{\mu}_3'} D^2_{\bar{\mu}_2'} D^2_{\bar{\mu}_1'}$ induced by the partition $\bar{\mu}'$ is pseudo-Anosov. To construct the train track, we will consider the train track $\tau_{\bar{\mu}}$ associated to the map $\phi_{\bar{\mu}}$ from the previous example. Consider the train track $\tau_{\bar{\mu}}$ and place an extra puncture between the punctures labeled 0 and 1 in the previous example. Relabel the punctures so that the labeling is as in Theorem 1.2; see Figure 16, left. Therefore, we have a train track on $S_{0,7}$ which does not have a branch around the puncture labeled 1,

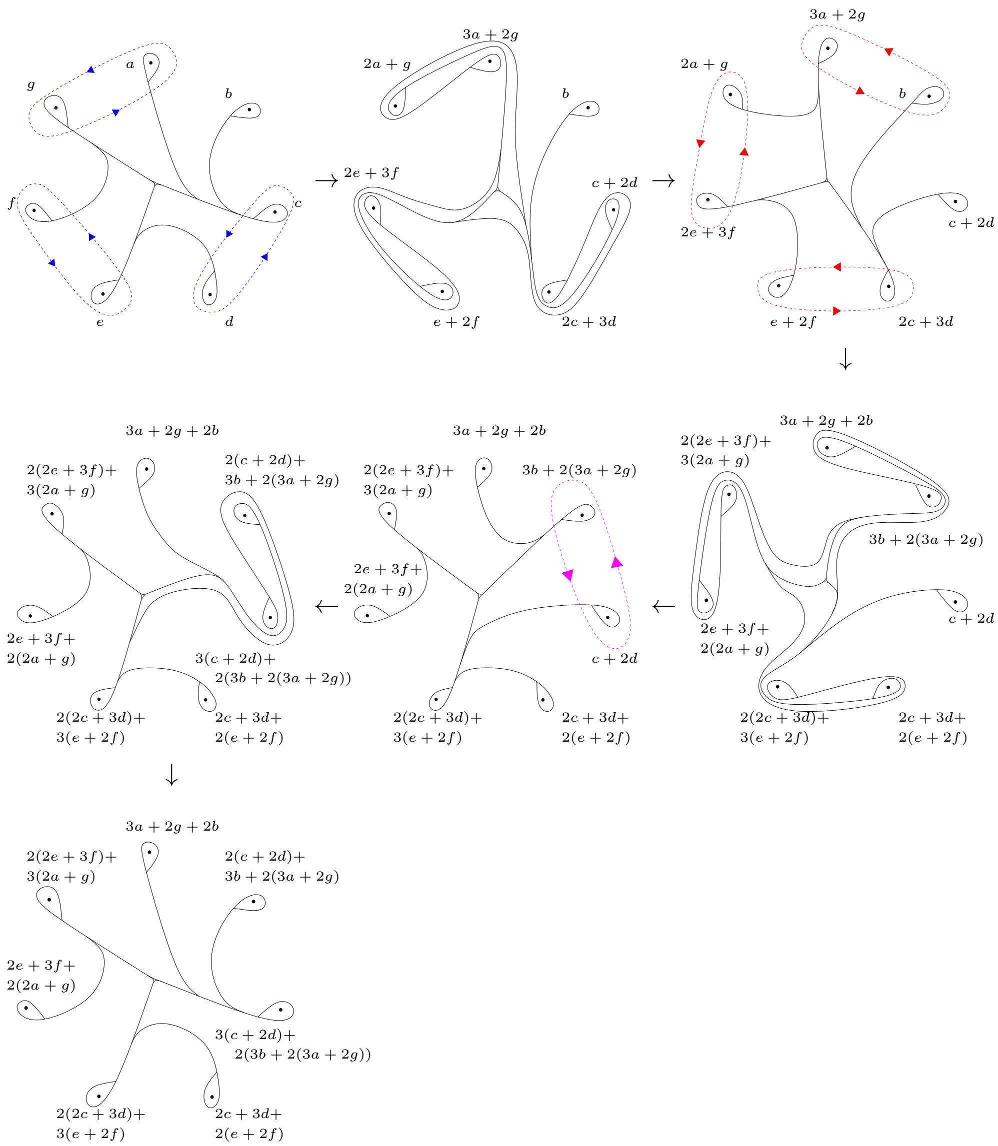


Figure 17: The train track $\phi'_{\bar{\mu}}(\tau'_{\bar{\mu}})$ is carried by $\tau'_{\bar{\mu}}$.

and the rest of the train track is as found in Example A.1 (up to relabeling). We then construct a branch around the puncture labeled 1 which will turn tangentially into the three valent pretrack, turning left towards the puncture labeled 2; see Figure 16, right.

The series of images in Figure 17 depict the train track $\tau_{\bar{\mu}'}$ and its images under successive applications of the Dehn twists associated to $\phi_{\bar{\mu}'}$.

Figure 17 shows that $\phi_{\bar{\mu}'}(\tau_{\bar{\mu}'})$ is indeed carried by $\tau_{\bar{\mu}'}$. By keeping track of the weights on $\tau_{\bar{\mu}'}$, we calculate that the induced action on the space of weights on $\tau_{\bar{\mu}'}$ is given by the matrix

$$D = \begin{pmatrix} 3 & 2 & 0 & 0 & 0 & 0 & 2 \\ 6 & 3 & 2 & 4 & 0 & 0 & 4 \\ 12 & 6 & 3 & 6 & 0 & 0 & 8 \\ 0 & 0 & 2 & 3 & 2 & 4 & 0 \\ 0 & 0 & 4 & 6 & 3 & 6 & 0 \\ 4 & 0 & 0 & 0 & 2 & 3 & 2 \\ 6 & 0 & 0 & 0 & 4 & 6 & 3 \end{pmatrix}.$$

The space of admissible weights on $\tau_{\bar{\mu}'}$ is the subset of $\mathbb{R}^7$ given by the positive real numbers a, b, c, d, e, f and g such that $c-b-a$, $e-d$ and $g-f$ are all positive and satisfy the triangle inequalities. The linear map described above preserves this subset. The square of the matrix D is strictly positive, which implies that the matrix is Perron–Frobenius. The top eigenvalue of this matrix is

$$5 + \tfrac{1}{3}\sqrt[3]{2916 - 12\sqrt{93}} + \left(\tfrac{2}{3}\right)^{2/3}\sqrt[3]{243 + \sqrt{93}},$$

which is associated to a unique irrational measured lamination F carried by $\tau'_{\bar{\mu}}$ which is fixed by $\phi'_{\bar{\mu}}$. As the train track is large, generic, and birecurrent, we may apply Lemma 2.4 to finish the proof that this map is pseudo-Anosov.

References

[1] **K Conrad**, *Irreducibility tests in $\mathbb{Q}[T]$*, preprint (2017) Available at https://kconrad.math.uconn.edu/blurbs/ringtheory/irredtestsoverQ.pdf

[2] **B Farb**, **D Margalit**, *A primer on mapping class groups*, Princeton Mathematical Series 49, Princeton Univ. Press (2012) MR Zbl

[3] **W J Harvey**, *Boundary structure of the modular group*, from "Riemann surfaces and related topics" (I Kra, B Maskit, editors), Ann. of Math. Stud. 97, Princeton Univ. Press (1981) 245–251 MR Zbl

[4] **J Hubbard**, **H Masur**, *Quadratic differentials and foliations*, Acta Math. 142 (1979) 221–274 MR Zbl

[5] **P Hubert**, **E Lanneau**, *Veech groups without parabolic elements*, Duke Math. J. 133 (2006) 335–346 MR Zbl

[6] **R Kenyon**, **J Smillie**, *Billiards on rational-angled triangles*, Comment. Math. Helv. 75 (2000) 65–108 MR Zbl

[7] **J Konvalina**, **V Matache**, *Palindrome-polynomials with roots on the unit circle*, C. R. Math. Acad. Sci. Soc. R. Can. 26 (2004) 39–44 MR Zbl

[8] **H A Masur**, **Y N Minsky**, *Geometry of the complex of curves, I: Hyperbolicity*, Invent. Math. 138 (1999) 103–149 MR Zbl

[9] **A Papadopoulos**, **R C Penner**, *A characterization of pseudo-Anosov foliations*, Pacific J. Math. 130 (1987) 359–377 MR Zbl

[10] **R C Penner**, *A construction of pseudo-Anosov homeomorphisms*, Trans. Amer. Math. Soc. 310 (1988) 179–197 MR Zbl

[11] **R C Penner**, **J L Harer**, *Combinatorics of train tracks*, Annals of Mathematics Studies 125, Princeton Univ. Press (1992) MR Zbl

[12] **K Rafi**, **Y Verberne**, *Geodesics in the mapping class group*, Algebr. Geom. Topol. 21 (2021) 2995–3017 MR Zbl

[13] **H Shin**, **B Strenner**, *Pseudo-Anosov mapping classes not arising from Penner's construction*, Geom. Topol. 19 (2015) 3645–3656 MR Zbl

[14] **W P Thurston**, *On the geometry and dynamics of diffeomorphisms of surfaces*, Bull. Amer. Math. Soc. 19 (1988) 417–431 MR Zbl

[15] **A Zorich**, *Flat surfaces*, from "Frontiers in number theory, physics, and geometry, I" (P Cartier, B Julia, P Moussa, P Vanhove, editors), Springer (2006) 437–583 MR Zbl

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

verberne.math@gmail.com

https://sites.google.com/view/yvonverberne

Received: 8 August 2019 Revised: 23 September 2021

Algebraic & Geometric Topology **23**:4 (2023) 1641–1692
DOI: 10.2140/agt.2023.23.1641
Published: 14 June 2023

Actions of solvable Baumslag–Solitar groups on hyperbolic metric spaces

CAROLYN R ABBOTT
ALEXANDER J RASMUSSEN

We give a complete list of the cobounded actions of solvable Baumslag–Solitar groups on hyperbolic metric spaces up to a natural equivalence relation. The set of equivalence classes carries a natural partial order first introduced by Abbott, Balasubramanya and Osin, and we describe the resulting poset completely. There are finitely many equivalence classes of actions, and each equivalence class contains the action on a point, a tree, or the hyperbolic plane.

20E06, 20E08, 20F16, 20F65

1 Introduction

The Baumslag–Solitar groups $\mathrm{BS}(m,n)$ are a classically studied family of groups defined by the particularly straightforward presentations $\langle a,t \mid ta^m t^{-1} = a^n\rangle$. In the case $m=n=1$, $\mathrm{BS}(1,1)$ is isomorphic to the abelian group $\mathbb{Z}^2$. In general, for $n \geq 2$, $\mathrm{BS}(1,n)$ is a nonabelian solvable group via the isomorphism $\mathrm{BS}(1,n) \cong \mathbb{Z}\big[\frac{1}{n}\big] \rtimes \mathbb{Z}$.

The group $\mathrm{BS}(1,n)$ admits several natural actions on hyperbolic metric spaces. The Cayley graph of $\mathrm{BS}(1,n)$ with respect to the generating set $\{a^{\pm 1}, t^{\pm 1}\}$ consists of a number of "sheets", each of which is quasi-isometric to the hyperbolic plane $\mathbb{H}^2$. The sheets are glued together along complements of horoballs in a pattern described by the $(n+1)$–regular tree. The result is pictured in Figure 1 for the case $n=2$. Collapsing each sheet down to a vertical geodesic line gives a projection from the Cayley graph to the $(n+1)$–regular tree. Moreover, the action of $\mathrm{BS}(1,n)$ on its Cayley graph permutes the fibers of this projection, so that we obtain an action of $\mathrm{BS}(1,n)$ on the $(n+1)$–regular tree. Similarly, the idea of "collapsing the sheets down to a single sheet" gives an action on the hyperbolic plane, although this idea is a bit harder to formalize.

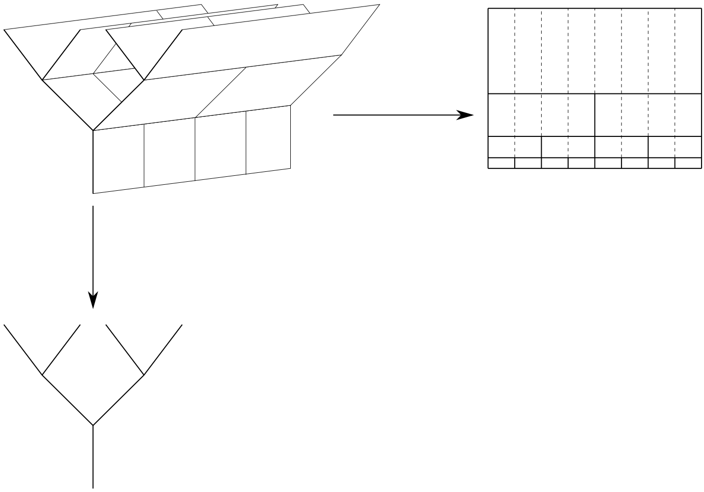


Figure 1: Two natural actions of the group $\mathrm{BS}(1,2)$ via projections.

Formally, the action of $\mathrm{BS}(1,n)$ on $\mathbb{H}^2$ is given by the representation

$$\mathrm{BS}(1,n) \to \mathrm{PSL}(2,\mathbb{R}),$$

where

$$a \mapsto \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \quad t \mapsto \begin{pmatrix} \sqrt{n} & 0 \\ 0 & 1/\sqrt{n} \end{pmatrix}.$$

Another way to obtain the natural action on the $(n+1)$–regular tree is by expressing $\mathrm{BS}(1,n)$ as an HNN extension over the subgroup $\langle a\rangle \cong \mathbb{Z}$ and considering the action of $\mathrm{BS}(1,n)$ on the Bass–Serre tree of the resulting one-edge graph of groups.

In addition to these actions, $\mathrm{BS}(1,n)$ admits an obvious homomorphism $\mathrm{BS}(1,n) \to \mathbb{Z}$ defined by $a \mapsto 0$ and $t \mapsto 1$. This defines an action of $\mathrm{BS}(1,n)$ on (the hyperbolic metric space) $\mathbb{R}$ via integral translations. This action may also be obtained by collapsing either the hyperbolic plane or the $(n+1)$–regular tree onto a vertical geodesic in a height-respecting manner and noting that $\mathrm{BS}(1,n)$ permutes the fibers of the resulting projection. Even more trivially, any group admits an action on a point (which is a hyperbolic metric space).

All of these actions are *cobounded* in the sense that the orbit of a point under the action admits the entire space as a bounded neighborhood. One may naturally wonder whether these four actions (the actions on a tree, the hyperbolic plane, the line, and a point) give a complete list of the nontrivial cobounded actions of $\mathrm{BS}(1,n)$ on hyperbolic metric spaces up to equivalence. We show that this is indeed the case *if* n is prime.

In the case that n is *not* prime, we show that $\mathrm{BS}(1,n)$ admits actions on certain other Bass–Serre trees which may be understood algebraically. For each divisor l of n, $\mathbb{Z}\big[\frac{1}{l}\big]$ is a subring of $\mathbb{Z}\big[\frac{1}{n}\big]$. We may form an ascending HNN extension of $\mathbb{Z}\big[\frac{1}{l}\big]$. Specifically, consider the one-edge graph of groups with vertex group $\mathbb{Z}\big[\frac{1}{l}\big]$ and edge group $\mathbb{Z}\big[\frac{1}{l}\big]$ which includes isomorphically onto $\mathbb{Z}\big[\frac{1}{l}\big]$ on one end and as the subgroup $n\mathbb{Z}\big[\frac{1}{l}\big]$ on the other end. It is not too hard to show that the fundamental group of this graph of groups is $\mathrm{BS}(1,n)$ and thus there is an action of $\mathrm{BS}(1,n)$ on the corresponding Bass–Serre tree. Considering these actions on Bass–Serre trees together with the canonical actions described in the above paragraph, we show that this gives a complete list of the cobounded hyperbolic actions of $\mathrm{BS}(1,n)$ (up to an equivalence relation, described below). Before stating these results precisely, we need to introduce some terminology.

1.1 Hyperbolic structures on groups

In groups that admit interesting actions on hyperbolic metric spaces, it is natural to wonder whether it is possible to describe all such actions explicitly. Unfortunately, this (slightly naive) goal is currently unattainable for almost all commonly studied groups. For instance, using the machinery of combinatorial horoballs introduced by Groves and Manning in [7], one may produce uncountably many *parabolic* actions of any countable group on hyperbolic metric spaces, ie actions with a fixed point on the boundary and all group elements acting elliptically or parabolically. For this reason, we restrict to considering *cobounded* actions on hyperbolic metric spaces, which are never parabolic. Moreover, we must use some notion of *equivalence* for the actions of a fixed group on different hyperbolic spaces, as it is quite easy to modify an action on a given hyperbolic space equivariantly to produce an action on a quasi-isometric space. The equivalence relation we consider, introduced by Abbott, Balasubramanya and Osin in [1], is roughly *equivariant quasi-isometry*. See Section 2 for more details.

Having restricted to cobounded actions up to equivalence, it is usually still quite difficult to describe explicitly all of the equivalence classes of actions of a given group on different hyperbolic metric spaces. For instance, in [1] Abbott, Balasubramanya and Osin considered the hyperbolic actions of *acylindrically hyperbolic* groups, a very wide class of groups all displaying some features of negative curvature. There they showed that any acylindrically hyperbolic group admits uncountably many distinct equivalence classes of actions on hyperbolic spaces.

But for groups which don't display strong features of negative curvature, it may be possible to give a complete list of their cobounded hyperbolic actions. For instance, in [1], Abbott, Balasubramanya and Osin gave a complete list of the equivalence classes of cobounded actions of $\mathbb{Z}^n$ on hyperbolic metric spaces. More recently, in [3], Balasubramanya gave a complete list of the actions of *lamplighter groups* on hyperbolic spaces. Our work draws inspiration from her strategy.

In these cases, it is possible to say even more about the actions of a fixed group on different hyperbolic metric spaces. We are interested in the question of when one action "retains more information" about the group than another. This question leads to a partial order on the set of equivalence classes of actions of a group G on hyperbolic spaces X, defined in [1]. Roughly, we say that $G \curvearrowright X$ *dominates* $G \curvearrowright Y$ when the action $G \curvearrowright Y$ may be obtained by equivariantly *coning off* certain subspaces of X. See Section 2.2 for the precise definition. This partial order descends to a partial order on equivalence classes of actions.

Hence, for a group G, the set of equivalence classes of actions of G on different hyperbolic metric spaces is a poset $\mathcal{H}(G)$. We will give a complete description of the poset $\mathcal{H}(\mathrm{BS}(1,n))$. Let $n = p_1^{n_1} p_2^{n_2} \cdots p_k^{n_k}$ be the prime factorization of n and let $\mathcal{K}_n$ be the poset $2^{\{1,\dots,k\}} \setminus \{\varnothing\}$, with the partial order given by inclusion.

Theorem 1.1 *For any $n \geq 2$, $\mathcal{H}(\mathrm{BS}(1,n))$ has the following structure: $\mathcal{H}_{\mathrm{qp}}(\mathrm{BS}(1,n))$, the subposet of quasiparabolic structures, consists of a copy of $\mathcal{K}_n$ and a single additional structure which is incomparable to every element of $\mathcal{K}_n$, and every quasiparabolic structure dominates a single lineal structure, which dominates a single elliptic structure; see Figure 2.*

Following [1], we say a group G is *$\mathcal{H}$–accessible* if $\mathcal{H}(G)$ contains a largest element. Otherwise, we say G is *$\mathcal{H}$–inaccessible*. The following result is immediate.

Corollary 1.2 $\mathrm{BS}(1,n)$ *is $\mathcal{H}$–inaccessible.*

Farb and Mosher prove in [5] that two solvable Baumslag–Solitar groups $\mathrm{BS}(1,m)$ and $\mathrm{BS}(1,n)$ with $m, n \geq 2$ are quasi-isometric if and only if they are commensurable, which occurs if and only if there exist integers $r, i, j > 0$ such that $m = r^i$ and $n = r^j$. In this case, we have $\mathcal{K}_n = \mathcal{K}_m$, which yields the following corollary.

Corollary 1.3 *If $\mathrm{BS}(1,m)$ and $\mathrm{BS}(1,n)$ are quasi-isometric, then the posets*

$$\mathcal{H}(\mathrm{BS}(1,m)) \quad \text{and} \quad \mathcal{H}(\mathrm{BS}(1,n))$$

are isomorphic.

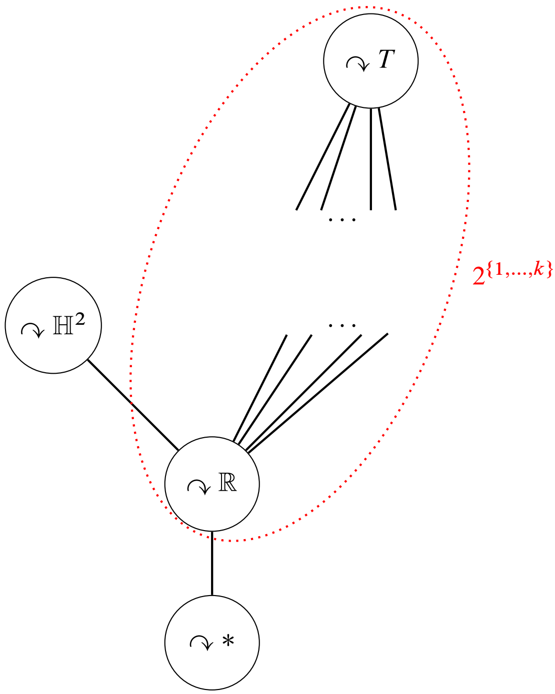


Figure 2: The poset $\mathcal{H}(\mathrm{BS}(1,n))$. The subposet circled is isomorphic to the power set $2^{\{1,\dots,k\}}$, which is a lattice.

Theorem 1.1 may be related to the algebraic description of Bass–Serre trees described above. Recall that for each divisor l of n there is a Bass–Serre tree of $\mathrm{BS}(1,n)$ corresponding to the subring $\mathbb{Z}\left[\frac{1}{l}\right]$ of $\mathbb{Z}\left[\frac{1}{n}\right]$. Note moreover that if the integers l and m both have the same prime divisors then in fact $\mathbb{Z}\left[\frac{1}{l}\right] = \mathbb{Z}\left[\frac{1}{m}\right]$. Thus the subring and the Bass–Serre tree only depend on the set of prime divisors of l. We obtain one Bass–Serre tree for each subset of $\{p_1, \dots, p_k\}$, and these correspond exactly to the subposet $\mathcal{K}_n$ of $\mathcal{H}(\mathrm{BS}(1,n))$.

1.2 About the proof

Any group action on a hyperbolic space falls into one of finitely many types (elliptic, parabolic, lineal, quasiparabolic, or general type) depending on the number of fixed points on the boundary and the types of isometries defined by various group elements. See Section 2 for precise definitions.

Since $\mathrm{BS}(1,n)$ is solvable, it contains no free subgroup and therefore — by the ping-pong lemma — any nonelliptic action of $\mathrm{BS}(1,n)$ on a hyperbolic metric space must have a fixed point on the boundary. Since we consider only cobounded actions, we see that $\mathrm{BS}(1,n)$ can have only lineal or quasiparabolic cobounded actions on hyperbolic metric spaces. Hence we are left to consider such actions.

We crucially use the fact that $\mathrm{BS}(1,n)$ can be written as a semidirect product $\mathbb{Z}\big[\frac{1}{n}\big]\rtimes_\alpha\mathbb{Z}$. In [4] Caprace, Cornulier, Monod and Tessera classified the lineal and quasiparabolic actions of certain groups $H\rtimes\mathbb{Z}$ in the language of *confining subsets* of H under the action of $\mathbb{Z}$; see Section 2.

Using techniques developed in [4], we show that the lineal and quasiparabolic actions of $\mathrm{BS}(1,n)$ naturally correspond to confining subsets of $\mathbb{Z}\big[\frac{1}{n}\big]$ under the actions of α and α^{-1}, and so we are led to try to classify such subsets. The confining subsets under the action of α^{-1} are straightforward to classify, and in fact they all correspond to (the equivalence class of) the action of $\mathrm{BS}(1,n)$ on $\mathbb{H}^2$. On the other hand, the classification of confining subsets under the action of α is more complicated. We show that such subsets correspond in a natural way to ideals in the ring of n–adic integers $\mathbb{Z}_n$. To see how such ideals arise, we consider confining subsets $Q\subset\mathbb{Z}\big[\frac{1}{n}\big]$ under the action of α and write elements $a\in Q$ in base n,

$$a=\pm a_r\dots a_0.a_{-1}\dots a_{-s},$$

allowing any number of zeros at the end of this expression. We then consider the set of all $\dots x_2x_1x_0\in\mathbb{Z}_n$ such that for any s, the sequence $x_sx_{s-1}\dots x_0$ appears as the sequence of digits to the right of the decimal point in some element of Q. That is, we require that there is a number of the form

$$a_r\dots a_0.x_s\dots x_0$$

in Q for arbitrarily large s. We show that the set of n–adic integers $\dots x_sx_{s-1}\dots x_0$ obtained in this manner is an ideal of $\mathbb{Z}_n$. It may be thought of as a kind of *limit set* of Q. We also show that this process can be reversed to associate confining subsets of $\mathbb{Z}[\frac{1}{n}]$ to ideals of $\mathbb{Z}_n$. With this correspondence in hand, we describe how all of the resulting actions are equivalent to the actions of $\mathrm{BS}(1,n)$ on certain Bass–Serre trees.

1.3 Other results and results in the literature

The poset described in Theorem 1.1 is interesting because of its asymmetry (when n is not a power of a prime). In [3] Balasubramanya described $\mathcal{H}(L_n)$, where L_n is the Lamplighter group $\mathbb{Z}/n\mathbb{Z}\wr\mathbb{Z}=\mathbb{Z}/n\mathbb{Z}\wr\langle t\rangle$. In this case, $\mathcal{H}(L_n)$ splits into two isomorphic subposets of quasiparabolic structures (corresponding to actions in which the fixed point of L_n is the attracting fixed point of t, and respectively the repelling fixed point of t) which each dominate a single lineal structure which in turn dominates an elliptic structure. Work of the authors in [2] shows that this structure also holds for semidirect products $\mathbb{Z}^2 \rtimes_\alpha \mathbb{Z}$ where $\alpha \in \mathrm{SL}(2,\mathbb{Z})$ is an Anosov matrix. It would be interesting to see what extra properties of a semidirect product $G \rtimes \mathbb{Z}$ are needed to ensure this kind of symmetry.

In [3] Balasubramanya considers hyperbolic actions of general wreath products

$$G \wr \mathbb{Z} = \left(\bigoplus_{n\in\mathbb{Z}} G \right) \rtimes \mathbb{Z}$$

and shows that $\mathcal{H}(G \wr \mathbb{Z})$ always contains two copies of the poset of subgroups of G. In the case $G = \mathbb{Z}/n\mathbb{Z}$ we have $G \wr \mathbb{Z} = L_n$ and this suffices to describe all of $\mathcal{H}(L_n)$.

In Section 5 we describe a general algebraic construction of quasiparabolic structures on semidirect products $H \rtimes \mathbb{Z}$. We show that in the case of $\mathbb{Z} \wr \mathbb{Z}$, this construction suffices to produce a countable chain of quasiparabolic structures.

Acknowledgements Abbott was partially supported by NSF award DMS-1803368. Rasmussen was partially supported by NSF award DMS-1610827. The authors thank Denis Osin for pointing out Corollary 1.3 and the referee for helpful comments.

2 Background

2.1 Actions on hyperbolic spaces

Given a metric space X, we denote by d_X the distance function on X. A map $f: X \to Y$ between metric spaces X and Y is a *quasi-isometric embedding* if there is a constant C such that for all $x, y \in X$,

$$\frac{1}{C} d_X(x,y) - C \le d_Y(f(x), f(y)) \le C d_X(x,y) + C.$$

If, in addition, Y is contained in the C–neighborhood of $f(X)$ then f is called a *quasi-isometry*. If a group G acts (by isometries) on X and Y, then a map $f: X \to Y$ is *coarsely G–equivariant* if for every $x \in X$ we have

$$\sup_{g\in G} d_Y(f(gx), gf(x)) < \infty.$$

We will assume that all actions are by isometries. The action of a group G on a metric space X is *cobounded* if there exists a bounded diameter subspace $B \subseteq X$ such that $X = \bigcup_{g\in G} gB$.

Given an action $G \curvearrowright X$ of G on a hyperbolic space, an element $g \in G$ is *elliptic* if it has bounded orbits; *loxodromic* if the map $\mathbb{Z} \to X$ given by $n \mapsto g^n \cdot x_0$ for some (equivalently, any) $x_0 \in X$ is a quasi-isometric embedding; and *parabolic* otherwise.

Any group action on a hyperbolic space falls into one of finitely many types depending on the number of fixed points on the boundary and the types of isometries defined by various group elements. This classification was described by Gromov in [6]: the action $G \curvearrowright X$ (where X is hyperbolic) is

- *elliptic* if G has a bounded orbit in X;
- *lineal* if G fixes two points in ∂X;
- *parabolic* if G fixes a unique point of ∂X and no element of G acts as a loxodromic isometry of X;
- *quasiparabolic* if G fixes a unique point of ∂X and at least one element of G acts as a loxodromic isometry; and
- *general type* if G doesn't fix any point of ∂X and at least one element of G acts as a loxodromic isometry.

2.2 Hyperbolic structures

Fix a group G. For any (possibly infinite) generating set S of G, let $\Gamma(G,S)$ be the Cayley graph of G with respect to the generating set S, and let $\|\cdot\|_S$ denote the word norm on G with respect to S. Given two generating sets S and T of a group G, we say T is *dominated by* S, written $T \preceq S$, if

$$\sup_{g \in S} \|g\|_T < \infty.$$

It is clear that $\preceq$ is a preorder on the set of generating sets of G and so induces an equivalence relation: $S \sim T$ if and only if $T \preceq S$ and $S \preceq T$. Let $[S]$ be the equivalence class of a generating set. Then the preorder $\preceq$ induces a partial order $\preccurlyeq$ on the set of all equivalence classes of generating sets of G via $[S] \preccurlyeq [T]$ if and only if $S \preceq T$.

Definition 2.1 Given a group G, the *poset of hyperbolic structures on* G is defined to be

$$\mathcal{H}(G) := \{[S] \mid G = \langle S \rangle \text{ and } \Gamma(G,S) \text{ is hyperbolic}\},$$

equipped with the partial order $\preccurlyeq$.

Notice that since hyperbolicity is a quasi-isometry invariant of geodesic metric spaces, the above definition is independent of the choice of representative of the equivalence class $[S]$. Every element $[S] \in \mathcal{H}(G)$ gives rise to a cobounded action on a hyperbolic space, namely $G \curvearrowright \Gamma(G, S)$. Moreover, given a cobounded action on a hyperbolic space $G \curvearrowright X$, a standard Schwarz–Milnor argument — see [1, Lemma 3.11] — provides a generating set S of G such that $\Gamma(G, S)$ is equivariantly quasi-isometric to X. We say that two actions $G \curvearrowright X$ and $G \curvearrowright Y$ are *equivalent* if there exists a coarsely G–equivariant quasi-isometry $X \to Y$. By [1, Proposition 3.12], there is a one-to-one correspondence between equivalence classes $[S] \in \mathcal{H}(G)$ and equivalence classes of cobounded actions $G \curvearrowright X$ with X hyperbolic.

We denote the set of equivalence classes of cobounded elliptic, lineal, quasiparabolic, and general-type actions by $\mathcal{H}_e$, $\mathcal{H}_\ell$, $\mathcal{H}_{\mathrm{qp}}$, and $\mathcal{H}_{\mathrm{gt}}$, respectively. Since parabolic actions cannot be cobounded, we have for any group G,

$$\mathcal{H}(G) = \mathcal{H}_e(G) \sqcup \mathcal{H}_\ell(G) \sqcup \mathcal{H}_{\mathrm{qp}}(G) \sqcup \mathcal{H}_{\mathrm{gt}}(G).$$

A lineal action of a group G on a hyperbolic space X is *orientable* if no element of G permutes the two limit points of G on ∂X. We denote the set of equivalence classes of orientable lineal actions of G by $\mathcal{H}_\ell^+(G)$.

The action of a group G on a hyperbolic metric space X is *focal* if it fixes a boundary point $\xi \in \partial X$ and if some element of G acts as a loxodromic isometry. If $[S] \in \mathcal{H}(G)$ is a focal action, then $[S] \in \mathcal{H}_\ell^+(G) \sqcup \mathcal{H}_{\mathrm{qp}}(G)$.

Quasicharacters A map $q\colon G \to \mathbb{R}$ is a *quasicharacter* (also called a *quasimorphism*) if there exists a constant D such that for all $g, h \in G$, we have $|q(gh)-q(g)-q(h)| \le D$. We say that q has *defect at most* D. If, in addition, the restriction of q to every cyclic subgroup is a homomorphism, then q is called a *pseudocharacter* (or *homogeneous quasimorphism*). Every quasicharacter q gives rise to a pseudocharacter ρ defined by $\rho(g) = \lim_{n\to\infty} \frac{1}{n} q(g^n)$; we call ρ the *homogenization of* q. Every pseudocharacter is constant on conjugacy classes. If q has defect at most D, then it is straightforward to check that $|q(g) - \rho(g)| \le D$ for all $g \in G$.

Let $G \curvearrowright X$ be an action on a hyperbolic space with a global fixed point $\xi \in \partial X$. For any sequence $\boldsymbol{x} = (x_n)$ in X converging to ξ and any fixed basepoint $s \in X$, we define the associated quasicharacter $q_{\boldsymbol{x}}\colon G \to \mathbb{R}$ as follows. For all $g \in G$,

$$q_{\boldsymbol{x}}(g) = \limsup_{n\to\infty} (d_X(gs, x_n) - d_X(s, x_n)). \tag{1}$$

Its homogenization $\rho_{\boldsymbol{x}}\colon G \to \mathbb{R}$ is the *Busemann pseudocharacter*. It is known that for any two sequences $\boldsymbol{x}$ and $\boldsymbol{y}$ converging to ξ, $\sup_{g\in G} |q_{\boldsymbol{x}}(g) - q_{\boldsymbol{y}}(g)| < \infty$, and thus we may drop the subscript $\boldsymbol{x}$ in $\rho_{\boldsymbol{x}}$. If ρ is a homomorphism, then the action $G \curvearrowright X$ is called *regular*.

Lemma 2.2 [4, Lemma 3.8] *Let $G \curvearrowright X$ be an action on a hyperbolic space with a global fixed point $\xi \in \partial X$. Then the (possibly empty) set of loxodromic isometries of the action is $\{g \in G \mid \rho(g) \neq 0\}$, and the set of those with attracting fixed point ξ is $\{g \in G \mid \rho(g) < 0\}$. In particular, the action of G is elliptic or parabolic if and only if $\rho \equiv 0$, and lineal or quasiparabolic otherwise.*

2.3 Confining subsets

Consider a group $G = H \rtimes_\alpha \mathbb{Z}$ where $\alpha \in \mathrm{Aut}(H)$ acts by $\alpha(h) = tht^{-1}$ for any $h \in H$, where t is a generator of $\mathbb{Z}$. Let Q be a *symmetric* subset of H. The following definition is from [4, Section 4].

Definition 2.3 The action of α is (*strictly*) *confining H into Q* if it satisfies the following three conditions:

(a) $\alpha(Q)$ is (strictly) contained in Q;

(b) $H = \bigcup_{k\geq 0} \alpha^{-k}(Q)$; and

(c) $\alpha^{k_0}(Q \cdot Q) \subseteq Q$ for some $k_0 \in \mathbb{Z}_{\geq 0}$.

Remark 2.4 The definition of confining subset given in [4] does not require symmetry of the subset $Q \subset H$. However, according to [4, Theorem 4.1], to classify regular quasiparabolic structures on a group it suffices to consider only confining subsets which are symmetric. See also Proposition 2.6 in this paper.

Remark 2.5 By the discussion after the statement of [4, Theorem 4.1], if there is a subset $Q \subseteq H$ such that the action of α is confining H into Q but not strictly confining, then $[Q \cup \{t^{\pm 1}\}] \in \mathcal{H}_\ell^+(G)$. If the action is strictly confining, then $[Q \cup \{t^{\pm 1}\}] \in \mathcal{H}_{\mathrm{qp}}(G)$.

We will focus primarily on describing subsets Q of H into which the action of α is (strictly) confining H. For brevity, we will refer to such Q as (*strictly*) *confining under the action of* α, or simply (*strictly*) *confining* if the action of α is understood.

To see an example of how confining subsets arise, it is useful to consider the action $\mathrm{BS}(1,n) \curvearrowright \mathbb{H}^2$ described in the introduction. In this action, the conjugates of t act as loxodromic isometries whereas the elements of $H = \mathbb{Z}\left[\frac{1}{n}\right]$ act as parabolic isometries.

Consider the subset $Q \subset H$ of isometries that translate a given point p (say i in the upper half plane model) by some bounded distance (say 1). If $g \in H$ then we consider the action of a conjugate $t^{-k} g t^k$ where $k \gg 0$. Considering the actions of t^k, g, and t^{-k} in turn we see that

- t^k first translates p vertically by a very large distance,
- g shifts $t^k p$ within the horocycle based at ∞ containing it by a very small distance,
- t^{-k} takes this horocycle isometrically back to the horocycle containing p.

In other words, $t^{-k} g t^k$ is a parabolic isometry which moves p by a much smaller distance than g itself does. In particular, if k is large enough, then $t^{-k} g t^k \in Q$. Furthermore, the infimal k with $t^{-k} g t^k \in Q$ depends only on how far g moves the original point p. Using these facts it is easy to see that Q is a confining subset of H.

Thus there is a correspondence of quasiparabolic structures on $\mathrm{BS}(1,n)$ and confining subsets of H. Precisely, we have the following result, which is a minor modification of [3, Theorems 4.4 and 4.5] and [4, Proposition 4.5]. It is proved in Section 4.1.

Proposition 2.6 *A hyperbolic structure $[T]$ is an element of $\mathcal{H}_{\mathrm{qp}}(\mathrm{BS}(1,n))$ if and only if there exists a symmetric subset $Q \subset \mathbb{Z}\left[\frac{1}{n}\right]$ which is strictly confining under the action of α or α^{-1} such that $[T] = [Q \cup \{t^{\pm 1}\}]$.*

2.4 *n*–adic integers

Definition 2.7 An *n–adic integer* is an infinite series

$$\sum_{i=0}^{\infty} a_i n^i,$$

where $a_i \in \{0, 1, \dots, n-1\}$. Such an element can also be written in its base n expansion as

$$\dots a_3 a_2 a_1 a_0.$$

We denote the set of n–adic integers by $\mathbb{Z}_n$.

We define operations of addition and multiplication on $\mathbb{Z}_n$, which gives it the structure of a ring.

Definition 2.8 Let $a = \sum_{i=0}^{\infty} a_i n^i$ and $b = \sum_{i=0}^{\infty} b_i n^i$ be elements of $\mathbb{Z}_n$. Then the sum $a+b$ is the n–adic integer $c = \sum_{i=0}^{\infty} c_i n^i$ defined inductively as follows. Let $c_0 = a_0 + b_0 \bmod n$ (we identify $\mathbb{Z}/n\mathbb{Z}$ with $\{0, \dots, n-1\}$) and $t_0 = \lfloor (a_0 + b_0)/n \rfloor$, so that $a_0 + b_0 = c_0 + t_0 n$. Assume that $c_0, \dots, c_{i-1}$ and $t_0, \dots, t_{i-1}$ have been defined, and let

$$c_i = a_i + b_i + t_{i-1} \bmod n \quad \text{and} \quad t_i = \left\lfloor \frac{a_i + b_i + t_{i-1}}{n} \right\rfloor,$$

so that $a_i + b_i = c_i + t_i n$.

The product ab is the n–adic number $d = \sum_{i=o}^{\infty} d_i n^i$ defined inductively as follows. Let $d_0 = a_0 b_0 \bmod n$, and let $s_0 = \lfloor a_0 b_0 / n \rfloor$, so that $a_0 b_0 = d_0 + s_0 n$. Assume that $d_1, \dots, d_{i-1}$ and $s_1, \dots, s_{i-1}$ have been defined, and let

$$d_i = \sum_{j=0}^{i} a_j b_{i-j} + s_{i-1} \bmod n \quad \text{and} \quad s_i = \left\lfloor \frac{1}{n} \sum_{j=0}^{i} a_j b_{n-j} + s_{i-1} \right\rfloor,$$

so that $\sum_{j=0}^{i} a_j b_{n-j} + s_{i-1} = d_i + s_i n$.

In the above operations, we think of t_i and s_i as the amounts that are "carried" at each step, analogous to how we carry digits when adding and multiplying in base 10. An element $a = \dots a_2 a_1 a_0$ of the ring $\mathbb{Z}_n$ is a unit if and only if a_0 is relatively prime to n.

Let

$$\varphi_l \colon \mathbb{Z}_n \to \mathbb{Z}/n^l\mathbb{Z} \tag{2}$$

be the ring homomorphism which identifies an element of $\mathbb{Z}_n$ with its l^{th} partial sum,

$$\varphi_l(a) = \varphi_l\left(\sum_{i=0}^{\infty} a_i n^i\right) = \sum_{i=0}^{l-1} a_i n^i.$$

These homomorphisms are compatible in the following sense. For any $k \le l$, let the map

$${}_nT_k^l \colon \mathbb{Z}/n^l\mathbb{Z} \to \mathbb{Z}/n^k\mathbb{Z}$$

be reduction modulo n^k. Then for any $a \in \mathbb{Z}_n$, we have ${}_nT_k^l(\varphi_l(a)) = \varphi_k(a)$.

In fact, any infinite sequence (a^i) of elements $a^i \in \mathbb{Z}/n^i\mathbb{Z}$ satisfying ${}_nT_k^l(a^l) = a^k$ defines an element $a \in \mathbb{Z}_n$. Namely, identify a^l with the unique representative b^l of its congruence class in the set $\{0, 1, \dots, n^l - 1\}$. We may write $b^l = \sum_{i=0}^{l-1} b_i^l n^i$. Then for any $i \ge 0$ and $l, k > i$ we have $b_i^l = b_i^k$, and hence we may write

$$a_0 = b_0^1 = b_0^2 = \cdots, \quad a_1 = b_1^2 = b_1^3 = \cdots, \quad \text{etc.}$$

Writing

$$a = \sum_{i=0}^{\infty} a_i n^i$$

we have $\varphi_i(a) = a^i$ for all i. In particular, this shows that $\mathbb{Z}_n$ is isomorphic to the inverse limit $\varprojlim \mathbb{Z}/n^i\mathbb{Z}$.

There is a metric d on $\mathbb{Z}_n$, called the *n–adic metric*, defined by $d(x,y) = n^{-q}$ if q is maximal such that n^q divides $x-y$, ie if the first q digits of $x-y$ are zero and the $(q+1)^{\text{st}}$ digit is nonzero.

Lemma 2.9 *With the topology coming from the n–adic metric, $\mathbb{Z}_n$ is a compact space.*

Proof Suppose we have an infinite sequence $\{x^i\}_{i \in \mathbb{Z}_{>0}}$ such that for each i, we have $x^i = \dots x_3^i x_2^i x_1^i x_0^i$. By the pigeon-hole principle, there is some $y_0 \in \{0,\dots,n-1\}$ such that $x_0^i = y_0$ for infinitely many i. The collection of these x^i forms a subsequence $\{x^{i_{0,j}}\}_{j \in \mathbb{Z}_{>0}}$. Repeating this construction iteratively, we have a sequence of subsequences $\{\{x^{i_{k,j}}\}_j\}_k$ and a number $y = \dots y_3 y_2 y_1 y_0 \in \mathbb{Z}_n$ such that for each k every element of $\{x^{i_{k,j}}\}_j$ agrees with y in its first $k+1$ digits. Moreover, $\{x^{i_{k+1,j}}\}_j$ is a subsequence of $\{x^{i_{k,j}}\}_j$. Thus the diagonal subsequence $\{x^{i_{k,k}}\}_k$ of $\{x^i\}_i$ converges to y. Consequently, $\mathbb{Z}_n$ is sequentially compact. Since $\mathbb{Z}_n$ is a metric space, it is also compact. □

We now consider the ring structure of $\mathbb{Z}_n$.

Lemma 2.10 *Let $n = p_1^{n_1} p_2^{n_2} \cdots p_k^{n_k}$. There is an isomorphism*

$$\mathbb{Z}_n \cong \mathbb{Z}_{p_1^{n_1}} \times \dots \times \mathbb{Z}_{p_k^{n_k}}. \tag{3}$$

Let us give a description of this isomorphism. First, suppose $n = n_1 n_2$, where n_1 and n_2 are relatively prime. To define a map on $\mathbb{Z}_n$, we use the identification of an element of $a \in \mathbb{Z}_m$ with the sequence $(\varphi_l(a)) \in \varprojlim \mathbb{Z}/m^l\mathbb{Z}$. Let

$$f : \mathbb{Z}_n \to \mathbb{Z}_{n_1} \times \mathbb{Z}_{n_2}$$

be defined by

$$f(a) = (x_l, y_l) = (\varphi_l(a) \bmod n_1^l, \varphi_l(a) \bmod n_2^l).$$

The sequence (x_l) (resp. (y_l)) satisfies ${}_{n_1}T_k^l(x_l) = x_k$ (resp. ${}_{n_2}T_k^l(y_l) = y_k$) for any $k \le l$, and so (x_l) (resp. (y_l)) determines a unique point in $\mathbb{Z}_{n_1}$ (resp. $\mathbb{Z}_{n_2}$). The fact that f is an isomorphism follows from the Chinese remainder theorem. The isomorphism (3) now follows by repeatedly applying f to distinct pairs of prime factors.

Moreover, there is an isomorphism $g\colon \mathbb{Z}_{p^j} \to \mathbb{Z}_p$ defined as follows. The number $g(a)$ is a with each coefficient expanded to $a_i = a_{i,j-1}p^{j-1} + \cdots + a_{i,1}p + a_{i,0}$, where $a_{i,k} \in \{0, \dots, p-1\}$. Composing g with the isomorphism (3) shows that, in fact, there is an isomorphism $\mathbb{Z}_n \to \mathbb{Z}_{p_1} \times \cdots \times Z_{p_k}$.

We use the isomorphism g solely to describe the ideals of $\mathbb{Z}_{p^j}$. The nonzero ideals of $\mathbb{Z}_p$ are exactly $p^i\mathbb{Z}_p = \bigl\{\sum_{j=i}^{\infty} a_j p^j \mid a_j \in \{0, \dots, p-1\}\bigr\}$. Using the above isomorphism, it is clear that the nonzero ideals of $\mathbb{Z}_{p^j}$ are exactly $g^{-1}(p^i\mathbb{Z}_p) = p^i\mathbb{Z}_{p^j}$.

We now give a technical description of when elements of $\mathbb{Z}_{p^j}$ and, more generally, $\mathbb{Z}_n$ are contained in a particular ideal. On a first reading the reader may want to skip Lemmas 2.11 and 2.13 and simply read Example 2.12 and Remark 2.14 instead.

The following lemma describes when an element $a \in \mathbb{Z}_{p^j}$ is contained in an ideal $p^i\mathbb{Z}_{p^j}$. By the above discussion, this occurs when the image of the element under the isomorphism g is contained in $g(p^i\mathbb{Z}_{p^j}) = p^i\mathbb{Z}_p$. An element $x \in \mathbb{Z}_p$ is in the ideal $p^i\mathbb{Z}_p$ exactly when $\varphi_i(x) \equiv 0 \bmod p^{i-1}$. Since g expands the coefficients of a, the definition of L in the statement of the lemma is simply the smallest positive integer such that the expansion of $\varphi_L(a)$ contains $\varphi_i(g(a))$. Equivalently, it is the smallest positive integer such that the expansion of the $p^{(L-1)j}$ term in a contains p^{i-1}, which is the largest power of p appearing in $\varphi_i(g(a))$. The largest power of p in the expansion of the $p^{(L-1)j}$ term of a is $p^{(L-1)j+j-1} = p^{Lj-1}$, and thus L is the smallest positive integer such that $Lj - 1 \geq i - 1$.

Lemma 2.11 *For any $a \in \mathbb{Z}_{p^j}$,*

$$a \in p^i\mathbb{Z}_{p^j} \iff \varphi_L(a) \equiv 0 \bmod p^i,$$

where $L = \lceil i/j \rceil$. Moreover,

$$a \in (0) \iff a = 0 \iff \varphi_s(a) = 0$$

for all s.

Proof For the first statement, notice that $a \in p^i\mathbb{Z}_{p^j}$ if and only if $g(a) \in p^i\mathbb{Z}_p$ if and only if $\varphi_i(g(a)) \equiv 0 \bmod p^i$ if and only if $\varphi_L(a) \equiv 0 \bmod p^i$. The first two "if and only if" statements are clear, while the last follows from the calculation

$$\begin{aligned}\varphi_L(a) &= a_{L-1}p^{(L-1)j} + \cdots + a_1p^j + a_0\\ &= a_{L-1,j-1}p^{Lj-1} + \cdots + a_{L-1,k}p^{i-1} + \cdots + a_{L-1,0}p^{(L-1)j}\\ &\qquad\qquad + a_{L-2,j-1}p^{(L-1)j-1} + \cdots + a_{0,1}p + a_{0,0}\\ &= a_{L-1,j-1}p^{Lj-1} + \cdots + a_{L-1,k+1}p^i + \varphi_i(g(a)),\end{aligned}$$

where $k \in \{0, \dots, j-1\}$ is such that $(L-1)j+k = i-1$. Now, if $\varphi_i(g(a)) \equiv 0 \bmod p^i$, then it is clear that $\varphi_L(a) \equiv 0 \bmod p^i$. Similarly, if $\varphi_L(a) \equiv 0 \bmod p^i$, then since $a_{L-1,j-1}p^{(L-1)j} + \dots + a_{L-1,k}p^i \equiv 0 \bmod p^i$, we must have $\varphi_i(g(a)) \equiv 0 \bmod p^i$.

The second statement is just the definition of the zero ideal. □

Example 2.12 We give explicit descriptions of two ideals in $\mathbb{Z}_{2^3}$.

(a) First consider the ideal $2\mathbb{Z}_{2^3}$. Then

$$a = \dots a_2 a_1 a_0 \in 2\mathbb{Z}_{2^3} \iff g(a) \in 2\mathbb{Z}_2 = \left\{ \sum_{i=1}^{\infty} b_i 2^i \;\middle|\; b_i \in \{0, \dots, p-1\} \right\}.$$

Therefore, the only restriction on the partial sums of a is that

$$\varphi_1(a) = a_0 = a_{0,2}2^2 + a_{0,1}2 + 0,$$

ie $a_0 \equiv 0 \bmod 2$.

(b) Next, consider the ideal $2^{13}\mathbb{Z}_{2^3}$. Then

$$a = \dots a_2 a_1 a_0 \in 2^{13}\mathbb{Z}_{2^3} \iff g(a) \in 2^{13}\mathbb{Z}_2 = \left\{ \sum_{i=13}^{\infty} b_i 2^i \;\middle|\; b_i \in \{0, \dots, p-1\} \right\}.$$

In this case, we must have

$$\begin{aligned} \varphi_5(a) &= a_4 2^{12} + a_3 2^9 + a_2 2^6 + a_1 2^3 + a_0 \\ &= a_{4,2} \cdot 2^{14} + a_{4,1} \cdot 2^{13} + 0 \cdot 2^{12} + 0 \cdot 2^{11} + 0 \cdot 2^{10} + \dots + 0 \cdot 2 + 0, \end{aligned}$$

ie $\varphi_5(a) \equiv 0 \bmod 2^{13}$. Note that this also shows that $\varphi_i(a) \equiv 0 \bmod 2^{3i-1}$ for all $i \le 4$ (for example, $\varphi_2(a) = a_1 2^3 + a_0 = 0 \cdot 2^5 + 0 \cdot 2^4 + 0 \cdot 2^3 + 0 \cdot 2^2 + 0 \cdot 2 + 0 \equiv 0 \bmod 2^5$).

Using the isomorphism in (3), any ideal $\mathfrak{a}$ of $\mathbb{Z}_n$ can be written as

$$\mathfrak{a} = \mathfrak{a}_1 \times \dots \times \mathfrak{a}_k,$$

where $\mathfrak{a}_i = p_i^{a_i}\mathbb{Z}_{p_i^{n_i}}$ for some a_i or $\mathfrak{a}_i = (0)$. The next lemma gives a precise description of when an element $a \in \mathbb{Z}_n$ is contained in an ideal $\mathfrak{a} = \mathfrak{a}_1 \times \dots \times \mathfrak{a}_k$. The conditions are similar to those in Lemma 2.11. For each i such that $\mathfrak{a}_i = p_i^{a_i}\mathbb{Z}_{p_i^{n_i}}$, we have a constant $L_i = \lceil a_i/n_i \rceil$ as in Lemma 2.11, and one needs only check a condition on a *single* partial sum, namely that $\varphi_{L_i}(a) \equiv 0 \bmod p_i^{a_i}$. On the other hand, for each i such that $\mathfrak{a}_i = (0)$, one needs to check that *every* partial sum of a satisfies an appropriate condition.

Lemma 2.13 *Let $\mathfrak{a} = \mathfrak{a}_1 \times \cdots \times \mathfrak{a}_k$, where for each i, $\mathfrak{a}_i = p_i^{a_i}\mathbb{Z}_{p_i^{n_i}}$ or $\mathfrak{a}_i = (0)$. For each i, let $L_i = \lceil a_i/n_i \rceil$. Then for any $a = \ldots a_3a_2a_1a_0 \in \mathbb{Z}_n$, $a \in \mathfrak{a}$ if and only if*

- $\varphi_s(a) \equiv 0 \bmod p_i^{sn_i}$ *for all s and all i such that $\mathfrak{a}_i = (0)$; and*
- $\varphi_s(a) \equiv 0 \bmod p_i^{a_i}$ *for any s such that $s = L_i$ for some i.*

Proof For any i such that $\mathfrak{a}_i = p_i^{a_i}\mathbb{Z}_{p_i^{n_i}}$, this follows immediately from the definition of the isomorphism

$$\mathbb{Z}_n \to \mathbb{Z}_{p_1^{n_1}} \times \cdots \times \mathbb{Z}_{p_k^{n_k}}$$

and Lemma 2.11. Fix i such that $\mathfrak{a}_i = (0)$. Then, by the isomorphism in (3), we have that any $a \in \mathbb{Z}_n$ can be written as $a = (a_1, \ldots, a_k)$ where, considering $\mathbb{Z}_n$ and $\mathbb{Z}_{p_i^{n_i}}$ as $\varprojlim \mathbb{Z}/n^l\mathbb{Z}$ and $\varprojlim \mathbb{Z}/p_i^{ln_i}\mathbb{Z}$ respectively, each a_i is given by the sequence $(\varphi_s(a) \bmod p_i^{sn_i})_{s=1}^{\infty}$. Now, $a_i \in (0) \subset \mathbb{Z}_{p_i^{n_i}}$ if and only if $g(a_i) \in (0) \subset \mathbb{Z}_{p_i}$. Recall that $g(a_i)$ is a_i with each coefficient expanded so that

$$\varphi_s(a) \bmod p_i^{sn_i} \equiv a_{s-1,n_i-1}\, p_i^{sn_i-1} + a_{s-1,n_i-2}\, p_i^{sn_i-2} + \cdots + a_{0,1}\, p_i + a_{0,0} \bmod p_i^{sn_i}.$$

From this we see that $g(a_i) \in (0) \subset \mathbb{Z}_{p_i}$ if and only if $\varphi_s(a) \equiv 0 \bmod p_i^k$ for all $1 \le k \le sn_i$ if and only if $\varphi_s(a) \equiv 0 \bmod p_i^{sn_i}$ for all s. □

Remark 2.14 We point out one particular case of this lemma which will be important in later sections: if some $\mathfrak{a}_i = (0)$, then $\varphi_1(a) = a_0 \equiv 0 \bmod p_i^{n_i}$.

We now describe a partial order on the set of ideals of $\mathbb{Z}_n$.

Definition 2.15 Define a relation $\le$ on ideals of $\mathbb{Z}_n$ by $\mathfrak{a} \le \mathfrak{b}$ if $n^k\mathfrak{a} \subseteq \mathfrak{b}$ for some k. Define an equivalence $\mathfrak{a} \sim \mathfrak{b}$ if $\mathfrak{a} \le \mathfrak{b}$ and $\mathfrak{b} \le \mathfrak{a}$.

Definition 2.16 An ideal $\mathfrak{a} = \mathfrak{a}_1 \times \cdots \times \mathfrak{a}_k$ is *full* if $\mathfrak{a}_j$ is either (0) or $\mathbb{Z}_{p_j^{n_j}}$ for every $j = 1, \ldots, k$.

Lemma 2.17 *For any n, there is a unique full ideal in each equivalence class of ideals of $\mathbb{Z}_n$.*

Proof Let $\mathfrak{a} = \mathfrak{a}_1 \times \cdots \times \mathfrak{a}_k$ be an ideal of $\mathbb{Z}_n$, and consider the full ideal $\mathfrak{b} = \mathfrak{b}_1 \times \cdots \times \mathfrak{b}_k$ where $\mathfrak{b}_i = (0)$ if and only if $\mathfrak{a}_i = (0)$. Recall that if $\mathfrak{b}_i \ne (0)$, then $\mathfrak{b}_i = \mathbb{Z}_{p_i^{n_i}}$. We claim that $\mathfrak{a} \sim \mathfrak{b}$. It is clear that $\mathfrak{a} \subseteq \mathfrak{b}$, and thus $\mathfrak{a} \le \mathfrak{b}$. For any i such that $\mathfrak{a}_i \ne (0)$, let $\mathfrak{a}_i = p^{a_i}\mathbb{Z}_{p_i^{n_i}}$, and let $A = \max_i\{a_i\}$. We will show that $n^A\mathfrak{b} \subset \mathfrak{a}$. We have

$$n^A\mathfrak{b} = n^A\mathfrak{b}_1 \times \cdots \times n^A\mathfrak{b}_k,$$

and, for each i,
$$n^A \mathfrak{b}_i = p_1^{An_1} \cdots p_k^{An_k} \mathbb{Z}_{p_i^{n_i}}.$$
For all $j \neq i$, the element $p_j^{An_j}$ is a unit in $\mathbb{Z}_{p_i^{n_i}}$, and thus $p_j^{n_j} \mathbb{Z}_{p_i^{n_i}} = \mathbb{Z}_{p_i^{n_i}}$. Therefore,
$$n^A \mathfrak{b}_i = p_i^{An_i} \mathbb{Z}_{p_i^{n_i}},$$
and since $A \geq a_i$ by definition, it follows that
$$n^A \mathfrak{b}_i = p_i^{An_i} \mathbb{Z}_{p_i^{n_i}} \subseteq p_i^{a_i} \mathbb{Z}_{p_i^{n_i}} = \mathfrak{a}_i.$$
Consequently,
$$n^A \mathfrak{b} \subseteq \mathfrak{a},$$
which implies that $\mathfrak{b} \leq \mathfrak{a}$. Therefore, $\mathfrak{a} \sim \mathfrak{b}$.

Suppose next that there are two distinct full ideals, $\mathfrak{b} = \mathfrak{b}_1 \times \cdots \times \mathfrak{b}_k$ and $\mathfrak{c} = \mathfrak{c}_1 \times \cdots \times \mathfrak{c}_k$ with $\mathfrak{b} \sim \mathfrak{a} \sim \mathfrak{c}$. Then $\mathfrak{b} \sim \mathfrak{c}$, which immediately implies that $\mathfrak{b}_i = (0)$ if and only if $\mathfrak{c}_i = (0)$. Indeed, if there is an i such that (without loss of generality) $\mathfrak{b}_i = (0)$ but $\mathfrak{c}_i \neq (0)$, then there is no power C of n such that $p_i^C \mathfrak{c}_i \subset \mathfrak{b}_i$, and so $\mathfrak{b} \not\sim \mathfrak{c}$, which is a contradiction. Thus by the definition of full ideals, we conclude that $\mathfrak{b} = \mathfrak{c}$. □

While we mostly work with n–adic integers in this paper, we will occasionally need the notion of an n–adic number as well.

Definition 2.18 An *n–adic number* is an infinite series
$$\sum_{i=m}^{\infty} a_i n^i$$
where $a_i \in \{0, 1, \dots, n-1\}$ and $m \in \mathbb{Z}$ can be positive, negative, or zero. If $m \geq 0$, then such an element is an n–adic integer. If $m = -\ell$ for $\ell \in \mathbb{Z}_{>0}$, then such an element can also be written in its base n expansion as
$$\dots a_3 a_2 a_1 a_0 . a_{-1} \dots a_{-\ell}.$$
We denote the set of n–adic numbers by $\mathbb{Q}_n$.

Letting $S = \{n, n^2, n^3, \dots\}$, we see that $\mathbb{Q}_n = S^{-1}\mathbb{Z}_n$ is the localization of $\mathbb{Z}_n$ at S. In particular, $\mathbb{Q}_n$ is a ring. If $n = p^k$ is a power of a prime, then $\mathbb{Z}_n$ is an integral domain and $\mathbb{Q}_n$ is its field of fractions. If n is not a power of a prime, however, then $\mathbb{Z}_n$ is not an integral domain and $\mathbb{Q}_n$ will not be a field.

The only property of n–adic numbers we will need is that one can define an n–adic absolute value on $\mathbb{Q}_n$.

Definition 2.19 Given $q = \sum_{i=m}^{\infty} a_i n^i \in \mathbb{Q}_n$ with $a_m \neq 0$, we define the *n–adic absolute value* of q to be

$$\|q\|_n = n^{-m}.$$

In particular, $q = \sum_{i=m}^{\infty} a_i n^i \in \mathbb{Z}_n$ if and only if $m \geq 0$ if and only if $\|q\|_n \leq 1$.

2.5 BS(1, *n*)

Fix $n = p_1^{n_1} p_2^{n_2} \dots p_k^{n_k}$, and recall that $\mathrm{BS}(1,n) = \langle a, t \mid tat^{-1} = a^n \rangle$. Let

$$\tau \colon \mathrm{BS}(1,n) \to \mathbb{Z}$$

be the homomorphism defined by $a \mapsto 0$ and $t \mapsto 1$. Then there is a short exact sequence

$$0 \to H \to \mathrm{BS}(1,n) \xrightarrow{\tau} \mathbb{Z} \to 0,$$

where $H := \ker(\tau) \cong \mathbb{Z}\big[\frac{1}{n}\big]$. This gives rise to an isomorphism

$$\mathrm{BS}(1,n) \cong \mathbb{Z}\big[\tfrac{1}{n}\big] \rtimes_\alpha \mathbb{Z},$$

where $\alpha(x) = n \cdot x$ for $x \in \mathbb{Z}\big[\frac{1}{n}\big]$. For the rest of this paper we will make the identifications

$$H = \mathbb{Z}\big[\tfrac{1}{n}\big] \quad \text{and} \quad \mathrm{BS}(1,n) = H \rtimes_\alpha \mathbb{Z} = \langle H, t \mid txt^{-1} = \alpha(x) \text{ for } x \in H \rangle.$$

In addition to the standard representation of elements of H as Laurent polynomials in n, we also represent elements by their n–ary expansion; eg $\frac{1}{n} = 0.1$ while

$$n + \frac{1}{n} + \frac{1}{n^4} = 10.1001.$$

We switch between these representations interchangeably.

Given an element $x = \pm x_k x_{k-1} \cdots x_2 x_1 x_0 . x_{-1} x_{-2} \cdots x_{-m} \in H$, we have $0 \leq x_i < n$ for all $-m \leq i \leq k$. We call $-m$ the *leading negative place of* x, which we denote by

$$p(x) = -m.$$

We call x_{-m} the *leading negative term of* x, which we denote by

$$c(x) = x_{-m}.$$

The automorphism α acts on H by multiplication by n, which has the effect of adding one to each index, so the i^{th} term of the image of x is the $(i+1)^{\text{st}}$ term of $\alpha(x)$. For example, $\alpha(21.021311) = 210.21311$ (here we are assuming that $n \geq 4$).

Lemma 2.20 *Let K be a group of the form*

$$K = K' \rtimes_\alpha \mathbb{Z} = \langle K', s \mid sxs^{-1} = \alpha(x) \text{ for } x \in K' \rangle.$$

Suppose $Q \subseteq K'$ is a subset such that $Q \cup \{s^{\pm 1}\}$ is a generating set of K and $\alpha(Q) \subset Q$. Then any element $w \in K$ can be written as

$$w = s^{-r} x_1 \dots x_m s^\ell$$

where $r, \ell \geq 0$, $x_i \in Q$ for all i, and $r + \ell + m = \|w\|_{Q \cup \{s^{\pm 1}\}}$. Moreover, if $w \in K'$ then $r = \ell$.

Proof Write w as a reduced word in $Q \cup \{s^{\pm 1}\}$. By using the relations $sx = \alpha(x)s$ and $xs^{-1} = s^{-1}\alpha(x)$ for $x \in Q$, we may move all copies of s in w to the right and all copies of s^{-1} to the left without increasing the word length of w in $Q \cup \{s^{\pm 1}\}$. The result is an expression of w as a reduced word,

$$w = s^{-r} x_1 \dots x_m s^\ell$$

where $r, \ell \geq 0$ and $x_i \in Q$ for all i. Since the word length of w has not changed, we have $r + \ell + m = \|w\|_{Q \cup \{s^{\pm 1}\}}$. The second statement is clear. □

3 Confining subsets of H

We first describe two particular subsets of $H = \mathbb{Z}\left[\frac{1}{n}\right]$ which are strictly confining under the action of α and α^{-1}, respectively.

Lemma 3.1 *The subset*

$$Q^+ = \{x \in H \mid x = \pm x_k x_{k-1} \dots x_2 x_1 x_0 \text{ for some } k \in \mathbb{N}\} = \mathbb{Z} \subset H \tag{4}$$

is strictly confining under the action of α. The subset

$$Q^- = \{x \in H \mid x = \pm 0.x_{-1} x_{-2} \dots x_{-m} \text{ for some } m \in \mathbb{N}\} \subset H \tag{5}$$

is strictly confining under the action of α^{-1}.

Proof We will verify that Definition 2.3 holds for Q^-; the proof for Q^+ is analogous. We have

$$\alpha^{-1}(Q^-) = \{x \in H \mid x = \pm 0.0 x_{-1} x_{-2} \dots x_{-m} \text{ for some } m \in \mathbb{Z}_{>0}\}.$$

Thus $\alpha^{-1}(Q^-) \subset Q^-$, so (a) holds. Moreover, it is clear that $\bigcup_{n \geq 0} \alpha^n(Q^-) = H$. Indeed, let

$$x = \pm x_k x_{k-1} \dots x_0 . x_{-1} x_{-2} \dots x_{-m}$$

be any element of H. Since $\alpha^{-(k+1)}(x) \in Q^-$, it follows that $x \in \alpha^{k+1}(Q^-)$, and thus (b) holds. Finally, let $x, y \in Q^-$. Then

$$x + y = \pm z_0.z_{-1} \dots z_{-m}$$

where each $z_i \in \{0, \dots, n-1\}$. Hence $\alpha^{-1}(x+y) \in Q^-$ and (c) holds with $k_0 = 1$. Thus Q^- is confining under the action of α^{-1}. To see that Q^- is *strictly* confining, note that $0.1 \in Q \setminus \alpha^{-1}(Q)$. □

The following lemma appears as [3, Lemma 4.9 and Corollary 4.10]; we include a proof here for completeness. Recall that given two (possibly infinite) generating sets S and T of a group G, we say $[S] = [T]$ if $\sup_{g \in S} \|g\|_T < \infty$ and $\sup_{h \in T} \|h\|_S < \infty$.

Lemma 3.2 *Suppose Q is a symmetric subset of H which is confining under the action of α. Let S be a symmetric subset of H such that there exists $K \in \mathbb{Z}_{\geq 0}$ with $\alpha^K(g) \in Q$ for all $g \in S$. Then*

$$\overline{Q} = Q \cup \bigcup_{i \geq 0} \alpha^i(S)$$

is confining under the action of α and

$$[Q \cup \{t^{\pm 1}\}] = [\overline{Q} \cup \{t^{\pm 1}\}].$$

We note that this lemma applies, for example, to all finite symmetric subsets S of H.

Proof First we prove that $\overline{Q}$ is confining under the action of α.

Conditions (a) and (b) of Definition 2.3 are clear (using that $Q \subseteq \overline{Q}$ for condition (b)). To see that condition (c) holds, note that for any $i \geq 0$, and any $g \in S$,

$$\alpha^K(\alpha^i(g)) = \alpha^i(\alpha^K(g)) \in \alpha^i(Q) \subseteq Q.$$

We also have $\alpha^K(g) \in Q$ for any $g \in \overline{Q}$. Hence, if $g, h \in \overline{Q}$ we have $\alpha^K(g) \in Q$ and $\alpha^K(h) \in Q$, and therefore

$$\alpha^{K+k_0}(g+h) = \alpha^{k_0}(\alpha^K(g) + \alpha^K(h)) \in \alpha^{k_0}(Q+Q) \subseteq Q \subseteq \overline{Q},$$

where k_0 is large enough that $\alpha^{k_0}(Q+Q) \subseteq Q$. Therefore (c) holds with constant $K + k_0$.

To see that $[Q \cup \{t^{\pm 1}\}] = [\overline{Q} \cup \{t^{\pm 1}\}]$, note first of all that $[\overline{Q} \cup \{t^{\pm 1}\}] \preccurlyeq [Q \cup \{t^{\pm 1}\}]$. On the other hand, by our above observation, $\overline{Q}$ is really just a finite union,

$$\overline{Q} = Q \cup \bigcup_{i=0}^{K-1} \alpha^i(S).$$

For each i between 0 and $K-1$ and each $g \in S$,

$$\alpha^i(g) = \alpha^{-(K-i)}(\alpha^K(g)) = t^{-(K-i)}\alpha^K(g)t^{(K-i)}$$

and $\alpha^K(g) \in Q$. Hence

$$\|\alpha^i(g)\|_{Q\cup\{t^{\pm 1}\}} \leq 2(K-i)+1 \leq 2K+1.$$

In other words, any element of $\overline{Q} \cup \{t^{\pm 1}\}$ has word length at most $2K+1$ with respect to $Q \cup \{t^{\pm 1}\}$, so $[Q \cup \{t^{\pm 1}\}] \preccurlyeq [\overline{Q} \cup \{t^{\pm 1}\}]$. □

Lemma 3.3 *For any $Q \subseteq H$ which is confining under the action of α, we have $[Q \cup \{t^{\pm 1}\}] \preccurlyeq [Q^+ \cup \{t^{\pm 1}\}]$.*

Proof We show that every element of $Q^+ = \mathbb{Z}$ has bounded word length with respect to $Q \cup \{t^{\pm 1}\}$. First, we apply Lemma 3.2 with $S = \{\pm 1\}$ to pass to $\overline{Q} \supset Q$ such that $\{\pm 1\} \subseteq \overline{Q}$ and $[Q \cup \{t^{\pm 1}\}] = [\overline{Q} \cup \{t^{\pm 1}\}]$.

We begin by showing that every element of $\mathbb{Z}_{>0} = \{1, 2, \dots\}$ has bounded word length with respect to $\overline{Q} \cup \{t^{\pm 1}\}$. Choose k_1 such that $\alpha^{k_1}(\overline{Q} + \overline{Q}) \subseteq \overline{Q}$. We actually initially prove that every element of $\alpha^{k_1}(\mathbb{Z}_{>0}) = \{n^{k_1}, 2n^{k_1}, 3n^{k_1}, \dots\}$ has bounded word length with respect to $\overline{Q} \cup \{t^{\pm 1}\}$. If every such element has word length at most L then every element of $\mathbb{Z}_{>0}$ has word length less than $L + n^{k_1}$ with respect to $\overline{Q} \cup \{t^{\pm 1}\}$ because such an element can be written as

$$an^{k_1} + \underbrace{1 + \cdots + 1}_{<n^{k_1} \text{ times}}, \quad \text{where } a \in \mathbb{Z}_{>0}$$

and $1 \in \overline{Q}$.

The proof of this weaker statement is by induction. First, note that $\alpha^{k_1}(1) = n^{k_1} \in \overline{Q}$. Hence every element of the set

$$\{n^{k_1}, 2n^{k_1}, 3n^{k_1}, \dots, n^{2k_1} = n^{k_1} \cdot n^{k_1}\}$$

has word length at most n^{k_1} with respect to $\overline{Q} \cup \{t^{\pm 1}\}$. Suppose for induction that every element of

$$\{n^{(l-1)k_1}, n^{(l-1)k_1} + n^{k_1}, n^{(l-1)k_1} + 2n^{k_1}, \dots, n^{lk_1}\} \\ = \{an^{k_1} \mid a \in \mathbb{Z}_{>0}\} \cap [n^{(l-1)k_1}, n^{lk_1}]$$

has word length at most n^{k_1} with respect to $\overline{Q} \cup \{t^{\pm 1}\}$. Enumerate the elements of this set as

$$x_0 = n^{(l-1)k_1}, x_1 = n^{(l-1)k_1} + n^{k_1}, \dots, x_s = n^{lk_1}.$$

Consider an element

$$y \in \{n^{lk_1}, n^{lk_1}+n^{k_1}, n^{lk_1}+2n^{k_1}, \dots, n^{(l+1)k_1}\} = \{an^{k_1} \mid a \in \mathbb{Z}_{>0}\} \cap [n^{lk_1}, n^{(l+1)k_1}].$$

Such an element y satisfies

$$n^{k_1} x_j \le y \le n^{k_1} x_{j+1} = n^{k_1}(x_j + n^{k_1}) = n^{k_1} x_j + n^{2k_1}$$

for some j. Hence

$$y = n^{k_1} x_j + a n^{k_1}, \tag{6}$$

where $0 \le a \le n^{k_1}$. Since x_j has word length at most n^{k_1}, we may write

$$x_j = g_1 + \dots + g_{n^{k_1}}$$

where all $g_i \in \overline{Q}$ and $g_i = 0$ for all $i > \|x_j\|_{\overline{Q} \cup \{t^{\pm 1}\}}$. Thus we can rewrite (6) as

$$\begin{aligned} y &= n^{k_1}(g_1 + \ldots + g_{n^{k_1}}) + n^{k_1}(\underbrace{1 + \dots + 1}_{a \le n^{k_1} \text{ times}}) \\ &= n^{k_1}(g_1+1) + n^{k_1}(g_2+1) + \dots + n^{k_1}(g_a+1) + n^{k_1}(g_{a+1}) + \dots + n^{k_1}(g_{n^{k_1}}). \end{aligned}$$

In this last sum, every term is an element of $\overline{Q}$, and there are n^{k_1} terms. Thus $\|y\|_{\overline{Q} \cup \{t^{\pm 1}\}} \le n^{k_1}$. This completes the induction.

We have shown so far that every element of $\mathbb{Z}_{>0}$ has bounded word length with respect to $\overline{Q}$. A completely analogous argument using multiples of $-n^{k_1}$ proves that every element of $\mathbb{Z}_{<0} = \{-1, -2, \dots\}$ has bounded word length with respect to $\overline{Q}$. Hence we have shown

$$[Q \cup \{t^{\pm 1}\}] = [\overline{Q} \cup \{t^{\pm 1}\}] \preccurlyeq [\mathbb{Z} \cup \{t^{\pm 1}\}] = [Q^+ \cup \{t^{\pm 1}\}]. \qquad \square$$

3.1 Subsets confining under the action of α and ideals of $\mathbb{Z}_n$

In this subsection, we describe the connections between subsets of H which are confining under the action of α and ideals of $\mathbb{Z}_n$.

3.1.1 From confining subsets to ideals We begin by describing a way to associate an ideal of $\mathbb{Z}_n$ to a symmetric subset Q of H which is confining under the action of α. We define

$$\mathcal{I}(Q) = \left\{ \dots x_2 x_1 x_0 \in \mathbb{Z}_n \;\middle|\; \begin{array}{l} \text{for any } t \ge 0, \exists\, a \in Q \text{ with } a = a_r \dots a_0.x_t \dots x_0 \\ \text{for some } a_r, \dots, a_0 \in \{0, \dots, n-1\} \end{array} \right\}. \tag{7}$$

That is, an element $\dots x_2x_1x_0$ is in $\mathcal{I}(Q)$ if for any t, there exists a positive element of Q whose fractional part is $0.x_t \dots x_0$. Note in particular that $\mathcal{I}(Q)$ is nonempty for any Q as above. To see this, first notice that Q always contains a positive *integer* $a = a_r \dots a_0$. We may equivalently write

$$a = a_r \dots a_0 . \underbrace{0 \dots 0}_{t \text{ times}}$$

and since t is arbitrary, this shows that 0 $(= \dots 000)$ is in $\mathcal{I}(Q)$.

Lemma 3.4 *The set $\mathcal{I}(Q) \subseteq \mathbb{Z}_n$ is closed.*

Proof Let $x \in \overline{\mathcal{I}(Q)}$ and write $x = \dots x_2x_1x_0$. Then for any $t \geq 0$, there exists $y \in \mathcal{I}(Q)$ with $y = \dots y_2y_1y_0$ and $y_i = x_i$ for $i \leq t$. By the definition of $\mathcal{I}(Q)$ there exists $a \in Q$ with $a = a_r \dots a_0 . y_t \dots y_0$. But then of course we also have $a = a_r \dots a_0 . x_t \dots x_0$. Since t is arbitrary, this implies that $x \in \mathcal{I}(Q)$. □

Lemma 3.5 *In the notation above, $\mathcal{I}(Q)$ is an ideal of $\mathbb{Z}_n$.*

Proof First we show that $\mathcal{I}(Q)$ is closed under addition. Let

$$x = \dots x_2x_1x_0,\ y = \dots y_1y_1y_0 \in \mathcal{I}(Q)$$

and

$$\begin{array}{r} \dots x_2\ x_1\ x_0 \\ + \dots y_2\ y_1\ y_0 \\ \hline \dots z_2\ z_1\ z_0 . \end{array}$$

Let k_0 be large enough that $\alpha^{k_0}(Q+Q) \subseteq Q$. By definition of $\mathcal{I}(Q)$, for any $t \geq 0$ there exist (positive numbers)

$$a = a_r \dots a_0 . x_{t+k_0} x_{t+k_0-1} \dots x_0,$$

$$b = b_s \dots b_0 . y_{t+k_0} y_{t+k_0-1} \dots y_0$$

in Q. We see immediately that $a+b$ is given by

$$c_u \dots c_0 . z_{t+k_0} z_{t+k_0-1} \dots z_0$$

for some $c_u, \dots, c_0 \in \{0, \dots, n-1\}$. This implies that

$$\alpha^{k_0}(a+b) = c_u \dots c_0 z_{t+k_0} z_{t+k_0-1} \dots z_{t+1} . z_t \dots z_0 \in Q.$$

Since t is arbitrary, this implies that $z \in \mathcal{I}(Q)$.

Now we show that $\mathcal{I}(Q)$ is closed under multiplication by elements of $\mathbb{Z}_n$. Let

$$x \in \mathcal{I}(Q) \quad \text{and} \quad p = \dots p_2 p_1 p_0 \in \mathbb{Z}_n.$$

For every $t \geq 0$,

$$p_t \dots p_1 p_0 \cdot x = \underbrace{x + \cdots + x}_{p_t \cdots p_1 p_0 \text{ times}} \in \mathcal{I}(Q)$$

by the above paragraph. Note that $p \cdot x$ is the limit of the sequence

$$\{p_t \dots p_0 \cdot x\}_{t=0}^{\infty} \subseteq \mathcal{I}(Q).$$

But by Lemma 3.4, $\mathcal{I}(Q)$ is closed, so this implies that $p \cdot x \in \mathcal{I}(Q)$ as well. $\square$

3.1.2 From ideals to confining subsets We next describe how to associate a subset of H which is confining under the action of α to an ideal of $\mathbb{Z}_n$. For any ideal $\mathfrak{b}$ of $\mathbb{Z}_n$, let

$$\mathcal{C}(\mathfrak{b}) = \left\{ (-1)^{\delta} x_r \dots x_0.x_{-1} \dots x_{-s} \in H \;\middle|\; \begin{array}{l} \delta \in \{0,1\} \text{ and } \exists\, b \in \mathfrak{b} \text{ with} \\ b = \dots c_2 c_1 x_{-1} \dots x_{-s} \text{ for} \\ \text{some } c_1, c_2, \dots \in \{0, \dots, n-1\} \end{array} \right\}. \tag{8}$$

Thus $\mathcal{C}(\mathfrak{b})$ is the set of elements of H whose fractional parts appear as the tail end of digits of some element of the ideal $\mathfrak{b}$.

Remark 3.6 Since $0 \in \mathfrak{b}$ for any ideal $\mathfrak{b}$, it follows that $\mathcal{C}(\mathfrak{b})$ must contain $\mathbb{Z}$.

Lemma 3.7 *In the notation above, $\mathcal{C}(\mathfrak{b})$ is confining under the action of α.*

Proof We will check the conditions of Definition 2.3.

Let

$$x = x_s \dots x_1 x_0.x_{-1} x_{-2} \dots x_{p(x)} \in \mathcal{C}(\mathfrak{b}).$$

By definition of $\mathcal{C}(\mathfrak{b})$, there is an element

$$b = \dots c_2 c_1 x_{-1} x_{-2} \dots x_{p(x)} \in \mathfrak{b},$$

and so

$$\alpha(x) = x_s \dots x_0 x_{-1}.x_{-2} \dots x_{p(x)} \in \mathcal{C}(\mathfrak{b}).$$

Thus $\alpha(\mathcal{C}(\mathfrak{b})) \subseteq \mathcal{C}(\mathfrak{b})$, and Definition 2.3(a) holds.

Since $\mathbb{Z} \subseteq \mathcal{C}(\mathfrak{b})$ by Remark 3.6, we have $\bigcup_{i=0}^{\infty} \alpha^{-i}(\mathcal{C}(\mathfrak{b})) = H$, and so Definition 2.3(b) holds.

Let $x, y \in \mathcal{C}(\mathfrak{b})$. We first deal with the case that x and y are both positive. We want to show that $x+y \in \mathcal{C}(\mathfrak{b})$. Let $x = x_r \dots x_0.x_{-1} \dots x_{p(x)}$, and $y = y_s \dots y_0.y_{-1} \dots y_{p(y)}$. By adding initial zeros, we may take $r = s$. We also assume without loss of generality that $p(x) \le p(y)$. Then $x + y = z$, where z is given by

$$\begin{array}{r l}
 & x_r \;\dots x_0\,.x_{-1}\;x_{-2}\dots x_{p(y)}\cdots\cdots\cdots x_{p(x)} \\
+ & y_r \;\dots y_0\,.y_{-1}\;y_{-2}\dots y_{p(y)}\;0\;\;\dots 0\;\;0 \\
\hline
 & z_t\;\dots\dots z_0\,.z_{-1}\;z_{-2}\;\dots z_{p(y)}\;\cdots\cdots\cdots z_{p(x)}
\end{array}$$

where here we've assumed without loss of generality that $r \ge s$ (the same argument works if $r < s$).

By the definition of $\mathcal{C}(\mathfrak{b})$, there exist $a, b \in \mathfrak{b}$ with $a = \dots x_{-1}x_{-2} \dots x_{p(x)}$ and $b = \dots y_{-1}y_{-2} \dots y_{p(y)}$. Since $\mathfrak{b}$ is an ideal,

$$n^{p(y)-p(x)}b = \dots y_{-1}y_{-2} \dots y_{p(y)}\ \underbrace{0 \dots\dots\dots 0}_{p(y)-p(x)\text{ times}} \in \mathfrak{b}$$

and

$$a + n^{p(y)-p(x)}b \in \mathfrak{b},$$

where $a + n^{p(y)-p(x)}b$ is given by

$$\begin{array}{r l}
 & \dots\; x_{-1}\;x_{-2}\dots\; x_{p(y)}\;\cdots\cdots\; x_{p(x)} \\
+ & \dots\; y_{-1}\;y_{-2}\dots\; y_{p(y)}\;0\;\;\;\dots\;0 \\
\hline
 & \dots\; z_{-1}\;z_{-2}\;\dots\; z_{p(y)}\;\cdots\cdots\; z_{p(x)}.
\end{array}$$

Therefore, $z = x + y \in \mathcal{C}(\mathfrak{b})$ by the definition of $\mathcal{C}(\mathfrak{b})$.

If $x, y \in \mathcal{C}(\mathfrak{b})$ are both negative, then we show in a completely analogous way that $x + y \in \mathcal{C}(\mathfrak{b})$.

We now consider the case that one of x and y is positive and the other is negative. By possibly multiplying $x + y$ by -1, we assume without loss of generality that $x = x_r \dots x_0.x_{-1} \dots x_{p(x)}$ and $y = -y_s \dots y_0.y_{-1} \dots y_{p(y)}$ with $x \ge |y|$ so that also $r \ge s$. Then $x + y = z$, where z is given by

$$\begin{array}{r l}
 & x_r \dots x_t \dots x_s \dots x_0\,.x_{-1}\;x_{-2}\dots x_{p(y)}\cdots\cdots x_{p(x)} \\
- & \qquad\qquad\quad y_s \dots y_0\,.y_{-1}\;y_{-2}\dots y_{p(y)}\;0\;\;\dots 0 \\
\hline
 & \quad z_t\;\dots z_s\;\dots z_0\,.z_{-1}\;z_{-2}\dots z_{p(y)}\cdots\cdots z_{p(x)}
\end{array}$$

(here we are assuming $p(x) \le p(y)$, but the argument is easily modified if $p(x) > p(y)$). By definition of $\mathcal{C}(\mathfrak{b})$, there exist elements $c, d \in \mathfrak{b}$ with $c = \dots c_2c_1x_{-1} \dots x_{p(x)}$ and

$d = \dots d_2 d_1 y_{-1} \dots y_{p(y)}$. Then $c - n^{p(y)-p(x)} d \in \mathfrak{b}$ is given by

$$\begin{array}{r}
\dots c_1 \;\; x_{-1} \dots x_{p(y)} \dots\dots x_{p(x)} \\
- \dots d_1 \;\; y_{-1} \dots y_{p(y)} \; 0 \;\; \dots 0 \\
\hline
\dots\dots z_{-1} \dots z_{p(y)} \dots\dots z_{p(x)}.
\end{array}$$

Hence we see that $z \in \mathcal{C}(\mathfrak{b})$, as desired.

By the above discussion, Definition 2.3(c) holds with $k_0 = 0$. We conclude that $\mathcal{C}(\mathfrak{b})$ is confining under the action of α. □

Lemma 3.8 *Let $Q \subseteq H$ be confining under the action of α. Then there exists $K > 0$ such that $\alpha^K(\mathbb{Z}) \subseteq Q$.*

Proof By Lemma 3.3, every element of $\mathbb{Z} = Q^+$ has uniformly bounded word length with respect to the generating set $Q \cup \{t^{\pm 1}\}$ of H. Consider an element $w \in \mathbb{Z}$. By Lemma 2.20 we may write w as a reduced word

$$w = t^{-r} x_1 \dots x_m t^r,$$

where $r \geq 0$ and $x_i \in Q$ for all i. This gives us

$$w = \alpha^{-r}(x_1) + \dots + \alpha^{-r}(x_m).$$

Since $\|w\|_{Q \cup \{t^{\pm 1}\}}$ is uniformly bounded, we have both r and m are uniformly bounded, say $r, m \leq R$. Hence

$$\alpha^R(w) = \alpha^{R-r}(x_1) + \dots + \alpha^{R-r}(x_m),$$

and $\alpha^{R-r}(x_i) \in Q$ for each i. Thus, $\alpha^R(w) \in Q^R$, where Q^R represents the words of length at most R in Q. Consequently, $\alpha^{Rk_0}(\alpha^R(w)) \in \alpha^{Rk_0}(Q^R) \subseteq Q$ where the last inclusion follows by Definition 2.3(c). Thus we see that $\alpha^K(\mathbb{Z}) \subseteq Q$, where $K = Rk_0 + R$. □

Lemma 3.9 *Let $Q \subseteq H$ be confining under the action of α. Then there exists $M > 0$ such that $\mathcal{C}(\mathcal{I}(Q)) \subseteq \alpha^{-M}(Q)$.*

Proof Let $a \in \mathcal{C}(\mathcal{I}(Q))$. Since Q and $\mathcal{C}(\mathcal{I}(Q))$ are symmetric, we may suppose that $a = a_r \dots a_0.a_{-1} \dots a_{-s}$ is positive. By definition of $\mathcal{C}(\mathcal{I}(Q))$ there exists an element

$$x = \dots x_2 x_1 a_{-1} \dots a_{-s} \in \mathcal{I}(Q).$$

Then by definition of $\mathcal{I}(Q)$, there exists an element

$$b = b_t \dots b_0.a_{-1} \dots a_{-s} \in Q.$$

We may add an integer c to b to obtain

$$c + b = a_r \dots a_0.a_{-1} \dots a_{-s} = a,$$

and by Lemma 3.8, we have $\alpha^K(c) \in Q$. Thus,

$$\alpha^K(a) = \alpha^K(c+b) = \alpha^K(c) + \alpha^K(b) \in Q + Q.$$

Let k_0 be large enough that $\alpha^{k_0}(Q+Q) \subseteq Q$. Then $\alpha^{K+k_0}(a) = \alpha^{k_0}(\alpha^K(a)) \in Q$ so the result holds with $M = K + k_0$. □

Recall that two ideals $\mathfrak{a}$ and $\mathfrak{b}$ in $\mathbb{Z}_n$ are equivalent (written $\mathfrak{a} \sim \mathfrak{b}$) if there exists a constant k such that $n^k \mathfrak{a} \subseteq \mathfrak{b}$ and $n^k \mathfrak{b} \subseteq \mathfrak{a}$; see Definition 2.15.

Lemma 3.10 *Let $\mathfrak{a}$ and $\mathfrak{b}$ be ideals of $\mathbb{Z}_n$ such that $\mathfrak{a} \sim \mathfrak{b}$. Then $\mathcal{C}(\mathfrak{a}) = \mathcal{C}(\mathfrak{b})$.*

Proof By definition, the ideal $\mathfrak{b}$ determines only the fractional parts of the elements in $\mathcal{C}(\mathfrak{b})$, and if $b = \dots b_2 b_1 b_0 \in \mathfrak{b}$, then there are elements of $\mathcal{C}(\mathfrak{b})$ with fractional part $0.b_k \dots b_0$ for each $k \geq 1$ and arbitrary integral part. From this description, it is clear that for any k, the elements

$$b = \dots b_2 b_1 b_0 \quad \text{and} \quad n^k b = \dots b_2 b_1 b_0 \underbrace{0 \dots 0}_{k \text{ times}}$$

define the same set of fractional parts of elements in $\mathcal{C}(\mathfrak{b})$. Since there exists k such that $n^k \mathfrak{b} \subset \mathfrak{a}$, we see that $\mathcal{C}(\mathfrak{b}) \subseteq \mathcal{C}(\mathfrak{a})$. By a symmetric argument, we also have $\mathcal{C}(\mathfrak{a}) \subseteq \mathcal{C}(\mathfrak{b})$. □

Lemma 3.11 *For any ideal $\mathfrak{a} \subseteq \mathbb{Z}_n$, the confining subset $\mathcal{C}(\mathfrak{a})$ is a subring of H.*

Proof It follows from the proof of Lemma 3.7 that $\mathcal{C}(\mathfrak{a})$ is closed under addition. Moreover, by definition it is closed under additive inverses, and so $\mathcal{C}(\mathfrak{a})$ is an additive subgroup of H. It also contains the multiplicative identity 1 by definition. It remains to be shown that it is closed under multiplication.

For this purpose it will be helpful to write elements of $\mathbb{Z}\left[\frac{1}{n}\right]$ in a slightly different form than their base n expansions. Given any element of $\mathbb{Z}\left[\frac{1}{n}\right]$ we may write it, for *any sufficiently large k*, as $(-1)^\delta(an^{-k} + x)$ where $x \in \mathbb{Z}$, $\delta \in \{\pm 1\}$, and

$$a = a_0 + a_1 n + \dots + a_{k-1} n^{k-1}$$

with each $a_i \in \{0, \dots, n-1\}$. That is, x is the integer part and an^{-k} is the fractional part.

In particular, if $u, v \in \mathcal{C}(\mathfrak{a})$ then we may write

$$u = (-1)^{\delta}(an^{-k} + x) \quad \text{and} \quad v = (-1)^{\epsilon}(bn^{-k} + y),$$

where $\delta, \epsilon \in \{0, 1\}$, $x, y \in \mathbb{Z}$,

$$a = a_0 + a_1 n + \ldots + a_{k-1} n^{k-1}, \quad b = b_0 + b_1 n + \ldots + b_{k-1} n^{k-1},$$

and the digits a_i agree with the first k digits of an element of $\mathfrak{a}$ and similarly for the b_i. This is to say that there are elements

$$a + zn^k, \quad b + wn^k \in \mathfrak{a} \quad \text{where } z, w \in \mathbb{Z}_n.$$

We aim to show that $uv \in \mathcal{C}(\mathfrak{a})$. We have

$$uv = (-1)^{\delta+\epsilon}(ab + ayn^k + bxn^k)n^{-2k} + xy.$$

Thus the fractional part of uv agrees with the first $2k$ digits of the integer

$$ab + ayn^k + bxn^k$$

(note that this integer may have arbitrarily many digits in base n). To show that $uv \in \mathcal{C}(\mathfrak{a})$, it suffices to show that the first $2k$ digits of $ab + ayn^k + bxn^k$ agree with the first $2k$ digits of some element of $\mathfrak{a}$.

To show this last fact we consider the elements $a + zn^k, b + wn^k \in \mathfrak{a}$. Since $\mathfrak{a}$ is an ideal it contains the element

$$(a+zn^k)(b+wn^k) - (a+zn^k)wn^k - (b+wn^k)zn^k + (a+zn^k)yn^k + (b+wn^k)xn^k.$$

Expanding this expression and canceling we have that

$$ab + ayn^k + bxn^k - zwn^{2k} + zyn^{2k} + wxn^{2k} \in \mathfrak{a}.$$

The first $2k$ digits of this element agree with the first $2k$ digits of $ab + ayn^k + bxn^k$. Thus $uv \in \mathcal{C}(\mathfrak{a})$ and the proof is complete. □

We now give a more concrete description of the subring $\mathcal{C}(\mathfrak{a})$, which will be useful in the following subsection. By Lemma 2.17, there is a *full* ideal $\mathfrak{b}$ of $\mathbb{Z}_n$ with $\mathfrak{a} \sim \mathfrak{b}$, and by Lemma 3.10 we have $\mathcal{C}(\mathfrak{a}) = \mathcal{C}(\mathfrak{b})$. Hence to describe $\mathcal{C}(\mathfrak{a})$ explicitly we may assume that $\mathfrak{a}$ itself is full. For ease of notation we may suppose that

$$\mathfrak{a} = \mathbb{Z}_{p_1^{n_1}} \times \cdots \times \mathbb{Z}_{p_r^{n_r}} \times 0 \times \cdots \times 0.$$

Set $l = p_1^{n_1} p_2^{n_2} \cdots p_r^{n_r}$ if $r > 0$ and $l = 1$ if $r = 0$.

Proposition 3.12 *We have $\mathcal{C}(\mathfrak{a}) = \mathbb{Z}\big[\frac{1}{l}\big]$ as a subring of $\mathbb{Z}\big[\frac{1}{n}\big]$.*

Proof First we show that $\mathbb{Z}\big[\frac{1}{l}\big] \subset \mathcal{C}(\mathfrak{a})$. Since $\mathcal{C}(\mathfrak{a})$ is a subring which contains $\mathbb{Z}$, it suffices to show that $\frac{1}{l} \in \mathcal{C}(\mathfrak{a})$. Set $q = p_{r+1}^{n_{r+1}} \cdots p_k^{n_k}$ so that $l = \frac{n}{q}$. Thus we need to show that $\frac{1}{l} = \frac{q}{n} \in \mathcal{C}(\mathfrak{a})$. From this equality we see that $\frac{1}{l}$ is $0.q$ in base n. Hence it suffices to show that $\mathfrak{a}$ contains an element of $\mathbb{Z}_n$ whose ones digit is q when written in base n.

We have the commutative diagram

$$\begin{tikzcd}
x \arrow[dddd, bend right, mapsto] & (q,\dots,q,0,\dots,0) \arrow[l, mapsto] \arrow[dddd, bend left, mapsto] \\
\rotatebox{90}{$\in$} & \rotatebox{90}{$\in$} \\
\mathbb{Z}_n \arrow[r, "\simeq"] \arrow[d, "\varphi_1"] & \mathbb{Z}_{p_1^{n_1}} \times \cdots \times \mathbb{Z}_{p_k^{n_k}} \arrow[d, "\varphi_1 \times \cdots \times \varphi_1"] \\
\mathbb{Z}/n\mathbb{Z} \arrow[r, "\simeq"] & \mathbb{Z}/p_1^{n_1}\mathbb{Z} \times \cdots \times \mathbb{Z}/p_k^{n_k}\mathbb{Z} \\
\rotatebox{90}{$\ni$} & \rotatebox{90}{$\ni$} \\
{[x]_n} & ([q]_{p_1^{n_1}},\dots,[q]_{p_r^{n_r}},0,\dots,0)
\end{tikzcd}$$

where the horizontal maps are isomorphisms, the vertical map $\varphi_1 \colon \mathbb{Z}_n \to \mathbb{Z}/n\mathbb{Z}$ on the left is the "reduction mod n" map defined in (2) which sends $a = \dots a_2 a_1 a_0 \in \mathbb{Z}_n$ to $[a]_n = a_0 \in \mathbb{Z}/n\mathbb{Z}$, and the vertical map on the right is the product of the "reduction mod $p_i^{n_i}$" maps $\varphi_1 \colon \mathbb{Z}_{p_i^{n_i}} \to \mathbb{Z}/p_i^{n_i}\mathbb{Z}$. Consider the unique element x of $\mathbb{Z}_n$ whose image in $\mathbb{Z}_{p_1^{n_1}} \times \cdots \times \mathbb{Z}_{p_k^{n_k}}$ is $(q,\dots,q,0,\dots,0)$, with exactly r nonzero entries. Note that x is necessarily an element of $\mathfrak{a}$; we will show that the ones digit of x is q. Applying φ_1 we obtain the element $[x]_n \in \mathbb{Z}/n\mathbb{Z}$, and applying the product of the maps φ_1 to $(q,\dots,q,0,\dots,0)$ yields the element $([q]_{p_1^{n_1}},\dots,[q]_{p_r^{n_r}},0,\dots,0)$. Since the diagram commutes, $[x]_n$ must be the unique element of $\mathbb{Z}/n\mathbb{Z}$ which maps to this element. As it is clear that $[q]_n \in \mathbb{Z}/n\mathbb{Z}$ also maps to this element, we must have $[x]_n = [q]_n$. Thus the ones digit of x is q, as desired.

Now we show that $\mathcal{C}(\mathfrak{a}) \subset \mathbb{Z}\big[\frac{1}{l}\big]$. Consider an element $(-1)^{\delta} x_r \dots x_0.a_s \dots a_1 a_0 \in \mathcal{C}(\mathfrak{a})$, where $\delta \in \{\pm 1\}$. By definition of $\mathcal{C}(\mathfrak{a})$ there is an element $a = \dots a_s \dots a_1 a_0 \in \mathfrak{a}$. Recall that $\mathfrak{a}$ is identified with $\mathbb{Z}_{p_1^{n_1}} \times \cdots \times \mathbb{Z}_{p_r^{n_r}} \times 0 \times \cdots \times 0$. We consider the analogous commutative diagram as above, but with vertical maps given by (products of) φ_{s+1}. Since $a \in \mathfrak{a}$, the $\mathbb{Z}/p_i^{n_i(s+1)}\mathbb{Z}$–component of the image of $\varphi_{s+1}(a)$ in $\mathbb{Z}/p_1^{n_1(s+1)}\mathbb{Z} \times \cdots \times \mathbb{Z}/p_k^{n_k(s+1)}\mathbb{Z}$ is 0 for each $i > r$. Note that

$$\varphi_{s+1}(a) = \varphi_{s+1}(a_s \dots a_1 a_0).$$

Thus, $a_s \dots a_0$ is divisible by $p_i^{n_i(s+1)}$ for each $i > r$ and therefore it is also divisible by $p_{r+1}^{n_{r+1}(s+1)} \cdots p_k^{n_k(s+1)} = q^{s+1}$. Write $a_s \dots a_0 = q^{s+1} y$ for some $y \in \mathbb{Z}$. We therefore have that

$$(-1)^\delta x_r \dots x_0.a_s \dots a_0 = (-1)^\delta \left(x_r \dots x_0 + \frac{q^{s+1} y}{n^{s+1}} \right) = (-1)^\delta \left(x_r \dots x_0 + \frac{y}{l^{s+1}} \right),$$

and so this element lies in $\mathbb{Z}\left[\frac{1}{l}\right]$. Since the element of $\mathcal{C}(\mathfrak{a})$ we started with was arbitrary, this shows that $\mathcal{C}(\mathfrak{a}) \subset \mathbb{Z}\left[\frac{1}{l}\right]$ as claimed. □

3.1.3 Actions on Bass–Serre trees In this subsection, we give an explicit geometric description of the action of $\mathrm{BS}(1,n)$ on the Cayley graph $\Gamma(\mathrm{BS}(1,n), \mathcal{C}(\mathfrak{a}) \cup \{t^{\pm 1}\})$ for any ideal $\mathfrak{a} \subseteq \mathbb{Z}_n$. We begin by considering a particular ascending HNN extension of $\mathcal{C}(\mathfrak{a})$,

$$G(\mathfrak{a}) = \langle \mathcal{C}(\mathfrak{a}), s \mid sxs^{-1} = \alpha(x) \text{ for } x \in \mathcal{C}(\alpha) \rangle.$$

Lemma 3.13 *For any ideal $\mathfrak{a} \subseteq \mathbb{Z}_n$,*

$$\mathrm{BS}(1,n) \cong G(\mathfrak{a}).$$

Proof By Remark 3.6, $\mathbb{Z} \subseteq \mathcal{C}(\mathfrak{a})$. There is a homomorphism $f \colon G(\mathfrak{a}) \to \mathrm{BS}(1,n)$ defined by

$$x \mapsto x \quad \text{for } x \in \mathcal{C}(\mathfrak{a}), \qquad s \mapsto t,$$

which is surjective because $\mathrm{BS}(1,n)$ is generated by $\mathbb{Z} \subseteq \mathcal{C}(\mathfrak{a})$ and t. We now show that f is injective. Let $g \in \ker(f)$. By Lemma 2.20 we can find an expression of g as a minimal length word in the generating set $\mathcal{C}(\mathfrak{a}) \cup \{s^{\pm 1}\}$ of the form

$$g = s^{-i}(x_1 + \cdots + x_w)s^j,$$

where $i, j \geq 0$ and $x_i \in \mathcal{C}(\mathfrak{a})$. By Lemma 3.11 we may write $x_1 + \cdots + x_w = x \in \mathcal{C}(\mathfrak{a})$, and the result is that $g = s^{-i} x s^j$. As g is in the kernel of f,

$$1 = f(g) = t^{-i} x t^j = \alpha^{-i}(x) t^{j-i}.$$

Since $\alpha^{-i}(x) \in H$, we obtain a contradiction unless $j = i$. In this case we have

$$f(g) = \alpha^{-i}(x) = 0$$

in H, and since α is an automorphism, $x = 0$. But then

$$g = s^{-i} 0 s^i = \alpha^{-i}(0) = 0$$

in $G(\mathfrak{a})$. □

Hence we have a description of $\mathrm{BS}(1,n)$ as an HNN extension over the additive subgroup $\mathcal{C}(\mathfrak{a}) \leq H$, and therefore an action of $\mathrm{BS}(1,n)$ on the standard Bass–Serre tree associated to this HNN extension. Denote this tree by $T(\mathfrak{a})$. For the statement of the next two results recall that *equivalence* of hyperbolic actions means equivalence up to coarsely equivariant quasi-isometry.

Proposition 3.14 *Let G be a group which may be expressed as an ascending HNN extension*

$$A*_A = \langle A, s \mid sas^{-1} = \varphi(a) \text{ for all } a \in A\rangle,$$

where A is a group and φ is an endomorphism of A. Then the action of G on the Bass–Serre tree associated to this HNN extension is equivalent to its action on $\Gamma(G, A \cup \{s^{\pm 1}\})$.

Before turning to the proof, we note one immediate corollary.

Corollary 3.15 *The action of $\mathrm{BS}(1,n)$ on $\Gamma(\mathrm{BS}(1,n), \mathcal{C}(\mathfrak{a}) \cup \{t^{\pm 1}\})$ is equivalent to the action of $\mathrm{BS}(1,n)$ on $T(\mathfrak{a})$.*

Proof of Proposition 3.14 We apply the standard Schwarz–Milnor lemma; see eg [1, Lemma 3.11].

Denote by T the Bass–Serre tree associated to this HNN extension, which may be described as follows. The vertices of T are the left cosets of A in G, and two cosets gA and hA are joined by an edge if

$$gA = hxsA \quad \text{or} \quad gA = hxs^{-1}A \quad \text{for some } x \in A.$$

Consider the vertex $v = A$ and the edge $E = [A, sA]$ containing v. Clearly we have $\bigcup_{g \in G} gE = T$. Hence by the Schwarz–Milnor lemma, the action of G on T is equivalent to the action of G on $\Gamma(G, S)$ where $S = \{g \in G \mid d(v, gv) \le 3\}$. Note that $A \subset S$ since it fixes the vertex v, and thus $[S \cup \{s^{\pm 1}\}] \preccurlyeq [A \cup \{s^{\pm 1}\}]$. We will show that also $[A \cup \{s^{\pm 1}\}] \preccurlyeq [S \cup \{s^{\pm 1}\}]$, which will prove the proposition.

By the description of the vertices of T as cosets of A, any vertex in the radius 3 neighborhood of v has one of the following forms:

- $xs^{\delta}v$ where $x \in A$ and $\delta \in \{\pm 1\}$;
- $x_1 s^{\delta_1} x_2 s^{\delta_2} v$ where $x_i \in A$ and $\delta_i \in \{\pm 1\}$ for $i = 1, 2$; or
- $x_1 s^{\delta_1} x_2 s^{\delta_2} x_3 s^{\delta_3} v$ where $x_i \in A$ and $\delta_i \in \{\pm 1\}$ for $i = 1, 2, 3$.

If $g \in S$, it therefore sends v to a vertex of one of the above three forms. We deal with the last case explicitly, showing that g has bounded word length in the generating set $A \cup \{s^{\pm 1}\}$. The other two cases are entirely analogous.

If $gv = x_1 s^{\delta_1} x_2 s^{\delta_2} x_3 s^{\delta_3} v$ then

$$(x_1 s^{\delta_1} x_2 s^{\delta_2} x_3 s^{\delta_3})^{-1} g \in \mathrm{Stab}_G(v) = A.$$

Hence

$$g = x_1 s^{\delta_1} x_2 s^{\delta_2} x_3 s^{\delta_3} y$$

for some $y \in A$, and this shows that g has word length at most 7 in the generating set $A \cup \{s^{\pm 1}\}$. □

As in the previous subsection, we may assume that $\mathfrak{a}$ is a full ideal, and for ease of notation we may suppose that

$$\mathfrak{a} = \mathbb{Z}_{p_1^{n_1}} \times \cdots \times \mathbb{Z}_{p_r^{n_r}} \times 0 \times \cdots \times 0.$$

We again set $l = p_1^{n_1} p_2^{n_2} \cdots p_r^{n_r}$ and $q = \frac{n}{l} = p_{r+1}^{n_{r+1}} \cdots p_k^{n_k}$. By Proposition 3.12, we have that $\mathcal{C}(\mathfrak{a}) = \mathbb{Z}\bigl[\frac{1}{l}\bigr]$ as a subring of $\mathbb{Z}\bigl[\frac{1}{n}\bigr]$. Thus the Bass–Serre tree $T(\mathfrak{a})$ is the Bass–Serre tree of the ascending HNN extension of $\mathbb{Z}\bigl[\frac{1}{l}\bigr]$, where one map from the edge group to the vertex group is the identity and the other is multiplication by n. Our final goal is to give a concrete geometric description of this tree and the associated action of $\mathrm{BS}(1,n)$.

We first describe an explicit action $\mathrm{BS}(1,n)$ on a tree $T'(\mathfrak{a})$ below. We will then show that $T(\mathfrak{a})$ and $T'(\mathfrak{a})$ are $\mathrm{BS}(1,n)$–equivariantly isomorphic.

Definition 3.16 Let $T'(\mathfrak{a})$ be the tree with the following vertices and edges.

- The vertices of $T'(\mathfrak{a})$ are identified with $\mathbb{Q}_q \times \mathbb{Z}$ up to an equivalence relation $\sim$.
- For pairs (x,h) and (x',h') in $\mathbb{Q}_q \times \mathbb{Z}$, we have $(x,h) \sim (x',h')$ if and only if $h = h'$ and $\|x - x'\|_q \le q^{-h}$, where $\|\cdot\|_q$ denotes the q–adic absolute value on $\mathbb{Q}_q$; see Definition 2.19.
- A vertex represented by $(x,h) \in \mathbb{Q}_q \times \mathbb{Z}$ is joined by an edge to the vertex represented by $(x, h+1)$.

We define an action of $\mathrm{BS}(1,n)$ on $T'(\mathfrak{a})$ as follows:

- If a denotes the normal generator 1 of $\mathbb{Z}\bigl[\frac{1}{n}\bigr] \le \mathrm{BS}(1,n)$, then the action of a on the vertices of $T'(\mathfrak{a})$ is given by
$$a\colon (x,h) \mapsto (x+1,h).$$
- The generator t acts on the vertices of $T'(\mathfrak{a})$ by
$$t\colon (x,h) \mapsto (nx, h+1).$$

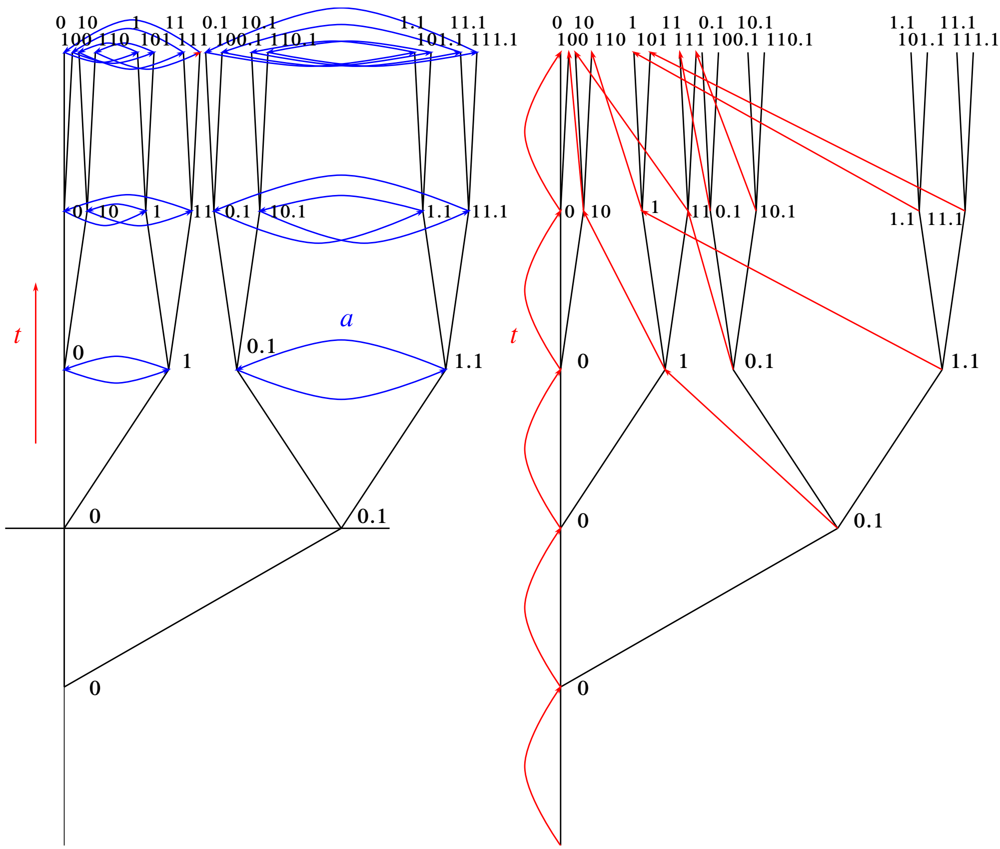


Figure 3: The actions $\mathrm{BS}(1,2) \curvearrowright T'(0)$, left, and $\mathrm{BS}(1,6) \curvearrowright T'(\mathbb{Z}_2 \times 0)$, right.

Here h is meant to indicate a "height" and the equivalence relation reflects the fact that the tree distinguishes between more q–adic numbers, the larger the parameter h is. The reader may check that the graph described above is indeed a tree and that the actions of a and t do indeed define an action of $\mathrm{BS}(1,n)$. In fact, $T'(\mathfrak{a})$ is just the regular $(q+1)$–valent tree. Before explaining why $T'(\mathfrak{a})$ is equivariantly isomorphic to $T(\mathfrak{a})$, we present two examples that the reader may find illustrative.

Example 3.17 First consider the group $\mathrm{BS}(1,2)$. The ring $\mathbb{Z}_2$ has two full ideals: (0) and $\mathbb{Z}_2$. If $\mathfrak{a} = \mathbb{Z}_2$, then $l = 2$ and $q = 1$, so the vertices of $T'(\mathbb{Z}_2)$ are identified with $\{0\} \times \mathbb{Z}$. Hence the action $\mathrm{BS}(1,2) \curvearrowright T'(\mathbb{Z}_2)$ is simply the standard action of $\mathrm{BS}(1,2)$ on the line by translations. If $\mathfrak{a} = (0)$, then $l = 1$ and $q = 2$, so the vertices of $T(0)$ are identified with $\mathbb{Q}_2 \times \mathbb{Z}$ up to the equivalence relation $\sim$. This results in the main Bass–Serre tree of $\mathrm{BS}(1,2)$ with the standard action, as shown in Figure 3, left. In the figure, vertices are labeled by 2–adic numbers. Heights are implicit in the figure, with vertices at the same height in the figure having the same height in $\mathbb{Z}$. The height $h=0$ is illustrated with a dotted line. The generator t acts loxodromically, shifting each vertex directly upward in the figure. For example, consider the vertex $(1.1,1) \in T'(0)$. We have $t(1.1,1)=(2\cdot 1.1, 1+1)=(11,2)$. The generator a acts elliptically, where the action is via a 2–adic odometer $x \mapsto x+1$ on $\mathbb{Q}_2$.

Example 3.18 Now consider the group $\mathrm{BS}(1,6)$. The ring $\mathbb{Z}_6$ has four full ideals: (0), $\mathbb{Z}_2 \times (0)$, $(0)\times\mathbb{Z}_3$, and $\mathbb{Z}_2\times\mathbb{Z}_3$. The trees $T'(0)$ and $T'(\mathbb{Z}_2\times\mathbb{Z}_3)$ are the main Bass–Serre tree and the standard action on the line, respectively, as in the previous example. We will describe the action $\mathrm{BS}(1,6) \curvearrowright T'((0)\times\mathbb{Z}_3)$. The action $\mathrm{BS}(1,6)\curvearrowright T'(\mathbb{Z}_2\times(0))$ may be described in a similar way.

If $\mathfrak{a}=(0)\times\mathbb{Z}_3$, then $l=3$ and $q=2$; hence the vertex set of the tree $T'((0)\times\mathbb{Z}_3)$ is identified with $\mathbb{Q}_2\times\mathbb{Z}$ up to the equivalence relation $\sim$. In particular, the tree $T'((0)\times\mathbb{Z}_3)$ for $\mathrm{BS}(1,6)$ is isomorphic to the main Bass–Serre tree for $\mathrm{BS}(1,2)$. However, the action is different. The generator a is still elliptic, acting as a 2–adic odometer on $\mathbb{Q}_2$, just as it did for $\mathrm{BS}(1,2)$. The generator t is also still loxodromic. However, t has a more complicated action than simply "shifting vertices directly upward". This is because t now acts as multiplication by 6 on $\mathbb{Q}_2$ while simultaneously increasing heights by 1, while in the previous example it acted as multiplication by 2 on $\mathbb{Q}_2$ while increasing heights by 1. For example, consider the same vertex $(1.1,1)$ as in the previous example, but now as a vertex of $T'((0)\times\mathbb{Z}_3)$. Recall that in the previous example, applying $t\in\mathrm{BS}(1,2)$ sent $(1.1,1)$ to $(11,2)$. Here, applying $t\in\mathrm{BS}(1,6)$ yields

$$t(1.1,1)=(6\cdot 1.1, 1+1)=(1001,2)\sim(1,2).$$

The action of t on certain vertices of the tree is illustrated via arrows in Figure 3.

Proposition 3.19 *The trees $T(\mathfrak{a})$ and $T'(\mathfrak{a})$ are $\mathrm{BS}(1,n)$–equivariantly isomorphic.*

Proof This is a standard exercise in Bass–Serre theory. The group $\mathrm{BS}(1,n)$ acts on $T'(\mathfrak{a})$ with a single orbit of vertices and a single orbit of edges. Note that the stabilizer of any vertex is contained in $\mathbb{Z}\left[\frac{1}{n}\right] \trianglelefteq \mathrm{BS}(1,n)$ since t acts loxodromically. Moreover, $\mathbb{Z}\left[\frac{1}{n}\right]$ permutes the vertices at any given height in the tree. For $y\in\mathbb{Z}\left[\frac{1}{n}\right]$ the action on vertices at height h is given explicitly by $(x,h)\mapsto(x+y,h)$. Hence $y\in\mathbb{Z}\left[\frac{1}{n}\right]$ fixes (the equivalence class of) the vertex (x,h) if and only if $\|y\|_q\le q^{-h}$. The additive subgroup of $\mathbb{Z}\left[\frac{1}{n}\right]$ consisting of elements of q–adic absolute value at most q^{-h} is given exactly by $q^h\mathbb{Z}\left[\frac{1}{l}\right]=n^h\mathbb{Z}\left[\frac{1}{l}\right]$. Choosing a vertex at height 0 and an adjacent vertex at height 1, the vertex stabilizers are $\mathbb{Z}\left[\frac{1}{l}\right]$ and $n\mathbb{Z}\left[\frac{1}{l}\right]$, respectively. Thus, the quotient graph of groups for $\mathrm{BS}(1,n)$ has a single vertex group $\mathbb{Z}\left[\frac{1}{l}\right]$. The edge group embeds isomorphically into $\mathbb{Z}\left[\frac{1}{l}\right]$ on one end and embeds as the subgroup $n\mathbb{Z}\left[\frac{1}{l}\right]$ on the other end. This is the same as the graph of groups defining $T(\mathfrak{a})$, so the two trees are equivariantly isomorphic. □

3.2 Subsets confining under the action of α^{-1}

Let Q^- be the subset defined in (5).

Lemma 3.20 *Let $Q \subseteq H$ be confining under the action of α^{-1}. Then $\alpha^{-k}(Q^-) \subseteq Q$ for some $k \geq 0$.*

Proof Let $k_0 \in \mathbb{Z}_{\geq 0}$ be large enough that $\alpha^{-k_0}(Q+Q) \subseteq Q$. We may suppose that $k_0 > 0$, for otherwise Q is a subgroup of H and it is easy to check that in fact $Q = H$. Since $H = \bigcup_{k\in\mathbb{Z}_{\geq 0}} \alpha^k(Q)$, we may also choose k_1 large enough that $\alpha^{-k_1}(a) = an^{-k_1} \in Q$ for any $a \in \{0, \dots, n-1\}$.

We claim that any number of the form

$$\sum_{i=0}^{r} a_i n^{-k_1-(i+1)k_0} \quad \text{with } a_i \in \{0,\dots,n-1\} \text{ for all } 0 \leq i \leq r$$

(for any $r \geq 0$) lies in Q. In other words, any number

$$0.\underbrace{0\dots\dots\dots0}_{k_1+k_0-1 \text{ times}}a_0\underbrace{0\dots\dots0}_{k_0-1 \text{ times}}a_1\underbrace{0\dots\dots0}_{k_0-1 \text{ times}}a_2\underbrace{0\dots\dots0}_{k_0-1 \text{ times}}\dots a_r$$

between 0 and 1 which may be written in base n with

- $k_1 + k_0 - 1$ zeros after the decimal point, and then
- $k_0 - 1$ zeros between any consecutive potentially nonzero digits

lies in Q. We will prove this by induction on r. The base case, when $r = 0$, follows since $a_0 n^{-k_1} \in Q$ for any $a_0 \in \{0,\dots,n-1\}$ and therefore $\alpha^{-k_0}(a_0 n^{-k_1}) = a_0 n^{-k_1-k_0} \in Q$ because Q is closed under the action of α^{-1}. Suppose that the claim is true for all $r < s$. Consider a number $x = \sum_{i=0}^{s} a_i n^{-k_1-(i+1)k_0}$ with $a_i \in \{0,\dots,n-1\}$ for all i. We may write

$$x = a_0 n^{-k_1-k_0} + \sum_{i=1}^{s} a_i n^{-k_1-(i+1)k_0}.$$

We have that $a_0 n^{-k_1} \in Q$ and by induction,

$$\sum_{i=1}^{s} a_i n^{-k_1-ik_0} = \sum_{i=0}^{s-1} a_{i+1} n^{-k_1-(i+1)k_0} \in Q.$$

Hence,

$$x = a_0 n^{-k_1-k_0} + \sum_{i=1}^{s} a_i n^{-k_1-(i+1)k_0} = \alpha^{-k_0}\left(a_0 n^{-k_1} + \sum_{i=1}^{s} a_i n^{-k_1-ik_0}\right) \in Q.$$

This proves the claim.

Now consider a number of the form $x = \sum_{i=0}^{r} a_i n^{-k_1-k_0-i}$. In other words,

$$x = 0.\underbrace{0\ldots\ldots\ldots0}_{k_1+k_0-1 \text{ times}} a_0 a_1 \cdots a_r.$$

We may write

$$x = \sum_{j=0}^{k_0-1} \sum_{\substack{i\in\mathbb{Z}_{\geq 0}\\ j+ik_0\leq r}} a_{j+ik_0} n^{-(k_1+(i+1)k_0+j)} = \sum_{j=0}^{k_0-1} x_j,$$

where

$$x_j = \sum_{\substack{i\in\mathbb{Z}_{\geq 0}\\ j+ik_0\leq r}} a_{j+ik_0} n^{-(k_1+(i+1)k_0+j)}.$$

In other words, we are writing

$$\begin{aligned} x = {} & 0.0\ \ldots 0a_0 00\ldots 00a_{k_0}00\ldots \\ & + 0.0\ \ldots 00a_1 0\ldots 000a_{1+k_0}0\ldots \\ & + 0.0\ \ldots 000a_2\ldots 0000a_{2+k_0}\ldots \\ & \quad\vdots \\ & + 0.0\ \ldots 0000\ \ldots 0a_{k_0-1}000\ldots \end{aligned}$$

with the summands being $x_0, x_1, x_2, \ldots, x_{k_0-1}$, respectively. Notice that

$$\begin{aligned} x_j &= \sum_{\substack{i\in\mathbb{Z}_{\geq 0}\\ j+ik_0\leq r}} a_{j+ik_0} n^{-(k_1+(i+1)k_0+j)} \\ &= \alpha^{-j}\left(\sum_{\substack{i\in\mathbb{Z}_{\geq 0}\\ j+ik_0\leq r}} a_{j+ik_0} n^{-(k_1+(i+1)k_0)}\right) = \alpha^{-j}(y_j), \end{aligned}$$

where

$$y_j = \sum_{\substack{i\in\mathbb{Z}_{\geq 0}\\ j+ik_0\leq r}} a_{j+ik_0} n^{-(k_1+(i+1)k_0)}$$

and, by our claim, $y_j \in Q$ for each $j \in \{0, \ldots, k_0 - 1\}$. Since Q is closed under the action of α^{-1}, we have $x_j \in Q$ for each j. Therefore $x \in Q^{k_0}$. Using that $\alpha^{-k_0}(Q+Q) \subseteq Q$, we have $\alpha^{-k_0^2}(x) = \alpha^{-k_0\cdot k_0}(x) \in Q$.

Finally, for any *positive* number $y = \sum_{i=0}^{r} a_i n^{-i-1} \in Q^-$,

$$\alpha^{-k_1-k_0+1}(y) = \sum_{i=0}^{r} a_i n^{-k_1-k_0-i}.$$

By our work in the last paragraph, this proves that

$$\alpha^{-k_0^2}(\alpha^{-k_1-k_0+1}(y)) \in Q.$$

In other words, $\alpha^{-k_0^2-k_0-k_1+1}(y) \in Q$.

We have shown that $\alpha^{-k}(y) \in Q$ for any *positive* $y \in Q^-$ where $k = k_0^2 + k_0 + k_1 - 1$. Since Q^- and Q are symmetric, this proves that for any *negative* $y \in Q^-$ we have $\alpha^{-k}(y) = -\alpha^{-k}(-y) \in Q$. That is, $\alpha^{-k}(Q^-) \subset Q$ where $k = k_0^2 + k_0 + k_1 - 1$. □

Proposition 3.21 *Let $Q \subseteq H$ be strictly confining under the action of α^{-1}. Then $[Q \cup \{t^{\pm 1}\}] = [Q^- \cup \{t^{\pm 1}\}]$.*

Proof Lemma 3.20, we have $[Q \cup \{t^{\pm 1}\}] \preccurlyeq [Q^- \cup \{t^{\pm 1}\}]$. Suppose that the inequality is strict. We will show then that $[Q \cup \{t^{\pm 1}\}] = [H \cup \{t^{\pm 1}\}]$ and this will contradict that Q is *strictly* confining.

By Lemmas 3.20 and 3.2, we may suppose that $Q^- \subseteq Q$. If there exists k large enough that $\alpha^{-k}(x) \in Q^-$ for all $x \in Q$, then we have $[Q \cup \{t^{\pm 1}\}] = [Q^- \cup \{t^{\pm 1}\}]$, as desired.

Otherwise, there exist numbers $x = a_r \dots a_0.a_{-1} \dots a_{p(x)} \in Q$ with r arbitrarily large and $a_r \neq 0$. We will prove in this case that $n^t \in Q$ for t arbitrarily large. This will complete the proof, for in this case, for any $a \in \{0, \dots, n-1\}$ we will also have $an^t \in Q$ for any t, by the standard arguments. Hence given a positive number $y = \sum_{i=-p}^{q} a_i n^i \in H$, we have $\sum_{i=-p}^{q} a_i n^{i+k_0(p+q+1)} \in Q^{p+q+1}$ and therefore

$$\alpha^{-(p+q+1)k_0}\left(\sum_{i=-p}^{q} a_i n^{i+k_0(p+q+1)}\right) = y \in Q.$$

Since Q is symmetric, this will prove that $H \subseteq Q$.

Consider a number $x = a_r \dots a_0.a_{-1} \dots a_{p(x)} \in Q$ with $r \gg 0$ (how large will be determined later in the proof). We clearly have $n^r \le x < n^{r+1}$. For any $s > 0$,

$$n^s x = a_r \dots a_0 \dots a_{-s}.a_{-s-1} \dots a_{p(x)} \ge n^{r+s} = 1\underbrace{0\dots\dots0}_{r+s \text{ times}}$$

(with the decimal point to the right of $a_{p(x)}$ if $s > p(x)$). We claim that there exists $c \in \{0, 1, \dots, n^s\}$ such that

$$cx = 1\underbrace{0\dots\dots0}_{s-2\text{ times}}b_r b_{r-1}\dots b_0.b_{-1}\dots b_{p(cx)}.$$

If this is not the case, then since $n^s x \geq n^{r+s}$ there exists some $c \in \{0, \dots, n^s\}$ with

$$cx < 1\underbrace{0\dots\dots\dots0}_{r+s-1\text{ times}} \quad \text{but} \quad (c+1)x \geq 1\underbrace{0\dots\dots0}_{s-3\text{ times}}1\underbrace{0\dots\dots0}_{r+1\text{ times}}.$$

From these inequalities we see that

$$x = (c+1)x - cx \geq 1\underbrace{0\dots\dots0}_{r+1\text{ times}} = n^{r+1},$$

which is a contradiction.

So choose $c \in \{0, 1, \dots, n^s\}$ with

$$cx = 1\underbrace{0\dots\dots0}_{s-2\text{ times}}b_r b_{r-1}\dots b_0.b_{-1}\dots b_{p(cx)}.$$

Assuming that $r > n^s k_0$,

$$\alpha^{-ck_0}(cx) = 1\underbrace{0\dots\dots0}_{s-2\text{ times}}b_r b_{r-1}\dots b_{ck_0}.b_{ck_0-1}\dots b_0 b_{-1}\dots b_{p(cx)} \in Q.$$

Since Q is closed under the action of α^{-1}, we then also have

$$1\underbrace{0\dots\dots0}_{s-2\text{ times}}.b_r b_{r-1}\dots b_{p(cx)} \in Q.$$

Since $Q^- \subseteq Q$, we then have

$$10\dots0 = n^{s-2} \in Q + Q$$

and therefore $n^{s-2-k_0} \in Q$. But since s is arbitrarily large, we then have $n^t \in Q$ for t arbitrarily large. □

3.2.1 The action on $\mathbb{H}^2$ There is a well-known action of $\mathrm{BS}(1,n)$ on $\mathbb{H}^2$ defined via the representation

$$\mathrm{BS}(1,n) \to \mathrm{PSL}(2,\mathbb{R}),$$

where

$$a \mapsto \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \quad t \mapsto \begin{pmatrix} n^{1/2} & 0 \\ 0 & n^{-1/2} \end{pmatrix}.$$

The restriction of this representation to $H = \langle t^k a t^{-k} \mid k \in \mathbb{Z} \rangle$ is given by

$$r \mapsto \begin{pmatrix} 1 & r \\ 0 & 1 \end{pmatrix}$$

where we identify H as a subset of $\mathbb{R}$ in the usual way.

Proposition 3.22 *The action of* $\mathrm{BS}(1,n)$ *on* $\mathbb{H}^2$ *is equivalent to the action of* $\mathrm{BS}(1,n)$ *on* $\Gamma(\mathrm{BS}(1,n), Q^- \cup \{t^{\pm 1}\})$.

Proof We will apply the Schwarz–Milnor lemma; see eg [1, Lemma 3.11].

We consider the upper half-plane model of $\mathbb{H}^2$. We show first that the orbit of i under $\mathrm{BS}(1,n)$ is $(\log(n)+1)$–dense in $\mathbb{H}^2$. For this, note first that the orbit of i under $\langle a\rangle$ is 1–dense in the horocycle $\{z \in \mathbb{H}^2 \mid \Im(z) = 1\}$. This follows easily from the fact that $d(i, ai) = d(i, 1+i) < 1$. Now, for any $k \in \mathbb{Z}$, t^k takes the horocycle $\{z \in \mathbb{H}^2 \mid \Im(z) = 1\}$ isometrically to the horocycle $\{z \in \mathbb{H}^2 \mid \Im(z) = n^k\}$. Hence the orbit of i is 1–dense in the horocycle $\{z \in \mathbb{H}^2 \mid \Im(z) = n^k\}$ for any $k \in \mathbb{Z}$. Moreover, for any $k \in \mathbb{Z}$, the distance between the horocycles $\{z \in \mathbb{H}^2 \mid \Im(z) = n^k\}$ and $\{z \in \mathbb{H}^2 \mid \Im(z) = n^{k+1}\}$ is exactly $\log(n)$. Hence, any $z \in \mathbb{H}^2$ has distance at most $\log(n)$ from a horocycle $\{z \in \mathbb{H}^2 \mid \Im(z) = n^k\}$ for some $k \in \mathbb{Z}$. By the triangle inequality, z has distance at most $\log(n) + 1$ from some point in the orbit of i.

This proves that $\bigcup_{g \in \mathrm{BS}(1,n)} gB_{\log(n)+1}(i) = \mathbb{H}^2$. Therefore by [1, Lemma 3.11], the action $\mathrm{BS}(1,n) \curvearrowright \mathbb{H}^2$ is equivalent to the action $\mathrm{BS}(1,n) \curvearrowright \Gamma(\mathrm{BS}(1,n), S)$ where

$$S = \{g \in \mathrm{BS}(1,n) \mid d_{\mathbb{H}^2}(i, gi) \le 2\log(n) + 3\}.$$

We will prove that $[S \cup \{t^{\pm 1}\}] = [Q^- \cup \{t^{\pm 1}\}]$, which will finish the proof.

Let $g \in S$. We may write $g = rt^k$ where $r \in H$ and $k \in \mathbb{Z}$. Observe that

$$d(i, gi) \ge |k| \log(n).$$

Since $g \in S$,

$$|k| \log(n) \le 2\log(n) + 3,$$

and hence

$$|k| \le \frac{2\log(n) + 3}{\log(n)} \le \frac{2\log(2) + 3}{\log(2)} < 7.$$

We have $gi = n^k i + r$, and hence

$$d(i, gi) = 2\operatorname{arcsinh}\left(\frac{1}{2}\sqrt{\frac{r^2 + (n^k - 1)^2}{n^k}}\right) \ge 2\operatorname{arcsinh}\left(\frac{1}{2}\sqrt{\frac{r^2}{n^7}}\right).$$

This defines an upper bound on $|r|$, and hence there exists a uniform $l > 0$ (that is, independent of r) such that $|\alpha^{-l}(r)| < 1$. For such an l, we have $\alpha^{-l}(r) \in Q^-$.

To summarize, we have $g = rt^k$, where $|k| < 7$ and $\alpha^{-l}(r) \in Q^-$. In other words, there exists $s \in Q^-$ such that

$$g = \alpha^l(s)t^k = t^l s t^{-l} t^k.$$

This proves that $\|g\|_{Q^- \cup \{t^{\pm 1}\}} \le 2l + |k| + 1 < 2l + 8$. Thus,

$$[Q^- \cup \{t^{\pm 1}\}] \preccurlyeq [S \cup \{t^{\pm 1}\}].$$

On the other hand, any element $s \in Q^-$ has $d(i, si) < 1 < 2\log(n) + 3$, so we automatically have $s \in S$. So $Q^- \cup \{t^{\pm 1}\} \subseteq S \cup \{t^{\pm 1}\}$ and this proves

$$[S \cup \{t^{\pm 1}\}] \preccurlyeq [Q^- \cup \{t^{\pm 1}\}]. \qquad \square$$

4 Hyperbolic structures of BS$(1, n)$

4.1 Quasiparabolic structures

Lemma 4.1 *The commutator subgroup of* $\mathrm{BS}(1,n)$ *has index* $n-1$ *in* H.

Proof The abelianization of $\mathrm{BS}(1,n)$ is given by the obvious homomorphism

$$\mathrm{BS}(1,n) = \langle a, t \mid tat^{-1} = a^n \rangle \to \langle a, t \mid [a,t] = 1, a = a^n \rangle = \mathbb{Z} \oplus \mathbb{Z}/((n-1)\mathbb{Z}).$$

The kernel of this homomorphism is $[\mathrm{BS}(1,n), \mathrm{BS}(1,n)]$, whereas H is the kernel of the composition

$$\mathrm{BS}(1,n) \to \mathbb{Z} \oplus \mathbb{Z}/((n-1)\mathbb{Z}) \to \mathbb{Z}.$$

The lemma follows easily from this. $\square$

The proof of Proposition 2.6 is a modification to the proofs of [3, Theorems 4.4 and 4.5] and [4, Proposition 4.5]. We recall the statement for the reader's convenience:

Proposition 2.6 *A hyperbolic structure* $[T]$ *is an element of* $\mathcal{H}_{\mathrm{qp}}(\mathrm{BS}(1,n))$ *if and only if there exists a symmetric subset* $Q \subset \mathbb{Z}\left[\frac{1}{n}\right]$ *which is strictly confining under the action of* α *or* α^{-1} *such that* $[T] = [Q \cup \{t^{\pm 1}\}]$.

Proof Given a strictly confining subset Q, a quasiparabolic structure is constructed in [4, Proposition 4.6].

It remains to prove the forward direction. Let $[T] \in \mathcal{H}_{\mathrm{qp}}(\mathrm{BS}(1,n))$. Fix a sequence $\boldsymbol{x} = (x_n)$ in $\Gamma(G,T)$, let $q_{\boldsymbol{x}} : G \to \mathbb{R}$ be the associated quasicharacter, and let $\rho : G \to \mathbb{R}$ be the Busemann pseudocharacter — see Section 2.2 for definitions. Since $\mathrm{BS}(1,n)$ is amenable, ρ is a homomorphism, and $\rho(h) = 0$ for all $h \in H$ by Lemma 4.1. Moreover, we must have $\rho(t) \neq 0$, or else $\rho(g) = 0$ for all $g \in \mathrm{BS}(1,n)$.

Claim 4.2 *There exist constants $r_0, n_0 \geq 0$ such that $Q = \bigcup_{i=1}^{n_0-1} \alpha^i (B(1, r_0) \cap H)$ is confining under the action of α or α^{-1}.*

Proof of claim As $[T] \in \mathcal{H}_{qp}(\mathrm{BS}(1,n))$, the group $\mathrm{BS}(1,n)$ fixes a single point $\xi \in \partial\Gamma(\mathrm{BS}(1,n), T)$. As $\rho(t) \neq 0$, t acts as a hyperbolic isometry of $\Gamma(G,T)$, and thus either t or t^{-1} has ξ as its repelling point. Assume without loss of generality that it is t; we will show that Q is strictly confining under the action of α. On the other hand, if we assume that t^{-1} has ξ as its repelling point, then an analogous proof will show that Q is strictly confining under the action of α^{-1}.

First note that the sequence $(1, t^{-1}, t^{-2}, \dots)$ defines a K–quasigeodesic ray in $\Gamma(G,T)$ for some K, and thus so does the sequence $(g, gt^{-1}, gt^{-2}, \dots)$ for any $g \in \mathrm{BS}(1,n)$. Recall that there is a constant r_0, depending only on the hyperbolicity constant of $\Gamma(G,T)$ and K such that any two K–quasigeodesic rays with the same endpoint on $\partial\Gamma(G,T)$ are eventually r_0–close to each other. In particular, if $\rho(g) \leq C$, then there is a constant $n_0 = n_0(d_T(1,g))$ such that

$$d_T(t^{-n}, gt^{-n}) \leq r_0 + C \tag{9}$$

for all $n \geq n_0$.

Fix $n_0 = n_0(r_0)$, and define

$$Q = \bigcup_{i=0}^{n_0-1} \alpha^i (B(1, r_0) \cap H).$$

Choose r_1 such that $Q \subseteq B(1, r_1)$. Note that such an r_1 exists since for any i and any $h \in B(1, r_0) \cap H$,

$$\|\alpha^i(h)\|_T = \|t^i h t^{-i}\|_T \leq r_0 + 2i\,\|t\|_T.$$

Let $n_1 = n_0(2r_1)$.

For any $h \in B(1, r_0) \cap H$, we have $d_T(1,h) \leq r_0$ and $\rho(h) = 0$, and so it follows from (9) that for all $n \geq n_0$,

$$d_T(\alpha^n(h), 1) = d_T(t^n h t^{-n}, 1) = d_T(ht^{-n}, t^{-n}) \leq r_0.$$

Therefore, $\alpha^n(h) \in B(1, r_0) \cap H$, and thus, for all $n \geq n_0$,

$$\alpha^n(B(1, r_0) \cap H) \subseteq B(1, r_0) \cap H.$$

We now check the conditions of Definition 2.3. Let $h \in Q$, so that $h \in \alpha^i(B(1, r_0) \cap H)$ for some $0 \leq i \leq n_0 - 1$. If $i < n_0 - 1$, then $\alpha(h) \in \alpha^{i+1}(B(1, r_0) \cap H) \subseteq Q$. On the other hand, if $i = n_0 - 1$, then $\alpha(h) \in \alpha^{n_0}(B(1,r_0) \cap H) \subseteq B(1,r_0) \cap H \subseteq Q$. Therefore condition (a) holds.

For any $h \in H$, there is a constant $n_h = n_0(d_T(1,h))$ such that

$$\alpha^{n_h}(h) \in B(1,r_0) \cap H \subseteq Q.$$

Therefore $H = \bigcup_{i=0}^{\infty} \alpha^{-i}(Q)$ and (b) holds.

Finally, $Q + Q \subseteq B(1, 2r_1)$, and so $\alpha^{n_1}(Q+Q) \subseteq B(1,r_0) \cap H \subseteq Q$, and (c) holds with constant n_1.

Therefore, Q is confining under the action of α, concluding the proof of the claim. □

Let $S = Q \cup \{t^{\pm 1}\}$. We will show that the map $\iota\colon (\mathrm{BS}(1,n), d_S) \to (\mathrm{BS}(1,n), d_T)$ is a quasi-isometry. Since $\sup_{s \in S} d_T(1,s) < \infty$, the map ι is Lipschitz. Thus it suffices to show that for any bounded subset $B \subseteq \Gamma(\mathrm{BS}(1,n), T)$, we have $\sup_{b \in B} d_S(1,b) < \infty$. Fix any $M > 0$ and let $B \subseteq \Gamma(\mathrm{BS}(1,n), T)$ be such that $d_T(1,b) \leq M$ for all $b \in B$. For each $b \in B$, we have $b = ht^k$ for some $h \in H$ and some k.

By the definition of $q_{\boldsymbol{x}}$ (see (1)), we have $q_{\boldsymbol{x}}(g) \leq d_T(1,g)$, and since there exists a constant D (the defect of $q_{\boldsymbol{x}}$) such that $|\rho(g) - q_{\boldsymbol{x}}(g)| \leq D$,

$$-d_T(1,g) - D \leq \rho(g) \leq d_T(1,g) + D.$$

Consequently ρ maps bounded subsets of $\Gamma(G,T)$ to bounded subsets of $\mathbb{R}$. Since $\rho(ht^n) = \rho(h) + n\rho(t) = n\rho(t)$, for all n, it follows that there is a constant K such that for any b with $d_T(1,b) = d_T(1,ht^k) \leq M$ we have $k \leq K$. This implies that

$$d_T(1,h) \leq d_T(1,ht^k) + k d_T(1,t) \leq M + K d_T(1,t).$$

Hence, by (9), there is a uniform N such that $\alpha^N(h) \in Q$ and thus $d_S(1,h) \leq 2N+1$. Therefore,

$$d_S(1,b) = d_S(1,ht^k) \leq d_S(1,h) + k \leq 2N + 1 + K.$$

Since the map ι is the identity map on vertices, it is clearly surjective, and therefore it is a quasi-isometry.

Finally, Q is strictly confining under the action of α. Indeed, if Q is confining but not strictly confining under the action of α, then $[Q \cup \{t^{\pm 1}\}] = [T] \in \mathcal{H}_\ell(\mathrm{BS}(1,n))$, which is a contradiction. □

Recall that given an ideal $\mathfrak{a}$ of $\mathbb{Z}_n$, there is an associated subset $\mathcal{C}(\mathfrak{a}) \subseteq H$ defined in (8) which is confining under the action of α. Recall also that $\mathfrak{a} = \mathfrak{a}_1 \times \cdots \times \mathfrak{a}_k$ is *full* if $\mathfrak{a}_j$ is either (0) or $\mathbb{Z}_{p_j}$ for every $j = 1, \ldots, k$; see Definition 2.16.

Lemma 4.3 *Let $\mathfrak{a}$ and $\mathfrak{b}$ be full ideals of $\mathbb{Z}_n$ such that $\mathfrak{a} \not\leq \mathfrak{b}$ and $\mathfrak{b} \not\leq \mathfrak{a}$. Then $[\mathcal{C}(\mathfrak{a}) \cup \{t^{\pm 1}\}]$ and $[\mathcal{C}(\mathfrak{b}) \cup \{t^{\pm 1}\}]$ are incomparable.*

Proof Since $\mathfrak{a} = \mathfrak{a}_1 \times \cdots \times \mathfrak{a}_k$ and $\mathfrak{b} = \mathfrak{b}_1 \times \cdots \times \mathfrak{b}_k$ are full ideals of $\mathbb{Z}_n$, there exist $1 \leq i, j \leq k$ such that $\mathfrak{a}_i = (0)$, $\mathfrak{a}_j = \mathbb{Z}_{p_j^{n_j}}$, $\mathfrak{b}_i = \mathbb{Z}_{p_i^{n_i}}$, and $\mathfrak{b}_j = (0)$.

Consider $\mathcal{C}(\mathfrak{a})$ and $\mathcal{C}(\mathfrak{b})$. For any $x = \pm x_r \ldots x_0 . x_{-1} \ldots x_{p(x)} \in \mathcal{C}(\mathfrak{a})$, there is an element $a = \ldots x_{-1} \ldots x_{p(x)} \in \mathfrak{a}$. Similarly, for any $y = \pm y_m \ldots y_0 . y_{-1} \ldots y_{p(y)} \in \mathcal{C}(\mathfrak{b})$, there is an element $b = \ldots y_{-1} \ldots y_{p(y)} \in \mathfrak{b}$. Since $\mathfrak{a}_i = (0)$,

$$x_{p(x)} \equiv 0 \bmod p_i^{n_i},$$

and, since $\mathfrak{b}_j = (0)$,

$$y_{p(y)} \equiv 0 \bmod p_j^{n_j}$$

by Lemma 2.13.

For any $K \geq 0$, choose $x \in \mathcal{C}(\mathfrak{a})$ such that $p(x) = -K$ and $c(x) \not\equiv 0 \bmod p_j^{n_j}$, which is possible since $\mathfrak{a}_j = \mathbb{Z}_{p_j^{n_j}}$. By Lemma 2.20 we can find an expression of x as a minimal length word in the generating set $\mathcal{C}(\mathfrak{b}) \cup \{t^{\pm 1}\}$ of the form

$$x = t^{-u}(g_1 + \cdots + g_w)t^u.$$

By Lemma 3.11 we may write $g_1 + \cdots + g_w = g \in \mathcal{C}(\mathfrak{b})$, and the result is that $x = \alpha^{-u}(g)$. Now, since $g \in \mathcal{C}(\mathfrak{b})$ and $\mathfrak{b}_j = (0)$, it follows that if $p(g) < 0$, then $c(g) \equiv 0 \bmod p_j^{n_j}$. But this implies that $c(x) = c(\alpha^{-u}(g)) \equiv 0 \bmod p_j^{n_j}$, which is a contradiction. Consequently, we must have $p(g) \geq 0$, which implies that $u \geq K$. Therefore

$$\|x\|_{\mathcal{C}(\mathfrak{b}) \cup \{t^{\pm 1}\}} \geq 2K + 1.$$

Since K was arbitrary, it follows that

$$\sup_{x \in \mathcal{C}(\mathfrak{a}) \cup \{t^{\pm 1}\}} \|x\|_{\mathcal{C}(\mathfrak{b}) \cup \{t^{\pm 1}\}} = \infty,$$

and thus

$$[\mathcal{C}(\mathfrak{a}) \cup \{t^{\pm 1}\}] \not\preccurlyeq [\mathcal{C}(\mathfrak{b}) \cup \{t^{\pm 1}\}].$$

A similar argument shows that

$$[\mathcal{C}(\mathfrak{a}) \cup \{t^{\pm 1}\}] \not\succcurlyeq [\mathcal{C}(\mathfrak{b}) \cup \{t^{\pm 1}\}]. \qquad \square$$

Lemma 4.4 *Let $\mathfrak{a}$ and $\mathfrak{b}$ be full ideals of $\mathbb{Z}_n$, and suppose $\mathfrak{a} \lneq \mathfrak{b}$. Then*

$$[\mathcal{C}(\mathfrak{a}) \cup \{t^{\pm 1}\}] \gnsim [\mathcal{C}(\mathfrak{b}) \cup \{t^{\pm 1}\}].$$

Proof Since $\mathfrak{a} \lneq \mathfrak{b}$, and $\mathfrak{a} = \mathfrak{a}_1 \times \cdots \times \mathfrak{a}_k$ and $\mathfrak{b} = \mathfrak{b}_1 \times \cdots \times \mathfrak{b}_k$ are both full ideals of $\mathbb{Z}_n$, we have $\mathfrak{a} \subsetneq \mathfrak{b}$. Therefore, there exists some $1 \le i \le k$ such that $\mathfrak{a}_i = (0)$ while $\mathfrak{b}_i = \mathbb{Z}_{p_i^{n_i}}$. Consequently, for all $a = \dots a_2 a_1 a_0 \in \mathfrak{a}$, we have $a_0 \equiv 0 \bmod p_i^{n_i}$ by Lemma 2.13.

We first show that $\mathcal{C}(\mathfrak{a}) \subseteq \mathcal{C}(\mathfrak{b})$. Let $x = \pm x_r \dots x_1 x_0 . x_{-1} \dots x_{p(x)} \in \mathcal{C}(\mathfrak{a})$. Then there exists some $a \in \mathfrak{a}$ such that $a = \dots x_{-1} \dots x_{p(x)}$. Since $\mathfrak{a} \subseteq \mathfrak{b}$, we have $a \in \mathfrak{b}$, and so by the definition of $\mathcal{C}(\mathfrak{b})$, it follows that $x \in \mathcal{C}(\mathfrak{b})$. Therefore $C(\mathfrak{a}) \cup \{t^{\pm 1}\} \subseteq \mathcal{C}(\mathfrak{b}) \cup \{t^{\pm 1}\}$, and

$$[\mathcal{C}(\mathfrak{a}) \cup \{t^{\pm 1}\}] \succcurlyeq [\mathcal{C}(\mathfrak{b}) \cup \{t^{\pm 1}\}].$$

Finally, an argument analogous to the proof of Lemma 4.3 shows that

$$[\mathcal{C}(\mathfrak{a}) \cup \{t^{\pm 1}\}] \neq [\mathcal{C}(\mathfrak{b}) \cup \{t^{\pm 1}\}]. \qquad \square$$

Recall that given a subset $Q \subseteq H$ which is confining under the action of α, there is an associated ideal $\mathcal{I}(Q)$ of $\mathbb{Z}_n$ defined by (7).

Lemma 4.5 *For any confining subset Q under the action of α,*

$$[Q \cup \{t^{\pm 1}\}] = [\mathcal{C}(\mathcal{I}(Q)) \cup \{t^{\pm 1}\}].$$

Proof By Lemma 3.9, we have that $\mathcal{C}(\mathcal{I}(Q)) \subseteq \alpha^{-M}(Q)$ for some M. Note that this implies that

$$[Q \cup \{t^{\pm 1}\}] \preccurlyeq [\mathcal{C}(\mathcal{I}(Q)) \cup \{t^{\pm 1}\}].$$

We will show that $[\mathcal{C}(\mathcal{I}(Q)) \cup \{t^{\pm 1}\}] \preccurlyeq [Q \cup \{t^{\pm 1}\}]$. Suppose this is not the case. Then $Q \not\subseteq \alpha^{-k}(\mathcal{C}(\mathcal{I}(Q)))$ for any k and hence there exist elements

$$a = a_r \dots a_0 . a_{-1} \dots a_{-s} \in Q$$

with

$$\inf\{k \mid \alpha^k(a) \in \mathcal{C}(\mathcal{I}(Q))\}$$

arbitrarily large. Let a be an element as above, let

$$\ell = \inf\{k \mid \alpha^k(a) \in \mathcal{C}(\mathcal{I}(Q))\},$$

and assume $\ell > k_0$, where k_0 is large enough that $\alpha^{k_0}(Q+Q) \subseteq Q$. Let $t \le s$ be largest with the property that there does not exist an element of the form

$$\dots a_{-t} \dots a_{-s} \in \mathcal{I}(Q).$$

Note that $t > \ell$, for otherwise we would have

$$\alpha^{\ell-1}(a) = a_r \dots a_0 a_{-1} \dots a_{-\ell+1}.a_{-\ell} \dots a_{-s} \in \mathcal{C}(\mathcal{I}(Q)),$$

contradicting the definition of ℓ as an infimum. We consider two cases:

(1) If $t = s$, then by definition there does not exist an element of $\mathcal{I}(Q)$ with one's digit a_{-s}. In other words, there does not exist an element of the form $\dots a_{-s}$ in $\mathcal{I}(Q)$.

(2) On the other hand suppose that $t < s$. Then by definition of t, there exists an element

$$x = \dots x_2 x_1 x_0 a_{-t-1} \dots a_{-s} \in \mathcal{I}(Q).$$

Let $y = \dots y_2 y_1 y_0 \in \mathcal{I}(Q)$ be the additive inverse of x. That is,

$$\begin{array}{rccc} & \dots x_0 & a_{-t-1} & \dots a_{-s} \\ + & \dots y_{s-t} & y_{s-t-1} & \dots y_0 \\ \hline & \dots 0 & 0 & \dots 0 \end{array}$$

By definition of $\mathcal{I}(Q)$, there exists an element

$$b = b_u \dots b_0.y_{s-1} y_{s-2} \dots y_0 \in Q.$$

We have $c = a + b \in Q + Q$, where $c = c_v \dots c_0.c_{-1} \dots c_{-t}$ is given by

$$\begin{array}{rcccccc} & & a_r \dots a_0. & a_{-1} & \dots a_{-t} & a_{-t-1} & \dots a_{-s} \\ + & b_u \dots & b_r \dots b_0. & y_{s-1} & \dots y_{s-t} & y_{s-t-1} & \dots y_0 \\ \hline c_v \dots & c_u \dots & c_r \dots c_0. & c_{-1} & \dots c_{-t} & 0 & \dots 0 \end{array}$$

(note, we are assuming in the above expression that $u \ge r$, but the case $u < r$ is identical). Therefore

$$\alpha^{k_0}(c) = c_v \dots c_0 c_{-1} \dots c_{-k_0}.c_{-k_0-1} \dots c_{-t} \in Q.$$

Note that there does not exist an element of $\mathcal{I}(Q)$ whose one's digit is c_{-t}. For suppose that $z = \dots z_2 z_1 z_0 c_{-t}$ is such an element. Then we also have

$$n^{s-t} z = \dots z_2 z_1 z_0 c_{-t} \underbrace{0 \dots\dots 0}_{s-t \text{ times}} \in \mathcal{I}(Q)$$

and hence $n^{s-t}z - y \in \mathcal{I}(Q)$ is given by

$$\begin{array}{rcr} \ldots c_{-t} & 0 & \ldots 0 \\ - \ldots y_{s-t} & y_{s-t-1} & \cdots y_0 \\ \hline \ldots a_{-t} & a_{-t-1} & \ldots a_{-s} \end{array}$$

contradicting the definition of t.

Taking $d = a$ in case (1) above or $d = \alpha^{k_0}(c) = c_v \ldots c_{-k_0}.c_{-k_0-1} \ldots c_{-t}$ in case (2) above, we have shown so far that there are elements $d_w \ldots d_0.d_{-1} \ldots d_{-u} \in Q$ with u arbitrarily large (at least $\ell - k_0$) and with the property that there is no element of the form $\ldots d_{-u}$ in $\mathcal{I}(Q)$. That is, there exists a sequence $u_i \to \infty$ and a sequence $\{d^i\}_{i=1}^{\infty} \subset Q$ with $p(d^i) = -u_i$, and $c(d^i) = d^i_{-u_i}$ with the property that there is no element of the form $\ldots d^i_{-u_i}$ in $\mathcal{I}(Q)$. Writing $d^i = d^i_{w_i} \ldots d^i_0.d^i_{-1} \ldots d^i_{-u_i}$ we may pass to a subsequence to assume that the sequence of integers $d^i_{-1} \ldots d^i_{-u_i} \in \mathbb{Z} \subseteq \mathbb{Z}_n$ converges to a number $\ldots e_2 e_1 e_0 \in \mathbb{Z}_n$.

We claim that $\ldots e_2 e_1 e_0 \in \mathcal{I}(Q)$. To prove the claim, note that given any $t \geq 0$ and all large enough i, the number $d^i_{-1} \ldots d^i_{-u_i}$ has

$$d^i_{-u_i} = e_0, \quad d^i_{-u_i+1} = e_1, \quad \ldots, \quad d^i_{-u_i+t} = e_t.$$

We have

$$\begin{aligned} \alpha^{u_i-t-1}(d^i) &= d^i_{w_i} \cdots d^i_0 d^i_{-1} \cdots d^i_{-u_i+t+1}.d^i_{-u_i+t} d^i_{-u_i+t-1} \cdots d^i_{-u_i} \\ &= d^i_{w_i} \cdots d^i_{-u_i+t+1}.e_t e_{t-1} \cdots e_0 \end{aligned}$$

for all such i. Since $\alpha^{u_i-t-1}(d^i) \in Q$ and t is arbitrary, this proves the claim. However, this is a contradiction, as we have $d^i_{-u_i} = e_0$ for all large enough i, while there does not exist an element of the form $\ldots d^i_{-u_i}$ in $\mathcal{I}(Q)$ for any i. This completes the proof. □

Define $\mathcal{J}_n$ to be the poset $2^{\{1,\ldots,k\}} \setminus \{\{1,\ldots,k\}\}$ with the partial order given by inclusion. Recall that $\mathcal{K}_n$ is the poset $2^{\{1,\ldots,k\}} \setminus \{\varnothing\}$. We have that $\mathcal{K}_n$ is isomorphic to the *opposite poset* of $\mathcal{J}_n$.

We now define a map

$$\Phi\colon \mathcal{J}_n \to \mathcal{H}_{\mathrm{qp}}(\mathrm{BS}(1,n))$$

as follows. Given $A \in \mathcal{J}_n$, let $\mathfrak{a} = \mathfrak{a}_1 \times \cdots \times \mathfrak{a}_k$ be the full ideal of $\mathbb{Z}_n$ defined by $\mathfrak{a}_i = (0)$ if and only if $i \in A$, and let

$$\Phi(A) = [\mathcal{C}(\mathfrak{a}) \cup \{t^{\pm 1}\}]. \tag{10}$$

Proposition 4.6 *The map $\Phi\colon \mathcal{J}_n \to \mathcal{H}_{qp}(\mathrm{BS}(1,n))$ defined in (10) is an order-reversing injective map. Hence Φ induces an injective homomorphism of posets*

$$\mathcal{K}_n \to \mathcal{H}_{qp}(\mathrm{BS}(1,n)).$$

Moreover, $H_{qp}(\mathrm{BS}(1,n))$ contains exactly one additional structure which is incomparable to every $[Y] \in \Phi(\mathcal{J}_n)$.

Proof Lemmas 4.3 and 4.4 show that the map Φ is an injective order-reversing map of posets.

By Proposition 2.6, if $[S] \in \mathcal{H}_{qp}(\mathrm{BS}(1,n))$, then there exists a $Q \subseteq H$ which is strictly confining under the action of α or α^{-1} and such that $[S] = [Q \cup \{t^{\pm 1}\}]$.

Fix $[S] \in H_{qp}(\mathrm{BS}(1,n))$ such that the corresponding subset Q is strictly confining under the action of α, and consider the ideal $\mathcal{I}(Q)$. By Lemma 4.5, $[S] = [\mathcal{C}(\mathcal{I}(Q)) \cup \{t^{\pm 1}\}]$. Moreover, by Lemma 2.17, there is a proper full ideal $\mathfrak{a} \sim \mathcal{I}(Q)$, and $\mathcal{C}(\mathcal{I}(Q)) = \mathcal{C}(\mathfrak{a})$ by Lemma 3.10. Thus $[S] = [\mathcal{C}(\mathfrak{a}) \cup \{t^{\pm 1}\}]$. Let $A = \{1 \le i \le k \mid \mathfrak{a}_i = (0)\} \subseteq \mathcal{J}_n$. Then $[S] = \Phi(A)$, and so every quasiparabolic structure whose associated subset is strictly confining under the action of α is in the image of Φ.

By Proposition 3.21, $\mathcal{H}_{qp}(\mathrm{BS}(1,n))$ has a single additional element, $[Q^- \cup \{t^{\pm 1}\}]$, where Q^-, defined in (5), is strictly confining under the action of α^{-1}. It remains to show that $[Q^- \cup \{t^{\pm 1}\}]$ is incomparable to all $[S] \in \mathcal{H}_{qp}(\mathrm{BS}(1,n)) \setminus \{[Q^- \cup \{t^{\pm 1}\}]\}$.

To see this last fact, note that the action $\mathrm{BS}(1,n) \curvearrowright \Gamma(\mathrm{BS}(1,n), Q^- \cup \{t^{\pm 1}\})$ is equivalent to the action $\mathrm{BS}(1,n) \curvearrowright \mathbb{H}^2$ by Proposition 3.22. Hence in this action, the common fixed point of all elements of $\mathrm{BS}(1,n)$ is the *attracting* fixed point of t. On the other hand, for $[S] \in \mathcal{H}_{qp}(\mathrm{BS}(1,n)) \setminus \{[Q^- \cup \{t^{\pm 1}\}]\}$, $\mathrm{BS}(1,n) \curvearrowright \Gamma(\mathrm{BS}(1,n), S)$ is equivalent to the action of $\mathrm{BS}(1,n)$ on one of the Bass–Serre trees described in Section 3.1.3. Hence in the action $\mathrm{BS}(1,n) \curvearrowright \Gamma(\mathrm{BS}(1,n), S)$, the common fixed point of all elements of $\mathrm{BS}(1,n)$ is the *repelling* fixed point of t. If we had, for example, $[Q^- \cup \{t^{\pm 1}\}] \preccurlyeq [S]$ then this would imply that every element of $\mathrm{BS}(1,n)$ would fix the repelling fixed point of t in $\partial\Gamma(\mathrm{BS}(1,n), Q^- \cup \{t^{\pm 1}\})$ as well as the attracting fixed point of t. This would imply that the action $\mathrm{BS}(1,n) \curvearrowright \Gamma(\mathrm{BS}(1,n), Q^- \cup \{t^{\pm 1}\})$ is lineal, which is a contradiction. □

4.2 Proof of Theorem 1.1

Proposition 4.6 gives a complete description of $\mathcal{H}_{qp}(\mathrm{BS}(1,n))$. We now turn our attention to other hyperbolic structures.

We first show that for any $n \geq 2$, $|\mathcal{H}_l(\mathrm{BS}(1,n))| = 1$. Consider an action $\mathrm{BS}(1,n) \curvearrowright X$ with X hyperbolic. Then every element of H must act elliptically or parabolically. For if $g \in H$ then tgt^{-1} and g have the same stable translation length. However, $tgt^{-1} = g^n$ has stable translation length equal to n times the stable translation length of g. This is only possible if the stable translation length of g is 0. Hence the induced action of H on X is either elliptic or parabolic. Since $H \trianglelefteq \mathrm{BS}(1,n)$, if $H \curvearrowright X$ is parabolic, then every element of $\mathrm{BS}(1,n)$ must fix the unique point in ∂X which is fixed by H; hence $\mathrm{BS}(1,n) \curvearrowright X$ is parabolic or quasiparabolic. In particular, if $\mathrm{BS}(1,n) \curvearrowright X$ is a lineal action then $H \curvearrowright X$ is elliptic. This shows that if $[S] \in \mathcal{H}_l(\mathrm{BS}(1,n))$ then $[S] \preccurlyeq [H \cup \{t^{\pm 1}\}]$. But by [1, Theorem 4.22], if $[A] \preccurlyeq [B]$ and both structures are lineal, then $[A] = [B]$. Therefore $[S] = [H \cup \{t^{\pm 1}\}]$.

Every quasiparabolic structure dominates the lineal structure defined by its Busemann quasimorphism. Since $|\mathcal{H}_l(\mathrm{BS}(1,n))| = 1$, it follows that every element of $\mathcal{H}_{\mathrm{qp}}(\mathrm{BS}(1,n))$ dominates this single lineal structure.

For any $n \geq 2$, $\mathrm{BS}(1,n)$ is solvable, and so contains no free subgroups. Thus by the ping-pong lemma, $H_{\mathrm{gt}}(\mathrm{BS}(1,n)) = \varnothing$.

Finally, for any group G, $\mathcal{H}_e(G)$ has a single element which is the smallest element in $\mathcal{H}(G)$, completing the proof of Theorem 1.1.

5 Generating confining subsets

Consider a group $G = H \rtimes \mathbb{Z}$ where $\mathbb{Z} = \langle t \rangle$ acts by $tht^{-1} = \alpha(h)$ for any $h \in H$, where $\alpha \in \mathrm{Aut}(H)$. In this section, we give a general method for constructing confining subsets of such groups.

Let S be a symmetric subset of H with the properties

(i) $\alpha(S) \subseteq S$,

(ii) $\bigcup_{n \geq 0} \alpha^{-n}(S)$ generates H.

Define a subset $Q^i \subseteq H$ by

$$Q^i = Q^i_0 \cup Q^i_1 \cup Q^i_2 \cup \cdots$$

where $Q^i_0 = S$ and

$$Q^i_{n+1} = Q^i_n \cup \alpha^i(Q^i_n \cdot Q^i_n)$$

for all $n \geq 0$. In other words, Q^i is the smallest subset of H containing S and with the property that $\alpha^i(Q^i \cdot Q^i) \subset Q^i$.

Proposition 5.1 *Q^i is a confining subset of H under the action of α.*

Proof First, we prove by induction that $\alpha(Q^i_n) \subseteq Q^i_n$ for all n. The base case $n = 0$ holds by point (i). Suppose for induction that $\alpha(Q^i_n) \subseteq Q^i_n$. Now if $x \in Q^i_{n+1}$ then $x \in Q^i_n$ or $x \in \alpha^i(Q^i_n \cdot Q^i_n)$. In the first case we have $\alpha(x) \in Q^i_n \subseteq Q^i_{n+1}$. Otherwise, we may write $x = \alpha^i(yz)$ where $y, z \in Q^i_n$. Then we have

$$\alpha(x) = \alpha^i(\alpha(y)\alpha(z))$$

and since $\alpha(y), \alpha(z) \in Q^i_n$ we have $\alpha(x) \in \alpha^i(Q^i_n \cdot Q^i_n) \subseteq Q^i_{n+1}$.

Since $\alpha(Q^i_n) \subset Q^i_n$ for all n, we have $\alpha(Q^i) \subset Q^i$.

Now to see that $\alpha^i(Q^i \cdot Q^i) \subseteq Q^i$ simply use the fact that $\alpha^i(Q^i_n \cdot Q^i_n) \subseteq Q^i_{n+1}$ and Q^i is the union of the sets Q^i_n for $n \geq 0$.

Finally, we prove that $H = \bigcup_{n\geq 0} \alpha^{-n}(Q^i)$. We use fact (ii) above and induction on the word length of an element $h \in H$ in the semigroup generating set $\bigcup_{n\geq 0} \alpha^{-n}(S)$. If h has word length one with respect to this generating set then $h = \alpha^{-n}(s)$ for some $s \in S$. Hence, $\alpha^n(h) = s \in Q^i_0$. Suppose for induction that if h has word length at most k, then $\alpha^n(h) \in Q^i$ for some $n \geq 0$. Let h have word length $k + 1$. Then we may write $h = \alpha^{-n}(s)h'$ where $s \in Q^i$ and h' has word length k. By induction there exists m such that $\alpha^m(h') \in Q^i$. We have

$$\alpha^{m+n}(h) = \alpha^m(s)\alpha^{m+n}(h').$$

Since $\alpha(Q^i) \subseteq Q^i$ and $\alpha^m(h') \in Q^i$, we have $\alpha^{m+n}(h') \in Q^i$. Similarly $\alpha^m(s) \in Q^i$. Choose r large enough such that $\alpha^m(s), \alpha^{m+n}(h') \in Q^i_r$. Then we have

$$\alpha^{m+n+i}(h) = \alpha^i(\alpha^m(s)\alpha^{m+n}(h')) \in \alpha^i(Q^i_r \cdot Q^i_r) \subseteq Q^i_{r+1}.$$

This completes the induction. □

By [4, Proposition 4.6], $[Q^i \cup \{t^{\pm 1}\}] \in \mathcal{H}_{\mathrm{qp}}(G) \cup \mathcal{H}_\ell(G)$ for all i. Clearly we have

$$[Q^1 \cup \{t^{\pm 1}\}] \preccurlyeq [Q^2 \cup \{t^{\pm 1}\}] \preccurlyeq [Q^3 \cup \{t^{\pm 1}\}] \preccurlyeq \cdots. \tag{11}$$

In general, there is no reason that any of these inequalities need be strict. However, for certain groups they are *all* strict. In the following proposition, we apply Proposition 5.1 to the lamplighter group $\mathbb{Z} \wr \mathbb{Z} \simeq \mathbb{Z}\left[x, \frac{1}{x}\right] \rtimes \mathbb{Z}$ to demonstrate this phenomenon.

Proposition 5.2 *Let $H = \mathbb{Z}\left[x, \frac{1}{x}\right]$ (the additive group of the Laurent polynomial ring over $\mathbb{Z}$) where t acts on H by $tp(x)t^{-1} = xp(x)$. Set $S = \{\pm 1, \pm x, \pm x^2, \dots\}$. Define Q^i using S as above. Then we have $[Q^i \cup \{t^{\pm 1}\}] \neq [Q^j \cup \{t^{\pm 1}\}]$ for $i < j$. Moreover, $[Q^i \cup \{t^{\pm 1}\}] \in \mathcal{H}_{qp}(H \rtimes \mathbb{Z})$ for each i.*

Before turning to the proof, we note that the existence of a countable chain of quasiparabolic structures for this group follows from [3, Theorem 1.4], which also gives additional information about the structure of $\mathcal{H}_{qp}(\mathbb{Z} \wr \mathbb{Z})$, though it does not give a complete description. In fact, one can see that such a chain exists by fixing any $m \in \mathbb{Z}_{>0}$ and considering the sequence of quotients

$$\mathbb{Z} \wr \mathbb{Z} \twoheadrightarrow \cdots \twoheadrightarrow (\mathbb{Z}/m^3\mathbb{Z}) \wr \mathbb{Z} \twoheadrightarrow (\mathbb{Z}/m^2\mathbb{Z}) \wr \mathbb{Z} \twoheadrightarrow (\mathbb{Z}/m\mathbb{Z}) \wr \mathbb{Z}.$$

By expressing each $(\mathbb{Z}/m^n\mathbb{Z}) \wr \mathbb{Z}$ as an HNN extension, we obtain a quasiparabolic action of $(\mathbb{Z}/m^n\mathbb{Z}) \wr \mathbb{Z}$ on the associated Bass–Serre tree. Since $\mathbb{Z} \wr \mathbb{Z}$ surjects onto $(\mathbb{Z}/m^n\mathbb{Z}) \wr \mathbb{Z}$, we see that $\mathbb{Z} \wr \mathbb{Z}$ acts on this Bass–Serre tree as well. From the sequence of quotients, it is clear that the collection of such actions will form a countable chain in $\mathcal{H}_{qp}(\mathbb{Z} \wr \mathbb{Z})$. This countable chain is equivalent to one constructed by Balasubramanya in [3, Theorem 1.4] associated to the nested subgroups

$$\cdots \leq m^n\mathbb{Z} \leq m^{n-1}\mathbb{Z} \leq \cdots \leq m\mathbb{Z} \leq \mathbb{Z}.$$

In contrast, we do not expect that the countable chain described in Proposition 5.2 corresponds to any chain constructed in [3].

Proof of Proposition 5.2 Note the following, which can be proven inductively:

(i) Q^i_r consists of polynomials in x (with no terms of negative degree).

(ii) If $p(x) \in Q^i_r \setminus Q^i_{r-1}$ then every term of $p(x)$ has degree at least ri.

(iii) The largest coefficient of a term of $p(x) \in Q^i_r$ is 2^r.

Hence we have the following table:

monomial	1	x	x^2	$\dots$	x^{i-1}	x^i	x^{i+1}	$\dots$	x^{2i-1}	x^{2i}	x^{2i+1}	$\dots$	x^{ri-1}	x^{ri}	x^{ri+1}	$\dots$
largest coeff.	1	1	1	$\dots$	1	2	2	$\dots$	2	4	4	$\dots$	2^{r-1}	2^r	2^r	$\dots$

Here the entry under x^k denotes the *largest absolute value* of the coefficient of the degree k term of any polynomial $p(x) \in Q^i$. A similar table holds for Q^j. In particular, we see that if $p(x) \in Q^j$ and $p(x)$ contains a term of degree k, then the absolute value of the coefficient of x^k is at most $2^{\lfloor k/j \rfloor}$.

Note, in particular, that the sequence

$$1, 2x^i = \alpha^i(1+1), 4x^{2i} = \alpha^i(2x^i + 2x^i), 8x^{3i} = \alpha^i(4x^{2i} + 4x^{2i}), \ldots$$

is contained in Q^i. Hence all have word length 1 in the generating set $Q^i \cup \{t\}$. We claim that the word length of $2^r x^{ri}$ in the generating set $Q^j \cup \{t\}$ goes to infinity as $r \to \infty$. Let $\|\cdot\|_j$ denote word length with respect to the generating set $Q^j \cup \{t^{\pm 1}\}$.

To prove the claim, write

$$2^r x^{ri} = g_1 \ldots g_n$$

where $g_1 \ldots g_n$ is a word in $Q^j \cup \{t^{\pm 1}\}$ of minimal length $n = \|2^r x^{ri}\|_j$ representing $2^r x^{ri}$. Each g_i is either t, t^{-1}, or a polynomial $p(x) \in Q^j$. By Lemma 2.20, we may rewrite $2^r x^{ri}$ as a word of length n of the form

$$2^r x^{ri} = t^{-k}(p_1(x) + \cdots + p_m(x))t^l.$$

Since the word on the right represents a polynomial, we must in fact have $k = l$ and we have

$$2^r x^{ri} = \alpha^{-k}(p_1(x) + \cdots + p_m(x)),$$

and therefore

$$p_1(x) + \cdots + p_m(x) = 2^r x^{ri+k}.$$

It follows that $n = 2k + m$. Since each $p_*(x)$ contains x^{ri+k} as a term with coefficient at most $2^{\lfloor (ri+k)/j \rfloor}$, we must have

$$m \geq \frac{2^r}{2^{\lfloor (ri+k)/j \rfloor}} = 2^{r - \lfloor (ri+k)/j \rfloor} \geq 2^{r-(ri+k)/j} = 2^{(1-i/j)r - k/j}.$$

So to bound n from below, it suffices to minimize $2k + m$ subject to the condition

$$m = 2^{(1-i/j)r-k/j}.$$

Rewriting $2k + m$ in terms of k yields

$$2k + 2^{(1-i/j)r-k/j}.$$

Defining a function $f(k) = 2k + 2^{(1-i/j)r-k/j}$, we see that f has a unique minimum at the unique zero of its derivative. The derivative with respect to k is

$$f'(k) = 2 - \frac{1}{j}\ln(2)2^{(1-i/j)r-k/j}.$$

Solving the equation $f'(k) = 0$ for k yields

$$k = (j-i)r - j\log_2\left(\frac{j}{\ln(2)}\right) - j.$$

Since $j-i>0$ we must have $k\to\infty$ as $r\to\infty$ and, in particular, $n\to\infty$ as $r\to\infty$. In other words, $\|2^r x^{ri}\|_j\to\infty$ as $r\to\infty$.

For the final sentence, simply note that each Q^i is strictly confining since $1\notin\alpha(Q^i)$ for any i. $\square$

The argument in the proof of Proposition 5.2 does not work for the wreath product $\mathbb{Z}/n\mathbb{Z}\wr\mathbb{Z}\cong\mathbb{Z}/n\mathbb{Z}\big[x,\frac{1}{x}\big]\rtimes\mathbb{Z}$ since the generator of $\mathbb{Z}/nZ$ doesn't have infinite order. Balasubramanya shows that $\mathbb{Z}/n\mathbb{Z}\wr\mathbb{Z}$ has only finitely many quasiparabolic structures [3, Theorem 1.4]; hence only finitely many of the inequalities in (11) can be strict.

References

[1] **C Abbott**, **S H Balasubramanya**, **D Osin**, *Hyperbolic structures on groups*, Algebr. Geom. Topol. 19 (2019) 1747–1835 MR Zbl

[2] **C R Abbott**, **A J Rasmussen**, *Largest hyperbolic actions and quasi-parabolic actions in groups*, preprint (2019) arXiv 1910.14157

[3] **S H Balasubramanya**, *Hyperbolic structures on wreath products*, J. Group Theory 23 (2020) 357–383 MR Zbl

[4] **P-E Caprace**, **Y Cornulier**, **N Monod**, **R Tessera**, *Amenable hyperbolic groups*, J. Eur. Math. Soc. 17 (2015) 2903–2947 MR Zbl

[5] **B Farb**, **L Mosher**, *A rigidity theorem for the solvable Baumslag–Solitar groups*, Invent. Math. 131 (1998) 419–451 MR Zbl

[6] **M Gromov**, *Hyperbolic groups*, from "Essays in group theory" (S M Gersten, editor), Math. Sci. Res. Inst. Publ. 8, Springer (1987) 75–263 MR Zbl

[7] **D Groves**, **J F Manning**, *Dehn filling in relatively hyperbolic groups*, Israel J. Math. 168 (2008) 317–429 MR Zbl

Department of Mathematics, Brandeis University
Waltham, MA, United States

Department of Mathematics, University of Utah
Salt Lake City, UT, United States

carolynabbott@brandeis.edu, rasmussen@math.utah.edu

Received: 11 December 2019 Revised: 14 April 2021

Algebraic & Geometric Topology **23**:4 (2023) 1693–1724
DOI: 10.2140/agt.2023.23.1693
Published: 14 June 2023

On the cohomology ring of symplectic fillings

ZHENGYI ZHOU

We consider symplectic cohomology twisted by sphere bundles, which can be viewed as an analogue of symplectic cohomology with local systems. Using the associated Gysin exact sequence, we prove the uniqueness of part of the ring structure on cohomology of fillings for those asymptotically dynamically convex manifolds with vanishing property considered by Zhou (Int. Math. Res. Not. 2020 (2020) 9717–9729 and J. Topol. 14 (2021) 112–182). In particular, for any simply connected $4n+1$–dimensional flexibly fillable contact manifold Y, we show that the real cohomology $H^*(W)$ is unique as a ring for any Liouville filling W of Y as long as $c_1(W)=0$. Uniqueness of real homotopy type of Liouville fillings is also obtained for a class of flexibly fillable contact manifolds.

53D40; 57R17

1 Introduction

It is conjectured that Liouville fillings of certain contact manifolds are unique. The first result along this line is that Liouville fillings of the standard contact 3–sphere are unique; see Gromov [13]. The dimension 4 case is special because of the intersection theory of J–holomorphic curves. For higher-dimensional cases only weaker assertions can be made so far. Eliashberg, Floer and McDuff [21] proved that any symplectically aspherical filling of the standard contact sphere of dimension ≥ 5 is diffeomorphic to a ball. Oancea and Viterbo [22] showed that $H_*(Y;\mathbb{Z}) \to H_*(W;\mathbb{Z})$ is surjective for a simply connected subcritically fillable contact manifold Y and any symplectically aspherical W. Barth, Geiges and Zehmisch [4] generalized the Eliashberg–Floer–McDuff theorem to the subcritically fillable case assuming Y is simply connected and of dimension ≥ 5. Roughly speaking, the method used to obtain the above results is finding a "homological foliation", which is hinted at by the splitting result of Cieliebak [8, Theorem 14.16] for subcritical domains.

On the other hand, contact manifolds considered above are asymptotically dynamically convex (ADC) in the sense of Lazarev [18], which is a much larger class of contact manifolds. Those contact manifolds admit only the trivial (contact DGA) augmentation due to degree reasons, hence one may expect that the filling is rigid in some sense. In [29; 30; 31], we studied fillings of such manifolds using Floer theories. Roughly speaking, a contact manifold is ADC if and only if the symplectic field theory (SFT) grading is positive (asymptotically). By the neck-stretching argument, such a condition is sufficient to prove invariance of many structures in [30; 31]. However, in many cases, the SFT gradings are greater than some positive integer k, which provides more room in the neck-stretching argument. The goal of this paper is trying to make use of this extra room and getting more information. In particular, we will study the ring structure of symplectic fillings. Throughout this paper, the default coefficient is $\mathbb{R}$. Our main theorem is the following, where we call a Liouville filling W of Y *topologically simple* if and only if $c_1(W)=0$ and $\pi_1(Y)\to\pi_1(W)$ is injective:

Theorem 1.1 *Let Y be a k–ADC contact manifold* (Definition 2.2) *with a topologically simple Liouville filling W_1 and $SH^*(W_1)=0$. Then for any topologically simple Liouville filling W_2, there is a linear isomorphism $\phi\colon H^*(W_1)\to H^*(W_2)$ preserving grading such that $\phi(\alpha\wedge\beta)=\phi(\alpha)\wedge\phi(\beta)$ for all $\alpha\in H^{2m}(W_1)$ with $2m\le k+1$.*

The main example where Theorem 1.1 can be applied is a flexibly fillable contact manifold Y^{2n-1}, which is $(n-3)$–ADC by Lazarev [18]. In particular, combining with [30, Corollary B], we have:

Corollary 1.2 *Let Y be a simply connected $4n+1$–dimensional flexibly fillable contact manifold with $c_1(Y)=0$. Then the real cohomology ring of Liouville fillings of Y with vanishing first Chern class is unique.*

Remark 1.3 By [30, Corollary B], manifolds considered in Theorem 1.1 have the property that $H^*(W;\mathbb{Z})\to H^*(Y;\mathbb{Z})$ is independent of topologically simple Liouville fillings. Therefore on the degree region where the restriction map is injective, we can infer the ring structure of the filling from the boundary. However, such a method cannot yield information when the product lands in a degree region where the restriction is not injective, for example, the middle degree for flexibly fillable manifolds. The method used in this paper asserts the uniqueness of product structure in those ambiguous regions.

There are also non-Weinstein examples to which Theorem 1.1 can be applied; see Section 2. If Y is subcritically fillable, then the π_1–injective condition can be dropped because Reeb orbits can be assumed to be contractible [18]. However, this case is covered by both [30, Corollary B] and [4, Theorem 1.2] along with the universal coefficient theorem.

In some cases, the knowledge of the cohomology ring is enough to determine the real homotopy type. In particular:

Corollary 1.4 *Let M be the product of $\mathbb{CP}^n$, $\mathbb{HP}^n$, S^{2n} and at most one copy of S^{2n+1} for $n \geq 1$, and let Y denote the contact boundary of the flexible cotangent bundle of M. Then the real homotopy type of a Liouville filling of Y is unique, as long as the Liouville filling has vanishing first Chern class.*

The method used in this paper is very different from the method used in [4; 20; 22], where they studied the moduli spaces of J–holomorphic curves in a partial compactification of W. The essential property needed for the partial compactification is that W splits as $V \times \mathbb{C}$ with V Weinstein. However, for many flexible critical Weinstein domains, such a splitting does not exist even in the topology category, eg flexible cotangent bundles T^*S^{2n} cannot be written as a complex line bundle over a manifold for $n > 1$.[1] Our method is based on symplectic cohomology and uses the index property of the contact boundary, hence we need to assume $c_1 = 0$. The strategy of the proof is to represent the cup product as a multiplication with an Euler class of a sphere bundle. Therefore we consider symplectic cohomology twisted by sphere bundles, which leads to Gysin exact sequences. The Gysin exact sequence associated to a k–sphere bundle uses moduli spaces of dimension up to k. We show that the Gysin exact sequence for a k–sphere bundle on the positive symplectic cohomology is independent of the filling by a neck-stretching argument, if the boundary is k–ADC. Then we can relate it to the regular Gysin sequence of the filling by the vanishing result in [29].

Remark 1.5 The reason for restricting to real coefficients is twofold. Firstly, it is not true that every class in $H^{2k}(M;\mathbb{Z})$ can be represented as the Euler class of an oriented vector bundle (see Walschap [25]) unless multiplied by a large integer (see Guijarro, Schick and Walschap [14]), which only depends on the degree and dimension. Secondly, the Gysin exact sequence is derived from the Morse–Bott framework developed in Zhou [28], which is defined over $\mathbb{R}$. In our case, one can get Gysin exact sequences in

[1] Assume otherwise. Then T^*S^{2n} can be written as a complex line bundle over some manifold V with boundary, and since $H^2(V;\mathbb{Z}) = 0$ when $n > 1$, the complex line bundle is necessarily trivial. Therefore $T^*S^{2n} = V \times \mathbb{D}$, where $\mathbb{D}$ is the unit disk in $\mathbb{C}$. On the other hand, $H^*(T^*S^{2n};\mathbb{Z}) \to H^*(\partial T^*S^{2n};\mathbb{Z})$ is not injective (in degree $2n$) but $H^*(V \times \mathbb{D};\mathbb{Z}) \to H^*(\partial(V \times \mathbb{D});\mathbb{Z})$ is always injective, hence we arrive at a contradiction.

$\mathbb{Z}$–coefficient, since our moduli spaces do not have isotropy or weight. For example, one can generalize the Morse–Bott construction in Hutchings and Nelson [17] to sphere bundles to prove a $\mathbb{Z}$–coefficient Gysin exact sequence.

Remark 1.6 By symplectic cohomology we mean the symplectic cohomology generated by *contractible* orbits. The role of topological simplicity of the filling is to guarantee that the symplectic cohomology of the filling is canonically graded by $\mathbb{Z}$ using any trivialization of $\det\xi$ on Y. From the SFT perspective, it is related to the fact that the augmentation from the filling is (canonically) graded by $\mathbb{Z}$. Since the ADC condition only asserts unique contact DGA augmentation with a $\mathbb{Z}$ grading, we can only hope for uniqueness for topologically simple fillings using the ADC condition; see also [30, Remark 3.6].

Organization of the paper

Section 2 reviews the contact geometry background and provides a list of examples where Theorem 1.1 applies. In Section 3, we define the symplectic cohomology of sphere bundles and prove the independence result when the boundary is k–ADC. We finish the proof of Theorem 1.1, its corollaries and applications in Section 4.

Acknowledgements

Zhou is supported by the National Science Foundation under Grant No. DMS-1638352. It is a great pleasure to acknowledge the Institute for Advanced Study for its warm hospitality. The author would like to thank the referee for many suggestions that improved this paper. This paper is dedicated to the memory of Chenxue.

2 Asymptotically dynamically convex manifolds

Let α be a contact form of (Y^{2n-1},ξ) and $D>0$. We use $\mathcal{P}^{<D}(Y,\alpha)$ to denote the set of contractible Reeb orbits of α with period smaller than D. Letting α_1 and α_2 be two contact forms of (Y,ξ), we write $\alpha_1\geq\alpha_2$ if $\alpha_1=f\alpha_2$ for $f\geq 1$. For a nondegenerate Reeb orbit γ, the degree is defined to be $\mu_{CZ}(\gamma)+n-3$, which is canonically defined in $\mathbb{Z}$ if $c_1(\xi)=0$ and γ is contractible.

Definition 2.1 [18, Definition 3.6] A contact manifold (Y,ξ) with $c_1(\xi)=0$ is asymptotically dynamically convex (ADC) if there exists a nonincreasing sequence of contact forms $\alpha_1 \geq \alpha_2 \geq \alpha_3 \geq \cdots$ for ξ and positive numbers $D_1 < D_2 < D_3 < \cdots$ going to infinity such that all elements of $\mathcal{P}^{<D_i}(Y, \alpha_i)$ are nondegenerate and have positive degree.

One important consequence of ADC is that the positive symplectic cohomology is independent of the filling W whenever $c_1(W) = 0$ and $\pi_1(Y) \to \pi_1(W)$ is injective [18, Proposition 3.8]. Moreover, many Floer theoretic properties of the filling are independent of fillings [30; 31]. Roughly speaking, ADC guarantees the 0–dimensional moduli spaces used in the definition of the positive symplectic cohomology are completely contained in the cylindrical end of the completion $\widehat{W}$, hence are independent of the filling. We consider sphere bundles over (positive) symplectic cohomology. The information for k–sphere bundles is encoded in moduli spaces with dimension up to k. In particular, the associated Gysin exact sequence depends on moduli spaces with dimension up to k. Therefore we need more positivity in the degree of Reeb orbits, so the following finer dynamical convexity is needed:

Definition 2.2 A contact manifold (Y, ξ) with $c_1(\xi) = 0$ is k–ADC if there exists a nonincreasing sequence of contact forms $\alpha_1 \geq \alpha_2 \geq \alpha_3 \geq \cdots$ for ξ and positive numbers $D_1 < D_2 < D_3 < \cdots$ going to infinity such that all elements of $\mathcal{P}^{<D_i}(Y, \alpha_i)$ are nondegenerate and have degree greater than k.

Similarly, we say a Liouville domain (W, λ) with $c_1(W) = 0$ is k–ADC if and only if there exist positive functions $f_1 \geq f_2 \geq \cdots$ and positive numbers $D_1 < D_2 < D_3 < \cdots$ going to infinity such that all contractible (in W) orbits of $(\partial W, f_i \lambda)$ of period up to D_i are nondegenerate and have degree greater than k.

In particular, $(k+1)$–ADC implies k–ADC, and 0–ADC is the usual ADC condition in [18]. The basic example of a k–ADC manifold is the standard contact sphere S^{2n-1}, which is $(2n-3)$–ADC. From this basic example, the following propositions yield many k–ADC manifolds.

Proposition 2.3 [18, Theorems 3.15, 3.17, and 3.18] *Let Y be a $(2n-3-k)$–ADC contact manifold. Then the attachment of an index $k \neq 2$ subcritical or flexible handle to Y^{2n-1} preserves the $(2n-3-k)$–ADC property. When $k = 2$, the same holds if the conditions in* [18, Theorem 3.17] *are met.*

Let V be a manifold with boundary. We define the Morse dimension $\dim_M V$ to be the minimal value of the maximal index of an exhausting Morse function on V.

Proposition 2.4 [30, Theorem 6.3] *Let V^{2n} be a Liouville domain with $c_1(V)=0$. Then $\partial(V\times\mathbb{C})$ is $(2n-1-\dim_M V)$–ADC.*

Proposition 2.5 [30, Theorem 6.19] *Let V and W be p– and q–ADC domains, respectively. Then $\partial(V\times W)$ is k–ADC, where*

$$k=\min\{p+q+4,\ p+\dim W-\dim_M W,\ q+\dim V-\dim_M V\}.$$

Example 2.6 Using the above three propositions, we can apply Theorem 1.1 to classes of contact manifolds:

(i) By Proposition 2.3, for $n\geq 3$, any $2n-1$–dimensional flexibly fillable contact manifold Y with $c_1(Y)=0$ is $(n-3)$–ADC.

(ii) By Proposition 2.4, if V is the $2n$–dimensional Liouville but not Weinstein domain constructed in [19], then $\partial(V\times\mathbb{C}^k)$ is $(2k-2)$–ADC.

(iii) By Proposition 2.5, products of any k–ADC domain for $k>0$ with an example from the above two classes are m–ADC for a suitable $m>0$. We can also attach a flexible handle afterwards.

In general, there are many more k–ADC contact manifolds of interest, eg cotangent bundles and links of terminal singularities. For certain cotangent bundles, symplectic cohomology is zero with an appropriate local system [2]. In general, symplectic cohomology in these cases is not zero, hence they are beyond our scope.

3 Symplectic cohomology and fiber bundles

In this section, we review some basic properties of symplectic cohomology associated to a Liouville domain [9; 23; 24]. Then we introduce the symplectic cohomology of sphere bundles and the associated Gysin exact sequences using the abstract Morse–Bott framework developed in [28].

3.1 Symplectic cohomology

3.1.1 Floer cochain complexes To a Liouville filling (W,λ) of the contact manifold (Y,ξ), one can associate the completion $(\widehat{W},d\hat{\lambda})=(W\cup_Y[1,\infty)_r\times Y,d\hat{\lambda})$, where $\hat{\lambda}=\lambda$ on W and $\hat{\lambda}=r(\lambda|_Y)$ on $[1,\infty)_r\times Y$. Let $H:S^1\times\widehat{W}\to\mathbb{R}$ be a Hamiltonian. Our convention for the Hamiltonian vector field is

$$\omega(\cdot,X_H)=dH.$$

Then symplectic cohomology is defined as the "Morse cohomology" of the symplectic action functional

$$\mathcal{A}_H(x) := -\int x^*\hat{\lambda} + \int H \circ x(t)\, dt, \tag{1}$$

for a Hamiltonian $H = r^2$ for $r \gg 0$ [23; 24]. Equivalently, one may define symplectic cohomology as the direct limit of the Hamiltonian Floer cohomology of $H = Dr$ for $r \gg 0$ as D goes to infinity. For simplicity, we will use the former construction and a special class of Hamiltonian in this paper. Let α be a nondegenerate contact form of the contact manifold (Y, ξ) and R_α its associated Reeb vector field. Then we define

$$\mathcal{S}(Y,\alpha) := \left\{ \int_\gamma \alpha \;\middle|\; \gamma \text{ is the periodic orbit of } R_\alpha \right\}.$$

Following [7], we can choose a smooth family of time-dependent Hamiltonians H_R for $R \in [0,1]$ as a careful perturbation of an autonomous Hamiltonian, such that the following hold:

(i) $H_R|_W$ is time independent C^2–small Morse for $R \neq 0$, and $H_0|_W = 0$.

(ii) There exists a sequence of nonempty open intervals $(a_0, b_0), (a_1, b_1), \dots$ with a_i and b_i converging to infinity and $a_0 = 1$ such that $H_R|_{Y\times(a_i,b_i)} = f_{i,R}(r)$ with $f''_{i,R} > 0$ and $f'_{i,R} \notin \mathcal{S}(Y,\alpha)$, and $\lim_i \min f'_{i,R} = \infty$.

(iii) H_R outside $r = b_0$ does not depend on R.

(iv) For $R \neq 0$, the periodic orbits of X_{H_R} are nondegenerate, and are either critical points of $H_R|_W$ or nonconstant orbits in $\partial W \times [b_i, a_{i+1}]$.

(v) There exist $0 < D_0 < D_1 < \cdots \to \infty$ such that all periodic orbits of X_{H_R} of action greater than $-D_i$ are contained in $W^i := \{r < a_i\}$.

(vi) $\partial_R H_R \leq 0$.

We use $\mathcal{C}(H_R)$ to denote the set of critical points of H_R on W and $\mathcal{P}^*(H)$ to denote the set of nonconstant *contractible* orbits of X_{H_R} outside W, which does not depend on R.

Remark 3.1 A few remarks regarding our choice of Hamiltonian are in order.

(i) We do not define symplectic cohomology of sphere bundles as an invariant, but rather use one model to infer topological information. Therefore we choose to work with one specific Hamiltonian.

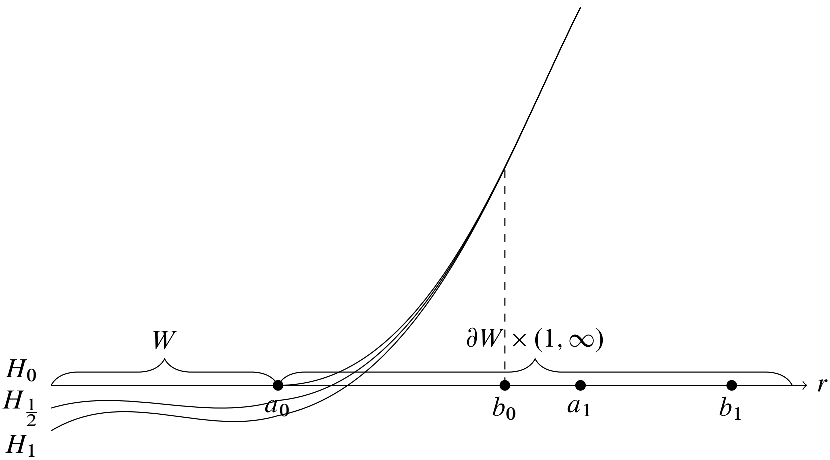


Figure 1: Graphs of H_R.

(ii) The requirement of H_R on interval (a_i, b_i) is for the purpose of the integrated maximum principle [1; 9]. In particular, with an admissible almost complex structure in Definition 3.2, any Floer cylinder asymptotic to orbits in W^i will be completely contained in W^i.

(iii) Ideally, we would like to work with H_0, where the neck-stretching argument will be cleaner. H_0 can be viewed as a "Morse–Bott" situation, which is used in [30]. We will use the nondegenerate Hamiltonian H_R for $R > 0$ to approximate H_0, because the relevant polyfolds are easier to construct and partially exist in the literature; see Remark 3.5.

(iv) The requirement $\partial_R H_R \leq 0$ ensures that the continuation map from H_{R_+} to H_{R_-} respects the action filtration for $R_+ > R_-$. The independence of H_R outside $r = b_0$ simplifies the continuation map for the positive symplectic cohomology to the identity map for different R.

For an admissible Hamiltonian H_R, there are infinitely many periodic orbits and they are not bounded in the r–coordinate. To guarantee the compactness of moduli spaces, we need to use the following almost complex structure so that the integrated maximum principle [1] can be applied.

Definition 3.2 An S^1–dependent almost complex structure J_t is admissible if the following hold:

(i) J_t is compatible with $d\hat{\lambda}$ on $\widehat{W}$.

(ii) J_t is cylindrical convex on $\partial W \times (a_i, b_i)$, that is, $\hat{\lambda} \circ J_t = dr$.

(iii) J_t is only required to be S^1–independent on W. We will often abbreviate J_t by J for simplicity.

The set of admissible almost complex structures is denoted by $\mathcal{J}(W)$.

Let $x, y \in \mathcal{C}(H_R) \cup \mathcal{P}^*(H)$ for $R > 0$ and J be an admissible almost complex structure. We use $\mathcal{M}_{x,y,H_R}$ to denote the compactified moduli space of solutions to the following equation modulo the $\mathbb{R}$–translation:

$$\partial_s u + J(\partial_t u - X_{H_R}) = 0, \qquad \lim_{s \to \infty} u = x, \qquad \lim_{s \to -\infty} u = y. \tag{2}$$

We will suppress H_R when there is no confusion. Then we have the following regularity result:

Proposition 3.3 *For any $R > 0$, there exists a subset $\mathcal{J}^R(W) \subset \mathcal{J}(W)$ of second Baire category such that the following hold*:

(i) *For all $x, y \in \mathcal{C}(H_R) \cup \mathcal{P}^*(H)$ the manifold $\mathcal{M}_{x,y}$ is compact and smooth with boundary and corners.*

(ii) $\partial \mathcal{M}_{x,z} = \bigcup_y \mathcal{M}_{x,y} \times \mathcal{M}_{y,z}$.

(iii) *$\mathcal{M}_{x,z}$ can be oriented so that the induced orientation of $\partial \mathcal{M}_{x,z}$ on $\mathcal{M}_{x,y} \times \mathcal{M}_{y,z}$ is given by the product orientation twisted by $(-1)^{\dim \mathcal{M}_{x,y}}$.*

(iv) *If $x \in \mathcal{C}(H_R)$ and $y \in \mathcal{P}^*(H)$, then $\mathcal{M}_{x,y} = \varnothing$.*

This proposition is folklore, although it is usually stated and proven for moduli spaces $\mathcal{M}_{x,y}$ with virtual dimensional smaller than or equal to 1. Since $\widehat{W}$ is exact and J can depend on $t \in S^1$, we have transversality for unbroken Floer trajectories. A more classical treatment to prove the first two claims is constructing compatible gluing maps for families of Floer trajectories. In the case of Lagrangian Floer theory, such a construction can be found in [3]. In the case of Morse theory, a more elementary approach can be used to give the compactified moduli spaces structures of manifolds with boundary and corners; see [27]. Another method is adopting the polyfold theory developed in [16]. In view of this, we make the following assumption.

Assumption 3.4 For any admissible almost complex structure J, there exists an M-polyfold construction for the symplectic cohomology moduli spaces. More precisely, for every $x, y \in \mathcal{C}(H_R) \cup \mathcal{P}^*(H)$, there exists a strong tame M-polyfold bundle $\mathcal{E}_{x,y} \to \mathcal{B}_{x,y}$ along with an oriented proper sc-Fredholm section $s_{x,y} \colon \mathcal{B}_{x,y} \to \mathcal{E}_{x,y}$ such that the following hold:

(i) $s_{x,y}^{-1}(0) = \mathcal{M}_{x,y}$, where $\mathcal{M}_{x,y}$ is the compact moduli space using J.

(ii) Classical transversality implies that $s_{x,y}$ is transverse and in general position.

(iii) The boundary of $\mathcal{B}_{x,z}$ is the union of products $\mathcal{B}_{x,y} \times \mathcal{B}_{y,z}$, over which the bundle and section have the same splitting.

Remark 3.5 Giving a detailed proof of Assumption 3.4 is not our goal. Symplectic cohomology is a special case of Hamiltonian Floer cohomology, whose polyfold construction was sketched in [26]. An alternative approach is using the full SFT polyfolds [12] as in [11]. In those constructions, the linearization in the polyfold and the linearization of the Floer equation modulo an $\mathbb{R}$–translation are the same. Then we have that classical transversality implies polyfold transversality, ie Assumption 3.4(ii) holds. We only use Assumption 3.4 to prove Proposition 3.3. In particular, we will not use any polyfold perturbation scheme but only the existence of polyfolds.

Proof of Proposition 3.3 To obtain the compactness of moduli spaces, in addition to including Floer breakings, we also need to rule out the possibility of a curve escaping to infinity. To this end, since we choose J to be cylindrical convex on $\partial W \times (a_i, b_i)$ where $H_R = f_{i,R}(r)$, we can apply the integrated maximum principle of Abouzaid and Seidel [1] to any $r \in (a_i, b_i)$; see also [9, Lemma 2.2] for the specific version of the integrated maximum principle we need here. We pick an admissible almost complex structure such that moduli spaces of unbroken Floer trajectories of any virtual dimension are cut out transversely. By Assumption 3.4, we have the M-polyfolds description of compactified moduli spaces as zero sets of sc-Fredholm sections. By Assumption 3.4(ii), those sc-Fredholm sections are cut out transversely. Then the M-polyfold implicit function theorem [16, Theorem 3.15][2] endows the compactified moduli spaces smooth structures of manifolds with boundary and corners. It is worth noting that we only need the existence of M-polyfolds with sc-Fredholm sections without evoking any abstract perturbation scheme. In particular, the first two claims hold. The claim on orientations follows from [28, Section 5.1.1]. If $\mathcal{M}_{x,y} \neq \varnothing$, then for energy reasons we have $\mathcal{A}_{H_R}(y) - \mathcal{A}_{H_R}(x) \geq 0$. Then the last claim follows from property (v) of H_R. □

The Hamiltonian Floer cochain complex is defined by counting the zero-dimensional moduli spaces $\mathcal{M}_{x,y}$. However, since we need to consider sphere bundles over the

[2]This theorem is stated for sections in good position. To obtain a decomposition of the boundary in the form of Proposition 3.3(ii) for sections in general position, one also needs [16, Theorem 4.3].

moduli spaces later, which is naturally a Morse–Bott situation, we need to introduce the Morse–Bott framework developed in [28]. To this purpose, we recall the concept of a *flow category*, which was first introduced in [10].

Definition 3.6 [28, Definition 2.9] A flow category is a small category $\mathcal{C}$ with:

(i) The object space $\mathrm{Obj}_{\mathcal{C}} = \bigsqcup_{i\in\mathbb{Z}} C_i$ is a disjoint union of closed manifolds C_i. The morphism space $\mathrm{Mor}_{\mathcal{C}} = \mathcal{M}$ is a manifold with boundary and corners. The source and target maps $s, t\colon \mathcal{M} \to C$ are smooth.

(ii) Let $\mathcal{M}_{i,j}$ denote $(s\times t)^{-1}(C_i \times C_j)$. Then $\mathcal{M}_{i,i} = C_i$, corresponding to the identity morphisms, and s and t restricted to $\mathcal{M}_{i,i}$ are identities. $\mathcal{M}_{i,j} = \varnothing$ for $j < i$, and $\mathcal{M}_{i,j}$ is a compact manifold with boundary and corners for $j > i$.

(iii) Let $s_{i,j}$ and $t_{i,j}$ denote $s|_{\mathcal{M}_{i,j}}$ and $t|_{\mathcal{M}_{i,j}}$. For every strictly increasing sequence $i_0 < i_1 < \cdots < i_k$,

$$\begin{aligned} t_{i_0,i_1} \times s_{i_1,i_2} \times t_{i_1,i_2} \times \cdots \times s_{i_{k-1},i_k} \colon \mathcal{M}_{i_0,i_1} \times \mathcal{M}_{i_1,i_2} \times \cdots \times \mathcal{M}_{i_{k-1},i_k} \\ \to C_{i_1} \times C_{i_1} \times C_{i_2} \times C_{i_2} \times \cdots \times C_{i_{k-1}} \times C_{i_{k-1}} \end{aligned}$$

is transverse to the submanifold $\Delta_{i_1} \times \cdots \times \Delta_{i_{k-1}}$, where Δ_{i_j} is the diagonal in $C_{i_j} \times C_{i_j}$. Therefore the fiber product

$$\begin{aligned} &\mathcal{M}_{i_0,i_1} \times_{i_1} \mathcal{M}_{i_1,i_2} \times_{i_2} \cdots \times_{i_{k-1}} \mathcal{M}_{i_{k-1},i_k} \\ &\qquad := (t_{i_0,i_1} \times s_{i_1,i_2} \times t_{i_1,i_2} \times \cdots \times s_{i_{k-1},i_k})^{-1} (\Delta_{i_1} \times \Delta_{i_2} \times \cdots \times \Delta_{i_{k-1}}) \\ &\qquad \subset \mathcal{M}_{i_0,i_1} \times \mathcal{M}_{i_1,i_2} \times \cdots \times \mathcal{M}_{i_{k-1},i_k} \end{aligned}$$

is a submanifold.

(iv) The composition $m\colon \mathcal{M}_{i,j} \times_j \mathcal{M}_{j,k} \to \mathcal{M}_{i,k}$ is a smooth map such that

$$m\colon \bigsqcup_{i<j<k} \mathcal{M}_{i,j} \times_j \mathcal{M}_{j,k} \to \partial \mathcal{M}_{i,k}$$

is a diffeomorphism up to zero-measure, ie m is a diffeomorphism from a full measure open subset to a full measure open subset.

In the case of Floer theory considered here, the object space is the set of critical points and the morphism space is the union of all compactified moduli spaces of Floer trajectories in addition to the identity morphisms. The source and target maps are evaluation maps at two ends and the composition is the concatenation of trajectories. The fiber product transversality is tautological, as both source and target maps map to 0–dimensional manifolds. If we label periodic orbits $\mathcal{C}(H_R)\cup\mathcal{P}^*(H)$ by integers so that $\mathcal{A}_{H_R}(x_i)\le\mathcal{A}_{H_R}(x_j)$ if and only if $i\le j$, then we have $\mathcal{M}_{x_i,x_j}=\varnothing$ if $i>j$. Moreover, we can require that x_i is a critical point of $H_R|_W$ if and only if $i\ge 0$. With such labels, Proposition 3.3 gives flow categories $\mathcal{C}^{R,J}$, $\mathcal{C}_0^{R,J}$ and $\mathcal{C}_+^{R,J}$:

$$\begin{aligned}
&\mathrm{Obj}(\mathcal{C}^{R,J}):=\{x_i\}, && \mathrm{Mor}(\mathcal{C}^{R,J}):=\{\mathcal{M}_{i,j}:=\mathcal{M}_{x_i,x_j}\};\\
&\mathrm{Obj}(\mathcal{C}_0^{R,J}):=\{x_i\}_{i\ge 0}, && \mathrm{Mor}(\mathcal{C}_0^{R,J}):=\{\mathcal{M}_{i,j}:=\mathcal{M}_{x_i,x_j}\}_{i,j\ge 0};\\
&\mathrm{Obj}(\mathcal{C}_+^{R,J}):=\{x_i\}_{i<0}, && \mathrm{Mor}(\mathcal{C}_+^{R,J}):=\{\mathcal{M}_{i,j}:=\mathcal{M}_{x_i,x_j}\}_{i,j<0}.
\end{aligned}$$

Moreover, $\mathcal{C}_0^{R,J}$ is a subflow category of $\mathcal{C}^{R,J}$ with quotient flow category $\mathcal{C}_+^{R,J}$ in the sense of [28, Proposition 3.38]. By considering only periodic orbits of action greater than $-D_i$, ie those contained in W^i, we have two subflow categories, $\mathcal{C}_{\le i}^{R,J}\subset\mathcal{C}^{R,J}$ and $\mathcal{C}_{+,\le i}^{R,J}\subset\mathcal{C}_+^{R,J}$. In particular $\mathcal{C}_{\le 0}^{R,J}=\mathcal{C}_0^{R,J}$. The orientation property of Proposition 3.3 implies that $\mathcal{C}^{R,J}$, $\mathcal{C}_0^{R,J}$ and $\mathcal{C}_+^{R,J}$, and the truncated versions $\mathcal{C}_{\le i}^{R,J}$ and $\mathcal{C}_{+,\le i}^{R,J}$ are *oriented flow categories* [28, Definition 2.15]. The main theorem of [28] is that for every oriented flow category $\mathcal{C}=\{C_i,\mathcal{M}_{i,j}\}$, one can associate to it a cochain complex $C^*(\mathcal{C})$ over $\mathbb{R}$ generated by $H^*(C_i;\mathbb{R})$, whose homotopy type is well defined. The one feature of the construction in [28] that we will use is the following.

Proposition 3.7 [28, Corollary 3.13] *Let $\mathcal{C}=\{C_i,\mathcal{M}_{i,j}\}$ be an oriented flow category. Assume $\dim C_i\le k$ for all i. Then the cochain complex $C^*(\mathcal{C})$ only depends on C_i and those $\mathcal{M}_{i,j}$ with $\dim\mathcal{M}_{i,j}\le 2k$.*

Remark 3.8 Roughly speaking, the part of the differential D from $H^*(C_i)$ to $H^*(C_{i+k})$ is defined by the composition $t_*\circ s^*$ through $C_i\xleftarrow{s}\mathcal{M}_{i,i+k}\xrightarrow{t}C_{i+k}$. However, since $\mathcal{M}_{i,i+k}$ is not closed, $t_*\circ s^*$ is not well defined on cohomology. In fact, after choosing representatives of $H^*(C_i)$ in $\Omega^*(C_i)$ (eg harmonic forms), the differential D for a Morse–Bott flow category is given by $t_*\circ s^*$ on $\mathcal{M}_{i,i+k}$, plus many correction terms from possible breakings of $\mathcal{M}_{i,i+k}$. Thus

$$\begin{aligned}\int_{C_{i+k}}D\alpha\wedge\gamma &= \pm\int_{\mathcal{M}_{i,i+k}}s^*\alpha\wedge t^*\gamma\\
&\quad+\lim_{n\to\infty}\sum_{0<j<k}\pm\int_{\mathcal{M}_{i,i+j}\times\mathcal{M}_{i+j,i+k}}s^*\alpha\wedge(t\times s)^*f_{i+j}^n\wedge t^*\gamma+\cdots,\end{aligned}\tag{3}$$

where α and γ are the chosen differential form representatives of elements in $H^*(C_i)$ and $H^*(C_{i+k})$, and f_{i+j}^n is a $\dim C_{i+j}-1$–form on $C_{i+j}\times C_{i+j}$. The suppressed terms are integrations on products $\mathcal{M}_{i,*}\times\cdots\times\mathcal{M}_{*,i+k}$ with more f_*^n inserted; see [28] for details. It is clear that Proposition 3.7 follows from (3). Although (3) only depends on $\mathcal{M}_{i,j}$ with $\dim \mathcal{M}_{i,j} \le 2k$, the proof that $D^2=0$ requires the existence of higher-dimensional ($\dim \le 4k+1$) moduli spaces.

We call a flow category Morse if and only if $\dim C_i = 0$ for all i, and Morse–Bott otherwise. In the Morse case considered in Proposition 3.3, since f_i^n has degree -1 ($f_i^n = 0$), the cochain complex associated to $\mathcal{C}^{R,J}$ is the usual Floer cochain complex generated by $\mathcal{C}(H_R) \cup \mathcal{P}^*(H)$ with differential solely contributed by zero-dimensional moduli spaces

$$\mathrm{D}x_i := \sum_j \Big(\int_{\mathcal{M}_{x_i,x_j}} 1\Big) x_j,$$

that is, we count those moduli spaces $\mathcal{M}_{x_i,x_j}$ of dimension 0. Similarly, we have cochain complexes $C^*(\mathcal{C}_0^{R,J})$ and $C^*(\mathcal{C}_+^{R,J})$, and a tautological short exact sequence of cochain complexes

$$0 \to C^*(\mathcal{C}_0^{R,J}) \to C^*(\mathcal{C}^{R,J}) \to C^*(\mathcal{C}_+^{R,J}) \to 0, \tag{4}$$

as well as the truncated versions. Moreover, we have

$$C^*(\mathcal{C}^{R,J}) = \varinjlim_i C^*(\mathcal{C}_{\le i}^{R,J}) \quad \text{and} \quad C^*(\mathcal{C}_+^{R,J}) = \varinjlim_i C^*(\mathcal{C}_{+,\le i}^{R,J}).$$

Since J is time-independent on W, the cochain complex $C^*(\mathcal{C}_0^{R,J})$ is the Morse cochain complex of W for the Morse–Smale pair $(H_R, g := \omega(\cdot, J\,\cdot))$. Hence we have $H^*(C^*(\mathcal{C}_0^{R,J})) = H^*(W)$. Moreover, $H^*(C^*(\mathcal{C}^{R,J}))$ is the symplectic cohomology $SH^*(W)$, and $H^*(C^*(\mathcal{C}_+^{R,J}))$ is the positive symplectic cohomology $SH_+^*(W)$; see [9; 23; 24] for a more detailed discussion on those invariants. Then (4) gives rise to the tautological long exact sequence

$$\cdots \to H^*(W) \to SH^*(W) \to SH_+^*(W) \to H^{*+1}(W) \to \cdots.$$

Remark 3.9 Since we only consider contractible orbits in domains with vanishing first Chern class, the Conley–Zehnder index is well defined in $\mathbb{Z}$ independent of all choices. Our grading convention follows [23]: $|x_i| := n - \mu_{CZ}(x_i)$, where μ_{CZ} is the Conley–Zehnder index. Such convention implies that if x_i is a critical point of the C^2–small Morse function $H_R|_W$, then $|x_i|$ equals the Morse index. The convention here differs from [24] by n.

3.2 Continuation maps

We will only consider a special class of continuation maps, namely homotopies of almost complex structures and homotopies of Hamiltonians between H_R for different R.

Let $\rho(s)$ be a smooth nondecreasing function such that $\rho(s) = 0$ for $s \ll 0$, and $\rho(s) = 1$ for $s \gg 0$. Given $0 < R_- \le R_+ \le 1$, we have a homotopy of Hamiltonians $H_{R_+,R_-} := H_{\rho(s)R_+ + (1-\rho(s))R_-} \colon \mathbb{R}_s \times S^1 \times \widehat{W} \to \mathbb{R}$. Then we have the following properties for H_{R_+,R_-}:

(i) $H_{R_+,R_-} = H_{R_-}$ for $s \ll 0$ and $H_{R_+,R_-} = H_{R_+}$ for $s \gg 0$.

(ii) $\partial_s H_{R_+,R_-} \le 0$.

(iii) H_{R_+,R_-} outside $r = b_0$ does not depend on s.

Then for $x \in \mathcal{C}(H_{R_+}) \cup \mathcal{P}^*(H)$ and $y \in \mathcal{C}(H_{R_-}) \cup \mathcal{P}^*(H)$, let J_s be a homotopy of admissible almost complex structures. We use $\mathcal{H}_{x,y}$ to denote the compactified moduli space of solutions to

$$\partial_s u + J_s(u - X_{H_{R_+,R_-}}) = 0, \qquad \lim_{s\to\infty} u = x, \qquad \lim_{s\to-\infty} u = y.$$

Then for generic choice of J_s, $\mathcal{H}_{x,y}$ is a manifold with boundary and corners by an analogue of Proposition 3.3. They give rise to a flow morphism in the following sense.

Definition 3.10 [28, Definition 3.18] An oriented flow morphism $\mathfrak{H} \colon \mathcal{C} \Rightarrow \mathcal{D}$ between oriented flow categories $\mathcal{C} := \{C_i, \mathcal{M}^C_{i,j}\}$ and $\mathcal{D} := \{D_i, \mathcal{M}^D_{i,j}\}$ is a family of compact oriented manifolds with boundary and corners $\{\mathcal{H}_{i,j}\}_{i,j\in\mathbb{Z}}$ such that the following hold:

(i) There exists $N \in \mathbb{Z}$ such that when $i - j > N$, we have $\mathcal{H}_{i,j} = \varnothing$.

(ii) There are two smooth maps $s \colon \mathcal{H}_{i,j} \to C_i$ and $t \colon \mathcal{H}_{i,j} \to D_j$.

(iii) For every $i_0 < i_1 < \cdots < i_k$ and $j_0 < \cdots < j_{m-1} < j_m$, the fiber product $\mathcal{M}^C_{i_0,i_1} \times_{i_1} \cdots \times_{i_k} \mathcal{H}_{i_k,j_0} \times_{j_0} \cdots \times_{j_{m-1}} \mathcal{M}^D_{j_{m-1},j_m}$ is cut out transversely.

(iv) There are smooth maps $m_L \colon \mathcal{M}^C_{i,j} \times_j \mathcal{H}_{j,k} \to \mathcal{H}_{i,k}$ and $m_R \colon \mathcal{H}_{i,j} \times_j \mathcal{M}^D_{j,k} \to \mathcal{H}_{i,k}$ such that

$$s \circ m_L(a,b) = s^C(a), \quad t \circ m_L(a,b) = t(b),$$
$$s \circ m_R(a,b) = s(a), \quad t \circ m_R(a,b) = t^D(b),$$

where map s^C is the source map for flow category $\mathcal{C}$ and map t^D is the target map for flow category $\mathcal{D}$.

(v) The map $m_L \cup m_R \colon \bigcup_j (\mathcal{M}^C_{i,j} \times_j \mathcal{H}_{j,k} \cup \mathcal{H}_{i,j} \times_j \mathcal{M}^D_{j,k}) \to \partial\mathcal{H}_{i,k}$ is a diffeomorphism up to zero measure.

(vi) The orientations of the $\mathcal{H}_{i,j}$ are compatible with orientations of C_i, D_i, $\mathcal{M}^C_{i,j}$ and $\mathcal{M}^D_{i,j}$ in the sense of [28, Definition 3.18(6)].

Therefore $\{\mathcal{H}_{x,y}\}$ defines an oriented flow morphism $\mathfrak{H}^{R_+,R_-}$ from $\mathcal{C}^{R_+,J_+}$ to $\mathcal{C}^{R_-,J_-}$. By [28, Theorem 3.21], flow morphisms induce cochain maps between the cochain complexes of the flow categories according to a formula similar to (3). Hence in our situation, $\mathfrak{H}^{R_+,R_-}$ is the geometric data required to define the continuation map. In the Morse case, the cochain map is defined by counting zero-dimensional moduli spaces in $\{\mathcal{H}_{x,y}\}$, which is indeed the classical continuation map. Since we have $\partial_s H_{R_+,R_-} \le 0$, then if $\mathcal{H}_{x,y} \neq \varnothing$, we have $\mathcal{A}_{H_{R_-}}(y) - \mathcal{A}_{H_{R_+}}(x) \ge 0$. Therefore the flow morphism $\mathfrak{H}^{R_+,R_-}$ preserves the action filtration, and in particular, the filtration induced by W^i. Hence we have the flow morphisms

$$\mathfrak{H}_0^{R_+,R_-} : \mathcal{C}_0^{R_+,J_+} \Rightarrow \mathcal{C}_0^{R_-,J_-}, \quad \mathfrak{H}_+^{R_+,R_-} : \mathcal{C}_+^{R_+,J_+} \Rightarrow \mathcal{C}_+^{R_-,J_-},$$
$$\mathfrak{H}_{\le i}^{R_+,R_-} : \mathcal{C}_{\le i}^{R_+,J_+} \Rightarrow \mathcal{C}_{\le i}^{R_-,J_-}, \quad \mathfrak{H}_{+,\le i}^{R_+,R_-} : \mathcal{C}_{+,\le i}^{R_+,J_+} \Rightarrow \mathcal{C}_{+,\le i}^{R_-,J_-}.$$

3.3 Sphere bundles and Gysin exact sequences

For any oriented k–sphere bundle $\pi : E \to W$ with k odd, there is an associated Gysin exact sequence

$$(5) \qquad \to H^i(W) \xrightarrow{\pi^*} H^i(E) \xrightarrow{\pi_*} H^{i-k}(W) \xrightarrow{\wedge(-e)} H^{i+1}(W) \to$$

Here π_* is integration along the fiber using the convention in [5, Section 6] and e is the Euler class of π; the extra sign is for consistency with [28, Proposition 6.24]. In this subsection, we consider sphere bundles over symplectic cohomology and deduce the associated Gysin exact sequences. This construction can be viewed as a higher-dimensional analogue of Floer cohomology with local systems. Gysin exact sequences in Floer theory were first considered by Bourgeois and Oancea [6], where the exact sequence arises from an S^1–bundle in the construction of S^1–equivariant symplectic homology. Fiber bundles over Floer theory were considered by Barraud and Cornea [3], where they considered the path–loop fibration. The smooth fiber bundles we consider are technically easier to deal with. The construction in [28] works as long as the moduli spaces support integration [15]. We first recall the concept of sphere bundles over flow categories:

Definition 3.11 [28, Definition 6.17] Let $\mathcal{C} = \{C_i, \mathcal{M}_{i,j}^C\}$ be an oriented flow category. An oriented k–sphere bundle over $\mathcal{C}$ is a flow category $\mathcal{E} = \{E_i, \mathcal{M}_{i,j}^E\}$ with functor $\pi : \mathcal{E} \to \mathcal{C}$ such that the following hold:

(i) π maps E_i to C_i and $\mathcal{M}_{i,j}^E$ to $\mathcal{M}_{i,j}^C$.

(ii) The maps $\pi : E_i \to C_i$ and $\pi : \mathcal{M}_{i,j}^E \to \mathcal{M}_{i,j}^C$ are oriented sphere bundles such that both bundle maps $s_{i,j}^E$ and $t_{i,j}^E$ preserve the orientation.

By [28, Proposition 6.18], an oriented k–sphere bundle $\mathcal{E}$ over an oriented flow category is an oriented flow category. The construction [28, Definition 3.8] assigns $\mathcal{E}$ to a cochain complex, and we have:

Proposition 3.12 [28, Theorem 6.19] *Let $\mathcal{E}$ be an oriented k–sphere bundle over an oriented flow category $\mathcal{C}$. Then we have a short exact sequence of cochain complexes*[3]

$$0 \to C^*(\mathcal{C}) \xrightarrow{\pi^*} C^*(\mathcal{E}) \xrightarrow{\pi_*} C^{*-k}(\mathcal{C}) \to 0.$$

It induces the Gysin exact sequence

$$\cdots \to H^*(\mathcal{C}) \to H^*(\mathcal{E}) \to H^{*-k}(\mathcal{C}) \to H^{*+1}(\mathcal{C}) \to \cdots. \tag{6}$$

Remark 3.13 Both π^* and π_* are induced by oriented flow morphisms, which are completely determined by $\mathcal{E}$. Here we give an explanation in the special case when $E \to C$ is an actual sphere bundle. The compact manifold C can be understood as a flow category whose object space is diffeomorphic to C and morphism space consists of only identity morphisms. Then E can be understood as a sphere bundle over the flow category C, π^* is given by the flow morphism $C \xleftarrow{s=\pi} E \xrightarrow{t=\mathrm{id}} E$ and π_* is given by the flow morphism $E \xleftarrow{s=\mathrm{id}} E \xrightarrow{t=\pi} C$. In particular, π^* is the composition $t_* \circ s^*$ from $C \xleftarrow{s=\pi} E \xrightarrow{t=\mathrm{id}} E$, which is indeed the pullback π^* on cohomology, and π_* is the composition $t_* \circ s^*$ from $E \xleftarrow{s=\mathrm{id}} E \xrightarrow{t=\pi} C$, which is the pushforward π_* on cohomology. In general, the underlying flow morphisms of π^* and π_* are induced from the identity flow morphism of $\mathcal{E}$ [28, Definition 3.23].

Remark 3.14 [28, Corollary 6.23] Assume $\mathcal{C} = \{C_i, \mathcal{M}_{i,j}\}$ is a Morse flow category, ie $\dim C_i = 0$. Then $H^*(\mathcal{E})$ and the Gysin exact sequence only depend on $\mathcal{M}^E_{i,j}$ with $\dim \mathcal{M}^E_{i,j} \le 2k$. In particular, we only use moduli spaces $\mathcal{M}_{i,j}$ of dimension up to k. The nontriviality of higher-dimensional moduli spaces $\mathcal{M}_{i,j}$ is the foundation of the existence of interesting sphere bundles. Although the formula only requires $\mathcal{M}^E_{i,j}$ of dimension up to $2k$, we need a priori the existence of the full flow category to guarantee the existence of Gysin sequences.

Remark 3.15 The Gysin exact sequence considered in [28] works for any Morse–Bott flow category $\mathcal{C}$. In the case considered here ($\mathcal{C}$ is Morse) it is possible to generalize the construction in [17] to the S^k case to get a $\mathbb{Z}$–coefficient Gysin exact sequence.

[3] To be more precise, we have a short exact sequence using certain choices in the construction. However, in the special case that $\mathcal{C}$ is Morse, the minimal construction in [28, Theorem 3.10], ie the one in Remark 3.8, gives the short exact sequence.

We call a Gysin exact sequence (6) trivial when the Euler part $H^*(\mathcal{C}) \to H^{*+k+1}(\mathcal{C})$ is zero. In the case we consider, a sphere bundle over the Liouville domain will induce a sphere bundle over the symplectic flow category.

Proposition 3.16 *Let W be a Liouville domain and $J \in \mathcal{J}^R(W)$. Let $\pi\colon E \to W$ be an oriented k–sphere bundle and P_γ the parallel transport along path γ for a fixed connection on E. Then we have oriented k–sphere bundles $\mathcal{E}^{R,J}$, $\mathcal{E}_0^{R,J}$, $\mathcal{E}_+^{R,J}$, $\mathcal{E}_{\le i}^{R,J}$ and $\mathcal{E}_{+,\le i}^{R,J}$ over $\mathcal{C}^{R,J}$, $\mathcal{C}_0^{R,J}$, $\mathcal{C}_+^{R,J}$, $\mathcal{C}_{\le i}^{R,J}$ and $\mathcal{C}_{+,\le i}^{R,J}$, respectively.*

Proof If $\mathcal{C} = \{x_i, \mathcal{M}_{i,j}\}$, then we define $E_i := E_{x_i(0)} \simeq S^k$ and $\mathcal{M}_{i,j}^E := \mathcal{M}_{i,j} \times E_i$. The structure maps are

$$s^E\colon \mathcal{M}_{i,j} \times E_i \to E_i \quad \text{given by } (u,v) \mapsto v,$$

$$t^E\colon \mathcal{M}_{i,j} \times E_i \to E_j \quad \text{given by } (u,v) \mapsto P_{u(-\cdot,0)}v,$$

$$m\colon (\mathcal{M}_{i,j} \times E_i) \times_{E_j} (\mathcal{M}_{j,k} \times E_j) \to \mathcal{M}_{i,k} \times E_i$$
$$\text{given by } (u_1, v, u_2, P_{u_1(-\cdot,0)}v) \mapsto (u_1, u_2, v).$$

It is direct to check that they form a category. The fiber product transversality follows since s^E and t^E are submersive. Because $E \to W$ is an oriented sphere bundle, we have that $E_i = E_{x_i(0)}$ is oriented and P_γ preserves the orientation. Hence $\mathcal{E}^{R,J} = \{E_i, \mathcal{M}_{i,j}^E\}$ is an oriented k–sphere bundle over $\mathcal{C}^{R,J}$. Similarly for other flow categories. □

Example 3.17 To further explain Remark 3.14, we can look at two flow categories: $\mathrm{Obj}_{\mathcal{C}_1}$ is set of two points $\{x_0, x_1\}$ with $\mathcal{M}_{0,1} = \varnothing$, and $\mathrm{Obj}_{\mathcal{C}_2} = \{x_0, x_1\}$ while $\mathcal{M}_{0,1} = S^1$. $\mathcal{C}_2$ can be viewed the flow category associated to the Morse theory of the height function on S^2. Then $\mathcal{C}_1$ does not admit any nontrivial S^n bundle; in particular, the associated Euler part is always trivial. Even though $\mathcal{C}_1$ and $\mathcal{C}_2$ have the same cohomology of rank 2, $\mathcal{C}_2$ admits a nontrivial S^1 bundle $\mathcal{E}_2 = \{S_0^1, S_1^1, \mathcal{M}_{0,1}^E = S^1 \times S^1\}$, where $\mathcal{M}_{0,1}^E$ is viewed as an S^1 bundle over the second factor S^1, which is viewed as $\mathcal{M}_{0,1}$. The structural maps are defined as $s^E\colon (\theta, t) \mapsto \theta$ and $t^E\colon (\theta, t) \mapsto \theta + t$. One may check the induced Gysin exact sequence has nontrivial Euler part. Indeed, $\mathcal{E}_2$ is the S^1 bundle induced from the Hopf fibration over S^2 using an appropriate parallel transport. This example shows that higher-dimensional moduli spaces are foundations for interesting fibrations.

Similarly, there is a notion of oriented sphere bundles over flow morphisms. Given two oriented k–sphere bundles $\mathcal{E} \to \mathcal{C}$ and $\mathcal{F} \to \mathcal{D}$, let $\mathfrak{H}\colon \mathcal{C} \Rightarrow \mathcal{D}$ be an oriented flow morphism. Then a k–sphere bundle $\mathfrak{P}$ over $\mathfrak{H}$ is defined as follows:

(i) $\mathfrak{P}=\{\mathcal{P}_{i,j}\}$ is a flow morphism from $\mathcal{E}$ to $\mathcal{F}$.

(ii) $\pi:\mathcal{P}_{i,j}\to\mathcal{H}_{i,j}$ is a k–sphere bundle such that s^P and t^P are bundles maps covering s^H and t^H.

(iii) $\pi:\mathcal{P}_{i,j}\to\mathcal{H}_{i,j}$ is an oriented bundle, and s^P and t^P preserve the orientation.

Given a sphere bundle $E\to W$ with a parallel transport, let $\mathfrak{H}$, $\mathfrak{H}_0$, $\mathfrak{H}_+$, $\mathfrak{H}_{\le i}$ and $\mathfrak{H}_{+,\le i}$ be the flow morphisms constructed from H_{R_+,R_-}. Then by the same construction as in Proposition 3.16, there are induced oriented sphere bundles $\mathfrak{P}$, $\mathfrak{P}_0$, $\mathfrak{P}_+$, $\mathfrak{P}_{\le i}$ and $\mathfrak{P}_{+,\le i}$ over them. Moreover, the parallel transport at two ends can be different. In this case, we need to fix a smooth family of connections $\{\xi_s\}$ such that ξ_s is the connection for the negative end for $s\ll 0$ and ξ_s is the connection for the positive end for $s\gg 0$. Then given a Floer solution $u(s,t)$ in the flow morphism for continuation maps, the structure maps for the sphere bundle are defined using the parallel transport with respect to xi_s over $u(s,0)$.

By [28, Proposition 6.27], sphere bundles over flow morphisms induce morphisms of Gysin sequences. We define $\mathcal{J}^R_{\le i}$ to be the set of almost complex structures such that the flow category $\mathcal{C}^{R,J}_{\le i}$ is defined. Given a sequence of real numbers $1>R_1>R_2>\cdots>0$ and a sequences of almost complex structures J_i such that $J_i\in\mathcal{J}^{R_i}_{\le i}(W)$, if we fix any oriented S^k bundle $E\to W$ along with a connection, then Proposition 3.12 induces the commutative diagram of exact sequences

$$\begin{array}{ccccccccc}
 & \downarrow & & \downarrow & & \downarrow & & \downarrow & \\
\to & \varinjlim_i H^{*+k}(\mathcal{E}_0^{R_i,J_i}) & \to & \varinjlim_i H^{*}(\mathcal{C}_0^{R_i,J_i}) & \to & \varinjlim_i H^{*+k+1}(\mathcal{C}_0^{R_i,J_i}) & \to & \varinjlim_i H^{*+k+1}(\mathcal{E}_0^{R_i,J_i}) & \to \\
 & \downarrow & & \downarrow & & \downarrow & & \downarrow & \\
\to & \varinjlim_i H^{*+k}(\mathcal{E}^{R_i,J_i}) & \to & \varinjlim_i H^{*}(\mathcal{C}^{R_i,J_i}) & \to & \varinjlim_i H^{*+k+1}(\mathcal{C}^{R_i,J_i}) & \to & \varinjlim_i H^{*+k+1}(\mathcal{E}^{R_i,J_i}) & \to \\
 & \downarrow & & \downarrow & & \downarrow & & \downarrow & \\
\to & \varinjlim_i H^{*+k}(\mathcal{E}_+^{R_i,J_i}) & \to & \varinjlim_i H^{*}(\mathcal{C}_+^{R_i,J_i}) & \to & \varinjlim_i H^{*+k+1}(\mathcal{C}_+^{R_i,J_i}) & \to & \varinjlim_i H^{*+k+1}(\mathcal{E}_+^{R_i,J_i}) & \to \\
 & \downarrow & & \downarrow & & \downarrow & & \downarrow &
\end{array}$$

Note that

$$\varinjlim_i H^*(\mathcal{C}_0^{R_i,J_i})=H^*(W),\quad \varinjlim_i H^*(\mathcal{C}^{R_i,J_i})=SH^*(W),$$
$$\varinjlim_i H^*(\mathcal{C}_+^{R_i,J_i})=SH^*_+(W).$$

We expect $\varinjlim_i H^*(\mathcal{E}^{R_i,J_i})$ and $\varinjlim_i H^*(\mathcal{E}_+^{R_i,J_i})$ are also well-defined objects, but this requires proving invariance under changing various defining data like H_R, R_i, J_i and the parallel transport P. In the Morse–Bott situation considered here, we need to use the flow-homotopy introduced in [28, Definition 3.29] to prove the invariance.

However, for the purpose of this paper, we do not need a well-defined Floer theory for the sphere bundle and are only interested in the Euler part. We will proceed with this version involving all specific choices. We will suppress the choice of parallel transport for simplicity, and only specify our choice when it matters.

Since the constant orbits part corresponds to the Morse theory on W, there the Gysin sequence should be the regular Gysin sequence.

Proposition 3.18 [28, Theorem 8.14] *The Gysin sequence*

$$\to \varinjlim_i H^{*+k}(\mathcal{E}_0^{R_i,J_i}) \to \varinjlim_i H^*(\mathcal{C}_0^{R_i,J_i}) \to \varinjlim_i H^{*+k+1}(\mathcal{C}_0^{R_i,J_i}) \to \varinjlim_i H^{*+k+1}(\mathcal{E}_0^{R_i,J_i}) \to$$

is the classical Gysin exact sequence (5) for $\pi : E \to W$.

By the Gysin exact sequence for symplectic cohomology $SH^*(W)$, we have the following vanishing result:

Proposition 3.19 *If $SH^*(W) = 0$ and E is an oriented sphere bundle over the Liouville domain W, then $\varinjlim_i H^*(\mathcal{E}^{R_i,J_i}) = 0$ for any defining data.*

3.4 Naturality

In the neck-stretching argument, we need to compare moduli spaces of two fillings, hence naturality is important. Moreover, we can only get the moduli spaces appearing in the counting matched up for two fillings, ie moduli spaces of dimension up to k. But to apply Proposition 3.12 we need the full flow category. In particular, it is possible that the higher-dimensional moduli spaces are not cut out transversely in the neck-stretching. In the following, we discuss those aspects in a similar way to [30].

Definition 3.20 $\mathcal{J}^{R,\le k}(W) \subset \mathcal{J}(W)$ is the set of admissible almost complex structures such that moduli spaces of H_R up to dimension k are cut out transversely. $\mathcal{J}_+^{R,\le k}(W)$ stands for the positive version, and $\mathcal{J}_{\le i}^{R,\le k}(W)$ and $\mathcal{J}_{+,\le i}^{R,\le k}(W)$ are the truncated versions.

All above sets are of second Baire category. Moreover, as a consequence of compactness, $\mathcal{J}_{(+),\le i}^{R,\le k}$ is open and dense. The following is a standard result in Floer theory:

Proposition 3.21 *Let $J_0 \in \mathcal{J}_{+,\le i}^{R_0,\le 0}(W)$ and $J_1 \in \mathcal{J}_{+,\le i+1}^{R_1,\le 0}(W)$ for $R_0 > R_1$. Then $H^*(\mathcal{C}_{+,\le i}^{R_0,J_0}) \to H^*(\mathcal{C}_{+,\le i+1}^{R_1,J_1})$, the continuation map, is independent of the homotopy of almost complex structures.*

We also recall the following result from [30]:

Proposition 3.22 [30, Lemma 2.15] *Let $J_s, s \in [0,1]$ be a smooth path in $\mathcal{J}(W)$ and R_s be a nonincreasing function in $(0,1]$ such that $J_s \in \mathcal{J}^{R_s,\le 0}_{+,\le i}(W)$. Then the continuation map $C^*(\mathcal{C}^{R_0,J_0}_{+,\le i}) \to C^*(\mathcal{C}^{R_1,J_1}_{+,\le i})$ is homotopic to the identity.*[4]

Note that we assume H_R stays the same outside $r = b_0$ for any R, meaning the generators for positive symplectic cohomology stay the same. The same argument of [30, Lemma 2.15] can be applied here for positive symplectic cohomology, even though we assume $H = 0$ on W in [30, Lemma 2.15].

Although the full flow category requires transversality for all moduli spaces, the Gysin sequence is well defined for almost complex structure of low regularity:

Proposition 3.23 *Let $E \to W$ be a k–sphere bundle. Then the Euler part of the Gysin exact sequence*

$$\to \varinjlim_i H^{*+k}(\mathcal{E}^{R_i,J_i}_{+,\le i}) \to \varinjlim_i H^*(\mathcal{C}^{R_i,J_i}_{+,\le i}) \to \varinjlim_i H^{*+k+1}(\mathcal{C}^{R_i,J_i}_{+,\le i}) \to \varinjlim_i H^{*+k+1}(\mathcal{E}^{R_i,J_i}_{+,\le i}) \to$$

is well defined for $J_i \in \mathcal{J}^{R_i,\le k}_{+,\le i}(W)$.

Proof We first prove the truncated Gysin sequence

$$\to H^{*+k}(\mathcal{E}^{R_i,J_i}_{+,\le i}) \to H^*(\mathcal{C}^{R_i,J_i}_{+,\le i}) \to H^{*+k+1}(\mathcal{C}^{R_i,J_i}_{+,\le i}) \to H^{*+k+1}(\mathcal{E}^{R_i,J_i}_{+,\le i}) \to$$

is defined. Since we can find an open neighborhood $\mathcal{U} \subset \mathcal{J}^{R_i,\le k}_{+,\le i}(W)$ of J_i, we have a universal moduli space $\bigcup_{J\in\mathcal{U}} M_{x,y,J}$, where $M_{x,y,J}$ is the moduli space of *unbroken* Floer trajectories using J in (2) for the positive symplectic cohomology for $x, y \subset W^i$. The universal moduli space is a Banach manifold and its projection to $\mathcal{U}$ is regular. For each $J \in \mathcal{U} \cap \mathcal{J}^{R_i}_{+,\le i}(W)$, we have a flow category with sphere bundle. We use d_J to denote the differential on the cochain complex of the sphere bundle. Moreover, d_J is well defined by (3) for $J \in \mathcal{U}$, even though d_J^2 may not be zero a priori unless $J \in \mathcal{U} \cap \mathcal{J}^{R_i}_{+,\le i}(W)$ since the integration (3) only depends on the full measure set $M_{x,y,J}$. We have that d_J varies continuously[5] over $\mathcal{U}$. Since $\mathcal{U} \cap \mathcal{J}^{R_i}_{+,\le i}(W)$ is dense in $\mathcal{U}$, we have $d_J^2 = 0$ for every $J \in \mathcal{U}$. As a consequence, the Gysin sequence is

[4] Note that generators are the same for (H_{R_0}, J_0) and (H_{R_1}, J_1), hence the identity map makes sense.

[5] The compactification $\mathcal{M}_{x,y}$ also varies continuously for $J \in \mathcal{U} \subset \mathcal{J}^{R,\le k}_{+,\le i}$ when $\dim \mathcal{M}_{x,y} \le k$.

defined for every $J \in \mathcal{U}$, in particular for J_i. Then by a similar argument, by finding $J_{i,i+1} \in \mathcal{J}^{R_i,R_{i+1},\le k}_{+,\le i}(W)$ we have a commutative diagram of the truncated Gysin sequence. This yields a Gysin sequence of the direct limit. Since we only need the well-definedness of the Euler part, the continuation map $H^*(\mathcal{C}^{R_i,J_i}_{+,\le i}) \to H^*(\mathcal{C}^{R_{i+1},J_{i+1}}_{+,\le i+1})$ is independent of the choice of $J_{i,i+1}$ by Proposition 3.21. □

Corollary 3.24 *Let $E \to W$ be a k–sphere bundle. Assume $J_s, s \in [0,1]$ is a smooth path in $\mathcal{J}(W)$ and R_s is a nonincreasing smooth function taking values in $(0,1]$ such that $J_s \in \mathcal{J}^{R_s,\le k}_{+,\le i}(W)$. Then the Euler parts of the Gysin exact sequences are commutative*:

$$\begin{CD} H^*(\mathcal{C}^{R_0,J_0}_{+,\le i}) @>>> H^{*+k+1}(\mathcal{C}^{R_0,J_0}_{+,\le i}) \\ @VVV @VVV \\ H^*(\mathcal{C}^{R_1,J_1}_{+,\le i}) @>>> H^{*+k+1}(\mathcal{C}^{R_1,J_1}_{+,\le i}) \end{CD} \tag{7}$$

Here the vertical arrows are the continuation maps, which are homotopic to the identity by Proposition 3.22.

Proof Assume in addition that $J_0 \in \mathcal{J}^{R_0}_{+,\le i}$ and $J_1 \in \mathcal{J}^{R_1}_{+,\le i}$. Then we can find a regular enough homotopy from J_0 to J_1 such that we have a flow morphism between the associated flow categories. The induced continuation map induces an isomorphism on the Euler parts of the Gysin exact sequences. By Proposition 3.22, the continuation map $H^*(\mathcal{C}^{R_0,J_0}_{+,\le i}) \to H^*(\mathcal{C}^{R_1,J_1}_{+,\le i})$ is the identity. Therefore the Euler parts of the Gysin sequences are the same for J_0 and J_1, since $\mathcal{J}^{R_*,\le k}_{+,\le i}(W)$ is open and contains $\mathcal{J}^{R_*}_{+,\le i}(W)$ as a dense set. Then the argument in Proposition 3.23 shows that the Euler part varies continuously with respect to J. □

Proposition 3.25 *For $J_i \in \mathcal{J}^{R_i,\le k}_{+,\le i}(W)$ we have the well-defined commutative diagram*

$$\begin{CD} \varinjlim_i H^*(\mathcal{C}^{R_i,J_i}_{+,\le i}) @>>> \varinjlim_i H^{*+k+1}(\mathcal{C}^{R_i,J_i}_{+,\le i}) \\ @VVV @VVV \\ H^{*+1}(W) @>{\wedge(-e(E))}>> H^{*+k+2}(W) \end{CD}$$

where the horizontal map is the Euler part, and the vertical map is the connecting map from the positive symplectic cohomology to the cohomology of the filling.

Proof Since $\mathcal{J}^{R_i,\le k}_{+,\le i}(W)$ is open, we can choose J_i' in a connected neighborhood of J_i in $\mathcal{J}^{R_i,\le k}_{+,\le i}(W)$ such that $J_i' \in \mathcal{J}^{R_i}_{\le i}(W)$. Then by Corollary 3.24 we have the commutative diagram

$$\begin{tikzcd}
\varinjlim_i H^*(\mathcal{C}_{+,\le i}^{R_i,J_i}) \arrow[r] \arrow[d, equal] & \varinjlim_i H^{*+k+1}(\mathcal{C}_{+,\le i}^{R_i,J_i}) \arrow[d, equal] \\
\varinjlim_i H^*(\mathcal{C}_{+,\le i}^{R_i,J_i'}) \arrow[r] \arrow[d] & \varinjlim_i H^{*+k+1}(\mathcal{C}_{+,\le i}^{R_i,J_i'}) \arrow[d] \\
H^{*+1}(W) \arrow[r, "\wedge(-e(E))"] & H^{*+k+2}(W)
\end{tikzcd}$$

By [30, Proposition 2.17], the vertical arrows in the bottom square do not depend on the choice of J_i'. □

3.5 Neck-stretching and independence of the positive Gysin sequence

Let (Y,α) be a k–ADC contact manifold with two topologically simple fillings W_1 and W_2. Note that $\widehat{W}_1$ and $\widehat{W}_2$ both contain the symplectization $(Y\times(0,\infty)_r, d(r\alpha))$. Since Y is k–ADC, there exist nested contact type surfaces $Y_i\subset Y\times(0,1)$ such that Y_i lies outside of Y_{i+1} and contractible Reeb orbits of contact form $r\alpha|_{Y_i}$ have the property that the degree is greater than k if the period is smaller than D_i.

We now define neck-stretching near Y_i. Assume domains of the form $Y_i\times[1-\epsilon_i,1+\epsilon_i]_{r_i}$ are disjoint for some small ϵ_i, where r_i is the coordinate determined by the Liouville vector field near Y_i such that $r_i|_{Y_i}=1$. Assume $J|_{Y_i\times[1-\epsilon_i,1+\epsilon_i]_{r_i}}=J_0$ where J_0 is independent of S^1 and r_i, and $J_0(r_i\partial_{r_i})=R_i$ and $J_0\xi_i=\xi_i$ where $\xi_i=\ker r\alpha|_{Y_i}$ and R_i is the associated Reeb vector field. Then we pick a family of diffeomorphisms $\phi_R\colon[(1-\epsilon_i)e^{1-1/R},(1+\epsilon_i)e^{1/R-1}]\to[1-\epsilon_i,1+\epsilon_i]$ for $R\in(0,1]$ such that $\phi_1=\mathrm{id}$ and ϕ_R near the boundary is linear with slope 1. Then the stretched almost complex

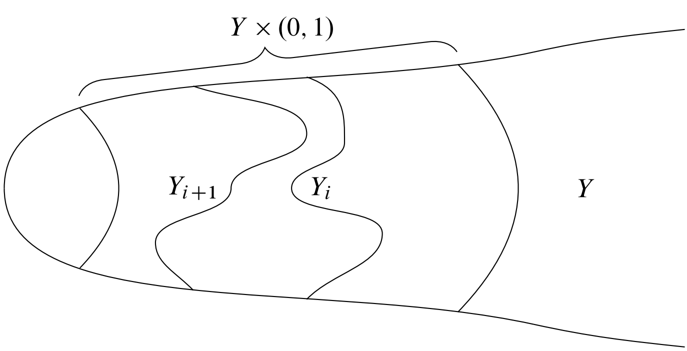


Figure 2: $Y_i\subset\widehat{W}_*$.

structure $NS_{i,R}(J)$ is defined to be J outside $Y_i \times [1-\epsilon_i, 1+\epsilon_i]$ and is $(\phi_R \times \mathrm{id})_* J_0$ on $Y_i \times [1-\epsilon_i, 1+\epsilon_i]$. Then $NS_{i,1}(J) = J$ and $NS_{i,0}(J)$ gives almost complex structures on the completions of the cobordism X_i between Y and Y_i, the filling of Y_i, and the symplectization $Y_i \times \mathbb{R}_+$.

Since we need to stretch along different contact surfaces, we assume the $NS_{i,R}(J)$ have the property that $NS_{i,R}(J)$ will modify the almost complex structure near Y_{i+1} to a cylindrical almost complex structure for R from 1 to $\frac{1}{2}$, and for $R \le \frac{1}{2}$ we only keep stretching along Y_i. We use $\mathcal{J}^{\le k}_{\mathrm{reg,SFT},\le i}(H_0)$ to denote the set of admissible regular J, ie almost complex structures satisfying Definition 3.2 on the completion of W outside Y_i and asymptotic (in a prescribed way as in the stretching process) to J_0 on the negative cylindrical end such that the following moduli space up to dimension k is cut out transversely:

$$\left\{ u : \mathbb{R} \times S^1 \backslash Z \to \widehat{X}_i \;\middle|\; \begin{array}{c} \partial_s u + J(\partial_t u - X_{H_0}) = 0, \\ \lim\limits_{s\to\infty} u = x, \quad \lim\limits_{s\to\infty} u = y, \\ \mathcal{A}_{H_0}(x), \mathcal{A}_{H_0}(y) > -D_i, \\ Z = \{z_1, \dots, z_I\}, \\ \lim\limits_{z\to z_j} u = \gamma_j \times \{-\infty\}, \forall 1 \le j \le I, \end{array} \right\} \Big/ \mathbb{R}.$$

Here γ_j is a Reeb orbit on Y_i and we write $\lim_{z\to z_j} u = \gamma_j \times \{-\infty\}$ if u is asymptotic to γ_j near the negative puncture $z_j \in \mathbb{R} \times S^1$. Then $\mathcal{J}^{\le k}_{\mathrm{reg,SFT},\le i}(H_0)$ is an open dense subset of all admissible almost complex structures on $\widehat{X}_i$. To compare moduli spaces for two Liouville fillings W_1 and W_2 we can assume that H_R outside Y_i is the same for W_1 and W_2 whenever $R \le \frac{1}{i}$. The following is simply a variant of [30, Proposition 3.12]:

Proposition 3.26 *With the setup above there exist admissible J^1_* and J^2_* on $\widehat{W}_*$ for $* = 1, 2$, and positive real numbers $\epsilon_1, \epsilon_2, \dots \le 1$ and $\delta_1, \delta_2, \dots \le 1$ with $\delta_i \le \frac{1}{i}$ such that the following hold*:

(i) *For $R < \epsilon_i$ and any $R' \in [0,1]$,*

$$NS_{i,R}(J^i_*) \in \mathcal{J}^{R\delta_i,\le k}_{+,\le i}(W_*) \quad \text{and} \quad NS_{i+1,R'}(NS_{i,R}(J^i_*)) \in \mathcal{J}^{R'R\delta_i,\le k}_{+,\le i}(W_*).$$

Moreover, all moduli spaces $\mathcal{M}_{x,y}$ of dimension up to k are the same for both W_1 and W_2, and contained outside Y_i for $x, y \in \mathcal{P}^*(H)$ with action at least $-D_i$.

(ii) *$J^{i+1}_* = NS_{i,\epsilon_i/2}(J^i_*)$ on W^i_* and $\delta_{i+1} = \frac{1}{2}\epsilon_i\delta_i$.*

Proof We prove the proposition by induction. Firstly, we set $\delta_1 = 1$. We then choose a J^1 such that $NS_{1,0}(J^1) \in \mathcal{J}^{\le k}_{\mathrm{reg,SFT},\le 1}(H_0)$. We will apply neck-stretching to J^1 at Y_1.

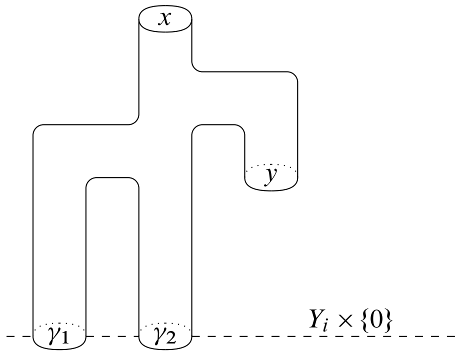


Figure 3: Moduli spaces for the definition of $\mathcal{J}^{\le k}_{\mathrm{reg},\mathrm{SFT},\le i}(H_0)$.

Note that we need to arrange the Hamiltonian converging to a constant near Y_i. We consider moduli space $\mathcal{M}_{x,y,H_R}$ with expected dimension at most k for $NS_{1,R}(J^1)$. Assume $\mathcal{M}_{x,y,H_R}$ is not contained outside Y_1 in the stretching process. Then a limit curve u outside Y_1 has one component by [9, Lemma 2.4].[6] Moreover, by the argument in [9, Lemma 2.4], u can only be asymptotic to Reeb orbits $\{\gamma_i\}_{i\in I}$ that are contractible in W_* on Y_1 with period smaller than D_1. Since W_* is topological simple, $\{\gamma_i\}_{i\in I}$ are contractible in Y_1. In particular, they all have well-defined $\mathbb{Z}$–valued Conley–Zehnder indices with SFT degree greater than k. The expected dimension of the moduli spaces of such u is $\operatorname{ind}(u)-1=|y|-|x|-\sum_{i\in I}(\mu_{CZ}(\gamma_i)+n-3)-1<|y|-|x|-1-k<0$. Since $NS_{1,0}(J^1)\in\mathcal{J}^{\le k}_{\mathrm{reg},\mathrm{SFT},\le 1}(H_0)$, we have that such a u is cut transversely. In particular, there is no such u as the expected dimension is negative. Then for $R\ll 1$, we have that $\mathcal{M}_{x,y,H_R}$ using $NS_{1,R}(J^1)$ is contained outside Y_1 whenever $\dim\mathcal{M}_{x,y,H_R}\le k$. Then $NS_{1,0}(J^1)\in\mathcal{J}^{\le k}_{\mathrm{reg},\mathrm{SFT},\le 1}(H_0)$ also implies that $NS_{1,R}(J^1)\in\mathcal{J}^{R,\le k}_{+,\le 1}(W_*)$ by the openness of transversality.

Next we will apply neck-stretching both at Y_1 and Y_2. By the same argument as above, for every $R'\in[0,1]$, we can find $\epsilon_{R'}>0$ and $\delta_{R'}>0$ such that for $\epsilon<\epsilon_{R'}$ and $|\delta-R'|<\delta_{R'}$,

- (i) $NS_{2,\delta}(NS_{1,\epsilon}(J^1))\in\mathcal{J}^{\delta\epsilon,\le k}_{+,\le 1}(W_*)$, and
- (ii) $\mathcal{M}_{x,y,H_{\epsilon\delta}}$ is contained outside Y_1 if the expected dimension is at most k.

Then compactness of $[0,1]_{R'}$ implies that there exists $\epsilon_1>0$ such that, for $R<\epsilon_i$ and any $R'\in[0,1]$, we have $NS_{1,R}(J^i_*)\in\mathcal{J}^{R,\le k}_{+,\le 1}(W_*)$ and $NS_{2,R'}(NS_{1,R}(J^1_*))\in\mathcal{J}^{R'R,\le k}_{+,\le 1}(W_*)$.

[6]Note that our symplectic action has the opposite sign compared to [9, Proposition 9.17].

Moreover, the moduli space $\mathcal{M}_{x,y,H_{R'R}}$ for $NS_{2,R'}(NS_{1,R}(J^1_*))$ is contained outside Y_1. We can certainly arrange ϵ_1 small enough so that $\delta_2 = \frac{1}{2}\epsilon_1\delta_1 = \frac{1}{2}\epsilon_1 \le \frac{1}{2}$. Since moduli spaces in Figure 3 for $NS_{2,0}(NS_{1,R}(J^1_*))$ must be contained outside Y_1 for x and y with action at least $-D_1$ when $R \ll 0$ by the same neck-stretching argument along Y_1, we may assume $NS_{2,0}(NS_{1,\epsilon_1/2}(J^1)) \in \mathcal{J}^{\le k}_{\mathrm{reg},\mathrm{SFT},\le 1}(H_0)$. Therefore we can perturb $NS_{1,\epsilon_1/2}(J^1) \in \mathcal{J}^{\epsilon_1/2,\le k}_{+,\le 1}(W_*)$ outside W^1_* near orbits in W^2_* to obtain J^2_* such that $NS_{2,0}(J^2) \in \mathcal{J}^{\le k}_{\mathrm{reg},\mathrm{SFT},\le 2}(H_0)$. This will not influence the previous regularity property for periodic orbits with action down to $-D_1$ by the integrated maximum principle. Then we can apply neck-stretching to J^2_* at Y_2 to obtain ϵ_2 with the desired properties and keep the induction going.

Since we require that H_R outside Y_i is the same for W_1 and W_2 whenever $R \le \frac{1}{i}$ and $\delta_i \le \frac{1}{i}$, it is clear that $\mathcal{M}_{x,y,H_{R'R\delta_i}}$ using $NS_{i+1,R'}(NS_{i,R}(J^i_*))$ can be identified for $R < \epsilon_i$ whenever the dimension is at most k and the action of x and y is greater than $-D_i$. This is because it is contained outside Y_i where all the geometric data are the same. □

Proposition 3.27 *Let Y be a k–ADC contact manifold with two topologically simple Liouville fillings W_1 and W_2. Then for $* = 1,2$, there exists a sequence of almost complex structures $\widetilde{J}^1_*, \widetilde{J}^2_*, \dots$ and positive numbers $1 > R_1 > R_2 > \cdots > 0$ such that for any oriented k–sphere bundles E_* over W_* with $E_1|_Y = E_2|_Y$, we have an isomorphism $\Phi\colon \varinjlim_i H(\mathcal{C}^{R_i,\widetilde{J}^i_1}_+) \simeq SH^*_+(W_1) \to SH^*_+(W_2) \simeq \varinjlim_i H(\mathcal{C}^{R_i,\widetilde{J}^i_2}_+)$ such that the following Euler part of the Gysin exact sequence commutes*:

$$\begin{array}{ccc} \varinjlim_i H^*(\mathcal{C}^{R_i,\widetilde{J}^i_1}_{+,\le i}) & \longrightarrow & \varinjlim_i H^{*+k+1}(\mathcal{C}^{R_i,\widetilde{J}^i_1}_{+,\le i}) \\ \downarrow{\scriptstyle\Phi} & & \downarrow{\scriptstyle\Phi} \\ \varinjlim_i H^*(\mathcal{C}^{R_i,\widetilde{J}^i_2}_{+,\le i}) & \longrightarrow & \varinjlim_i H^{*+k+1}(\mathcal{C}^{R_i,\widetilde{J}^i_2}_{+,\le i}) \end{array}$$

Proof Using the almost complex structures from Proposition 3.26, we define $\widetilde{J}^i_*$ to be $NS_{i,\epsilon_i/2}(J^i_*)$ for $* = 1,2$. By Proposition 3.26, $\widetilde{J}^i_* \in \mathcal{J}^{\epsilon_i\delta_i/2,\le k}_{+,\le i}(W_*) = \mathcal{J}^{\delta_{i+1},\le k}_{+,\le i}(W_*)$. Therefore by Proposition 3.23, the direct limit of the following commutative sequence computes the Euler part of the Gysin exact sequence:

$$\begin{array}{ccccc} H^*(\mathcal{C}^{\delta_2,\widetilde{J}^1_*}_{+,\le 1}) & \longrightarrow & H^*(\mathcal{C}^{\delta_3,\widetilde{J}^2_*}_{+,\le 2}) & \longrightarrow & \cdots \\ \downarrow & & \downarrow & & \\ H^{*+k+1}(\mathcal{C}^{\delta_2,\widetilde{J}^1_*}_{+,\le 1}) & \longrightarrow & H^{*+k+1}(\mathcal{C}^{\delta_3,\widetilde{J}^2_*}_{+,\le 2}) & \longrightarrow & \cdots \end{array}$$

We first show that the continuation map $H^*(\mathcal{C}_{+,\le i}^{\delta_{i+1},\tilde{J}^i_*}) \to H^*(\mathcal{C}_{+,\le i+1}^{\delta_{i+2},\tilde{J}^{i+1}_*})$ is naturally identified for $*=1,2$. Note that the continuation map is decomposed into continuation maps

$$\Xi\colon H^*(\mathcal{C}_{+,\le i}^{\delta_{i+1},\tilde{J}^i_*}) \to H^*(\mathcal{C}_{+,\le i}^{\delta_{i+2},NS_{i+1,\epsilon_{i+1}/2}(\tilde{J}^i_*)})$$

and

$$\Psi\colon H^*(\mathcal{C}_{+,\le i}^{\delta_{i+2},NS_{i+1,\epsilon_{i+1}/2}(\tilde{J}^i_*)}) \to H^*(\mathcal{C}_{+,\le i+1}^{\delta_{i+2},\tilde{J}^{i+1}_*}).$$

Then Ξ is the identity by Proposition 3.22 using the regular homotopy $NS_{i+1,s}(\tilde{J}^i_*)$ for $s \in \left[\frac{1}{2}\epsilon_{i+1}, 1\right]$. Since J^{i+1}_* is the same as $\tilde{J}^i_*$ inside W^i, $NS_{i+1,\epsilon_{i+1}/2}(\tilde{J}^i_*)$ is $\tilde{J}^{i+1}_*$ inside W^i. Then the integrated maximum principle implies that Ψ is the composition

$$H^*(\mathcal{C}_{+,\le i}^{\delta_{i+2},NS_{i+1,\epsilon_{i+1}/2}(\tilde{J}^i_*)}) \xrightarrow{=} H^*(\mathcal{C}_{+,\le i}^{\delta_{i+2},\tilde{J}^{i+1}_*}) \xrightarrow{\subseteq} H^*(\mathcal{C}_{+,\le i+1}^{\delta_{i+2},\tilde{J}^{i+1}_*}),$$

which is the same for $*=1,2$. Therefore all the horizontal arrows in the diagram can be identified for both W_1 and W_2. We still need to identify the vertical arrow, ie the Euler part of Gysin sequence. For $\mathcal{C}_{+,\le i}^{\delta_{i+1},\tilde{J}^i_*}$, we pick the parallel transport outside Y_i such that they are identified for $*=1,2$, which is possible since $E|_{\partial W_1} = E|_{\partial W_2}$. Since the Euler part only requires $\mathcal{M}_{x,y}$ with $\dim \mathcal{M}_{x,y} \le k$ and parallel transport over them, Proposition 3.26(i) implies that whole diagram can be identified for $*=1,2$. Then Proposition 3.23 completes the proof □

Remark 3.28 Using that the almost complex structure satisfies the condition here and is close to the condition in [30, Theorem A], the isomorphism in Proposition 3.27 also yields the identification of the map $\delta_\partial\colon SH^\bullet_+(W_*) \to H^{\bullet+1}(Y)$ for $*=1,2$.

4 Proof of the main theorem and applications

Our method of proving Theorem 1.1 is to represent even degree cohomology classes as Euler classes of sphere bundles. The following result explains which class can be realized as the Euler class of a sphere bundle.

Theorem 4.1 [14, Theorem 4.1] *Given $k, m \in \mathbb{N}$, let $K(\mathbb{Z},2k)^m$ be the m–skeleton of the Eilenberg–Mac Lane space $K(\mathbb{Z},2k)$, with inclusion $i\colon K(\mathbb{Z},2k)^m \hookrightarrow K(\mathbb{Z},2k)$. Then there is an integer $N(k,m) > 0$ and an oriented $2k$–dimensional vector bundle $\xi_{k,m}$ over $K(\mathbb{Z},2k)^m$ with $e(\xi_{k,m}) = N(k,m)\cdot i^*u$, where u is the generator of $H^{2k}(K(\mathbb{Z},2k);\mathbb{Z})$.*

As a corollary of Theorem 4.1, let W be a manifold of dimension $2n$ and $\alpha \in H^{2k}(W;\mathbb{Z})$. Then α is uniquely represented by the homotopy class of a map $f_\alpha : M \to K(\mathbb{Z},2k)^{2n+1}$, so the Euler class of $f_\alpha^*\xi_{k,2n+1}$ is $N(k,2n+1)\cdot\alpha$. We first obtain the following proposition, which may have some independent interest:

Proposition 4.2 *Let Y be a k–ADC manifold with a topologically simple Liouville filling W_1 such that $SH^*(W_1)=0$. Then for any other topologically simple Liouville filling W_2, we have that $H^{2m}(W_2) \to H^{2m}(Y)$ is injective for $2m \le k+1$.*

Proof Note that the abelian group $H^{2m}(W_2;\mathbb{Z})$ has a noncanonical decomposition into free and torsion parts $H^{2m}(W_2;\mathbb{Z}) = H^{2m}_{\mathrm{free}}(W_2;\mathbb{Z}) \oplus H^{2m}_{\mathrm{tor}}(W_2;\mathbb{Z})$. To prove the injectivity of $H^{2m}(W_2) \to H^{2m}(Y)$ for real cohomology, it suffices to show that for any decomposition $H^{2m}_{\mathrm{free}}(W_2;\mathbb{Z}) \to H^{2m}(Y;\mathbb{Z})$ is an injection for $2m \le k+1$. Assume otherwise, so there is an element $\alpha \in H^{2m}_{\mathrm{free}}(W_2;\mathbb{Z}) \subset H^{2m}(W_2;\mathbb{Z})$ such that $\alpha|_Y = 0$ in $H^{2m}(Y;\mathbb{Z})$. By Theorem 4.1, there exist $N \in \mathbb{N}$, a $2m$–dimensional vector bundle $\xi_{m,2n+1}$ over the $2n+1$–skeleton $K(\mathbb{Z},2m)^{2n+1}$ of the Eilenberg–Mac Lane space $K(\mathbb{Z},2m)$, and $f_\alpha : W_2 \to K(\mathbb{Z},2m)^{2n+1}$ such that the Euler class of $E_2 := f_\alpha^*\xi_{m,2n+1}$ is $N\alpha$. Since $\alpha|_Y = 0$ in $H^{2m}(Y;\mathbb{Z})$, we have that $f_\alpha|_Y : Y \to K(\mathbb{Z},2m)^{2n+1}$ is contractible. Hence $E_2|_Y$ is a trivial bundle. Let $E_1 \to W_1$ be the trivial sphere bundle. Therefore by Proposition 3.27, there exist almost complex structures $J^1_*, J^2_*, \dots$ and $1 > R_1 > R_2 > \cdots > 0$ such the Euler part for the positive symplectic cohomology for $E_1 \to W_1$ and $E_2 \to W_2$ can be identified. By [30, Corollary B], we have $SH^*(W_2)=0$. Then Proposition 3.25 implies the commutative diagram

$$\begin{tikzcd}
H^{*+1}(W_1) \arrow[r, "0"] & H^{*+k+2}(W_1) \\
\varinjlim H^*(\mathcal{C}^{R_i,J^i_1}_{+,\le i}) \arrow[r] \arrow[u, "\simeq"] \arrow[d, "\simeq"] & H^{*+k+1}(\mathcal{C}^{R_i,J^i_1}_{+,\le i}) \arrow[u, "\simeq"] \arrow[d, "\simeq"] \\
\varinjlim H^*(\mathcal{C}^{R_i,J^i_2}_{+,\le i}) \arrow[r] \arrow[d, "\simeq"] & H^{*+k+1}(\mathcal{C}^{R_i,J^i_2}_{+,\le i}) \arrow[d, "\simeq"] \\
H^{*+1}(W_2) \arrow[r, "\wedge(-N\alpha)"] & H^{*+k+2}(W_2)
\end{tikzcd}$$

We arrive at a contradiction, since $N\alpha \neq 0$. □

Proposition 4.2 says that if we have extra room in the positivity of the SFT degree and also the vanishing of symplectic cohomology, then $H^*(W) \to H^*(Y)$ is necessarily injective for low even degrees. For example, if Y^{2n-1} is a flexibly fillable contact manifold, then Y is $(n-3)$–ADC [18]. In this case, we have that $H^*(W) \to H^*(Y)$ is always injective for even degree with $* \le n-2$. Note that such a property also follows from [30, Corollary B]: $H^*(W) \to H^*(Y)$ is independent of fillings and for $* < n-2$ we have that $H^*(W) \to H^*(Y)$ is an isomorphism for Weinstein fillings. However, Proposition 4.2 holds for very different reasons. Note that we do not assume $H^{2m}(W_1) \to H^{2m}(Y)$ is injective in Proposition 4.2.

Proof of Theorem 1.1 By Proposition 3.27, we can pick $1 > R_1 > R_2 > \cdots > 0$ and $J^1_*, J^2_*, \ldots$ such that the Euler part of the positive symplectic cohomology for W_1 and W_2 can be identified as long as $E_1|_Y = E_2|_Y$. Since

$$SH^*(W_1) = SH^*(W_2) = 0,$$

we can define ϕ to be the composition

$$H^*(W_1) \xrightarrow{\simeq} \varinjlim_i H^{*-1}(\mathcal{C}^{R_i,J^i_1}_{+,\le i}) \xrightarrow{\simeq} \varinjlim_i H^{*-1}(\mathcal{C}^{R_i,J^i_2}_{+,\le i}) \xrightarrow{\simeq} H^*(W_2).$$

In other words, ϕ is the identification in [30, Corollary B] such that

$$\begin{tikzcd} H^*(W_1) \arrow[rr, "\phi"] \arrow[dr] & & H^*(W_2) \arrow[dl] \\ & H^*(Y) & \end{tikzcd} \tag{8}$$

is commutative. In particular, $\phi(1) = 1$[7] and ϕ is actually induced from an isomorphism $\phi_{\mathbb{Z}}$ for $\mathbb{Z}$–coefficient cohomology. We pick an element $\alpha_1 \ne 0 \in H^{2k}(W_1;\mathbb{Z})$ for $2k \le n-2$. By [30, Corollary B], let $\alpha_2 = \phi_{\mathbb{Z}}(\alpha_1) \in H^{2k}(W_2;\mathbb{Z})$. Then we have $\alpha_2|_Y = \alpha_1|_Y \in H^*(Y;\mathbb{Z})$ by the $\mathbb{Z}$–coefficient version of (8), [30, Corollary B]. By Theorem 4.1, there exist $N \in \mathbb{N}$ and a bundle $\xi_{k,2n+1}$ such that $E_* := f^*_{\alpha_*}\xi_{k,2n+1}$ is a vector bundle over W_* with Euler class $N\alpha_*$ for $* = 1, 2$, and where the map $f_{\alpha_*}: W_* \to K(\mathbb{Z}, 2k)^{2n+1}$ represents α_*. Since $\alpha_2|_Y = \alpha_1|_Y \in H^*(Y;\mathbb{Z})$, we have that $f_{\alpha_1}|_Y$ is homotopic to $f_{\alpha_2}|_Y$. As a consequence, we have $E_1|_Y = E_2|_Y$, $e(E_1) = N\alpha_1$ and $e(E_2) = N\alpha_2$. Then by the same argument as in Proposition 4.2, we have the

[7] Without [30, Corollary B], $\phi(1) = \pm 1$ can already be obtained by grading.

commutative diagram

$$\begin{tikzcd}
H^{*+1}(W_1) \arrow[r, "\wedge(-N\alpha_1)"] \arrow[ddd, bend right, "\phi"'] & H^{*+k+2}(W_1) \arrow[ddd, bend left, "\phi"] \\
\varinjlim H^*(\mathcal{C}_{+,\le i}^{R_i,J_1^i}) \arrow[r] \arrow[u, "\simeq"'] \arrow[d, "\simeq"] & H^{*+k+1}(\mathcal{C}_{+,\le i}^{R_i,J_1^i}) \arrow[u, "\simeq"'] \arrow[d, "\simeq"] \\
\varinjlim H^*(\mathcal{C}_{+,\le i}^{R_i,J_2^i}) \arrow[r] \arrow[d, "\simeq"] & H^{*+k+1}(\mathcal{C}_{+,\le i}^{R_i,J_2^i}) \arrow[d, "\simeq"] \\
H^{*+1}(W_2) \arrow[r, "\wedge(-N\alpha_2)"] & H^{*+k+2}(W_2)
\end{tikzcd}$$

So $\phi(N\alpha_1 \wedge \beta) = N\alpha_2 \wedge \phi(\beta)$. Since $\phi(1) = 1$, it follows that $\phi(N\alpha_1) = N\alpha_2$ and $\phi(N\alpha_1 \wedge \beta) = N\alpha_2 \wedge \phi(\beta) = \phi(N\alpha_1) \wedge \phi(\beta)$. $\square$

Remark 4.3 Combining the argument in this paper with [30], one can prove that the following commutative diagram for a k sphere bundle E is independent of fillings and extensions of $E|_Y$ to W as long as Y is k–ADC:

$$\begin{tikzcd}
SH_+^*(W) \arrow[r, "\delta_\partial"] \arrow[d, "e(E)"] & H^{*+1}(Y) \arrow[d, "e(E)"] \\
SH_+^{*+k+1}(W) \arrow[r, "\delta_\partial"] & H^{*+k+2}(Y)
\end{tikzcd}$$

As a corollary, $\operatorname{im}\delta_\partial$ is closed under multiplication by the Euler class $e(E|_Y)$. Then the argument of Theorem 1.1 implies that $\operatorname{im}\delta_\partial$ is closed under multiplication by even elements of degree at most $k+1$ in $\operatorname{im}\delta_\partial$. Note that $\operatorname{im}\delta_\partial$ is an interesting invariant of ADC manifolds and can be used to define obstructions to Weinstein fillability.

Theorem 1.1 can be applied to examples listed in Example 2.6; the major class would be flexibly fillable contact manifolds. In the following, we list several cases where the whole real cohomology ring is unique. For simplicity, we only consider simply connected contact manifolds. Note that the following corollary includes Corollary 1.2:

Corollary 4.4 *Let Y be a simply connected flexibly fillable contact manifold satisfying one the following conditions*:

(i) *Y is $4n+1$–dimensional for $n \ge 1$.*

(ii) *Y is $4n+3$–dimensional for $n \geq 1$, and the flexible filling W has the property that for every $\alpha \wedge \beta \in H^{2n+2}(W)$ with $\deg(\alpha)$ and $\deg(\beta)$ odd, then α or β can be decomposed into a nontrivial product.*

Then $H^*(W)$ as a ring is unique for any Liouville filling W with $c_1(W) = 0$.

Proof By Theorem 1.1 and [30, Corollary B], the ring structure on $H^*(W)$ is unique if one of the factors is of even degree at most $\frac{1}{2}\dim W - 2$, or if the degree of the product is at most $\frac{1}{2}\dim W - 1$. When $\dim Y = 4n+1$, if the degree of the product is in the undetermined region, ie $\frac{1}{2}\dim W = 2n+1$, then one of the factors must be of even degree. If Y is simply connected then $H^1(W)$ is 0 by [30, Corollary B]. As a consequence, the other odd degree factor must have degree at least 3. Therefore the even degree factor has degree at most $\frac{1}{2}\dim W - 3$. In particular, all products fall in the above two cases. Therefore the ring structure is unique. In case (ii), the undetermined case is when the product has degree $2n+2$. If the product is from two classes of even degree, then we can apply Theorem 1.1. If the product is from two classes of odd degree, then by assumption one of them can be reduced to a nontrivial product. Since the ring structure in that degree is unique, the decomposition exists for any other filling. Therefore the product can be rewritten as a product of two even elements. Hence the ring structure is unique. □

Proof of Corollary 1.4 By Corollary 4.4, the real cohomology ring of the filling is unique. Moreover, it is straightforward to verify that the cohomology ring of products of $\mathbb{CP}^n$, $\mathbb{HP}^n$, S^{2n} and at most one copy of S^{2n+1} for $n \geq 1$ have unique minimal models. By [30, Theorem E], any exact filling of $\partial(\mathrm{Flex}(T^*M))$ with vanishing first Chern class is necessarily simply connected, in which case the real homotopy type is determined by the minimal model by [5, Section 19]. □

Theorem 1.1 can only be applied when there is one filling with vanishing symplectic cohomology. In some cases, symplectic cohomology vanishes with nontrivial local systems [2]. Here we only give one special example in such a case.

Proposition 4.5 *Assume W is a Liouville filling of $Y := \partial T^*\mathbb{CP}^n$ for $n \geq 3$ odd, which is $(2n-4)$–ADC. If $c_1(W) = 0$ and $H^2(W;\mathbb{Z}) \to H^2(Y;\mathbb{Z})$ is not zero, then the real cohomology ring $H^*(W)$ is isomorphic to $H^*(T^*\mathbb{CP}^n)$.*

Proof Since $\mathbb{CP}^n$ is spin for n odd, by [30, Theorem D] there is a local system on both W and $T^*\mathbb{CP}^n$ such that they are the same on Y and the twisted symplectic cohomology vanishes for both W and $T\mathbb{CP}^n$. Then we can apply the same argument of Theorem 1.1 to the case with local systems to finish the proof. □

References

[1] **M Abouzaid**, **P Seidel**, *An open string analogue of Viterbo functoriality*, Geom. Topol. 14 (2010) 627–718 MR Zbl

[2] **P Albers**, **U Frauenfelder**, **A Oancea**, *Local systems on the free loop space and finiteness of the Hofer–Zehnder capacity*, Math. Ann. 367 (2017) 1403–1428 MR Zbl

[3] **J-F Barraud**, **O Cornea**, *Lagrangian intersections and the Serre spectral sequence*, Ann. of Math. 166 (2007) 657–722 MR Zbl

[4] **K Barth**, **H Geiges**, **K Zehmisch**, *The diffeomorphism type of symplectic fillings*, J. Symplectic Geom. 17 (2019) 929–971 MR Zbl

[5] **R Bott**, **L W Tu**, *Differential forms in algebraic topology*, Graduate Texts in Math. 82, Springer (1982) MR Zbl

[6] **F Bourgeois**, **A Oancea**, *An exact sequence for contact- and symplectic homology*, Invent. Math. 175 (2009) 611–680 MR Zbl

[7] **F Bourgeois**, **A Oancea**, *Symplectic homology, autonomous Hamiltonians, and Morse–Bott moduli spaces*, Duke Math. J. 146 (2009) 71–174 MR Zbl

[8] **K Cieliebak**, **Y Eliashberg**, *From Stein to Weinstein and back*, American Mathematical Society Colloquium Publications 59, Amer. Math. Soc., Providence, RI (2012) MR Zbl

[9] **K Cieliebak**, **A Oancea**, *Symplectic homology and the Eilenberg–Steenrod axioms*, Algebr. Geom. Topol. 18 (2018) 1953–2130 MR Zbl

[10] **R L Cohen**, **J D S Jones**, **G B Segal**, *Floer's infinite-dimensional Morse theory and homotopy theory*, from "The Floer memorial volume" (H Hofer, C H Taubes, A Weinstein, E Zehnder, editors), Progr. Math. 133, Birkhäuser, Basel (1995) 297–325 MR Zbl

[11] **B Filippenko**, **K Wehrheim**, *A polyfold proof of the Arnold conjecture*, Selecta Math. 28 (2022) art. id. 11 MR Zbl

[12] **J Fish**, **H Hofer**, *Applications of polyfold theory*, *II*: *The polyfolds of symplectic field theory*, in preparation

[13] **M Gromov**, *Pseudo holomorphic curves in symplectic manifolds*, Invent. Math. 82 (1985) 307–347 MR Zbl

[14] **L Guijarro**, **T Schick**, **G Walschap**, *Bundles with spherical Euler class*, Pacific J. Math. 207 (2002) 377–391 MR Zbl

[15] **H Hofer**, **K Wysocki**, **E Zehnder**, *Integration theory on the zero sets of polyfold Fredholm sections*, Math. Ann. 346 (2010) 139–198 MR Zbl

[16] **H Hofer**, **K Wysocki**, **E Zehnder**, *Polyfold and Fredholm theory*, Ergebnesse der Math. (3) 72, Springer (2021) MR Zbl

[17] **M Hutchings**, **J Nelson**, *Axiomatic S^1 Morse–Bott theory*, Algebr. Geom. Topol. 20 (2020) 1641–1690 MR Zbl

[18] **O Lazarev**, *Contact manifolds with flexible fillings*, Geom. Funct. Anal. 30 (2020) 188–254 MR Zbl

[19] **P Massot**, **K Niederkrüger**, **C Wendl**, *Weak and strong fillability of higher dimensional contact manifolds*, Invent. Math. 192 (2013) 287–373 MR Zbl

[20] **D McDuff**, *The structure of rational and ruled symplectic 4–manifolds*, J. Amer. Math. Soc. 3 (1990) 679–712 MR Zbl

[21] **D McDuff**, *Symplectic manifolds with contact type boundaries*, Invent. Math. 103 (1991) 651–671 MR Zbl

[22] **A Oancea**, **C Viterbo**, *On the topology of fillings of contact manifolds and applications*, Comment. Math. Helv. 87 (2012) 41–69 MR Zbl

[23] **A F Ritter**, *Topological quantum field theory structure on symplectic cohomology*, J. Topol. 6 (2013) 391–489 MR Zbl

[24] **P Seidel**, *A biased view of symplectic cohomology*, from "Current developments in mathematics, 2006" (B Mazur, T Mrowka, W Schmid, R Stanley, S-T Yau, editors), International, Somerville, MA (2008) 211–253 MR Zbl

[25] **G Walschap**, *The Euler class as a cohomology generator*, Illinois J. Math. 46 (2002) 165–169 MR Zbl

[26] **K Wehrheim**, *Fredholm notions in scale calculus and Hamiltonian Floer theory*, preprint (2012) arXiv 1209.4040 To appear in J. Symplectic Geom.

[27] **K Wehrheim**, *Smooth structures on Morse trajectory spaces, featuring finite ends and associative gluing*, from "Proceedings of the Freedman Fest" (R Kirby, V Krushkal, Z Wang, editors), Geom. Topol. Monogr. 18, Geom. Topol. Publ., Coventry (2012) 369–450 MR Zbl

[28] **Z Zhou**, *Morse–Bott cohomology from homological perturbation theory*, preprint (2019) arXiv 1902.06587 To appear in Algebr. Geom. Topol.

[29] **Z Zhou**, *Vanishing of symplectic homology and obstruction to flexible fillability*, Int. Math. Res. Not. 2020 (2020) 9717–9729 MR Zbl

[30] **Z Zhou**, *Symplectic fillings of asymptotically dynamically convex manifolds, I*, J. Topol. 14 (2021) 112–182 MR Zbl

[31] **Z Zhou**, *Symplectic fillings of asymptotically dynamically convex manifolds, II: k–dilations*, Adv. Math. 406 (2022) art. id. 108522 MR Zbl

Morningside Center of Mathematics, Chinese Academy of Sciences
Beijing, China

zhyzhou@amss.ac.cn

https://sites.google.com/view/zhengyizhou/

Received: 1 August 2020 Revised: 30 April 2021

Algebraic & Geometric Topology **23**:4 (2023) 1725–1786
DOI: 10.2140/agt.2023.23.1725
Published: 14 June 2023

A model structure for weakly horizontally invariant double categories

LYNE MOSER
MARU SARAZOLA
PAULA VERDUGO

We construct a model structure on the category DblCat of double categories and double functors, whose trivial fibrations are the double functors that are surjective on objects, full on horizontal and vertical morphisms, and fully faithful on squares; and whose fibrant objects are the weakly horizontally invariant double categories.

We show that the functor $\mathbb{H}^{\simeq}\colon 2\mathrm{Cat} \to \mathrm{DblCat}$, a more homotopical version of the usual horizontal embedding $\mathbb{H}$, is right Quillen and homotopically fully faithful when considering Lack's model structure on 2Cat. In particular, $\mathbb{H}^{\simeq}$ exhibits a levelwise fibrant replacement of $\mathbb{H}$. Moreover, Lack's model structure on 2Cat is right-induced along $\mathbb{H}^{\simeq}$ from the model structure for weakly horizontally invariant double categories.

We also show that this model structure is monoidal with respect to Böhm's Gray tensor product. Finally, we prove a Whitehead theorem characterizing the weak equivalences with fibrant source as the double functors which admit a pseudoinverse up to horizontal pseudonatural equivalence.

18D20, 18N10, 18N40

1 Introduction

This paper aims to study and compare the homotopy theories of two related types of 2–dimensional categories: 2–*categories* and *double categories*. While 2–categories consist of objects, morphisms, and 2–morphisms, double categories admit two types of morphisms between objects — *horizontal* and *vertical* morphisms — and their 2–morphisms are given by *squares*. In particular, a 2–category $\mathcal{A}$ can always be seen as a horizontal double category $\mathbb{H}\mathcal{A}$ with only trivial vertical morphisms. This assignment $\mathbb{H}$ gives a full embedding of 2–categories into double categories.

The category 2Cat of 2–categories and 2–functors admits a model structure, constructed by Lack in [13; 14]. In this model structure, the weak equivalences are the biequivalences; the trivial fibrations are the 2–functors which are surjective on objects, full on morphisms, and fully faithful on 2–morphisms; and all 2–categories are fibrant. Moreover, Lack gives a characterization of the cofibrant objects as the 2–categories whose underlying category is free. With this well-established model structure at hand, we raise the question of whether there is a homotopy theory for double categories which contains that of 2–categories.

Several model structures for double categories were first constructed by Fiore and Paoli in [4], and by Fiore, Paoli and Pronk in [5], but the homotopy theory of 2–categories does not embed in any of these homotopy theories for double categories. The first positive answer to this question is given by the authors in [16], and further related results appear in work in progress by Campbell [2]. In [16], we construct a model structure on the category DblCat of double categories and double functors that is right-induced from two copies of Lack's model structure on 2Cat; its weak equivalences are called the *double biequivalences*, and like most of the model structure, they exhibit a strong horizontal bias.

As a consequence, this model structure is very well behaved with respect to the horizontal embedding $\mathbb{H}$: the functor $\mathbb{H}\colon 2\mathrm{Cat} \to \mathrm{DblCat}$ is both left and right Quillen, and Lack's model structure is both left- and right-induced along it. In particular, this says that Lack's model structure on 2Cat is created by $\mathbb{H}$ from the model structure on DblCat of [16]. Moreover, the functor $\mathbb{H}$ is homotopically fully faithful, and it embeds the homotopy theory of 2–categories into that of double categories in a reflective and coreflective way.

As it was constructed with a pronounced horizontal bias, this model structure is unsurprisingly not well behaved with respect to the vertical direction. For example, trivial fibrations, which are full on horizontal morphisms, are only surjective on vertical morphisms, and the free double category on two composable vertical morphisms is not cofibrant, as opposed to its horizontal analogue. In particular, this prevents the model structure from being monoidal with respect to the Gray tensor product for double categories defined by Böhm in [1].

Additionally, the model structure of [16] is not compatible with the first-named author's nerve construction from double categories to double $(\infty, 1)$–categories in [15]. Since all objects of this model structure on DblCat are fibrant, while the nerve of a double category is in general not fibrant, we see that the nerve functor fails to be right Quillen. In fact, the double categories whose nerve is fibrant are precisely the *weakly horizontally invariant* ones. This condition requires that every vertical morphism in the double category can be lifted along horizontal equivalences at its source and target; see Definition 2.10.

The aim of this paper is to provide a new model structure on DblCat, whose trivial fibrations behave symmetrically with respect to the horizontal and vertical directions, and whose fibrant objects are the weakly horizontally invariant double categories. We achieve this by adding the inclusion $\mathbb{1} \sqcup \mathbb{1} \to \mathbb{V}\mathbb{2}$ of the two endpoints into the vertical morphism to the class of cofibrations of the model structure in [16]. In particular, by making this inclusion into a cofibration, the trivial fibrations will now be given by the double functors that are surjective on objects, full on horizontal *and* vertical morphisms, and fully faithful on squares. The existence of this model structure was independently noticed at roughly the same time by Campbell [2].

As a referee pointed out, this change in the generating cofibrations requires us to enlarge the class of weak equivalences, since now the class of double functors that are both cofibrations and double biequivalences is not closed under pushouts, and therefore cannot be the class of trivial cofibrations in a model structure. Instead, we find that the weak equivalences of the desired model structure can be described as the double functors which induce a double biequivalence between fibrant replacements.

Theorem A *There is a model structure on* DblCat, *in which the trivial fibrations are the double functors which are surjective on objects*, *full on horizontal and vertical morphisms*, *and fully faithful on squares*; *and the fibrant objects are the weakly horizontally invariant double categories.*

This new model structure on DblCat takes care of the issues posed above. Namely, it is compatible with the double $(\infty, 1)$–categorical nerve construction of [15], and it is moreover monoidal, as we prove in Theorem 7.8.

Theorem B *The model structure on* DblCat *of Theorem A is monoidal with respect to Böhm's Gray tensor product.*

While the horizontal embedding $\mathbb{H}\colon 2\mathrm{Cat} \to \mathrm{DblCat}$ remains a left Quillen and homotopically fully faithful functor between Lack's model structure and our new model structure, it is not right Quillen anymore. Indeed, the horizontal double category

$\mathbb{H}\mathcal{A}$ associated to a 2–category $\mathcal{A}$ is typically not weakly horizontally invariant; see Remark 6.4.

Instead, we consider a more homotopical version of the horizontal embedding given by the functor $\mathbb{H}^{\simeq}\colon 2\mathrm{Cat} \to \mathrm{DblCat}$. It sends a 2–category $\mathcal{A}$ to the double category $\mathbb{H}^{\simeq}\mathcal{A}$, whose underlying horizontal 2–category is still $\mathcal{A}$, but whose vertical morphisms are given by the adjoint equivalences of $\mathcal{A}$. In particular, the inclusion $\mathbb{H}\mathcal{A} \to \mathbb{H}^{\simeq}\mathcal{A}$ is a weak equivalence, as shown in Theorem 6.5, and therefore exhibits $\mathbb{H}^{\simeq}\mathcal{A}$ as a fibrant replacement of $\mathbb{H}\mathcal{A}$ in the model structure for weakly horizontally invariant double categories.

In Theorem 6.6, we prove that $\mathbb{H}^{\simeq}$ is a right Quillen functor, and that the derived counit is levelwise a biequivalence in 2Cat; therefore, $\mathbb{H}^{\simeq}$ embeds the homotopy theory of 2–categories into that of weakly horizontally invariant double categories in a reflective way. Furthermore, we show in Theorem 6.8 that $\mathbb{H}^{\simeq}$ not only preserves, but also reflects weak equivalences and fibrations.

Theorem C *The adjunction*

$$\begin{tikzcd} 2\mathrm{Cat} \arrow[r, bend right, "\mathbb{H}^{\simeq}"'] \arrow[r, phantom, "\perp"] & \mathrm{DblCat} \arrow[l, bend right, "L^{\simeq}"'] \end{tikzcd}$$

is a Quillen pair between Lack's model structure on 2Cat *and the model structure on* DblCat *of Theorem A. Moreover, the derived counit of this adjunction is levelwise a biequivalence, and Lack's model structure on* 2Cat *is right-induced along* $\mathbb{H}^{\simeq}$ *from the model structure on* DblCat.

We also show in Theorem 6.1 that the identity functor from our new model structure on DblCat to the one of [16] is right Quillen and homotopically fully faithful. This implies that, unsurprisingly, the homotopy theory of weakly horizontally invariant double categories is embedded into the homotopy theory for double categories developed in [16].

To summarize, we have a triangle of right Quillen and homotopically fully faithful functors

$$\begin{tikzcd} & 2\mathrm{Cat} \arrow[dl, "\mathbb{H}^{\simeq}"'] \arrow[dr, "\mathbb{H}"] \arrow[d, phantom, "\simeq \Swarrow"] & \\ \mathrm{DblCat}_{\mathsf{whi}} \arrow[rr, "\mathrm{id}"'] & & \mathrm{DblCat} \end{tikzcd}$$

filled by a natural transformation which is levelwise a weak equivalence.

Finally, in a similar vein to Grandis's result [6, Theorem 4.4.5], we obtain a Whitehead theorem characterizing the weak equivalences with fibrant source in our model structure as the double functors which admit a pseudoinverse up to horizontal pseudonatural equivalences; see Theorem 8.1. Indeed, the weakly horizontally invariant condition on a double category is a 2–categorical analogue of the horizontally invariant condition of [6, Theorem and Definition 4.1.7].

Theorem D (Whitehead theorem for double categories) *Let $\mathbb{A}$ and $\mathbb{B}$ be double categories such that $\mathbb{A}$ is weakly horizontally invariant. Then a double functor $F\colon \mathbb{A} \to \mathbb{B}$ is a double biequivalence if and only if there is a pseudodouble functor $G\colon \mathbb{B} \to \mathbb{A}$ together with horizontal pseudonatural equivalences $\mathrm{id}_{\mathbb{A}} \simeq GF$ and $FG \simeq \mathrm{id}_{\mathbb{B}}$.*

This result implies that the weak equivalences between fibrant objects in the model structure for weakly horizontally invariant double categories resemble the biequivalences between 2–categories.

Outline

In Section 2, we recall some notations and definitions of double category theory introduced in [16]. We also introduce weakly horizontally invariant double categories and the homotopical horizontal embedding functor $\mathbb{H}^{\simeq}\colon \mathrm{2Cat} \to \mathrm{DblCat}$. Then, in Section 3, we give the main features of the model structure on DblCat. In particular, we describe the cofibrations, trivial fibrations, and weak equivalences. The proof of the existence of this model structure uses several technical results presented in Section 4 and is completed in Section 5. After establishing the model structure, we study in Section 6 its relation with the model structure on DblCat of [16] and with Lack's model structure on 2Cat. In Section 7, we prove that it is monoidal with respect to the Gray tensor product for double categories. The last section, Section 8, is devoted to the proof of the Whitehead theorem for double categories.

Acknowledgements

The authors would like to thank tslil clingman for sharing LaTeX commands which greatly simplify the drawing of diagrams. We would also like to thank Rune Haugseng, Viktoriya Ozornova and Jérôme Scherer for interesting discussions related to the subject of this paper, and Martina Rovelli, Hadrian Heine and Yuki Maehara for useful answers to our questions. We also acknowledge several discussions with Alexander Campbell regarding ours and his work on the topic.

In a first version of this paper, the authors had proposed a model structure with the same cofibrations and fibrant objects, but with double biequivalences as the class of weak equivalences. We are thankful for the careful reading of a referee, who alerted us to this mistake and provided an example showing that the proposed trivial cofibrations were not stable under pushout. Since then, the paper has gone through a thorough rewriting, and we are grateful to Jérôme Scherer and Denis-Charles Cisinski for helpful suggestions regarding the construction of our new class of weak equivalences. We are also grateful to the referee who read this second version and provided detailed feedback to improve the exposition.

During the realization of this work, Moser was supported by the Swiss National Science Foundation under the project P1ELP2_188039 and the Max Planck Institute for Mathematics.

2 Double categorical preliminaries

We introduce in this section the concepts and notations that will be used throughout this paper; for a more detailed treatment of 2–categories and double categories, we refer the reader to [11] and [6], respectively. We denote by 2Cat the category of 2–categories and 2–functors, and by DblCat the category of double categories and double functors. We will use the fact that these categories are locally presentable and hence both complete and cocomplete. For 2Cat this is given as a special case of [12, Section 4]; the statement for DblCat can be found in the proof of [5, Theorem 4.1].

To fix notation, we first recall the definition of a double category.

Definition 2.1 A *double category* $\mathbb{A}$ consists of

(i) objects $A, B, C, \dots,$

(ii) horizontal morphisms $a\colon A \to B$ with composition denoted by ba,

(iii) vertical morphisms $u\colon A \mathrel{\mathrlap{\rightarrow}\mkern1mu\bullet}\mkern-3mu\rightarrow A'$ with composition denoted by vu,

(iv) squares (or cells) $\alpha\colon (u \,{}^{a}_{b}\, v)$ of the form

$$\begin{tikzcd} A \arrow[r, "a"] \arrow[d, "u"', "\bullet" marking] \arrow[dr, phantom, "\alpha"] & B \arrow[d, "v", "\bullet" marking] \\ A' \arrow[r, "b"'] & B' \end{tikzcd}$$

with both horizontal composition along their vertical boundaries and vertical composition along their horizontal boundaries, and

(v) horizontal identities $\mathrm{id}_A\colon A \to A$ and vertical identities $e_A\colon A \to\!\!\!\!\!\bullet\,\, A$ for each object A, vertical identity squares $e_a\colon (\mathrm{id}_A\, {}^a_a\, \mathrm{id}_B)$ for each horizontal morphism $a\colon A \to B$, horizontal identity squares $\mathrm{id}_u\colon (u\, {}^{\mathrm{id}_A}_{\mathrm{id}_{A'}}\, u)$ for each vertical morphism $u\colon A \to\!\!\!\!\!\bullet\,\, A'$, and identity squares $\square_A = \mathrm{id}_{e_A} = e_{\mathrm{id}_A}$ for each object A,

such that all compositions are unital and associative, and such that the horizontal and vertical compositions of squares satisfy the interchange law.

Let us also recall the following notation.

Notation 2.2 We write $\mathbb{H}\colon \mathrm{2Cat} \to \mathrm{DblCat}$ for the horizontal embedding, which sends a 2–category $\mathcal{A}$ to the double category $\mathbb{H}\mathcal{A}$ with the same objects as $\mathcal{A}$, the morphisms of $\mathcal{A}$ as its horizontal morphisms, only trivial vertical morphisms, and the 2–morphisms of $\mathcal{A}$ as its squares. This functor has a right adjoint $\boldsymbol{H}\colon \mathrm{DblCat} \to \mathrm{2Cat}$ that sends a double category $\mathbb{A}$ to its underlying horizontal 2–category $\boldsymbol{H}\mathbb{A}$ obtained by forgetting the vertical morphisms of $\mathbb{A}$. Note that $\boldsymbol{H}\mathbb{H} = \mathrm{id}_{\mathrm{2Cat}}$.

Similarly, there is a vertical embedding $\mathbb{V}\colon \mathrm{2Cat} \to \mathrm{DblCat}$ which also admits a right adjoint $\boldsymbol{V}\colon \mathrm{DblCat} \to \mathrm{2Cat}$ extracting from a double category its underlying vertical 2–category.

2.1 Weak horizontal invertibility in a double category

We recall the notions of weak horizontal invertibility for horizontal morphisms and squares introduced in [16, Section 2]. These notions were independently developed by Grandis and Paré in [8, Section 2], where the weakly horizontally invertible squares are called *equivalence cells*, and they rely on the notion of (adjoint) equivalences in a 2–category which we now recall.

Definition 2.3 A morphism $a\colon A \to C$ in a 2–category $\mathcal{A}$ is an *equivalence* if there is a morphism $c\colon C \to A$ together with 2–isomorphisms $\eta\colon \mathrm{id}_A \Rightarrow ca$ and $\epsilon\colon ac \Rightarrow \mathrm{id}_C$ in $\mathcal{A}$. It is an *adjoint equivalence* if the 2–isomorphisms η and ϵ further satisfy the following triangle identities:

$$
\begin{tikzcd}
A \arrow[d, "a"'] \arrow[dr, equal] & \\
C \arrow[r, "c"] \arrow[dr, equal] \arrow[u, phantom, shift left=3, "\cong\Swarrow\eta"] & A \arrow[d, "a"] \\
& C \arrow[u, phantom, shift right=3, "\cong\Swarrow\epsilon"]
\end{tikzcd}
=
\begin{tikzcd}
A \arrow[d, "a"] \\
C
\end{tikzcd}
\qquad\qquad
\begin{tikzcd}
C \arrow[r, "c"] \arrow[dr, equal] & A \arrow[d, "a"] \arrow[dr, equal] & \\
& C \arrow[r, "c"'] \arrow[u, phantom, shift right=3, "\cong\Swarrow\epsilon"] & A \arrow[u, phantom, shift right=3, "\cong\Swarrow\eta"]
\end{tikzcd}
=
C \xrightarrow{c} A
$$

Definition 2.4 A horizontal morphism $a\colon A \to B$ in a double category $\mathbb{A}$ is a *horizontal* (*adjoint*) *equivalence* if it is an (adjoint) equivalence in the underlying horizontal 2–category $\boldsymbol{H}\mathbb{A}$. We write $a\colon A \xrightarrow{\simeq} B$.

For the next definition, we remind the reader that the category DblCat is cartesian closed, and we denote its internal hom double category by $[-,-]$. In particular, we consider the functor $\boldsymbol{H}[\mathbb{V}2,-]\colon \mathrm{DblCat} \to 2\mathrm{Cat}$, where $\mathbb{V}2$ is the free double category on a vertical morphism. See [16, Definition 2.11] for an explicit description.

Definition 2.5 A square $\alpha\colon (u \,{}^{a}_{a'}\, w)$ in a double category $\mathbb{A}$ is *weakly horizontally invertible* if it is an equivalence in the 2–category $\boldsymbol{H}[\mathbb{V}2,\mathbb{A}]$. In other words, if there is a square $\gamma\colon (w \,{}^{c}_{c'}\, u)$ in $\mathbb{A}$ together with four vertically invertible squares η, η', ϵ, and ϵ' as in the following pasting equalities:

$$
\begin{tikzcd}
A \arrow[rr, equal] \arrow[d, equal] & \arrow[d, phantom, "\eta\ \cong"] & A \arrow[d, equal] \\
A \arrow[r, "a"] \arrow[d, "u"'] \arrow[r, phantom, shift right=6, "\alpha"] & C \arrow[r, "c"] \arrow[d, "w"'] \arrow[r, phantom, shift right=6, "\gamma"] & A \arrow[d, "u"] \\
A' \arrow[r, "a'"'] & C' \arrow[r, "c'"'] & A'
\end{tikzcd}
=
\begin{tikzcd}
A \arrow[rr, equal] \arrow[d, "u"'] & \arrow[d, phantom, "\mathrm{id}_u"] & A \arrow[d, "u"] \\
A' \arrow[rr, equal] \arrow[d, equal] & \arrow[d, phantom, "\eta'\ \cong"] & A' \arrow[d, equal] \\
A' \arrow[r, "a'"'] & C' \arrow[r, "c'"'] & A'
\end{tikzcd}
$$

$$
\begin{tikzcd}
C \arrow[r, "c"] \arrow[d, equal] & A \arrow[r, "a"] \arrow[d, phantom, "\epsilon\ \cong"] & C \arrow[d, equal] \\
C \arrow[rr, equal] \arrow[d, "w"'] & \arrow[d, phantom, "\mathrm{id}_w"] & C \arrow[d, "w"] \\
C' \arrow[rr, equal] & & C'
\end{tikzcd}
=
\begin{tikzcd}
C \arrow[r, "c"] \arrow[d, "w"'] \arrow[r, phantom, shift right=6, "\gamma"] & A \arrow[r, "a"] \arrow[d, "u"'] \arrow[r, phantom, shift right=6, "\alpha"] & C \arrow[d, "w"] \\
C' \arrow[r, "c'"'] \arrow[d, equal] & A' \arrow[r, "a'"'] \arrow[d, phantom, "\epsilon'\ \cong"] & C' \arrow[d, equal] \\
C' \arrow[rr, equal] & & C'
\end{tikzcd}
$$

We call γ a *weak inverse* of α, and we denote weakly horizontally invertible squares by decorating the square with the symbol $\simeq$.

Remark 2.6 In particular, the horizontal boundaries a and a' of a weakly horizontally invertible square α as above are horizontal equivalences witnessed by the data (a, c, η, ϵ) and $(a', c', \eta', \epsilon')$. We call them the *horizontal equivalence data* of α. Moreover, if (a, c, η, ϵ) and $(a', c', \eta', \epsilon')$ are both horizontal adjoint equivalences, we call them the *horizontal adjoint equivalence data* of α.

Remark 2.7 A horizontal equivalence can always be promoted to a horizontal *adjoint* equivalence, since the corresponding result holds for 2–categories; see for example [19, Lemma 2.1.11]. Similarly, a weakly horizontally invertible square can always be promoted to one with horizontal *adjoint* equivalence data.

The next result ensures that the weak inverse of a weakly horizontally invertible square is unique with respect to fixed horizontal adjoint equivalences.

Lemma 2.8 [15, Lemma A.1.1] *Given a weakly horizontally invertible square $\alpha\colon (u \,{}^{a}_{a'}\, w)$ and two horizontal adjoint equivalences (a, c, η, ϵ) and $(a', c', \eta', \epsilon')$ in a double category $\mathbb{A}$, there is a unique weak inverse $\gamma\colon (w \,{}^{c}_{c'}\, u)$ of α with respect to these horizontal adjoint equivalences.*

2.2 Double biequivalences and weakly horizontally invariant double categories

The weak equivalences of the desired model structure for double categories rely on *double biequivalences*, which are the weak equivalences of the model structure on DblCat constructed in [16].

Definition 2.9 Let $\mathbb{A}$ and $\mathbb{B}$ be double categories. A double functor $F\colon \mathbb{A} \to \mathbb{B}$ is a (*horizontal*) *double biequivalence* if it is:

(db1) (Horizontally) biessentially surjective on objects: for every object $B \in \mathbb{B}$, there is an object $A \in \mathbb{A}$ together with a horizontal equivalence $B \xrightarrow{\simeq} FA$ in $\mathbb{B}$.

(db2) Essentially full on horizontal morphisms: for every pair of objects $A, C \in \mathbb{A}$ and every horizontal morphism $b\colon FA \to FC$ in $\mathbb{B}$, there is a horizontal morphism $a\colon A \to C$ together with a vertically invertible square in $\mathbb{B}$ of the form

$$\begin{tikzcd} FA \arrow[r, "b"] \arrow[d, equal] \arrow[dr, phantom, "\cong"] & FC \arrow[d, equal] \\ FA \arrow[r, "Fa"'] & FC \end{tikzcd}$$

(db3) (Horizontally) biessentially surjective on vertical morphisms: for every vertical morphism $v\colon B \to B'$ in $\mathbb{B}$, there is a vertical morphism $u\colon A \to A'$ in $\mathbb{A}$ together with a weakly horizontally invertible square in $\mathbb{B}$ of the form

$$\begin{tikzcd} B \arrow[r, "\simeq"] \arrow[d, "v"'] \arrow[dr, phantom, "\simeq"] & FA \arrow[d, "Fu"] \\ B' \arrow[r, "\simeq"'] & FA' \end{tikzcd}$$

(db4) Fully faithful on squares: for every pair of horizontal morphisms $a\colon A \to C$ and $a'\colon A' \to C'$ in $\mathbb{A}$, every pair of vertical morphisms $u\colon A \to A'$ and

$w\colon C \xrightarrow{\bullet} C'$ in $\mathbb{A}$, and every square β in $\mathbb{B}$ as depicted below left, there is a unique square α in $\mathbb{A}$ as depicted below right such that $\beta = F\alpha$:

$$\begin{tikzcd} FA \arrow[r, "Fa"] \arrow[d, "Fu"', "\bullet" marking] \arrow[dr, phantom, "\beta"] & FC \arrow[d, "Fw", "\bullet" marking] \\ FA' \arrow[r, "Fa'"'] & FC' \end{tikzcd} \qquad \begin{tikzcd} A \arrow[r, "a"] \arrow[d, "u"', "\bullet" marking] \arrow[dr, phantom, "\alpha"] & C \arrow[d, "w", "\bullet" marking] \\ A' \arrow[r, "a'"'] & C' \end{tikzcd}$$

We also introduce the notion of *weakly horizontally invariant* double categories, which will form the class of fibrant objects in our model structure. This is a 2–categorical analogue of the notion of *horizontally invariant* double categories, introduced by Grandis and Paré in [7, Section 2.4] as the double categories whose vertical morphisms are transferable along horizontal isomorphisms.

Definition 2.10 A double category $\mathbb{A}$ is *weakly horizontally invariant* if, for every diagram in $\mathbb{A}$ as depicted below left, where a and a' are horizontal equivalences, there is a vertical morphism $u\colon A \xrightarrow{\bullet} A'$ together with a weakly horizontally invertible square in $\mathbb{A}$ as depicted below right:

$$\begin{tikzcd} A \arrow[r, "a", "\simeq"'] & C \arrow[d, "w", "\bullet" marking] \\ A' \arrow[r, "\simeq", "a'"'] & C' \end{tikzcd} \qquad \begin{tikzcd} A \arrow[r, "a", "\simeq"'] \arrow[d, "u"', "\bullet" marking] \arrow[dr, phantom, "\simeq"] & C \arrow[d, "w", "\bullet" marking] \\ A' \arrow[r, "\simeq", "a'"'] & C' \end{tikzcd}$$

Example 2.11 One can easily check that the (flat) double category $\mathbb{R}\mathrm{elSet}$ of relations of sets is weakly horizontally invariant. More relevantly, this class of double categories also contains the double categories of quintets $\mathbb{Q}\mathcal{A}$ and of adjunctions $\mathbb{A}\mathrm{dj}\mathcal{A}$ built from any 2–category $\mathcal{A}$. A precise description of these double categories can be found in [6, Section 3.1]; in fact, the reader may check that all examples presented in that section are weakly horizontally invariant.

Remark 2.12 The horizontal double category $\mathbb{H}\mathcal{A}$ associated to a 2–category $\mathcal{A}$ is typically not weakly horizontally invariant. To see this, consider the horizontal double category $\mathbb{H}E_{\mathrm{adj}}$, where E_{adj} denotes the free-living adjoint equivalence. Then there is no vertical morphism in $\mathbb{H}E_{\mathrm{adj}}$ filling the diagram

$$\begin{tikzcd} 0 \arrow[r, "\simeq"] & 1 \arrow[d, equal, "\bullet" marking] \\ 1 \arrow[r, equal] & 1 \end{tikzcd}$$

as $\mathbb{H}E_{\mathrm{adj}}$ only contains trivial vertical morphisms, which shows that it is not weakly horizontally invariant.

Using the same reasoning, one can show that the horizontal double category $\mathbb{H}\mathcal{A}$ associated to a 2–category $\mathcal{A}$ is weakly horizontally invariant if and only if there is no adjoint equivalence in $\mathcal{A}$.

2.3 The homotopical horizontal embedding

Since all 2–categories are fibrant, Remark 2.12 implies that the functor

$$\mathbb{H}\colon 2\mathrm{Cat} \to \mathrm{DblCat}$$

is not right Quillen with respect to the desired model structure for weakly horizontally invariant double categories. Instead, we need to consider a more homotopical version of the horizontal embedding $\mathbb{H}$, which provides a levelwise fibrant replacement for $\mathbb{H}$.

Definition 2.13 The *homotopical horizontal embedding* is defined as the functor $\mathbb{H}^{\simeq}\colon 2\mathrm{Cat} \to \mathrm{DblCat}$ that sends a 2–category $\mathcal{A}$ to the double category $\mathbb{H}^{\simeq}\mathcal{A}$ having the same objects as $\mathcal{A}$, the morphisms of $\mathcal{A}$ as horizontal morphisms, one vertical morphism for each adjoint equivalence $(u, u^\sharp, \eta, \epsilon)$ in $\mathcal{A}$, and squares

$$\begin{tikzcd} A \arrow[r, "a"] \arrow[d, "{(u,u^\sharp,\eta,\epsilon)}"', "\bullet" marking] \arrow[dr, phantom, "\alpha"] & C \arrow[d, "{(w,w^\sharp,\eta',\epsilon')}", "\bullet" marking] \\ A' \arrow[r, "a'"'] & C' \end{tikzcd}$$

given by the 2–morphisms $\alpha\colon wa \Rightarrow a'u$ in $\mathcal{A}$. Compositions are induced by compositions of morphisms and 2–morphisms in $\mathcal{A}$. Although a vertical morphism always contains the whole data of an adjoint equivalence, we often denote it by its left adjoint u.

Remark 2.14 Every vertical morphism in the double category $\mathbb{H}^{\simeq}\mathcal{A}$ is a *vertical equivalence*, ie an equivalence in the underlying vertical 2–category.

The functor $\mathbb{H}^{\simeq}$ admits a left adjoint.

Proposition 2.15 *The functor $\mathbb{H}^{\simeq}$ is part of an adjunction*

$$\begin{tikzcd} \mathrm{DblCat} \arrow[r, bend left, "L^{\simeq}"] \arrow[r, phantom, "\perp"] & 2\mathrm{Cat} \arrow[l, bend left, "\mathbb{H}^{\simeq}"] \end{tikzcd}$$

Proof Consider the full subcategory $\Delta_{\leq 3}$ of the simplex category Δ on the objects [0], [1], [2], and [3]. Then the category DblCat can be seen as the full subcategory of $\mathrm{Set}^{\Delta_{\leq 3}^{\mathrm{op}}\times\Delta_{\leq 3}^{\mathrm{op}}}$ on the objects $X\in\mathrm{Set}^{\Delta_{\leq 3}^{\mathrm{op}}\times\Delta_{\leq 3}^{\mathrm{op}}}$ whose component sets $X_{i,j}$ for $2\leq i,\ j\leq 3$ are obtained as certain limits over the sets $X_{0,0}$ of objects, $X_{1,0}$ of horizontal morphisms, $X_{0,1}$ of vertical morphisms, and $X_{1,1}$ of squares; eg $X_{2,1}\cong X_{1,1}\times_{X_{0,1}}X_{1,1}$.

The strategy of the proof is to show that there is an adjunction

$$\begin{tikzcd} \mathrm{Set}^{\Delta_{\leq 3}^{\mathrm{op}}\times\Delta_{\leq 3}^{\mathrm{op}}} \arrow[r, bend left, "L^{\simeq}"] \arrow[r, phantom, "\perp"] & 2\mathrm{Cat} \arrow[l, bend left, "R"] \end{tikzcd}$$

such that the right adjoint R lands in $\mathrm{DblCat}\subseteq\mathrm{Set}^{\Delta_{\leq 3}^{\mathrm{op}}\times\Delta_{\leq 3}^{\mathrm{op}}}$ and agrees with the functor $\mathbb{H}^{\simeq}$; then the adjunction above must restrict to an adjunction $L^{\simeq}\dashv\mathbb{H}^{\simeq}$, as desired. We now proceed to prove these claims.

We define a functor $\ell^{\simeq}\colon\Delta_{\leq 3}\times\Delta_{\leq 3}\to 2\mathrm{Cat}$ by giving its values on the subcategory spanned by $[0,0]$, $[1,0]$, $[0,1]$, and $[1,1]$ and setting its values on $[i,j]$ for $2\leq i,\ j\leq 3$ in such a way that ℓ preserves colimits; eg $\ell^{\simeq}[2,1]=\ell^{\simeq}[1,1]\amalg_{\ell^{\simeq}[0,1]}\ell^{\simeq}[1,1]$. We set $\ell^{\simeq}[0,0]=\mathbb{1}$, $\ell^{\simeq}[1,0]=\mathbb{2}$, $\ell^{\simeq}[0,1]=E_{\mathrm{adj}}$, and $\ell^{\simeq}[1,1]=\mathcal{A}$, where E_{adj} is the 2–category containing an adjoint equivalence, and $\mathcal{A}$ is the 2–category generated by morphisms, adjoint equivalences, and 2–morphisms

$$\begin{tikzcd} A \arrow[r] \arrow[d, "\wr"'] \arrow[dr, phantom, "\overset{\alpha}{\swarrow}"] & C \arrow[d, "\wr"] \\ A' \arrow[r] & C' \end{tikzcd}$$

with the obvious images of the morphisms in $\Delta_{\leq 3}\times\Delta_{\leq 3}$ between these objects.

By considering the left Kan extension along the Yoneda embedding

$$\begin{tikzcd} \Delta_{\leq 3}\times\Delta_{\leq 3} \arrow[r, "\ell^{\simeq}"] \arrow[d, "y"'] & 2\mathrm{Cat} \arrow[dl, dashed, bend left, "R"] \\ \mathrm{Set}^{\Delta_{\leq 3}^{\mathrm{op}}\times\Delta_{\leq 3}^{\mathrm{op}}} \arrow[ur, dashed, "L^{\simeq}"] \arrow[ur, phantom, "\perp"] & \end{tikzcd}$$

we obtain a functor $L^{\simeq}$, which admits a right adjoint $R\colon 2\mathrm{Cat}\to\mathrm{Set}^{\Delta_{\leq 3}^{\mathrm{op}}\times\Delta_{\leq 3}^{\mathrm{op}}}$ given by $R(\mathcal{B})_{i,j}=2\mathrm{Cat}(\ell^{\simeq}[i,j],\mathcal{B})$, for all $0\leq i,\ j\leq 3$ and all 2–categories $\mathcal{B}$.

By definition of $\ell^{\simeq}$, the image of R lands in the subcategory DblCat. In particular, the adjunction $L^{\simeq} \dashv R$ restricts to an adjunction

$$\begin{tikzcd} \mathrm{DblCat} \arrow[r, bend left, "L^{\simeq}"] \arrow[r, phantom, "\perp"] & 2\mathrm{Cat} \arrow[l, bend left, "R"] \end{tikzcd}$$

Note that the representables at $[0,0]$, $[1,0]$, $[0,1]$, and $[1,1]$ in $\mathrm{Set}^{\Delta^{\mathrm{op}}_{\le 3}\times\Delta^{\mathrm{op}}_{\le 3}}$ coincide with the double categories $\mathbb{1}$, $\mathbb{H}\mathbb{2}$, $\mathbb{V}\mathbb{2}$, and $\mathbb{H}\mathbb{2}\times\mathbb{V}\mathbb{2}$, so that we have

$$L^{\simeq}(\mathbb{1}) = \mathbb{1}, \quad L^{\simeq}(\mathbb{H}\mathbb{2}) = \mathbb{2}, \quad L^{\simeq}(\mathbb{V}\mathbb{2}) = E_{\mathrm{adj}}, \quad L^{\simeq}(\mathbb{H}\mathbb{2}\times\mathbb{V}\mathbb{2}) = \mathcal{A}.$$

It remains to show that $R = \mathbb{H}^{\simeq}$. For this, it is enough to show that the sets of objects, horizontal morphisms, vertical morphisms, and squares of $R\mathcal{B}$ and $\mathbb{H}^{\simeq}\mathcal{B}$ coincide, for every 2–category $\mathcal{B}$. This is indeed the case, as for $\mathbb{A} \in \{\mathbb{1}, \mathbb{H}\mathbb{2}, \mathbb{V}\mathbb{2}, \mathbb{H}\mathbb{2}\times\mathbb{V}\mathbb{2}\}$,

$$\mathrm{DblCat}(\mathbb{A}, R\mathcal{B}) \cong 2\mathrm{Cat}(L^{\simeq}\mathbb{A}, \mathcal{B}) \cong \mathrm{DblCat}(\mathbb{A}, \mathbb{H}^{\simeq}\mathcal{B}),$$

where the first isomorphism holds by the universal property of the adjunction $L^{\simeq} \dashv R$ and the second by the definition of $\mathbb{H}^{\simeq}\mathcal{B}$. □

Remark 2.16 One can show that $L^{\simeq}$ admits the following, more explicit, description. Given a double category $\mathbb{A}$, $L^{\simeq}\mathbb{A}$ is the 2–category with the same objects as $\mathbb{A}$, a morphism for each horizontal morphism in $\mathbb{A}$, and a morphism for each vertical morphism in $\mathbb{A}$ which we formally make into an adjoint equivalence; ie we also add a formal inverse morphism, and the two necessary 2–isomorphisms. Aside from these formal 2–morphisms added to create the adjoint equivalences, we also have a 2–morphism $u'a \Rightarrow cu$ for each square in $\mathbb{A}$ of the form $\alpha\colon (u \,{}^{a}_{c}\, u')$.

Furthermore, the composite in $L^{\simeq}\mathbb{A}$ of two morphisms coming from horizontal morphisms in $\mathbb{A}$ is given by their composite in $\mathbb{A}$, and the composite in $L^{\simeq}\mathbb{A}$ of two adjoint equivalences coming from vertical morphisms in $\mathbb{A}$ is given by the adjoint equivalence induced by their composite in $\mathbb{A}$, while two morphisms with one coming from a horizontal morphism and one coming from a vertical morphism compose freely. Similar holds for the 2–morphisms.

Remark 2.17 The functor $\mathbb{H}^{\simeq}$ is not a left adjoint since it does not preserve colimits. To see this, consider the span of 2–categories $\mathcal{B} \leftarrow \mathcal{A} \rightarrow \mathcal{C}$. We set $\mathcal{A}$ to be the 2–category with two objects 0 and 1, and freely generated by two morphisms $f\colon 0 \to 1$ and $g\colon 1 \to 0$ and two 2–morphisms $\eta\colon \mathrm{id}_0 \Rightarrow gf$ and $\epsilon\colon fg \Rightarrow \mathrm{id}_1$. Then let $\mathcal{B}$ be the category obtained from $\mathcal{A}$ by inverting the 2–morphism η, and $\mathcal{C}$ be the category obtained from $\mathcal{A}$ by inverting the 2–morphism ϵ. The pushout $\mathcal{B} \sqcup_{\mathcal{A}} \mathcal{C}$ contains an equivalence (f, g, η, ϵ), and hence the double category $\mathbb{H}^{\simeq}(\mathcal{B} \sqcup_{\mathcal{A}} \mathcal{C})$ contains a vertical morphism induced by the corresponding adjoint equivalence. However, the double categories $\mathbb{H}^{\simeq}\mathcal{A}$, $\mathbb{H}^{\simeq}\mathcal{B}$, and $\mathbb{H}^{\simeq}\mathcal{C}$ do not have nontrivial vertical morphisms since there are no equivalences in $\mathcal{A}$, $\mathcal{B}$, or $\mathcal{C}$, and hence their pushout $\mathbb{H}^{\simeq}\mathcal{A} \sqcup_{\mathbb{H}^{\simeq}\mathcal{B}} \mathbb{H}^{\simeq}\mathcal{C}$ does not contain nontrivial vertical morphisms. This shows that $\mathbb{H}^{\simeq}$ does not preserve pushouts.

3 The model structure

Just as there are nerve constructions embedding categories into $(\infty, 1)$–categories, and 2–categories into $(\infty, 2)$–categories, in [15] the first-named author constructs a double categorical nerve, embedding double categories into double $(\infty, 1)$–categories. As the latter admit a natural model structure when considered as double Segal spaces, we expect to have a model structure on DblCat making this nerve into a right Quillen functor.

Since the double categories whose nerve is fibrant are precisely the weakly horizontally invariant ones, this suggests that such a model structure should have the weakly horizontally invariant double categories as its class of fibrant objects. Moreover, since the cofibrations of the model structure for double $(\infty, 1)$–categories are the monomorphisms and the inclusion $\mathbb{1} \sqcup \mathbb{1} \to \mathbb{V}\mathbb{2}$ is the image of a monomorphism under the left adjoint of the nerve, it should be added to the class of cofibrations of the model structure on DblCat of [16]; this allows us to characterize the trivial fibrations as the double functors which are surjective on objects, full on horizontal and vertical morphisms, and fully faithful on squares.

A first attempt to keep the double biequivalences — which were shown by the authors to be the class of weak equivalences in a model structure on DblCat in [16] — as the weak equivalences of this new model structure proves unsuccessful. Indeed, the resulting class of trivial cofibrations would not be closed under pushouts; the double functor $j_{\mathbb{A}}$ of Example 3.23 is an example of a pushout of such a trivial cofibration (namely, of the inclusion $\mathbb{H}E_{\mathrm{adj}} \to \mathbb{H}^{\simeq}E_{\mathrm{adj}}$) that is not a double biequivalence.

Instead, we identify the weak equivalences as the double functors which induce a double biequivalence between weakly horizontally invariant replacements.

Since many technical results, presented in Section 4, are needed to prove the existence of such a model structure, its proof is delayed to Section 5.

3.1 Weak factorization systems

We first recall the definition and basic results about weak factorization systems which will be used throughout the paper.

Notation 3.1 Let $\mathcal{M}$ be a category and $\mathcal{C}$ be a class of morphisms in $\mathcal{M}$. We write ${}^{\boxslash}\mathcal{C}$ (resp. $\mathcal{C}^{\boxslash}$) for the class of morphisms in $\mathcal{M}$ that have the left (resp. right) lifting property with respect to all morphisms in $\mathcal{C}$.

Definition 3.2 A *weak factorization system* $(\mathcal{L},\mathcal{R})$ in a category $\mathcal{M}$ consists of two classes $\mathcal{L}$ and $\mathcal{R}$ of morphisms in $\mathcal{M}$ such that $\mathcal{L} = {}^{\boxslash}\mathcal{R}$ and $\mathcal{R} = \mathcal{L}^{\boxslash}$, and every morphism f in $\mathcal{M}$ can be factored as $f = rl$ with $l \in \mathcal{L}$ and $r \in \mathcal{R}$.

Remark 3.3 Recall that given a weak factorization system $(\mathcal{L},\mathcal{R})$, both classes contain isomorphisms and are closed under composition and retracts. Furthermore, the left class $\mathcal{L}$ is closed under coproducts, pushouts, and transfinite compositions, and the right class $\mathcal{R}$ is closed under products and pullbacks, as explained for example in [18, Lemma 11.1.4] and the comment immediately below that lemma.

The following argument will be useful when proving that a certain map belongs to the left or right class of a weak factorization system; its proof can be found, for example, in [10, Lemma 1.1.9].

Remark 3.4 (retract argument) Consider a factorization $f = rl$ of a map f in a category $\mathcal{M}$. If f has the left lifting property with respect to r, then f is a retract of l. Dually, if f has the right lifting property with respect to l, then f is a retract of r.

Weak factorization systems are often generated by a set. To introduce this notion, we recall the following terminology.

Notation 3.5 Let $\mathcal{I}$ be a set of morphisms in a cocomplete category $\mathcal{M}$. Then a morphism in $\mathcal{M}$ is:

(i) $\mathcal{I}$*–injective* if it has the right lifting property with respect to every morphism in $\mathcal{I}$. The class of all such morphisms is denoted by $\mathcal{I}\text{–inj} := \mathcal{I}^{\boxslash}$.

(ii) An *$\mathcal{I}$–cofibration* if it has the left lifting property with respect to every $\mathcal{I}$–injective morphism. The class of all such morphisms is denoted by $\mathcal{I}-\mathrm{cof} := {}^{\boxslash}(\mathcal{I}^{\boxslash})$.

(iii) A *relative $\mathcal{I}$–cell complex* if it is a transfinite composition of pushouts of morphisms in $\mathcal{I}$. The class of all such morphisms is denoted by $\mathcal{I}-\mathrm{cell}$.

Remark 3.6 Recall that every $\mathcal{I}$–cofibration can be obtained as a retract of a relative $\mathcal{I}$–cell complex with the same source, and that in a locally presentable category $\mathcal{M}$, the pair $(\mathcal{I}-\mathrm{cof}, \mathcal{I}-\mathrm{inj})$ forms a weak factorization system, for any set of morphisms $\mathcal{I}$ in $\mathcal{M}$.

Definition 3.7 A weak factorization system $(\mathcal{L}, \mathcal{R})$ in $\mathcal{M}$ is said to be *generated by a set $\mathcal{I}$ of morphisms* if $(\mathcal{L}, \mathcal{R}) = (\mathcal{I}\text{–cof}, \mathcal{I}\text{–inj})$.

3.2 Trivial fibrations, cofibrations, and cofibrant objects

We now identify a set $\mathcal{I}_w$ of double functors such that the $\mathcal{I}_w$–injective morphisms are precisely the trivial fibrations we seek.

Notation 3.8 We denote by $\mathbb{1}$ the terminal double category, by $\mathbb{2}$ the free (2–)category on a morphism, by $\mathbb{S} = \mathbb{H}\mathbb{2} \times \mathbb{V}\mathbb{2}$ the free double category on a square, by $\delta\mathbb{S}$ its boundary, and by $\mathbb{S}_2$ the free double category on two squares with the same boundary.

Let $\mathcal{I}_w$ denote the set containing the following double functors:

(i) the unique map $I_1 \colon \varnothing \to \mathbb{1}$,

(ii) the inclusion $I_2 \colon \mathbb{1} \sqcup \mathbb{1} \to \mathbb{H}\mathbb{2}$,

(iii) the inclusion $I_3 \colon \mathbb{1} \sqcup \mathbb{1} \to \mathbb{V}\mathbb{2}$,

(iv) the inclusion $I_4 \colon \delta\mathbb{S} \to \mathbb{S}$,

(v) the double functor $I_5 \colon \mathbb{S}_2 \to \mathbb{S}$ sending the two nontrivial squares in $\mathbb{S}_2$ to the nontrivial square of $\mathbb{S}$.

Proposition 3.9 *A double functor $F \colon \mathbb{A} \to \mathbb{B}$ is in $\mathcal{I}_w$–*inj *if and only if it is surjective on objects, full on horizontal and vertical morphisms, and fully faithful on squares.*

Proof This is obtained directly from a close inspection of the right lifting properties with respect to the double functors in $\mathcal{I}_w$. □

Remark 3.10 It is straightforward to check that any double functor in $\mathcal{I}_w$–inj is a double biequivalence; see Definition 2.9.

The class of $\mathcal{I}_w$–cofibrations admits a nice characterization in terms of their underlying horizontal and vertical functors. We denote by $U\colon 2\mathrm{Cat}\to\mathrm{Cat}$ the functor sending a 2–category to its underlying category, where Cat is the category of categories and functors.

Theorem 3.11 *A double functor $F\colon \mathbb{A}\to\mathbb{B}$ is in $\mathcal{I}_w$–cof if and only if its underlying horizontal and vertical functors, $U\mathbf{H}F$ and $U\mathbf{V}F$, have the left lifting property with respect to surjective on objects and full functors.*

Proof The proof works as in [16, Proposition 4.7], with the evident modifications for the vertical direction. □

Remark 3.12 An equivalent characterization of functors which have the left lifting property with respect to surjective on objects and full functors can be found in [13, Corollary 4.12]. These are the functors $F\colon\mathcal{A}\to\mathcal{B}$ which are injective on objects, faithful, and such that there are functors $I\colon\mathcal{B}\to\mathcal{C}$ and $R\colon\mathcal{C}\to\mathcal{B}$ with $RI=\mathrm{id}_{\mathcal{B}}$, where the category $\mathcal{C}$ is obtained from the image of F by freely adjoining objects and then freely adjoining morphisms between specified objects.

In particular, we can see that a double functor in $\mathcal{I}_w$–cof is injective on objects, and faithful on horizontal and vertical morphisms.

Using the characterization mentioned in Remark 3.12, we can see that the cofibrant objects in the desired model structure are precisely the double categories whose underlying horizontal and vertical categories are free.

Corollary 3.13 *A double category $\mathbb{A}$ is such that the unique map $\varnothing\to\mathbb{A}$ is in $\mathcal{I}_w$–cof if and only if its underlying horizontal and vertical categories $U\mathbf{H}\mathbb{A}$ and $U\mathbf{V}\mathbb{A}$ are free.*

Proof The proof works as in [16, Proposition 4.9], with the evident modifications for the vertical direction. □

3.3 Weakly horizontally invariant replacements and weak equivalences

Our next goal is to introduce the class of weak equivalences; these will be the double functors that induce a double biequivalence between weakly horizontally invariant replacements. To construct a weakly horizontally invariant double category from a double category $\mathbb{A}$, we attach $\mathbb{H}^{\simeq}E_{\mathrm{adj}}$–data freely to every horizontal adjoint equivalence in $\mathbb{A}$, where the 2–category E_{adj} is the free-living adjoint equivalence $\{0 \xrightarrow{\simeq} 1\}$. Since this will be a key notion throughout the paper, let us first describe the double category $\mathbb{H}^{\simeq} E_{\mathrm{adj}}$.

Description 3.14 The double category $\mathbb{H}^{\simeq} E_{\mathrm{adj}}$ is generated by the data of a horizontal adjoint equivalence (f, g, η, ϵ), vertical morphisms u and v, and weakly horizontally invertible squares α and γ,

$$\begin{tikzcd} 0 \arrow[r, "f", "\simeq"'] \arrow[d, "u"', "\bullet" marking] \arrow[dr, phantom, "\alpha \simeq"] & 1 \arrow[d, equal, "\bullet" marking] \\ 1 \arrow[r, equal] & 1 \end{tikzcd} \qquad \begin{tikzcd} 1 \arrow[r, "g", "\simeq"'] \arrow[d, "v"', "\bullet" marking] \arrow[dr, phantom, "\gamma \simeq"] & 0 \arrow[d, equal, "\bullet" marking] \\ 0 \arrow[r, equal] & 0 \end{tikzcd}$$

where u and v are induced by the adjoint equivalences (f, g, η, ϵ) and $(g, f, \epsilon^{-1}, \eta^{-1})$, respectively, and the squares α and γ are induced by the identity 2–morphisms at f and g, respectively.

In particular, note that $\mathbb{H}^{\simeq} E_{\mathrm{adj}}$ also contains vertically invertible squares η and ϵ given by the unit and counit of the adjoint equivalence (f, g, η, ϵ), as well as horizontally invertible squares α' and γ' which are the weak inverses of α and γ, respectively.

Furthermore, we can compose these to form weakly horizontally invertible squares β and δ,

$$\begin{tikzcd} 0 \arrow[r, equal] \arrow[d, equal, "\bullet" marking] \arrow[dr, phantom, "\beta \simeq"] & 0 \arrow[d, "u", "\bullet" marking] \\ 0 \arrow[r, "\simeq", "f"'] & 1 \end{tikzcd} = \begin{tikzcd} 0 \arrow[rr, equal] \arrow[d, equal, "\bullet" marking] \arrow[drr, phantom, "\eta \ \cong"] & & 0 \arrow[d, equal, "\bullet" marking] \\ 0 \arrow[r, "f", "\simeq"'] \arrow[d, equal, "\bullet" marking] \arrow[dr, phantom, "e_f"] & 1 \arrow[r, "g", "\simeq"'] \arrow[d, equal, "\bullet" marking] \arrow[dr, phantom, "\alpha' \simeq"] & 0 \arrow[d, "u", "\bullet" marking] \\ 0 \arrow[r, "\simeq", "f"'] & 1 \arrow[r, equal] & 1 \end{tikzcd} \qquad \begin{tikzcd} 1 \arrow[r, equal] \arrow[d, equal, "\bullet" marking] \arrow[dr, phantom, "\delta \simeq"] & 1 \arrow[d, "v", "\bullet" marking] \\ 1 \arrow[r, "\simeq", "g"'] & 0 \end{tikzcd} = \begin{tikzcd} 1 \arrow[rr, equal] \arrow[d, equal, "\bullet" marking] \arrow[drr, phantom, "\epsilon^{-1} \cong"] & & 1 \arrow[d, equal, "\bullet" marking] \\ 1 \arrow[r, "g", "\simeq"'] \arrow[d, equal, "\bullet" marking] \arrow[dr, phantom, "e_g"] & 0 \arrow[r, "f", "\simeq"'] \arrow[d, equal, "\bullet" marking] \arrow[dr, phantom, "\gamma' \simeq"] & 1 \arrow[d, "v", "\bullet" marking] \\ 1 \arrow[r, "\simeq", "g"'] & 0 \arrow[r, equal] & 0 \end{tikzcd}$$

and we can similarly construct their weak inverses β' and δ'.

Note that the horizontal composite of β with α is the vertical identity square e_f at f, and the vertical composite of β with α is the horizontal identity square id_u at u. In other words, this says that (f, u, α, β) is the data of an *orthogonal companion pair*; see [6, Section 4.1.1]. On the other hand, the horizontal composite of α' with β' is the vertical identity square e_g at g, and the vertical composite of β' with α' is the horizontal identity square id_u at u. In other words, this says that (g, u, α', β') is the data of an *orthogonal adjoint pair*; see [6, Section 4.1.2]. Similarly, (g, v, γ, δ) is the data of an orthogonal companion pair, and (f, v, γ', δ') is the data of an orthogonal adjoint pair.

Finally, one can check that the vertical morphisms (u, v) form a vertical adjoint equivalence, ie an adjoint equivalence in the underlying vertical 2–category $\boldsymbol{V}\mathbb{H}^{\simeq}E_{\mathrm{adj}}$, with unit η' given by the vertical composite of β with γ', and counit ϵ' given by the vertical composite of δ' with α. In particular, all the squares in $\mathbb{H}^{\simeq}E_{\mathrm{adj}}$ are also *weakly vertically invertible* — the transposed notion of weakly horizontally invertible — with vertical weak inverses given by the obvious squares.

Notation 3.15 There is an inclusion $J_4\colon \mathbb{H}E_{\mathrm{adj}} \to \mathbb{H}^{\simeq}E_{\mathrm{adj}}$ which sends the horizontal adjoint equivalence in $\mathbb{H}E_{\mathrm{adj}}$ to the horizontal adjoint equivalence (f, g, η, ϵ) in $\mathbb{H}^{\simeq}E_{\mathrm{adj}}$.

Remark 3.16 By uniqueness of weak inverses with respect to fixed horizontal adjoint equivalence data of Lemma 2.8, we can see that a double functor $G\colon \mathbb{H}^{\simeq}E_{\mathrm{adj}} \to \mathbb{A}$ is completely determined by its value on the horizontal adjoint equivalence (f, g, η, ϵ) and the squares α, γ in $\mathbb{H}^{\simeq}E_{\mathrm{adj}}$.

We are now ready to construct a functorial weakly horizontally invariant replacement $(-)^{\mathrm{whi}}\colon \mathrm{DblCat} \to \mathrm{DblCat}^{\mathbb{2}}$.

Construction 3.17 Let $\mathbb{A}$ be a double category and let $\mathrm{HorEq}(\mathbb{A})$ denote the set of all horizontal adjoint equivalence data in $\mathbb{A}$. Each horizontal adjoint equivalence (a, c, η, ϵ) in $\mathbb{A}$ defines a double functor $\mathbb{H}E_{\mathrm{adj}} \to \mathbb{A}$, and we define $\mathbb{A}^{\mathrm{whi}}$ as the pushout below left:

$$\begin{tikzcd}
\bigsqcup_{\mathrm{HorEq}(\mathbb{A})} \mathbb{H}E_{\mathrm{adj}} \arrow[r] \arrow[d, "\bigsqcup_{\mathrm{HorEq}(\mathbb{A})} J_4"'] & \mathbb{A} \arrow[d, "j_{\mathbb{A}}"] \\
\bigsqcup_{\mathrm{HorEq}(\mathbb{A})} \mathbb{H}^{\simeq}E_{\mathrm{adj}} \arrow[r] & \mathbb{A}^{\mathrm{whi}} \arrow[ul, phantom, very near start, "\ulcorner"]
\end{tikzcd}
\qquad
\begin{tikzcd}
\mathbb{A} \arrow[r, "F"] \arrow[d, "j_{\mathbb{A}}"'] & \mathbb{B} \arrow[d, "j_{\mathbb{B}}"] \\
\mathbb{A}^{\mathrm{whi}} \arrow[r, "F^{\mathrm{whi}}"'] & \mathbb{B}^{\mathrm{whi}}
\end{tikzcd}$$

This extends naturally to a functor $(-)^{\mathrm{whi}}\colon \mathrm{DblCat} \to \mathrm{DblCat}^{\mathbb{2}}$. In particular, it sends a double category $\mathbb{A}$ to the double functor $j_{\mathbb{A}}\colon \mathbb{A} \to \mathbb{A}^{\mathrm{whi}}$ and a double functor $F\colon \mathbb{A} \to \mathbb{B}$ to a commutative square in DblCat as depicted above right.

Remark 3.18 The double functor $j_{\mathbb{A}}\colon \mathbb{A} \to \mathbb{A}^{\mathrm{whi}}$ is the identity on underlying horizontal categories and it is fully faithful on squares for every double category $\mathbb{A}$, since it is a pushout of coproducts of the double functor $J_4\colon \mathbb{H}E_{\mathrm{adj}} \to \mathbb{H}^{\simeq}E_{\mathrm{adj}}$. Hence a double functor $F\colon \mathbb{A} \to \mathbb{B}$ coincides with $F^{\mathrm{whi}}\colon \mathbb{A}^{\mathrm{whi}} \to \mathbb{B}^{\mathrm{whi}}$ on underlying horizontal categories.

Remark 3.19 The construction $j_{\mathbb{A}}\colon \mathbb{A} \to \mathbb{A}^{\mathrm{whi}}$ adds $\mathbb{H}^{\simeq}E_{\mathrm{adj}}$–data in $\mathbb{A}^{\mathrm{whi}}$ to each horizontal adjoint equivalence (a, c, η, ϵ) in $\mathbb{A}$, as detailed in Description 3.14. In particular, we can see that two vertical morphisms u and v were freely added in $\mathbb{A}^{\mathrm{whi}}$ for each equivalence (a, c, η, ϵ), as well as weakly horizontally invertible squares as in Description 3.14. We henceforth say that the morphisms u and v were *added using the horizontal adjoint equivalence data* (a, c, η, ϵ) *in* $\mathbb{A}$.

As claimed, the double category $\mathbb{A}^{\mathrm{whi}}$ is indeed weakly horizontally invariant.

Proposition 3.20 *For every double category* $\mathbb{A}$, *the double category* $\mathbb{A}^{\mathrm{whi}}$ *is weakly horizontally invariant.*

Proof Let $a\colon A \xrightarrow{\simeq} C$ and $a'\colon A' \xrightarrow{\simeq} C'$ be horizontal equivalences in $\mathbb{A}$ and $w\colon C \twoheadrightarrow C'$ be a vertical morphism in $\mathbb{A}^{\mathrm{whi}}$. By construction of $\mathbb{A}^{\mathrm{whi}}$, we have vertical morphisms $u\colon A \twoheadrightarrow C$ and $v\colon C' \twoheadrightarrow A'$ in $\mathbb{A}^{\mathrm{whi}}$ together with weakly horizontally invertible squares α and δ in $\mathbb{A}^{\mathrm{whi}}$:

$$\begin{tikzcd} A \arrow[r, "a", "\simeq"'] \arrow[d, "u"'] \arrow[dr, phantom, "\alpha\ \simeq"] & C \arrow[d, equal] \\ C \arrow[r, equal] & C \end{tikzcd} \qquad\qquad \begin{tikzcd} C' \arrow[r, equal] \arrow[d, "v"'] \arrow[dr, phantom, "\delta\ \simeq"] & C' \arrow[d, equal] \\ A' \arrow[r, "\simeq", "a'"'] & C' \end{tikzcd}$$

Then the composite of vertical morphisms $vwu\colon A \twoheadrightarrow A'$ together with the weakly horizontally invertible square given by the vertical composite of the squares α, id_w, and δ gives the desired lift. □

The foresight that $(-)^{\mathrm{whi}}$ will give a fibrant replacement in our desired model structure (as we show in Corollary 5.4) and that the double biequivalences will precisely be the weak equivalences between fibrant objects (proved in Proposition 5.5) motivates us to define our weak equivalences as the double functors inducing double biequivalences between weakly horizontally invariant replacements.

Definition 3.21 We define $\mathcal{W}$ to be the class of double functors $F\colon \mathbb{A} \to \mathbb{B}$ such that the induced double functor $F^{\mathrm{whi}}\colon \mathbb{A}^{\mathrm{whi}} \to \mathbb{B}^{\mathrm{whi}}$ is a double biequivalence.

Remark 3.22 Since double biequivalences are the weak equivalences in the model structure on DblCat of [16, Theorem 3.18], they satisfy 2-out-of-3 and are closed under retracts. As a consequence, the class $\mathcal{W}$ also has these properties, as the replacement $(-)^{\mathrm{whi}}$ is functorial.

Although double biequivalences are more tractable than our proposed weak equivalences, the passage to this bigger class is truly needed. Indeed, the class of double functors that are both in $\mathcal{I}_w$–cof and double biequivalences is not closed under pushouts, and thus cannot be the class of trivial cofibrations in a model structure.

Example 3.23 Let $\mathbb{A}$ be the double category generated by the data

$$\begin{tikzcd} A \arrow[d, "v"', "\bullet" marking] & \\ A' \arrow[r, "a", "\simeq"'] & B' \arrow[d, "w", "\bullet"' marking] \\ & B'' \end{tikzcd}$$

where a is a horizontal adjoint equivalence. As we will see in Corollary 5.4, the double functor $j_{\mathbb{A}}\colon \mathbb{A} \to \mathbb{A}^{\mathrm{whi}}$ given by Construction 3.17 is a weak equivalence in $\mathcal{W}$. However, it is not a double biequivalence. To see this, note that a vertical morphism $u\colon A' \mathrel{\bullet\!\!\to} B'$ is freely added in $\mathbb{A}^{\mathrm{whi}}$. Then the composite $wuv\colon A \mathrel{\bullet\!\!\to} B''$ in $\mathbb{A}^{\mathrm{whi}}$ does not admit a lift along $j_{\mathbb{A}}$ as required by (db3) of Definition 2.9, as the only objects horizontally equivalent to A and B'' in $\mathbb{A}^{\mathrm{whi}}$ are themselves through the horizontal identities, and there are no vertical morphisms from A to B'' in $\mathbb{A}$.

However, as we now show, double biequivalences are contained in $\mathcal{W}$. The reverse inclusion does not hold, but, as we will see in Proposition 5.5, a weak equivalence whose source is a weakly horizontally invariant double category is a double biequivalence.

We use the following technical lemma to prove that double biequivalences are contained in our class of weak equivalences.

Lemma 3.24 *Let $F\colon \mathbb{A} \to \mathbb{B}$ be a double biequivalence. Then for every vertical morphism $v\colon B \mathrel{\bullet\!\!\to} B'$ in $\mathbb{B}^{\mathrm{whi}}$ which is a composite of freely added vertical morphisms along the double functor $j_{\mathbb{B}}\colon \mathbb{B} \to \mathbb{B}^{\mathrm{whi}}$, and every pair of horizontal equivalences $b\colon FA \xrightarrow{\simeq} B$ and $b'\colon FA' \xrightarrow{\simeq} B'$ in $\mathbb{B}$, there is a vertical morphism $u\colon A \mathrel{\bullet\!\!\to} A'$ in $\mathbb{A}^{\mathrm{whi}}$ together with a weakly horizontally invertible square in $\mathbb{B}^{\mathrm{whi}}$ of the form*

$$\begin{tikzcd} FA \arrow[r, "b", "\simeq"'] \arrow[d, "F^{\mathrm{whi}}u"', "\bullet" marking] \arrow[dr, phantom, "\beta\ \simeq"] & B \arrow[d, "v", "\bullet"' marking] \\ FA' \arrow[r, "\simeq", "b'"'] & B' \end{tikzcd}$$

Proof First note that there is a horizontal adjoint equivalence (f, g, η, ϵ) in $\mathbb{B}$ and a weakly horizontally invertible square α in $\mathbb{B}^{\mathrm{whi}}$ of the form

$$\begin{tikzcd}
B \arrow[r, "f", "\simeq"'] \arrow[d, "v"', "\bullet" marking] \arrow[dr, phantom, "\alpha\ \simeq"] & B' \arrow[d, equal, "\bullet" marking] \\
B' \arrow[r, equal] & B'
\end{tikzcd}$$

obtained by composing the corresponding weakly horizontally invertible squares for each freely added vertical morphism appearing in the decomposition of v. Let $(b', d', \eta', \epsilon')$ be a choice of horizontal adjoint equivalence data for b'. Since F satisfies (db2) and (db4) of Definition 2.9, there is a horizontal equivalence $a\colon A \xrightarrow{\simeq} A'$ in $\mathbb{A}$ together with a vertically invertible square ψ in $\mathbb{B}$ of the form

$$\begin{tikzcd}
FA \arrow[r, "b", "\simeq"'] \arrow[d, equal, "\bullet" marking] & B \arrow[r, "f", "\simeq"'] \arrow[d, phantom, "\psi\ \cong"] & B' \arrow[r, "d'", "\simeq"'] & FA' \arrow[d, equal, "\bullet" marking] \\
FA \arrow[rrr, "\simeq", "Fa"'] & & & FA'
\end{tikzcd}$$

Let $u\colon A \xrightarrow{\bullet} A'$ be a vertical morphism in $\mathbb{A}^{\mathrm{whi}}$ freely added using horizontal adjoint equivalence data for a. We get a weakly horizontally square β, as desired,

$$\begin{tikzcd}
FA \arrow[r, "b"] \arrow[d, "F^{\mathrm{whi}}u"', "\bullet" marking] \arrow[dr, phantom, "\beta\ \simeq"] & B \arrow[d, "v", "\bullet" marking] \\
FA' \arrow[r, "b'"'] & B'
\end{tikzcd}
=
\begin{tikzcd}
FA \arrow[r, "b"] \arrow[d, equal, "\bullet" marking] \arrow[dr, phantom, "e_b"] & B \arrow[rrrr, equal] \arrow[d, equal, "\bullet" marking] & & \arrow[d, phantom, "\eta\ \cong"] & & B \arrow[d, equal, "\bullet" marking] \\
FA \arrow[r, "b"] \arrow[d, equal, "\bullet" marking] \arrow[dr, phantom, "e_b"] & B \arrow[r, "f"] \arrow[d, equal, "\bullet" marking] \arrow[dr, phantom, "e_f"] & B' \arrow[rr, equal] \arrow[d, equal, "\bullet" marking] \arrow[drr, phantom, "(\epsilon')^{-1}\cong"] & & B' \arrow[r, "g"] \arrow[d, equal, "\bullet" marking] \arrow[dr, phantom, "e_g"] & B \arrow[d, equal, "\bullet" marking] \\
FA \arrow[r, "b"'] \arrow[d, equal, "\bullet" marking] & B \arrow[r, "f"'] \arrow[d, phantom, "\psi\ \cong"] & B' \arrow[r, "d'"] & FA' \arrow[r, "b'"] \arrow[d, equal, "\bullet" marking] \arrow[dr, phantom, "e_{b'}"] & B' \arrow[r, "g"] \arrow[d, equal, "\bullet" marking] \arrow[dr, phantom, "e_g"] & B \arrow[d, equal, "\bullet" marking] \\
FA \arrow[rrr, "Fa"'] \arrow[d, "F^{\mathrm{whi}}u"', "\bullet" marking] & \arrow[d, phantom, "F^{\mathrm{whi}}\bar{\alpha}\ \simeq"] & & FA' \arrow[r, "b'"] \arrow[d, equal, "\bullet" marking] \arrow[dr, phantom, "e_{b'}"] & B' \arrow[r, "g"] \arrow[d, equal, "\bullet" marking] \arrow[dr, phantom, "\alpha'\ \simeq"] & B \arrow[d, "v", "\bullet" marking] \\
FA' \arrow[rrr, equal] & & & FA' \arrow[r, "b'"'] & B' \arrow[r, equal] & B'
\end{tikzcd}$$

where $\bar{\alpha}$ is the weakly horizontally invertible square in $\mathbb{A}^{\mathrm{whi}}$ that was freely added with u (see Description 3.14), and α' is the weak inverse of the square α. $\square$

Proposition 3.25 *Every double biequivalence is in $\mathcal{W}$.*

Proof Let $F\colon \mathbb{A} \to \mathbb{B}$ be a double biequivalence; we show that F^{whi} satisfies (db1)–(db4) of Definition 2.9. Since F and F^{whi} agree on underlying horizontal categories by Remark 3.18, and F satisfies (db1)–(db2), so does F^{whi}. Moreover, since $j_{\mathbb{A}}$, $j_{\mathbb{B}}$, and F are fully faithful on squares and $F^{\mathrm{whi}} j_{\mathbb{A}} = j_{\mathbb{B}} F$, we have that F^{whi} is also fully faithful on squares, ie it satisfies (db4). Finally, since every vertical morphism in $\mathbb{B}^{\mathrm{whi}}$ can be decomposed as an alternate composite of vertical morphisms in $\mathbb{B}$ and of composites of freely added vertical morphisms, the fact that F^{whi} satisfies (db3) follows from (db3) for F and Lemma 3.24. $\square$

3.4 The model structure

By taking cofibrations as the $\mathcal{I}_w$–cofibrations and weak equivalences as the double functors in $\mathcal{W}$, we obtain the desired model structure on DblCat. The relevant classes of morphisms, as well as an outline of the proof with shortcuts to the corresponding results, is provided below; the technical details are deferred to Section 5.

Theorem 3.26 *There is a model structure* $(\mathcal{C}, \mathcal{F}, \mathcal{W})$ *on* DblCat *such that*

(i) *the class $\mathcal{C}$ of cofibrations is given by $\mathcal{C} := \mathcal{I}_w$–*cof, *where $\mathcal{I}_w$ is the set described in Notation 3.8;*

(ii) *the class $\mathcal{W}$ of weak equivalences is as described in Definition 3.21;*

(iii) *the class $\mathcal{F}$ of fibrations is given by $\mathcal{F} := (\mathcal{C} \cap \mathcal{W})^{\boxslash}$; and*

(iv) *the fibrant objects are the weakly horizontally invariant double categories.*

Proof We follow the definition of model structure presented in [17, Definition 2.1]. By Remark 3.22, we know that the class $\mathcal{W}$ of weak equivalences satisfies the 2-out-of-3 property. Furthermore, by Proposition 5.1, we have that $\mathcal{F} \cap \mathcal{W} = \mathcal{I}_w$–inj, and hence the pair $(\mathcal{C}, \mathcal{F} \cap \mathcal{W}) = (\mathcal{I}_w\text{–cof}, \mathcal{I}_w\text{–inj})$ is the weak factorization system generated by the set $\mathcal{I}_w$ of Notation 3.8. The fact that the pair $(\mathcal{C} \cap \mathcal{W}, \mathcal{F})$ forms a weak factorization system is the content of Theorem 5.6 and Corollary 5.7. We present in Theorem 5.2 the desired characterization of fibrant objects. $\square$

4 $\mathcal{J}_w$–cofibrations and $\mathcal{J}_w$–injective double functors

As we saw in the previous section, our proposed classes of cofibrations and of trivial fibrations can be constructed from a generating set $\mathcal{I}_w$, and admit concise descriptions.

Unfortunately, a nice description of the proposed fibrations and trivial cofibrations is not available in general. To prove that these classes of double functors form a weak factorization system, we introduce an auxiliary weak factorization system ($\mathcal{J}_w$–cof, $\mathcal{J}_w$–inj) generated by a set $\mathcal{J}_w$ of double functors. Aside from admitting a simple description, the $\mathcal{J}_w$–injective double functors contain our proposed fibrations, and agree with these when we restrict to double functors with weakly horizontally invariant target; in particular, they can be used to identify our fibrant objects.

This section is largely technical, and the reader willing to trust our claims is encouraged to jump ahead to Section 5.

Let us first introduce the set $\mathcal{J}_w$.

Notation 4.1 Let $\mathcal{J}_w$ denote the set containing the following double functors:

(i) either inclusion $J_1 \colon \mathbb{1} \to \mathbb{H}E_{\mathrm{adj}}$, where the 2–category E_{adj} is the free-living adjoint equivalence;

(ii) either inclusion $J_2 \colon \mathbb{H}\mathbb{2} \to \mathbb{H}C_{\mathrm{inv}}$, where the 2–category C_{inv} is the free-living 2–isomorphism;

(iii) the inclusion $J_3 \colon \mathbb{W}^- \to \mathbb{W}$, where the double category $\mathbb{W}$ is the free-living weakly horizontally invertible square with horizontal adjoint equivalence data, and $\mathbb{W}^-$ is its double subcategory where we remove one of the vertical morphisms:

$$\mathbb{W} = \begin{tikzcd} 0 \arrow[r, "\simeq"] \arrow[d, "\bullet" marking] \arrow[dr, phantom, "\simeq"] & 1 \arrow[d, "\bullet" marking] \\ 0' \arrow[r, "\simeq"'] & 1' \end{tikzcd} \qquad \mathbb{W}^- = \begin{tikzcd} 0 \arrow[r, "\simeq"] & 1 \arrow[d, "\bullet" marking] \\ 0' \arrow[r, "\simeq"'] & 1' \end{tikzcd}$$

Remark 4.2 It is straightforward from the characterization of $\mathcal{I}_w$–cofibrations given in Theorem 3.11 and using Remark 3.12 that the double functors J_1, J_2, and J_3 are in $\mathcal{I}_w$–cof, and from Definition 2.9 that they are double biequivalences. In particular, by Proposition 3.25, this implies that they are trivial cofibrations in our proposed model structure on DblCat.

4.1 $\mathcal{J}_w$–injective double functors

By studying what it means to have the right lifting property with respect to the double functors in $\mathcal{J}_w$, we can characterize the $\mathcal{J}_w$–injective double functors as follows.

Proposition 4.3 *A double functor $F\colon \mathbb{A} \to \mathbb{B}$ is in $\mathcal{J}_w$–inj if and only if it satisfies the following conditions*:

(df1) *For every object $C \in \mathbb{A}$ and every horizontal equivalence $b\colon B \xrightarrow{\simeq} FC$ in $\mathbb{B}$, there is a horizontal equivalence $a\colon A \xrightarrow{\simeq} C$ in $\mathbb{A}$ such that $b = Fa$.*

(df2) *For every horizontal morphism $c\colon A \to C$ in $\mathbb{A}$ and every vertically invertible square β in $\mathbb{B}$ as depicted below left, there is a vertically invertible square α in $\mathbb{A}$ as depicted below right such that $\beta = F\alpha$:*

$$\begin{tikzcd} FA \arrow[r, "b"] \arrow[d, equal, "\bullet" marking] \arrow[dr, phantom, "\beta\ \cong"] & FC \arrow[d, equal, "\bullet" marking] \\ FA \arrow[r, "Fc"'] & FC \end{tikzcd} \qquad \begin{tikzcd} A \arrow[r, "a"] \arrow[d, equal, "\bullet" marking] \arrow[dr, phantom, "\alpha\ \cong"] & C \arrow[d, equal, "\bullet" marking] \\ A \arrow[r, "c"'] & C \end{tikzcd}$$

(df3) *For every diagram in $\mathbb{A}$ as depicted below left, where a and a' are horizontal equivalences, and every weakly horizontally invertible square β in $\mathbb{B}$ as depicted below middle, there is a weakly horizontally invertible square α in $\mathbb{A}$ as depicted below right such that $\beta = F\alpha$:*

$$\begin{tikzcd} A \arrow[r, "a", "\simeq"'] & C \arrow[d, "w", "\bullet" marking] \\ A' \arrow[r, "\simeq", "a'"'] & C' \end{tikzcd} \qquad \begin{tikzcd} FA \arrow[r, "Fa", "\simeq"'] \arrow[d, "v"', "\bullet" marking] \arrow[dr, phantom, "\beta\ \simeq"] & FC \arrow[d, "Fw", "\bullet" marking] \\ FA' \arrow[r, "\simeq", "Fa'"'] & FC' \end{tikzcd} \qquad \begin{tikzcd} A \arrow[r, "a", "\simeq"'] \arrow[d, "u"', "\bullet" marking] \arrow[dr, phantom, "\alpha\ \simeq"] & C \arrow[d, "w", "\bullet" marking] \\ A' \arrow[r, "\simeq", "a'"'] & C' \end{tikzcd}$$

Proof This is obtained directly from a close inspection of the right lifting properties with respect to the double functors in $\mathcal{J}_w$. □

As a consequence, we can use the class $\mathcal{J}_w$–inj to identify the weakly horizontally invariant double categories; see Definition 2.10.

Corollary 4.4 *A double category $\mathbb{A}$ is weakly horizontally invariant if and only if the unique double functor $\mathbb{A} \to \mathbb{1}$ is in $\mathcal{J}_w$–*inj.

The following result tells us that every $\mathcal{J}_w$–injective double functor has the right lifting property with respect to $J_4\colon \mathbb{H}E_{\mathrm{adj}} \to \mathbb{H}^{\simeq}E_{\mathrm{adj}}$, which is useful when proving that $\mathcal{J}_w$–injective double functors with weakly horizontally invariant targets are fibrations.

Proposition 4.5 *Let $F\colon \mathbb{A} \to \mathbb{B}$ be a double functor in $\mathcal{J}_w$–*inj. *Then F is in $\{J_4\}$–*inj, *where $J_4\colon \mathbb{H}E_{\mathrm{adj}} \to \mathbb{H}^{\simeq}E_{\mathrm{adj}}$ is the inclusion of Notation 3.15.*

Proof Consider a commutative square in DblCat of the form

$$\begin{tikzcd} \mathbb{H}E_{\mathrm{adj}} \arrow[r, "{(a,c,\eta,\epsilon)}"] \arrow[d, "J_4"'] & \mathbb{A} \arrow[d, "F"] \\ \mathbb{H}^{\simeq}E_{\mathrm{adj}} \arrow[r, "G"'] \arrow[ur, dashed, "L"] & \mathbb{B} \end{tikzcd}$$

where $a\colon A \xrightarrow{\simeq} C$ is a horizontal adjoint equivalence with data (a,c,η,ϵ); we want to find a lift L as depicted. The images under G of the weakly horizontally invertible squares $\alpha,\gamma \in \mathbb{H}^{\simeq}E_{\mathrm{adj}}$ from Description 3.14 are as in the two leftmost diagrams below:

$$\begin{tikzcd} FA \arrow[r, "Fa", "\simeq"'] \arrow[d, "Gu"'] \arrow[dr, phantom, "G\alpha\ \simeq"] & FC \arrow[d, equal] \\ FC \arrow[r, equal] & FC \end{tikzcd}
\qquad
\begin{tikzcd} FC \arrow[r, "Fc", "\simeq"'] \arrow[d, "Gv"'] \arrow[dr, phantom, "G\gamma\ \simeq"] & FA \arrow[d, equal] \\ FA \arrow[r, equal] & FA \end{tikzcd}
\qquad
\begin{tikzcd} A \arrow[r, "a", "\simeq"'] \arrow[d, "\bar{u}"'] \arrow[dr, phantom, "\bar{\alpha}\ \simeq"] & C \arrow[d, equal] \\ C \arrow[r, equal] & C \end{tikzcd}
\qquad
\begin{tikzcd} C \arrow[r, "c", "\simeq"'] \arrow[d, "\bar{v}"'] \arrow[dr, phantom, "\bar{\gamma}\ \simeq"] & A \arrow[d, equal] \\ A \arrow[r, equal] & A \end{tikzcd}$$

By (df3) of Proposition 4.3, there are weakly horizontally invertible squares $\bar{\alpha}$ and $\bar{\gamma}$ in $\mathbb{A}$, as in the two rightmost diagrams above, such that $F\bar{\alpha} = G\alpha$ and $F\bar{\gamma} = G\gamma$. Finally, by Remark 3.16, the data (a,c,η,ϵ), $\bar{\alpha}$, and $\bar{\gamma}$ determine a unique double functor $L\colon \mathbb{H}^{\simeq}E_{\mathrm{adj}} \to \mathbb{A}$ which gives the desired lift. Note that we indeed have $G = FL$ since their images on the generating data of $\mathbb{H}^{\simeq}E_{\mathrm{adj}}$ coincide. □

Remark 4.6 This result, together with Corollary 4.4, guarantees that for every weakly horizontally invariant double category $\mathbb{A}$, the double functor $\mathbb{A} \to \mathbb{1}$ is in $\{J_4\}$–inj.

Next, we show that the double functors which are $\mathcal{J}_w$–injective and double biequivalences are precisely the ones that are $\mathcal{I}_w$–injective.

Proposition 4.7 *A double functor $F\colon \mathbb{A} \to \mathbb{B}$ is $\mathcal{J}_w$–injective and a double biequivalence if and only if it is $\mathcal{I}_w$–injective.*

Proof Since $\mathcal{J}_w \subseteq \mathcal{I}_w$–cof by Remark 4.2, $\mathcal{I}_w$–inj $= \mathcal{I}_w$–$\mathrm{cof}^{\boxslash} \subseteq \mathcal{J}_w^{\boxslash} = \mathcal{J}_w$–inj. Furthermore, by Remark 3.10, a double functor in $\mathcal{I}_w$–inj is in particular a double biequivalence, which shows the converse statement.

Now suppose that F is $\mathcal{J}_w$–injective and a double biequivalence. We prove that F is $\mathcal{I}_w$–injective using Proposition 3.9. It is straightforward to see that F is surjective on objects, full on horizontal morphisms, and fully faithful on squares using (db1), (db2) and (db4) of Definition 2.9 and (df1)–(df2) of Proposition 4.3. To prove that F is full on vertical morphisms, let A, A' be objects in $\mathbb{A}$, and $v\colon FA \twoheadrightarrow FA'$ be a vertical morphism in $\mathbb{B}$. Since F satisfies (db3), there is a vertical morphism $w\colon C \twoheadrightarrow C'$ in $\mathbb{A}$ together with a weakly horizontally invertible square β in $\mathbb{B}$ as depicted below left:

$$\begin{tikzcd} FA \arrow[r, "b", "\simeq"'] \arrow[d, "v"'] \arrow[dr, phantom, "\beta\ \simeq"] & FC \arrow[d, "Fw"] \\ FA' \arrow[r, "\simeq", "b'"'] & FC' \end{tikzcd} \qquad \begin{tikzcd} A \arrow[r, "a", "\simeq"'] \arrow[d, "u"'] \arrow[dr, phantom, "\alpha\ \simeq"] & C \arrow[d, "w"] \\ A' \arrow[r, "\simeq", "a'"'] & C' \end{tikzcd}$$

Since F is full on horizontal morphisms and fully faithful on squares, there are horizontal equivalences $a\colon A \xrightarrow{\simeq} C$ and $a'\colon A' \xrightarrow{\simeq} C'$ in $\mathbb{A}$ such that $b = Fa$ and $b' = Fa'$. Then, by (df3), there is a weakly horizontally invertible square α in $\mathbb{A}$ as depicted above right such that $\beta = F\alpha$; in particular, $v = Fu$. This completes the proof. $\square$

4.2 $\mathcal{J}_w$–cofibrations and double biequivalences

We now focus on the $\mathcal{J}_w$–cofibrations. First, we show that they are cofibrations in our proposed model structure, which additionally satisfy the requirements of a double biequivalence except for condition (db3) on vertical morphisms.

Proposition 4.8 *Let $J\colon \mathbb{A} \to \mathbb{B}$ be a double functor in $\mathcal{J}_w$–*cell*. Then the functor J*

(i) *is injective on objects, and faithful on horizontal and vertical morphisms;*

(ii) *satisfies* (db1), (db2) *and* (db4) *of Definition 2.9.*

Proof Since $\mathcal{J}_w \subseteq \mathcal{I}_w$–cof by Remark 4.2, we have that $\mathcal{J}_w$–cell $\subseteq \mathcal{I}_w$–cof; hence J is injective on objects, and faithful on horizontal and vertical morphisms, by Remark 3.12.

Now, since objects can only be added along $J_1\colon \mathbb{1} \to \mathbb{H}E_{\mathrm{adj}}$, ie with a horizontal equivalence to an object which was already there, J satisfies (db1). Similarly, as horizontal morphisms can only be added along $J_2\colon \mathbb{H}2 \to C_{\mathrm{inv}}$, we can check that J satisfies (db2). Finally note that J satisfies (db4), since taking pushouts along J_1, J_2, and J_3 does not create new squares within an existing boundary, nor does it identify squares. $\square$

When the source of a $\mathcal{J}_w$–cofibration is a weakly horizontally invariant double category, we can further show that (db3) of Definition 2.9 is satisfied, and hence that every such $\mathcal{J}_w$–cofibration is a double biequivalence.

Proposition 4.9 *Let $J\colon \mathbb{A} \to \mathbb{B}$ be a double functor in $\mathcal{J}_w$–*cof *such that $\mathbb{A}$ is weakly horizontally invariant. Then J is a double biequivalence.*

Proof We first prove the case when $J \in \mathcal{J}_w$–cell. By Proposition 4.8, we have that J satisfies (db1), (db2) and (db4) of Definition 2.9; it remains to show (db3). Let λ be an ordinal and let $\mathbb{X}\colon \lambda \to \mathrm{DblCat}$ be a transfinite composition of pushouts of double functors in $\mathcal{J}_w$ such that J is the composite

$$J\colon \mathbb{A} \cong J(\mathbb{A}) = \mathbb{X}_0 \xrightarrow{\iota_0} \operatorname*{colim}_{\mu<\lambda} \mathbb{X}_\mu = \mathbb{B}.$$

Let $v\colon B \to B'$ be a vertical morphism in $\mathbb{B}$. We use transfinite induction to show that there is a vertical morphism $u\colon A \to A'$ in $\mathbb{A}$ and a weakly horizontally invertible square $\beta\colon (Ju \overset{\simeq}{\underset{\simeq}{\Rightarrow}} v)$ in $\mathbb{B}$. This amounts to showing that our statement holds for the base case $\lambda = 0$, for any successor ordinal $\mu + 1 < \lambda$, and for any limit ordinal $\kappa < \lambda$.

If $v \in \mathbb{X}_0 = J(\mathbb{A})$, then there is a vertical morphism $u\colon A \to A'$ in $\mathbb{A}$ such that $Ju = v$ and we can take $\beta = \mathrm{id}_{Ju}$.

Now suppose that $v \in \mathbb{X}_{\mu+1}$ for a successor ordinal $\mu + 1 < \lambda$. If $v \in \mathbb{X}_\mu$, then we are done by induction. Otherwise, by definition of $\mathbb{X}$, the double category $\mathbb{X}_{\mu+1}$ is a pushout along J_3 as depicted below left,

$$\begin{tikzcd} \mathbb{W}^- \arrow[r, "{(w,d,d')}"] \arrow[d, "J_3"'] & \mathbb{X}_\mu \arrow[d, "i_\mu"] \\ \mathbb{W} \arrow[r, "\delta"'] & \mathbb{X}_{\mu+1} \arrow[ul, phantom, "\ulcorner", very near start] \end{tikzcd} \qquad \begin{tikzcd} D \arrow[r, "d", "\simeq"'] \arrow[d, "w"'] \arrow[dr, phantom, "\delta\ \simeq"] & Y \arrow[d, "\overline{w}"] \\ D' \arrow[r, "\simeq", "d'"'] & Y' \end{tikzcd}$$

where w is a vertical morphism in $\mathbb{X}_\mu$, d and d' are horizontal equivalences in $\mathbb{X}_\mu$, and δ is a weakly horizontally invertible square in $\mathbb{B}$, as depicted above right. Then the vertical morphism $v \in \mathbb{X}_{\mu+1}$ is a composite of vertical morphisms in $\mathbb{X}_\mu$ and the freely added vertical morphism $\overline{w}$. We prove that the result holds for a composite of the form $v = v_1 \overline{w} v_0$ with $v_0\colon B \to Y$ and $v_1\colon Y' \to B'$ two vertical morphisms in $\mathbb{X}_\mu$; the general case where $\overline{w}$ appears several times in the decomposition of v proceeds similarly.

By induction, since v_0, v_1, and w are in $\mathbb{X}_\mu$, there are vertical morphisms u_0, u_1, and t in $\mathbb{A}$, and weakly horizontally invertible squares β_0, β_1 and φ in $\mathbb{B}$, as depicted below:

$$\begin{tikzcd} JA \arrow[r, "b_0", "\simeq"'] \arrow[d, "Ju_0"'] \arrow[dr, phantom, "\beta_0\ \simeq"] & B \arrow[d, "v_0"] \\ JC \arrow[r, "\simeq", "b_0'"'] & Y \end{tikzcd} \qquad \begin{tikzcd} JC' \arrow[r, "b_1", "\simeq"'] \arrow[d, "Ju_1"'] \arrow[dr, phantom, "\beta_1\ \simeq"] & Y' \arrow[d, "v_1"] \\ JA' \arrow[r, "\simeq", "b_1'"'] & B' \end{tikzcd} \qquad \begin{tikzcd} JX \arrow[r, "f", "\simeq"'] \arrow[d, "Jt"'] \arrow[dr, phantom, "\varphi\ \simeq"] & D \arrow[d, "w"] \\ JX' \arrow[r, "\simeq", "f'"'] & D' \end{tikzcd}$$

Let (df, g, η, ϵ) and $(d'f', g', \eta', \epsilon')$ be horizontal adjoint equivalence data in $\mathbb{B}$ for the composites df and $d'f'$. Since J satisfies (db2) and (db4), there are horizontal equivalences $a\colon C \xrightarrow{\simeq} X$ and $a'\colon C' \xrightarrow{\simeq} X'$ in $\mathbb{A}$ together with vertically invertible squares ψ and ψ' in $\mathbb{B}$ as in the two leftmost squares below:

$$
\begin{tikzcd}
JC \arrow[r, "b_0'", "\simeq"'] \arrow[d, equal, "\bullet"] & Y \arrow[r, "g", "\simeq"'] \arrow[d, phantom, "\psi\ \cong"] & JX \arrow[d, equal, "\bullet"] \\
JC \arrow[rr, "\simeq", "Ja"'] & & JX
\end{tikzcd}
\qquad
\begin{tikzcd}
JC' \arrow[r, "b_1", "\simeq"'] \arrow[d, equal, "\bullet"] & Y' \arrow[r, "g'", "\simeq"'] \arrow[d, phantom, "\psi'\ \cong"] & JX' \arrow[d, equal, "\bullet"] \\
JC' \arrow[rr, "\simeq", "Ja'"'] & & JX'
\end{tikzcd}
\qquad
\begin{tikzcd}
C \arrow[r, "a", "\simeq"'] \arrow[d, "\bar{u}"', "\bullet"] \arrow[dr, phantom, "\alpha\ \simeq"] & X \arrow[d, "t", "\bullet"'] \\
C' \arrow[r, "\simeq", "a'"'] & X
\end{tikzcd}
$$

Then, as $\mathbb{A}$ is weakly horizontally invariant, there is a vertical morphism $\bar{u}\colon C \twoheadrightarrow C'$ and a weakly horizontally invertible square α in $\mathbb{A}$ as depicted above right. Setting $u := u_1\bar{u}u_0\colon A \twoheadrightarrow A'$ and considering the pasting of squares in $\mathbb{B}$

$$
\begin{tikzcd}
JA \arrow[r, "b_0"] \arrow[d, "Ju_0"', "\bullet"] \arrow[dr, phantom, "\beta_0\ \simeq"] & B \arrow[rrr, equal] \arrow[d, "v_0", "\bullet"'] \arrow[drrr, phantom, "\mathrm{id}_{v_0}"] & & & B \arrow[d, "v_0", "\bullet"'] \\
JC \arrow[r, "b_0'"'] \arrow[d, equal, "\bullet"] \arrow[dr, phantom, "e_{b_0'}"] & Y \arrow[rrr, equal] \arrow[d, equal, "\bullet"] \arrow[drrr, phantom, "\epsilon^{-1}\ \cong"] & & & Y \arrow[d, equal, "\bullet"] \\
JC \arrow[r, "b_0'"'] \arrow[d, equal, "\bullet"] & Y \arrow[r, "g"'] \arrow[d, phantom, "\psi\ \cong"] & JX \arrow[r, "f"] \arrow[d, equal, "\bullet"] \arrow[dr, phantom, "e_f"] & D \arrow[r, "d"] \arrow[d, equal, "\bullet"] \arrow[dr, phantom, "e_d"] & Y \arrow[d, equal, "\bullet"] \\
JC \arrow[rr, "Ja"] \arrow[d, "J\bar{u}"', "\bullet"] & \arrow[d, phantom, "J\alpha\ \simeq"] & JX \arrow[r, "f"] \arrow[d, "Jt"', "\bullet"] \arrow[dr, phantom, "\varphi\ \simeq"] & D \arrow[r, "d"] \arrow[d, "w"', "\bullet"] \arrow[dr, phantom, "\delta\ \simeq"] & Y \arrow[d, "\overline{w}", "\bullet"'] \\
JC' \arrow[rr, "Ja'"'] \arrow[d, equal, "\bullet"] & \arrow[d, phantom, "(\psi')^{-1}\cong"] & JX' \arrow[r, "f'"'] \arrow[d, equal, "\bullet"] \arrow[dr, phantom, "e_{f'}"] & D' \arrow[r, "d'"'] \arrow[d, equal, "\bullet"] \arrow[dr, phantom, "e_{d'}"] & Y' \arrow[d, equal, "\bullet"] \\
JC' \arrow[r, "b_1"] \arrow[d, equal, "\bullet"] \arrow[dr, phantom, "e_{b_1}"] & Y' \arrow[r, "g'"] \arrow[d, equal, "\bullet"] \arrow[drrr, phantom, "\epsilon'\ \cong"] & JX' \arrow[r, "f'"'] & D' \arrow[r, "d'"'] & Y' \arrow[d, equal, "\bullet"] \\
JC' \arrow[r, "b_1"] \arrow[d, "Ju_1"', "\bullet"] \arrow[dr, phantom, "\beta_1\ \simeq"] & Y' \arrow[rrr, equal] \arrow[d, "v_1", "\bullet"'] \arrow[drrr, phantom, "\mathrm{id}_{v_1}"] & & & Y' \arrow[d, "v_1", "\bullet"'] \\
JA' \arrow[r, "b_1'"'] & B' \arrow[rrr, equal] & & & B'
\end{tikzcd}
$$

we obtain a weakly horizontally invertible square of the desired form between the vertical morphisms $Ju = (Ju_1)(J\bar{u})(Ju_0)$ and $v = v_1\overline{w}v_0$.

Finally, if $v \in \mathbb{X}_\kappa = \operatorname{colim}_{\mu<\kappa} \mathbb{X}_\mu$ for a limit ordinal $\kappa < \lambda$, there is an ordinal $\mu < \kappa$ such that $v \in \mathbb{X}_\mu$, and we are done by induction. This shows (db3) for J, and proves that J is a double biequivalence.

Now if $J\colon \mathbb{A} \to \mathbb{B}$ is in $\mathcal{J}_w$–cof, then it is a retract of a double functor $K\colon \mathbb{A} \to \mathbb{C}$ in $\mathcal{J}_w$–cell, whose source is also the weakly horizontally invariant double category $\mathbb{A}$. By the first part of the proof, the double functor K is a double biequivalence, and therefore so is J. □

4.3 Fibrations and $\mathcal{J}_w$–injective double functors

To conclude this section, we prove our claim that a double functor whose target is weakly horizontally invariant is a fibration precisely when it is $\mathcal{J}_w$–injective. We start by showing that the class of fibrations is included in $\mathcal{J}_w$–inj.

Lemma 4.10 *We have that $\mathcal{F} \subseteq \mathcal{J}_w$–inj.*

Proof Since every double functor in $\mathcal{J}_w$ is a double biequivalence by Remark 4.2, it is in $\mathcal{W}$ by Proposition 3.25. This, together with Remark 4.2, implies that $\mathcal{J}_w \subseteq \mathcal{C} \cap \mathcal{W}$. Therefore $\mathcal{F} = (\mathcal{C} \cap \mathcal{W})^{\boxslash} \subseteq \mathcal{J}_w^{\boxslash} = \mathcal{J}_w$–inj, which concludes the proof. □

For the converse inclusion, we will use the next incremental lemmas, which ultimately ensure that the weakly horizontally invariant replacement of a trivial cofibration is a $\mathcal{J}_w$–cofibration.

Lemma 4.11 *Let $I\colon \mathbb{A} \to \mathbb{B}$ be a double functor in $\mathcal{C} = \mathcal{I}_w$–cof which is fully faithful on squares. Then the induced double functor $I^{\mathrm{whi}}\colon \mathbb{A}^{\mathrm{whi}} \to \mathbb{B}^{\mathrm{whi}}$ is in $\mathcal{C}$.*

Proof We show that I^{whi} is in $\mathcal{I}_w$–cof by using Theorem 3.11. Since the double functors I and I^{whi} coincide on underlying horizontal categories by Remark 3.18, and $I \in \mathcal{I}_w$–cof, the functor $\boldsymbol{UH}I = \boldsymbol{UH}I^{\mathrm{whi}}$ has the left lifting property with respect to surjective on objects and full functors. It remains to prove that $\boldsymbol{UV}I^{\mathrm{whi}}$ satisfies this lifting property.

Let $P\colon \mathcal{X} \to \mathcal{Y}$ be a surjective on objects and full functor, and consider a commutative square as below left:

$$\begin{tikzcd} \boldsymbol{UV}\mathbb{A}^{\mathrm{whi}} \arrow[r, "F"] \arrow[d, "\boldsymbol{UV}I^{\mathrm{whi}}"'] & \mathcal{X} \arrow[d, "P"] \\ \boldsymbol{UV}\mathbb{B}^{\mathrm{whi}} \arrow[r, "G"'] \arrow[ur, dashed, "L"] & \mathcal{Y} \end{tikzcd} \qquad \begin{tikzcd} \boldsymbol{UV}\mathbb{A} \arrow[r, "F \circ \boldsymbol{UV}j_{\mathbb{A}}"] \arrow[d, "\boldsymbol{UV}I"'] & \mathcal{X} \arrow[d, "P"] \\ \boldsymbol{UV}\mathbb{B} \arrow[r, "G \circ \boldsymbol{UV}j_{\mathbb{B}}"'] \arrow[ur, dashed, "\hat{L}"] & \mathcal{Y} \end{tikzcd}$$

Recall that the category $UV\mathbb{B}^{\mathrm{whi}}$ is obtained from $UV\mathbb{B}$ by freely adding a morphism $v_{\underline{b}}\colon B \to B'$ for each horizontal adjoint equivalence $\underline{b} = (b, d, \eta, \epsilon)$ in $\mathbb{B}$; see Construction 3.17. Hence the data of a functor $L\colon UV\mathbb{B}^{\mathrm{whi}} \to \mathcal{X}$ is equivalent to the data of a functor $\widehat{L}\colon UV\mathbb{B} \to \mathcal{X}$ together with a choice of morphism $Lv_{\underline{b}}\colon \widehat{L}B \to \widehat{L}B'$ for each $v_{\underline{b}}\colon B \to B'$. Therefore, to construct a functor $L\colon UV\mathbb{B}^{\mathrm{whi}} \to \mathcal{X}$ as depicted, it is enough to define L on the subcategory $UV\mathbb{B}$ and on each $v_{\underline{b}}\colon B \to B'$ in such a way that $PL = G$ and $L(UVI^{\mathrm{whi}}) = F$.

Since I is in $\mathcal{I}_w$–cof, Theorem 3.11 tells us that UVI has the left lifting property with respect to P, and hence there is a lift $\widehat{L}\colon UV\mathbb{B} \to \mathcal{X}$ in the diagram above right; we define L to be $\widehat{L}$ on the subcategory $UV\mathbb{B}$.

Now, consider $v_{\underline{b}}\colon B \mathrel{\mathrlap{\rightarrow}\bullet\!\!\rightarrow} B'$ for a given horizontal adjoint equivalence $\underline{b} = (b, d, \eta, \epsilon)$. Since I is injective on objects and faithful on horizontal morphisms by Remark 3.12, and fully faithful on squares by assumption, there is at most one horizontal adjoint equivalence $\underline{a} = (a, c, \eta', \epsilon')$ in $\mathbb{A}$ such that $I\underline{a} = \underline{b}$. If there is such an $\underline{a}$, then there is a unique vertical morphism $u_{\underline{a}}$ in $\mathbb{A}^{\mathrm{whi}}$ (freely added using $\underline{a}$) such that $I^{\mathrm{whi}}(u_{\underline{a}}) = v_{\underline{b}}$; set $Lv_{\underline{b}} = Fu_{\underline{a}}$. If there is no such $\underline{a}$, then $v_{\underline{b}}$ is not in the image of I^{whi}. By fullness of P, we can then choose a morphism $w\colon LB \to LB'$ in $\mathcal{X}$ such that $Pw = Gv_{\underline{b}}$, and set $Lv_{\underline{b}} = w$. □

Lemma 4.12 *If $I\colon \mathbb{A} \to \mathbb{B}$ is a double functor in $\mathcal{C} \cap \mathcal{W}$, then $I^{\mathrm{whi}}\colon \mathbb{A}^{\mathrm{whi}} \to \mathbb{B}^{\mathrm{whi}}$ is in $\mathcal{J}_w$–cof.*

Proof First recall that, since $I \in \mathcal{W}$, the double functor $I^{\mathrm{whi}}\colon \mathbb{A}^{\mathrm{whi}} \to \mathbb{B}^{\mathrm{whi}}$ is a double biequivalence by definition. Next, consider a factorization $I^{\mathrm{whi}} = PJ$ with $J \in \mathcal{J}_w$–cof and $P \in \mathcal{J}_w$–inj. As $\mathbb{A}^{\mathrm{whi}}$ is weakly horizontally invariant, Proposition 4.9 ensures J is also a double biequivalence; then, by 2-out-of-3, so is P. Hence P is both $\mathcal{J}_w$–injective and a double biequivalence, and therefore it is $\mathcal{I}_w$–injective by Proposition 4.7.

Now, since I is in $\mathcal{W}$, it is fully faithful on squares, and so it follows from Lemma 4.11 that I^{whi} is in $\mathcal{I}_w$–cof. Then I^{whi} has the left lifting property with respect to $P \in \mathcal{I}_w$–inj, so, by the retract argument (see Remark 3.4), it is a retract of $J \in \mathcal{J}_w$–cell and hence is itself in $\mathcal{J}_w$–cof. □

Finally, we prove that every $\mathcal{J}_w$–injective double functor with weakly horizontally invariant target has the right lifting property with respect to every trivial cofibration I, by using its lifting property against the weakly horizontally invariant replacement I^{whi}.

Proposition 4.13 *Let $P\colon \mathbb{A} \to \mathbb{B}$ be a double functor with $\mathbb{B}$ weakly horizontally invariant. Then P is in $\mathcal{F} = (\mathcal{C} \cap \mathcal{W})^{\boxslash}$ if and only if P is in $\mathcal{J}_w$–inj.*

Proof If P is in $\mathcal{F}$, then P is in $\mathcal{J}_w$–inj by Lemma 4.10.

Now suppose that P is in $\mathcal{J}_w$–inj. We show that P has the right lifting property with respect to every double functor in $\mathcal{C} \cap \mathcal{W}$, ie it is in $\mathcal{F}$. Let $I\colon \mathbb{C} \to \mathbb{D}$ be a double functor in $\mathcal{C} \cap \mathcal{W}$ and consider a commutative square in DblCat as below; we want to find a lift $L\colon \mathbb{D} \to \mathbb{A}$ as pictured:

$$\begin{tikzcd} \mathbb{C} \arrow[r, "F"] \arrow[d, "I"'] & \mathbb{A} \arrow[d, "P"] \\ \mathbb{D} \arrow[r, "G"'] \arrow[ur, dashed, "L"] & \mathbb{B} \end{tikzcd}$$

Since $\mathbb{B}$ is weakly horizontally invariant, there is a lift in the diagram below left by Remark 4.6, which yields a double functor $\widehat{G}\colon \mathbb{D}^{\mathrm{whi}} \to \mathbb{B}$ as in the diagram below right, given by the universal property of the pushout:

$$\begin{tikzcd} \bigsqcup_{\mathrm{HorEq}(\mathbb{D})} \mathbb{H}E_{\mathrm{adj}} \arrow[r] \arrow[d, "\bigsqcup_{\mathrm{HorEq}(\mathbb{D})} J_4"'] & \mathbb{D} \arrow[r, "G"] & \mathbb{B} \\ \bigsqcup_{\mathrm{HorEq}(\mathbb{D})} \mathbb{H}^{\simeq}E_{\mathrm{adj}} \arrow[urr, dashed, "K"'] & & \end{tikzcd} \qquad \begin{tikzcd} \bigsqcup_{\mathrm{HorEq}(\mathbb{D})} \mathbb{H}E_{\mathrm{adj}} \arrow[r] \arrow[d, "\bigsqcup_{\mathrm{HorEq}(\mathbb{D})} J_4"'] & \mathbb{D} \arrow[d, "j_{\mathbb{D}}"] \arrow[ddr, bend left, "G"] & \\ \bigsqcup_{\mathrm{HorEq}(\mathbb{D})} \mathbb{H}^{\simeq}E_{\mathrm{adj}} \arrow[r] \arrow[drr, bend right, "K"'] & \mathbb{D}^{\mathrm{whi}} \arrow[dr, dashed, "\widehat{G}"] & \\ & & \mathbb{B} \end{tikzcd}$$

Now, since $P \in \mathcal{J}_w$–inj, by Proposition 4.5 there is a lift in the commutative diagram below left, which in turns yields a double functor $\widehat{F}\colon \mathbb{C}^{\mathrm{whi}} \to \mathbb{A}$ as in the diagram below right, given by the universal property of the pushout:

$$\begin{tikzcd} \bigsqcup_{\mathrm{HorEq}(\mathbb{C})} \mathbb{H}E_{\mathrm{adj}} \arrow[r] \arrow[d, "\bigsqcup_{\mathrm{HorEq}(\mathbb{C})} J_4"'] & \mathbb{C} \arrow[r, "F"] & \mathbb{A} \arrow[d, "P"] \\ \bigsqcup_{\mathrm{HorEq}(\mathbb{C})} \mathbb{H}^{\simeq}E_{\mathrm{adj}} \arrow[urr, dashed, "K'"'] \arrow[rr, "K \circ (\bigsqcup_I \mathrm{id})"'] & & \mathbb{B} \end{tikzcd} \qquad \begin{tikzcd} \bigsqcup_{\mathrm{HorEq}(\mathbb{C})} \mathbb{H}E_{\mathrm{adj}} \arrow[r] \arrow[d, "\bigsqcup_{\mathrm{HorEq}(\mathbb{C})} J_4"'] & \mathbb{C} \arrow[d, "j_{\mathbb{C}}"] \arrow[ddr, bend left, "F"] & \\ \bigsqcup_{\mathrm{HorEq}(\mathbb{C})} \mathbb{H}^{\simeq}E_{\mathrm{adj}} \arrow[r] \arrow[drr, bend right, "K'"'] & \mathbb{C}^{\mathrm{whi}} \arrow[dr, dashed, "\widehat{F}"] & \\ & & \mathbb{A} \end{tikzcd}$$

Here $\bigsqcup_I \mathrm{id}\colon \bigsqcup_{\mathrm{HorEq}(\mathbb{C})} \mathbb{H}^{\simeq}E_{\mathrm{adj}} \to \bigsqcup_{\mathrm{HorEq}(\mathbb{D})} \mathbb{H}^{\simeq}E_{\mathrm{adj}}$ is the double functor induced by the action of I on $\mathrm{HorEq}(\mathbb{C})$. By construction of $\widehat{F}$ and $\widehat{G}$, and using the universal property of the pushout for $\mathbb{C}^{\mathrm{whi}}$, we have that the following diagram commutes:

$$\begin{tikzcd}
\mathbb{C} \arrow[r, "j_{\mathbb{C}}"] \arrow[d, "I"'] \arrow[rr, bend left, "F"] & \mathbb{C}^{\mathrm{whi}} \arrow[r, "\widehat{F}"] \arrow[d, "I^{\mathrm{whi}}"'] & \mathbb{A} \arrow[d, "P"] \\
\mathbb{D} \arrow[r, "j_{\mathbb{D}}"'] \arrow[rr, bend right, "G"'] & \mathbb{D}^{\mathrm{whi}} \arrow[r, "\widehat{G}"'] \arrow[ur, dashed, "\widehat{L}"] & \mathbb{B}
\end{tikzcd}$$

Since $I^{\mathrm{whi}} \in \mathcal{J}_w$–cof by Lemma 4.12 as $I \in \mathcal{C} \cap \mathcal{W}$, there is a lift $\widehat{L}$ in the right-hand square of the diagram above, and the composite $L := \widehat{L}\, j_{\mathbb{D}}$ gives the desired lift. □

5 Proof of Theorem 3.26

We now use the technical results of Section 4 to prove the remaining claims in Theorem 3.26. Namely, we show that the pairs $(\mathcal{C}, \mathcal{F} \cap \mathcal{W})$ and $(\mathcal{C} \cap \mathcal{W}, \mathcal{F})$ form weak factorization systems, and identify the fibrant objects as the weakly horizontally invariant double categories.

Since by definition we have that $\mathcal{C} = \mathcal{I}_w$–cof, in order to prove that $(\mathcal{C}, \mathcal{F} \cap \mathcal{W})$ is a weak factorization system it suffices to show that $\mathcal{F} \cap \mathcal{W} = \mathcal{I}_w$–inj; this is the content of the following result.

Proposition 5.1 *We have that* $\mathcal{F} \cap \mathcal{W} = \mathcal{I}_w$–inj.

Proof Since $\mathcal{C} \cap \mathcal{W} \subseteq \mathcal{C} = \mathcal{I}_w$–cof, it follows that $\mathcal{I}_w$–inj $= \mathcal{I}_w$–cof$^{\boxslash} \subseteq (\mathcal{C} \cap \mathcal{W})^{\boxslash} = \mathcal{F}$. Moreover, every double functor in $\mathcal{I}_w$–inj is a double biequivalence by Remark 3.10, and these are in $\mathcal{W}$ by Proposition 3.25; hence $\mathcal{I}_w$–inj $\subseteq \mathcal{W}$.

For the inclusion $\mathcal{F} \cap \mathcal{W} \subseteq \mathcal{I}_w$–inj, note that every double functor P in $\mathcal{F} \cap \mathcal{W}$ factors as $P = QI$ with $I \in \mathcal{C} = \mathcal{I}_w$–cof and $Q \in \mathcal{I}_w$–inj. Since $Q \in \mathcal{W}$ by the above inclusion, and $P \in \mathcal{W}$ by assumption, we get that $I \in \mathcal{W}$ by 2-out-of-3; hence $I \in \mathcal{C} \cap \mathcal{W}$. Therefore, since $P \in \mathcal{F} = (\mathcal{C} \cap \mathcal{W})^{\boxslash}$ has the right lifting property with respect to I, by the retract argument (see Remark 3.4) we have that P is a retract of Q and hence is also in $\mathcal{I}_w$–inj. □

Before moving on to the next factorization system, we focus on the fibrant objects. Aside from obtaining the desired characterization as the weakly horizontally invariant double categories, we see that the weakly horizontally invariant replacements $j_{\mathbb{A}} \colon \mathbb{A} \to \mathbb{A}^{\mathrm{whi}}$ of Construction 3.17 are trivial cofibrations, and hence fibrant replacements in our model structure.

Theorem 5.2 *A double category $\mathbb{A}$ is fibrant if and only if it is weakly horizontally invariant.*

Proof We recall from Corollary 4.4 that a double category $\mathbb{A}$ is weakly horizontally invariant if and only if the double functor $\mathbb{A} \to \mathbb{1}$ is in $\mathcal{J}_w$–inj. Since $\mathbb{1}$ is weakly horizontally invariant, by Proposition 4.13 this holds if and only if $\mathbb{A} \to \mathbb{1}$ is in $\mathcal{F}$, ie $\mathbb{A}$ is fibrant. □

Proposition 5.3 *Let $\mathbb{A}$ be a weakly horizontally invariant double category. Then the double functor $j_{\mathbb{A}} \colon \mathbb{A} \to \mathbb{A}^{\mathrm{whi}}$ is a double biequivalence.*

Proof By construction, $j_{\mathbb{A}} \colon \mathbb{A} \to \mathbb{A}^{\mathrm{whi}}$ is a double functor in $\{J_4\}$–cof (recall Construction 3.17). Since $\mathcal{J}_w$–inj $\subseteq \{J_4\}$–inj by Proposition 4.5,

$$\{J_4\}\text{–cof} = {}^{\boxslash}\{J_4\}\text{–inj} \subseteq {}^{\boxslash}\mathcal{J}_w\text{–inj} = \mathcal{J}_w\text{–cof}.$$

Hence $j_{\mathbb{A}}$ is a $\mathcal{J}_w$–cofibration with weakly horizontally invariant source, and thus a double biequivalence by Proposition 4.9. □

Corollary 5.4 *The double functor $j_{\mathbb{A}} \colon \mathbb{A} \to \mathbb{A}^{\mathrm{whi}}$ is in $\mathcal{C} \cap \mathcal{W}$. In particular, this exhibits $\mathbb{A}^{\mathrm{whi}}$ as a fibrant replacement of $\mathbb{A}$.*

Proof Since $\mathcal{J}_w$–inj $\subseteq \{J_4\}$–inj by Proposition 4.5, and $\mathcal{J}_w$–cof $\subseteq \mathcal{I}_w$–cof by Remark 4.2, we have that J_4 is in $\mathcal{I}_w$–cof $= \mathcal{C}$. Hence so is $j_{\mathbb{A}}$, as it is constructed as a pushout of coproducts of J_4. The fact that $j_{\mathbb{A}}$ is in $\mathcal{W}$ follows from the relation $(j_{\mathbb{A}})^{\mathrm{whi}} = j_{\mathbb{A}^{\mathrm{whi}}}$ and the fact that the latter is a double biequivalence by Proposition 5.3. The second statement then follows from Theorem 5.2. □

We can also prove, using Proposition 5.3, that every weak equivalence with fibrant source is a double biequivalence. In particular, this implies that while our weak equivalences are more general, when restricted to the fibrant double categories they agree with the weak equivalences of the model structure on DblCat of [16]: the double biequivalences. As we will see in Section 8, the weak equivalences with fibrant source also admit a familiar description in terms of pseudoinverses.

Proposition 5.5 *Let $F\colon \mathbb{A} \to \mathbb{B}$ be a double functor with $\mathbb{A}$ weakly horizontally invariant. Then F is in $\mathcal{W}$ if and only if F is a double biequivalence.*

Proof If F is a double biequivalence, then F is in $\mathcal{W}$ by Proposition 3.25.

Now suppose that F is in $\mathcal{W}$, ie that $F^{\mathrm{whi}}\colon \mathbb{A}^{\mathrm{whi}} \to \mathbb{B}^{\mathrm{whi}}$ is a double biequivalence. We prove that F satisfies (db1)–(db4) of Definition 2.9. Since F and F^{whi} coincide on underlying horizontal categories by Remark 3.18 and $j_{\mathbb{B}}$ is fully faithful on squares, F satisfies (db1)–(db2) as F^{whi} does so. Moreover, since $j_{\mathbb{A}}$, $j_{\mathbb{B}}$, and F^{whi} are fully faithful on squares and $F^{\mathrm{whi}} j_{\mathbb{A}} = j_{\mathbb{B}} F$, we have that F satisfies (db4).

It remains to prove (db3). Since $\mathbb{A}$ is weakly horizontally invariant, Proposition 5.3 guarantees that $j_{\mathbb{A}}\colon \mathbb{A} \to \mathbb{A}^{\mathrm{whi}}$ is a double biequivalence; hence so is the composite $F^{\mathrm{whi}} j_{\mathbb{A}}\colon \mathbb{A} \to \mathbb{B}^{\mathrm{whi}}$. Then, given a vertical morphism $v\colon B \twoheadrightarrow B'$ in $\mathbb{B}$, there is a vertical morphism $u\colon A \twoheadrightarrow A'$ in $\mathbb{A}$ and a weakly horizontally invertible square β in $\mathbb{B}^{\mathrm{whi}}$ as depicted below left:

$$
\begin{tikzcd}
FA \arrow[r, "b", "\simeq"'] \arrow[d, "F^{\mathrm{whi}} j_{\mathbb{A}} u = j_{\mathbb{B}} Fu"'] \arrow[dr, phantom, "\beta \ \simeq"] & B \arrow[d, "j_{\mathbb{B}} v"] \\
FA' \arrow[r, "\simeq", "b'"'] & B'
\end{tikzcd}
\qquad\qquad
\begin{tikzcd}
FA \arrow[r, "b", "\simeq"'] \arrow[d, "Fu"'] \arrow[dr, phantom, "\beta' \ \simeq"] & B \arrow[d, "v"] \\
FA' \arrow[r, "\simeq", "b'"'] & B'
\end{tikzcd}
$$

By fully faithfulness on squares of $j_{\mathbb{B}}$, we get a weakly horizontally invertible square β' in $\mathbb{B}$ as depicted above right, which shows (db3). $\square$

We are now ready to finish the proof of Theorem 3.26 by showing that the classes of trivial cofibrations and fibrations form a weak factorization system. We first show that every double functor can be factored as a trivial cofibration followed by a fibration.

Theorem 5.6 *Every double functor $F\colon \mathbb{A} \to \mathbb{B}$ can be factored as $F = RI$ with $I \in \mathcal{C} \cap \mathcal{W}$ and $R \in \mathcal{F}$.*

Proof Given a double functor $F\colon \mathbb{A} \to \mathbb{B}$, we factor F^{whi} as

$$
\begin{tikzcd}
\mathbb{A}^{\mathrm{whi}} \arrow[rr, "F^{\mathrm{whi}}"] \arrow[dr, "J"'] & & \mathbb{B}^{\mathrm{whi}} \\
& \mathbb{C} \arrow[ur, "P"'] &
\end{tikzcd}
$$

where $J \in \mathcal{J}_w$–cof and $P \in \mathcal{J}_w$–inj. As $\mathbb{B}^{\mathrm{whi}}$ is weakly horizontally invariant, by Proposition 4.13 we have that $P \in \mathcal{F}$, and hence $\mathbb{C}$ is also weakly horizontally invariant.

Define $\mathbb{D}$ to be the pullback of P along $j_{\mathbb{B}}$ as in the diagram

$$\begin{tikzcd}
\mathbb{A} \arrow[dr, dashed, "K"] \arrow[drr, bend left, "F"] \arrow[dd, "j_{\mathbb{A}}"'] & & \\
& \mathbb{D} \arrow[r, "P'"] \arrow[d, "\pi"'] \arrow[dr, phantom, "\lrcorner", very near start] & \mathbb{B} \arrow[d, "j_{\mathbb{B}}"] \\
\mathbb{A}^{\mathrm{whi}} \arrow[r, "J"'] & \mathbb{C} \arrow[r, "P"'] & \mathbb{B}^{\mathrm{whi}}
\end{tikzcd}$$

Then there is a unique double functor $K\colon \mathbb{A} \to \mathbb{D}$ making the above diagram commute. To prove the result, it suffices to show that K is in $\mathcal{W}$. Indeed, assume that this is the case and factor K as $K = QI$ with $I \in \mathcal{I}_w$–cof $= \mathcal{C}$ and $Q \in \mathcal{I}_w$–inj $= \mathcal{F} \cap \mathcal{W}$, where the latter equality holds by Proposition 5.1. As $K, Q \in \mathcal{W}$, we have $I \in \mathcal{C} \cap \mathcal{W}$ by 2-out-of-3. Hence, as $F = P'K$, this gives a factorization of F as $F = RI$ with $I \in \mathcal{C} \cap \mathcal{W}$ and $R := P'Q \in \mathcal{F}$, as desired.

As J is in $\mathcal{W}$ by Proposition 3.25, $j_{\mathbb{A}}$ is in $\mathcal{W}$ by Corollary 5.4, and $\pi K = J j_{\mathbb{A}}$, in order to prove that K is in $\mathcal{W}$, by 2-out-of-3 it is enough to show that π is in $\mathcal{W}$. For this, we construct a double functor $\hat{\pi}\colon \mathbb{D}^{\mathrm{whi}} \to \mathbb{C}$ such that $\pi = \hat{\pi} j_{\mathbb{D}}$ and then show that $\hat{\pi}$ is a double biequivalence; this implies that $\pi \in \mathcal{W}$ by 2-out-of-3.

Let $T := \mathrm{HorEq}(\mathbb{D}) \setminus K(\mathrm{HorEq}(\mathbb{A}))$. As $\mathbb{C}$ is weakly horizontally invariant, by Remark 4.6 there is a lift L in the diagram

$$\begin{tikzcd}
\Big(\bigsqcup_{\mathrm{HorEq}(\mathbb{A})} \mathbb{H}^{\simeq} E_{\mathrm{adj}}\Big) \sqcup \Big(\bigsqcup_T \mathbb{H} E_{\mathrm{adj}}\Big) \arrow[r] \arrow[d, "(\bigsqcup_K \mathrm{id}) \sqcup (\bigsqcup_T J_4)"'] & \mathbb{A}^{\mathrm{whi}} \sqcup \mathbb{D} \arrow[r, "J \sqcup \pi"] & \mathbb{C} \\
\bigsqcup_{\mathrm{HorEq}(\mathbb{D})} \mathbb{H}^{\simeq} E_{\mathrm{adj}} \arrow[urr, dashed, "L"'] & &
\end{tikzcd}$$

This yields a double functor $\hat{\pi}\colon \mathbb{D}^{\mathrm{whi}} \to \mathbb{C}$, given by the universal property of the pushout, as depicted below:

$$\begin{tikzcd}
\bigsqcup_{\mathrm{HorEq}(\mathbb{D})} \mathbb{H} E_{\mathrm{adj}} \arrow[r] \arrow[d, "\bigsqcup_{\mathrm{HorEq}(\mathbb{D})} J_4"'] & \mathbb{D} \arrow[d, "j_{\mathbb{D}}"] \arrow[ddr, bend left, "\pi"] & \\
\bigsqcup_{\mathrm{HorEq}(\mathbb{D})} \mathbb{H}^{\simeq} E_{\mathrm{adj}} \arrow[r] \arrow[drr, bend right, "L"'] & \mathbb{D}^{\mathrm{whi}} \arrow[ul, phantom, "\ulcorner", very near start] \arrow[dr, dashed, "\hat{\pi}"'] & \\
& & \mathbb{C}
\end{tikzcd}$$

We finally show $\hat{\pi}$ satisfies (db1)–(db4). First note that π is fully faithful on squares as it is a pullback of $j_{\mathbb{B}}$ which satisfies this condition. Hence $\hat{\pi}$ also satisfies (db4), since $\hat{\pi} j_{\mathbb{D}} = \pi$. By Proposition 4.9, we know that $J\colon \mathbb{A}^{\mathrm{whi}} \to \mathbb{C}$ in $\mathcal{J}_w$–cof is a double biequivalence, and so (db1)–(db3) for $\hat{\pi}$ follow from the fact that J satisfies (db1)–(db3) and that $\hat{\pi} K^{\mathrm{whi}} = J$, by construction. □

As a direct consequence of this result, we get that the trivial cofibrations are precisely the double functors which have the left lifting property with respect to all fibrations. This concludes the proof of the existence of the model structure.

Corollary 5.7 *We have that $\mathcal{C} \cap \mathcal{W} = {}^{\boxslash}\mathcal{F}$.*

Proof By definition of $\mathcal{F}$, we already know that $\mathcal{C} \cap \mathcal{W} \subseteq {}^{\boxslash}\mathcal{F}$. The reverse inclusion follows from Theorem 5.6, the retract argument (Remark 3.4), and the fact that $\mathcal{C} \cap \mathcal{W}$ is closed under retracts. □

Remark 5.8 This shows that $\mathcal{J}_w$–cof $\subseteq \mathcal{C} \cap \mathcal{W}$. Indeed, we have that $\mathcal{F} \subseteq \mathcal{J}_w$–inj by Lemma 4.10, and hence $\mathcal{J}_w$–cof $= {}^{\boxslash}\mathcal{J}_w$–inj $\subseteq {}^{\boxslash}\mathcal{F} = \mathcal{C} \cap \mathcal{W}$.

6 Quillen pairs

Having constructed a new model structure on DblCat, it is natural to wonder how it compares to the one defined by the authors in [16]. We settle this question by showing that the identity functor induces a Quillen pair between our two model structures, but not a Quillen equivalence.

We then devote the rest of the section to comparing our model structure on DblCat to Lack's model structure on 2Cat; see [13; 14] for more details. As in [16], the horizontal embedding $\mathbb{H}\colon 2\mathrm{Cat} \to \mathrm{DblCat}$ is a left Quillen and homotopically fully faithful functor, but it is no longer right Quillen as it does not preserve fibrant objects. Instead, this role is now played by its more homotopical version $\mathbb{H}^{\simeq}\colon 2\mathrm{Cat} \to \mathrm{DblCat}$, which is also homotopically fully faithful. Furthermore, the double category $\mathbb{H}^{\simeq}\mathcal{A}$ associated to a 2–category $\mathcal{A}$ is weakly horizontally invariant and provides a fibrant replacement for $\mathbb{H}\mathcal{A}$.

First, we show that the identity adjunction embeds the homotopy theory of weakly horizontally invariant double categories into that of double categories.

Theorem 6.1 *The identity adjunction*

$$\begin{tikzcd} \mathrm{DblCat}_{\mathrm{whi}} \arrow[r, bend right, "\mathrm{id}"'] \arrow[r, phantom, "\perp"] & \mathrm{DblCat} \arrow[l, bend right, "\mathrm{id}"'] \end{tikzcd}$$

is a Quillen pair between the model structure on DblCat *for weakly horizontally invariant double categories of Theorem 3.26 and the one of* [16, Theorem 3.18]. *Moreover, the derived counit is levelwise a weak equivalence.*

Proof The set $\mathcal{I}'$ of generating cofibrations introduced in [16, Proposition 4.3] for the model structure on DblCat constructed therein can be described as the set $\mathcal{I}_w$ where the inclusion $I_3\colon \mathbb{1} \sqcup \mathbb{1} \to \mathbb{V}\mathbb{2}$ is replaced by the unique map $\varnothing \to \mathbb{V}\mathbb{2}$. Since the latter is also in $\mathcal{I}_w$–cof, it follows that $\mathcal{I}'$–cof $\subseteq \mathcal{I}_w$–cof, and hence $\mathrm{id}\colon \mathrm{DblCat} \to \mathrm{DblCat}_{\mathrm{whi}}$ preserves cofibrations. Furthermore, by Proposition 3.25, we have that the class of double biequivalences — which is precisely the class of weak equivalences for the model structure on DblCat of [16] — is contained in the class $\mathcal{W}$ of weak equivalences in $\mathrm{DblCat}_{\mathrm{whi}}$, and hence $\mathrm{id}\colon \mathrm{DblCat} \to \mathrm{DblCat}_{\mathrm{whi}}$ also preserves weak equivalences. This shows that the identity adjunction is a Quillen pair.

It remains to show that the derived counit is levelwise a weak equivalence in $\mathrm{DblCat}_{\mathrm{whi}}$. Let $\mathbb{A}$ be a fibrant double category in $\mathrm{DblCat}_{\mathrm{whi}}$. Then the component of the derived counit at $\mathbb{A}$ is given by the cofibrant replacement $q_{\mathbb{A}}\colon \mathbb{A}^c \to \mathbb{A}$ in the model structure on DblCat of [16]. In particular, the double functor $q_{\mathbb{A}}$ is a double biequivalence, and hence a weak equivalence in $\mathrm{DblCat}_{\mathrm{whi}}$ by Proposition 3.25. □

However, the identity adjunction does not induce a Quillen equivalence between the two model structures on DblCat, as shown in the following remark.

Remark 6.2 The derived unit of the identity adjunction above is not a levelwise double biequivalence. To see this, recall the double category $\mathbb{A}$ described in Example 3.23. By [16, Proposition 4.9], $\mathbb{A}$ is cofibrant in the model structure on DblCat of [16]. Then the component of the derived unit at $\mathbb{A}$ is given by a fibrant replacement of $\mathbb{A}$ in $\mathrm{DblCat}_{\mathrm{whi}}$, and hence we can consider the weakly horizontally invariant replacement $j_{\mathbb{A}}\colon \mathbb{A} \to \mathbb{A}^{\mathrm{whi}}$ given in Construction 3.17. In particular, as shown in Example 3.23, this is not a double biequivalence.

As a direct consequence of the above theorem, and the fact that $\mathbb{H} \dashv \boldsymbol{H}$ is a Quillen pair between Lack's model structure on 2Cat and the model structure on DblCat of [16], we get that $\mathbb{H} \dashv \boldsymbol{H}$ is also a Quillen pair between 2Cat and the model structure on DblCat introduced in this paper. Moreover, the derived unit is levelwise a biequivalence, and so the functor $\mathbb{H}$ is homotopically fully faithful.

Theorem 6.3 *The adjunction*

$$\begin{tikzcd} \mathrm{DblCat} \arrow[r, phantom, "\perp"] \arrow[r, bend right, "\boldsymbol{H}"'] & 2\mathrm{Cat} \arrow[l, bend right, "\mathbb{H}"'] \end{tikzcd}$$

is a Quillen pair between the model structure of Theorem 3.26 and Lack's model structure. Moreover, the derived unit is levelwise a biequivalence.

Proof The fact that this is a Quillen pair follows directly from Theorem 6.1 and [16, Proposition 6.1]. To show that the derived unit is levelwise a biequivalence, let $\mathcal{A}$ be a cofibrant 2–category. The component of the derived unit at $\mathcal{A}$ is given by the underlying horizontal 2–functor of a fibrant replacement $\boldsymbol{H} j_{\mathbb{H}\mathcal{A}} \colon \mathcal{A} = \boldsymbol{H}\mathbb{H}\mathcal{A} \to \boldsymbol{H}(\mathbb{H}\mathcal{A})^f$ of the horizontal double category $\mathbb{H}\mathcal{A}$ in DblCat. In particular, if we consider the fibrant replacement given in Construction 3.17, it does not change the underlying horizontal 2–category of $\mathbb{H}\mathcal{A}$ by Remark 3.18. Hence $\boldsymbol{H} j_{\mathbb{H}\mathcal{A}}$ is an identity, and in particular a biequivalence. □

As opposed to the case where DblCat is endowed with the model structure of [16] — see [16, Theorem 6.2] — the horizontal embedding is not right Quillen when we consider our new model structure.

Remark 6.4 The functor $\mathbb{H}$ is not right Quillen as, for example, the horizontal double category $\mathbb{H}E_{\mathrm{adj}}$ is not weakly horizontally invariant, where E_{adj} denotes the free-living adjoint equivalence, as shown in Remark 2.12. Since every 2–category is fibrant, this implies that $\mathbb{H}$ does not preserve fibrant objects.

This shortcoming of the horizontal embedding $\mathbb{H}$ can be remedied by instead considering the homotopical horizontal embedding $\mathbb{H}^{\simeq} \colon 2\mathrm{Cat} \to \mathrm{DblCat}$ of Definition 2.13. As we will see, the adjunction $L^{\simeq} \dashv \mathbb{H}^{\simeq}$ of Proposition 2.15 is compatible with the model structures considered, making the functor $\mathbb{H}^{\simeq}$ right Quillen. As a first step towards this, we show that $\mathbb{H}^{\simeq}$ provides a levelwise fibrant replacement of $\mathbb{H}$ in our model structure on DblCat.

Theorem 6.5 *Let $\mathcal{A}$ be a 2–category. Then the double category $\mathbb{H}^{\simeq}\mathcal{A}$ is weakly horizontally invariant and the inclusion $J_{\mathcal{A}} \colon \mathbb{H}\mathcal{A} \to \mathbb{H}^{\simeq}\mathcal{A}$ is a double biequivalence. In particular, this exhibits $\mathbb{H}^{\simeq}\mathcal{A}$ as a fibrant replacement of $\mathbb{H}\mathcal{A}$ in the model structure on* DblCat *of Theorem 3.26.*

Proof For the first statement, we have by [15, Lemma A.2.3] that a weakly horizontally invertible square $\alpha\colon (u\,{}^{a}_{a'}\,w)$ in $\mathbb{H}^{\simeq}\mathcal{A}$ corresponds to a 2–isomorphism $\alpha\colon wa \Rightarrow a'u$ in $\mathcal{A}$, where (a, c, η, ϵ) and $(a', c', \eta', \epsilon')$ are equivalences in $\mathcal{A}$. In particular, given a boundary in $\mathbb{H}^{\simeq}\mathcal{A}$ as below left,

$$\begin{tikzcd} A \arrow[r, "a", "\simeq"'] & C \arrow[d, "w", "\bullet"' marking] \\ A' \arrow[r, "\simeq", "a'"'] & C' \end{tikzcd} \qquad \begin{tikzcd} A \arrow[r, "a", "\simeq"'] \arrow[d, "u"', "\bullet" marking] \arrow[dr, phantom, "\alpha\ \simeq"] & C \arrow[d, "w", "\bullet"' marking] \\ A' \arrow[r, "\simeq", "a'"'] & C' \end{tikzcd}$$

there is an equivalence $u := c'wa\colon A \xrightarrow{\simeq} A'$ and a 2–isomorphism

$$\alpha := (\epsilon')^{-1}wa\colon wa \cong a'u$$

in $\mathcal{A}$, which provides a square as desired, depicted above right. This shows that $\mathbb{H}^{\simeq}\mathcal{A}$ is weakly horizontally invariant.

For the second statement, recall that $\boldsymbol{H}\,\mathbb{H}\mathcal{A} = \mathcal{A} = \boldsymbol{H}\,\mathbb{H}^{\simeq}\mathcal{A}$, and thus the inclusion $J_{\mathcal{A}}\colon \mathbb{H}\mathcal{A} \to \mathbb{H}^{\simeq}\mathcal{A}$ is the identity on underlying horizontal 2–categories; this shows that $J_{\mathcal{A}}$ satisfies (db1), (db2) and (db4) of Definition 2.9. It remains to show (db3). Let $u\colon A \to\!\!\!\!\bullet\ A'$ be a vertical morphism in $\mathbb{H}^{\simeq}\mathcal{A}$, ie an adjoint equivalence $u\colon A \xrightarrow{\simeq} A'$ in $\mathcal{A}$. Then the square

$$\begin{tikzcd} A \arrow[r, equal] \arrow[d, equal, "J_{\mathcal{A}}(e_A)"', "\bullet" marking] \arrow[dr, phantom, "\simeq"] & A \arrow[d, "u", "\bullet"' marking] \\ A \arrow[r, "\simeq", "u"'] & A' \end{tikzcd}$$

induced by the identity at u gives a weakly horizontally invertible square in $\mathbb{H}^{\simeq}\mathcal{A}$ as required. This shows that $J_{\mathcal{A}}$ is a double biequivalence. The second statement then follows from Proposition 3.25. □

We now show that the double functor $\mathbb{H}^{\simeq}$ is right Quillen, and moreover, that it is homotopically fully faithful.

Theorem 6.6 *The adjunction*

$$\begin{tikzcd} \mathrm{2Cat} \arrow[r, bend right, "\mathbb{H}^{\simeq}"'] \arrow[r, phantom, "\perp"] & \mathrm{DblCat} \arrow[l, bend right, "L^{\simeq}"'] \end{tikzcd}$$

is a Quillen pair between the model structure of Theorem 3.26 and Lack's model structure. Moreover, the derived counit is levelwise a weak equivalence.

Proof We show that $\mathbb{H}^{\simeq}\colon \mathrm{2Cat} \to \mathrm{DblCat}$ preserves fibrations and trivial fibrations. Let $F\colon \mathcal{A} \to \mathcal{B}$ be a Lack fibration in 2Cat. Since $\mathbb{H}^{\simeq}\mathcal{B}$ is weakly horizontally invariant (see Theorem 6.5), Proposition 4.13 guarantees that $\mathbb{H}^{\simeq}F$ is a fibration in DblCat if and only if it is $\mathcal{J}_w$–injective. Hence, we need to prove that $\mathbb{H}^{\simeq}F$ satisfies (df1)–(df3) of Proposition 4.3.

First note that $\mathbb{H}^{\simeq}F$ satisfies (df1)–(df2) by definition of F being a fibration in 2Cat. It remains to prove (df3). Consider a diagram in $\mathbb{H}^{\simeq}\mathcal{A}$ as below left, together with a weakly horizontally invertible square β in $\mathbb{H}^{\simeq}\mathcal{B}$, as depicted below right,

$$
\begin{tikzcd}
A \arrow[r, "a", "\simeq"'] & C \arrow[d, "w", "\bullet"'] \\
A' \arrow[r, "a'", "\simeq"'] & C'
\end{tikzcd}
\qquad\qquad
\begin{tikzcd}
FA \arrow[r, "Fa", "\simeq"'] \arrow[d, "v"', "\bullet"] \arrow[dr, phantom, "\beta\ \simeq"] & FC \arrow[d, "Fw", "\bullet"'] \\
FA' \arrow[r, "\simeq", "Fa'"'] & FC'
\end{tikzcd}
$$

ie a 2–isomorphism $\beta\colon (Fw)(Fa) \Rightarrow (Fa')v$ in $\mathcal{B}$ by [15, Lemma A.2.3]. Let (c', a', η, ϵ) be a choice of adjoint equivalence data for a', and let δ be the 2–isomorphism in $\mathcal{B}$ given by the pasting below left:

$$
\begin{tikzcd}
FA \arrow[r, "Fa", "\simeq"'] \arrow[d, "v"', "\wr"] \arrow[dr, phantom, "\overset{\beta}{\swarrow}\cong"] & FC \arrow[d, "Fw", "\wr"'] \\
FA' \arrow[r, dashed, "Fa'"] \arrow[dr, equal] \arrow[r, phantom, shift right=4ex, "\overset{F\epsilon}{\swarrow}\cong"] & FC' \arrow[d, "Fc'", "\wr"'] \\
& FA'
\end{tikzcd}
\qquad\qquad
\begin{tikzcd}
A \arrow[rr, "a", "\simeq"'] \arrow[dd, "u"', "\wr"] \arrow[ddrr, phantom, "\overset{\bar\alpha}{\swarrow}\cong"] & & C \arrow[d, "w", "\wr"'] \\
& & C' \arrow[dll, "c'"'] \arrow[d, equal] \arrow[dl, phantom, "\cong\overset{\eta}{\swarrow}"] \\
A' \arrow[rr, "\simeq", "a'"'] & & C'
\end{tikzcd}
$$

Since F is a fibration in 2Cat, there is an equivalence $u\colon A \xrightarrow{\simeq} A'$ in $\mathcal{A}$ and a 2–isomorphism $\bar\alpha\colon c'wa \cong u$ in $\mathcal{A}$ such that $\delta = F\bar\alpha$. We set $\alpha\colon wa \cong a'u$ to be the pasting above right; by the triangle identities for (η, ϵ), we get that $\beta = F\alpha$ as desired. This proves that $\mathbb{H}^{\simeq}F$ is a fibration in DblCat.

Now suppose that $F\colon \mathcal{A} \to \mathcal{B}$ is a trivial fibration in 2Cat. By definition, we directly see that $\mathbb{H}^{\simeq}F$ is surjective on objects, full on horizontal morphisms, and fully faithful on squares. Fullness on vertical morphisms for $\mathbb{H}^{\simeq}F$ follows from the fact that a lift of an adjoint equivalence by a biequivalence is also an adjoint equivalence. Hence $\mathbb{H}^{\simeq}F$ is a trivial fibration in DblCat by Proposition 3.9, and this shows that $\mathbb{H}^{\simeq}$ is right Quillen.

It remains to show that the derived counit is levelwise a biequivalence. Let $\mathcal{A}$ be a 2–category, and let $q_{\mathbb{H}^{\simeq}\mathcal{A}}\colon (\mathbb{H}^{\simeq}\mathcal{A})^c \to \mathbb{H}^{\simeq}\mathcal{A}$ denote the cofibrant replacement of $\mathbb{H}^{\simeq}\mathcal{A}$ constructed as follows. The double category $(\mathbb{H}^{\simeq}\mathcal{A})^c$ has the same objects as $\mathcal{A}$; it has a copy $\bar{a}$ for each morphism a in $\mathcal{A}$, and horizontal morphisms in $(\mathbb{H}^{\simeq}\mathcal{A})^c$ are given by free composites of $\bar{a}$'s; it has a copy $\bar{u}$ for each adjoint equivalence u in $\mathcal{A}$, and vertical morphisms in $(\mathbb{H}^{\simeq}\mathcal{A})^c$ are given by free composites of $\bar{u}$'s; and squares in $(\mathbb{H}^{\simeq}\mathcal{A})^c$ are given by squares of $\mathbb{H}^{\simeq}\mathcal{A}$ whose boundaries are the actual composites in $\mathbb{H}^{\simeq}\mathcal{A}$ of the representative of the free composites.

Then, by studying the data of the 2–category $L^{\simeq}(\mathbb{H}^{\simeq}\mathcal{A})^c$, we can see that the derived counit at $\mathcal{A}$

$$L^{\simeq}(\mathbb{H}^{\simeq}\mathcal{A})^c \xrightarrow{L^{\simeq}q_{\mathbb{H}^{\simeq}\mathcal{A}}} L^{\simeq}\mathbb{H}^{\simeq}\mathcal{A} \xrightarrow{\epsilon_{\mathcal{A}}} \mathcal{A}$$

is a trivial fibration in 2Cat as it is surjective on objects, full on morphisms, and fully faithful on 2–morphisms. □

Remark 6.7 The components of the derived unit of the adjunction $L^{\simeq} \dashv \mathbb{H}^{\simeq}$ are not weak equivalences in DblCat. Indeed, since every 2–category is fibrant, we know that the counit and the derived counit agree on cofibrant double categories. Then, if we consider the component $\eta_{\mathbb{V}\mathbb{2}}\colon \mathbb{V}\mathbb{2} \to \mathbb{H}^{\simeq}L^{\simeq}\mathbb{V}\mathbb{2} = \mathbb{H}^{\simeq}E_{\mathrm{adj}}$ of the unit at the cofibrant double category $\mathbb{V}\mathbb{2}$, we see that $\mathbb{H}^{\simeq}E_{\mathrm{adj}}$ has nontrivial horizontal morphisms, given by the adjoint equivalence created by $L^{\simeq}$ from the unique vertical morphism of $\mathbb{V}\mathbb{2}$, while $\mathbb{V}\mathbb{2}$ does not. Therefore $\eta_{\mathbb{V}\mathbb{2}}$ is not a double biequivalence, as it does not satisfy (db2). Then, since $\mathbb{V}\mathbb{2}$ is weakly horizontally invariant, Proposition 5.5 implies that $\eta_{\mathbb{V}\mathbb{2}}$ is not a weak equivalence in DblCat.

While Theorem 6.6 implies that $\mathbb{H}^{\simeq}\colon 2\mathrm{Cat} \to \mathrm{DblCat}$ preserves weak equivalences and fibrations, the following result says that it further reflects these classes of double functors. Hence the model structure on 2Cat is completely determined from our model structure on DblCat through its image under $\mathbb{H}^{\simeq}$.

Theorem 6.8 *Lack's model structure on* 2Cat *is right-induced along the adjunction*

$$\begin{tikzcd} 2\mathrm{Cat} \arrow[r, bend right, "\mathbb{H}^{\simeq}"'] \arrow[r, phantom, "\perp"] & \mathrm{DblCat} \arrow[l, bend right, "L^{\simeq}"'] \end{tikzcd}$$

from the model structure on DblCat *of Theorem 3.26.*

Proof We need to show that a 2–functor F is a fibration (resp. biequivalence) in 2Cat if and only if $\mathbb{H}^{\simeq}F$ is a fibration (resp. weak equivalence) in DblCat. Since $\mathbb{H}^{\simeq}$ is right Quillen, we know it preserves fibrations and trivial fibrations. Moreover, since all 2–categories are fibrant, by Ken Brown's lemma — see [10, Lemma 1.1.12] — the functor $\mathbb{H}^{\simeq}$ preserves all weak equivalences. Therefore, if F is a fibration (resp. biequivalence), then $\mathbb{H}^{\simeq}F$ is a fibration (resp. weak equivalence).

If $\mathbb{H}^{\simeq}F$ is a fibration in DblCat, then by Proposition 4.13 it is $\mathcal{J}_w$–injective, since its target is weakly horizontally invariant by Theorem 6.5. Hence, conditions (df1)–(df2) of Proposition 4.3 for $\mathbb{H}^{\simeq}F$ say that F is a fibration in 2Cat.

Finally, if $\mathbb{H}^{\simeq}F$ is a weak equivalence in DblCat, then by Proposition 5.5 it is a double biequivalence, since its source is weakly horizontally invariant. By (db1) and (db2) of Definition 2.9, we have that F is biessentially surjective on objects and essentially full on morphisms. Fully faithfulness on 2–morphisms follows from applying (db4) of Definition 2.9 to squares with trivial vertical boundaries. Hence F is a biequivalence. □

Finally, we can use the above result to deduce that Lack's model structure on 2Cat is also left-induced from our model structure on DblCat along the horizontal embedding $\mathbb{H}$.

Theorem 6.9 *Lack's model structure on* 2Cat *is left-induced along the adjunction*

$$\begin{tikzcd} \mathrm{DblCat} \arrow[r, bend right, "\boldsymbol{H}"'] \arrow[r, phantom, "\perp"] & \mathrm{2Cat} \arrow[l, bend right, "\mathbb{H}"'] \end{tikzcd}$$

from the model structure on DblCat *of Theorem 3.26.*

Proof We need to show that a 2–functor $F\colon \mathcal{A}\to\mathcal{B}$ is a cofibration (resp. biequivalence) in 2Cat if and only if $\mathbb{H}F$ is a cofibration (resp. weak equivalence) in DblCat.

By Theorem 3.11, the double functor $\mathbb{H}F$ is a cofibration if and only if its underlying functors $U\boldsymbol{H}\mathbb{H}F$ and $U\boldsymbol{V}\mathbb{H}F$ have the left lifting property with respect to all surjective on objects and full functors. Since $U\boldsymbol{V}\mathbb{H}F$ trivially satisfies this condition, this holds if and only if $UF = U\boldsymbol{H}\mathbb{H}F$ has the mentioned lifting property. By [13, Lemma 4.1], this is equivalent to saying that F is a cofibration.

Finally, since $\mathbb{H}^{\simeq}\mathcal{A}$ and $\mathbb{H}^{\simeq}\mathcal{B}$ are fibrant replacements of $\mathbb{H}\mathcal{A}$ and $\mathbb{H}\mathcal{B}$ in DblCat by Theorem 6.5, we have that $\mathbb{H}F$ is a weak equivalence if and only if $\mathbb{H}^{\simeq}F$ is a weak equivalence. By Theorem 6.8, this is the case if and only if F is a biequivalence. □

7 Compatibility with the Gray tensor product

We now explore the monoidality of the model structure on DblCat constructed in this paper. Although a similar argument as the one in [16, Remark 7.1] quickly shows that our model structure is not monoidal with respect to the cartesian product, in this section we prove that it is monoidal when we instead consider the Gray tensor product for double categories introduced by Böhm in [1]. This resembles the case of Lack's model structure on 2Cat, which is monoidal with respect to the Gray tensor product of 2–categories, and improves upon the model structure on DblCat of [16], which is only 2Cat–enriched.

The Gray tensor product $\otimes_{\mathrm{Gr}}\colon \mathrm{DblCat}\times\mathrm{DblCat}\to\mathrm{DblCat}$ endows the category DblCat with a symmetric monoidal structure, as shown in [1, Section 3]. We first give an explicit description of the Gray tensor product of two double categories.

Description 7.1 The Gray tensor product $\mathbb{A}\otimes_{\mathrm{Gr}}\mathbb{X}$ of two double categories $\mathbb{A}$ and $\mathbb{X}$ can be described as the double category given by the following data:

(i) The objects are pairs (A,X) of objects $A\in\mathbb{A}$ and $X\in\mathbb{X}$.

(ii) Two kinds of generating horizontal morphisms: pairs $(a,X)\colon (A,X)\to(C,X)$, where $a\colon A\to C$ is a horizontal morphism in $\mathbb{A}$ and X is an object in $\mathbb{X}$, which compose as in $\mathbb{A}$; and pairs $(A,x)\colon(A,X)\to(A,Z)$, where A is an object in $\mathbb{A}$ and $x\colon X\to Z$ is a horizontal morphism in $\mathbb{X}$, which compose as in $\mathbb{X}$.

(iii) Similarly, the generating vertical morphisms are given by pairs (u,X) and (A,t) with A and X being objects of $\mathbb{A}$ and $\mathbb{X}$ respectively, and u and t being vertical morphisms of $\mathbb{A}$ and $\mathbb{X}$ respectively.

(iv) There are six kinds of generating squares: the ones determined by a square $\alpha\colon(u\,{}^{a}_{a'}\,w)$ in $\mathbb{A}$ and an object $X\in\mathbb{X}$ as shown below left, the ones given by an object $A\in\mathbb{A}$ and a square $\chi\colon(t\,{}^{x}_{x'}\,v)$ in $\mathbb{X}$ as below right,

$$\begin{tikzcd}
(A,X) \arrow[r, "{(a,X)}"] \arrow[d, "{(u,X)}"', "\bullet" marking] \arrow[dr, phantom, "{(\alpha,X)}"] & (C,X) \arrow[d, "{(w,X)}", "\bullet" marking] \\
(A',X) \arrow[r, "{(a',X)}"'] & (C',X)
\end{tikzcd}
\qquad
\begin{tikzcd}
(A,X) \arrow[r, "{(A,x)}"] \arrow[d, "{(A,t)}"', "\bullet" marking] \arrow[dr, phantom, "{(A,\chi)}"] & (A,Z) \arrow[d, "{(A,v)}", "\bullet" marking] \\
(A,X') \arrow[r, "{(A,x')}"'] & (A,Z')
\end{tikzcd}$$

the squares determined by a horizontal morphism a in $\mathbb{A}$ and a vertical morphism t in $\mathbb{X}$ as displayed below left, and the ones given by a horizontal morphism x in $\mathbb{X}$ and a vertical morphism u in $\mathbb{A}$ as below right,

$$\begin{tikzcd} (A,X) \arrow[r, "{(a,X)}"] \arrow[d, "{(A,t)}"'] \arrow[dr, phantom, "{(a,t)}"] & (C,X) \arrow[d, "{(C,t)}"] \\ (A,X') \arrow[r, "{(a,X')}"'] & (C,X') \end{tikzcd} \qquad \begin{tikzcd} (A,X) \arrow[r, "{(A,x)}"] \arrow[d, "{(u,X)}"'] \arrow[dr, phantom, "{(u,x)}"] & (A,Z) \arrow[d, "{(u,Z)}"] \\ (A',X) \arrow[r, "{(A',x)}"'] & (A',Z) \end{tikzcd}$$

vertically invertible squares determined by horizontal morphisms a in $\mathbb{A}$ and x in $\mathbb{X}$,

$$\begin{tikzcd} (A,X) \arrow[r, "{(a,X)}"] \arrow[d, equal] & (C,X) \arrow[r, "{(C,x)}"] \arrow[d, phantom, "{(a,x)\,\Vert\wr}"] & (C,Z) \arrow[d, equal] \\ (A,X) \arrow[r, "{(A,x)}"'] & (A,Z) \arrow[r, "{(a,Z)}"'] & (C,Z) \end{tikzcd}$$

and horizontally invertible squares given by vertical morphisms u in $\mathbb{A}$ and t in $\mathbb{X}$,

$$\begin{tikzcd} (A,X) \arrow[r, equal] \arrow[d, "{(u,X)}"'] & (A,X) \arrow[d, "{(A,t)}"] \\ (A',X) \arrow[r, phantom, "\underset{(u,t)}{\cong}"] \arrow[d, "{(A',t)}"'] & (A,X') \arrow[d, "{(u,X')}"] \\ (A',X') \arrow[r, equal] & (A',X') \end{tikzcd}$$

subject to conditions which are equivalent to requiring that the projection double functor $\Pi_{\mathbb{A},\mathbb{X}} \colon \mathbb{A} \otimes_{\mathrm{Gr}} \mathbb{X} \to \mathbb{A} \times \mathbb{X}$ is fully faithful on squares.

Remark 7.2 The cartesian product of two double categories is obtained by taking the product of the sets of objects, horizontal morphisms, vertical morphisms, and squares, respectively. The projection double functor $\Pi_{\mathbb{A},\mathbb{X}} \colon \mathbb{A} \otimes_{\mathrm{Gr}} \mathbb{X} \to \mathbb{A} \times \mathbb{X}$ sends the squares of the form (a,x) and (u,t) to the identity squares $(\mathrm{id}_a, \mathrm{id}_x) = \mathrm{id}_{(a,x)}$ and $(e_u, e_t) = e_{(u,t)}$, and acts as the identity on the remaining generators. In particular, it is straightforward from this description that $\Pi_{\mathbb{A},\mathbb{X}}$ is functorial in $\mathbb{A}$ and $\mathbb{X}$. Note that the squares of the form $(a,t) := (\mathrm{id}_a, e_t)$ and $(u,x) := (e_u, \mathrm{id}_x)$ are not identity squares in the product $\mathbb{A} \times \mathbb{X}$ even though they come from identity squares in $\mathbb{A}$ and $\mathbb{X}$.

We can show that the projection $\Pi_{\mathbb{A},\mathbb{X}} \colon \mathbb{A} \otimes_{\mathrm{Gr}} \mathbb{X} \to \mathbb{A} \times \mathbb{X}$ is a trivial fibration in our model structure on DblCat.

Lemma 7.3 *The projection double functor* $\Pi_{\mathbb{A},\mathbb{X}} \colon \mathbb{A} \otimes_{\mathrm{Gr}} \mathbb{X} \to \mathbb{A} \times \mathbb{X}$ *is a trivial fibration, for all double categories* $\mathbb{A}$ *and* $\mathbb{X}$.

Proof We use the characterization of trivial fibrations of Proposition 3.9. Since $\Pi_{\mathbb{A},\mathbb{X}}$ is the identity on objects, it is clearly surjective on objects. Given a horizontal morphism $(a,x)\colon (A,X)\to(C,Z)$ in $\mathbb{A}\times\mathbb{X}$, the composite

$$(A,X)\xrightarrow{(a,X)}(C,X)\xrightarrow{(C,x)}(C,Z)$$

of horizontal morphisms in $\mathbb{A}\otimes_{\mathrm{Gr}}\mathbb{X}$ is sent by $\Pi_{\mathbb{A},\mathbb{X}}$ to (a,x), which shows that $\Pi_{\mathbb{A},\mathbb{X}}$ is full on horizontal morphisms. Similarly, one can show that $\Pi_{\mathbb{A},\mathbb{X}}$ is full on vertical morphisms. Fully faithfulness on squares holds by Description 7.1(iv). $\square$

We now show that $\mathbb{A}^{\mathrm{whi}}\times\mathbb{X}^{\mathrm{whi}}$ gives a fibrant replacement for $\mathbb{A}\times\mathbb{X}$, where $(-)^{\mathrm{whi}}$ is the weakly horizontally invariant replacement of Construction 3.17.

Lemma 7.4 *Let $\mathbb{A}$ and $\mathbb{X}$ be double categories. Then $j_{\mathbb{A}}\times j_{\mathbb{X}}\colon \mathbb{A}\times\mathbb{X}\to\mathbb{A}^{\mathrm{whi}}\times\mathbb{X}^{\mathrm{whi}}$ provides a fibrant replacement for the double category $\mathbb{A}\times\mathbb{X}$.*

Proof First, note that $\mathbb{A}^{\mathrm{whi}}\times\mathbb{X}^{\mathrm{whi}}$ is fibrant, as fibrant objects are closed under products. Now consider the commutative triangle

$$\begin{tikzcd}
 & \mathbb{A}\times\mathbb{X} \arrow[dl, "j_{\mathbb{A}\times\mathbb{X}}"'] \arrow[dr, "j_{\mathbb{A}}\times j_{\mathbb{X}}"] & \\
(\mathbb{A}\times\mathbb{X})^{\mathrm{whi}} \arrow[rr, "(\pi_{\mathbb{A}}^{\mathrm{whi}},\pi_{\mathbb{X}}^{\mathrm{whi}})"'] & & \mathbb{A}^{\mathrm{whi}}\times\mathbb{X}^{\mathrm{whi}}
\end{tikzcd}$$

where the bottom map is induced by the projections

$$\pi_{\mathbb{A}}\colon \mathbb{A}\times\mathbb{X}\to\mathbb{A} \quad\text{and}\quad \pi_{\mathbb{X}}\colon\mathbb{A}\times\mathbb{X}\to\mathbb{X}.$$

Since $j_{\mathbb{A}\times\mathbb{X}}$ is a weak equivalence by Corollary 5.4, to prove that $j_{\mathbb{A}}\times j_{\mathbb{X}}$ is a weak equivalence it suffices to show that $(\pi_{\mathbb{A}}^{\mathrm{whi}},\pi_{\mathbb{X}}^{\mathrm{whi}})$ is a weak equivalence; we use Proposition 3.9 to prove that it is in fact a trivial fibration.

One can see that $(\pi_{\mathbb{A}}^{\mathrm{whi}},\pi_{\mathbb{X}}^{\mathrm{whi}})$ is the identity on underlying horizontal categories, and that it is fully faithful on squares since $j_{\mathbb{A}\times\mathbb{X}}$ and $j_{\mathbb{A}}\times j_{\mathbb{X}}$ are so. Finally, by studying the weakly horizontally invariant replacements, we can see that it is also full on vertical morphisms. Indeed, all the vertical morphisms that were freely added to $\mathbb{A}^{\mathrm{whi}}\times\mathbb{X}^{\mathrm{whi}}$ from the image of $\mathbb{A}\times\mathbb{X}$ were also freely added to $(\mathbb{A}\times\mathbb{X})^{\mathrm{whi}}$ from the image of $\mathbb{A}\times\mathbb{X}$. $\square$

Mirroring the proof in [13, Section 7], we show that the cartesian product and the Gray tensor product of a weak equivalence with an identity is also a weak equivalence.

Remark 7.5 Given a double biequivalence $F\colon \mathbb{A}\to\mathbb{B}$ and a double category $\mathbb{X}$, the product $F\times\mathrm{id}_{\mathbb{X}}\colon \mathbb{A}\times\mathbb{X}\to\mathbb{B}\times\mathbb{X}$ is a double biequivalence. Indeed, it is straightforward to see that (db1)–(db4) of Definition 2.9 hold for $F\times\mathrm{id}_{\mathbb{X}}$ since they do for F.

Proposition 7.6 *Let $F\colon \mathbb{A}\to\mathbb{B}$ be a weak equivalence in the model structure on* DblCat *of Theorem 3.26. Then, for every double category $\mathbb{X}$, the induced double functors*

$$F\times\mathrm{id}_{\mathbb{X}}\colon \mathbb{A}\times\mathbb{X}\to\mathbb{B}\times\mathbb{X}\quad\textit{and}\quad F\otimes_{\mathrm{Gr}}\mathrm{id}_{\mathbb{X}}\colon \mathbb{A}\otimes_{\mathrm{Gr}}\mathbb{X}\to\mathbb{B}\otimes_{\mathrm{Gr}}\mathbb{X}$$

are also weak equivalences in DblCat.

Proof First note that the weakly horizontally invariant replacement F^{whi} is a double biequivalence, since F is a weak equivalence. Hence, by Remark 7.5, the double functor $F^{\mathrm{whi}}\times\mathrm{id}_{\mathbb{X}^{\mathrm{whi}}}\colon \mathbb{A}^{\mathrm{whi}}\times\mathbb{X}^{\mathrm{whi}}\to\mathbb{B}^{\mathrm{whi}}\times\mathbb{X}^{\mathrm{whi}}$ is also a double biequivalence. Since $\mathbb{A}^{\mathrm{whi}}\times\mathbb{X}^{\mathrm{whi}}$ and $\mathbb{B}^{\mathrm{whi}}\times\mathbb{X}^{\mathrm{whi}}$ are fibrant replacements for $\mathbb{A}\times\mathbb{X}$ and $\mathbb{B}\times\mathbb{X}$ by Lemma 7.4, this shows that $F\times\mathrm{id}_{\mathbb{X}}$ is a weak equivalence by 2-out-of-3.

For the statement regarding the Gray tensor product, we know by Lemma 7.3 that the double functors $\Pi_{\mathbb{A},\mathbb{X}}$ and $\Pi_{\mathbb{B},\mathbb{X}}$ are trivial fibrations. Since the diagram

$$\begin{tikzcd}
\mathbb{A}\otimes_{\mathrm{Gr}}\mathbb{X} \arrow[r, "F\otimes_{\mathrm{Gr}}\mathrm{id}_{\mathbb{X}}"] \arrow[d, "\Pi_{\mathbb{A},\mathbb{X}}"'] & \mathbb{B}\otimes_{\mathrm{Gr}}\mathbb{X} \arrow[d, "\Pi_{\mathbb{B},\mathbb{X}}"] \\
\mathbb{A}\times\mathbb{X} \arrow[r, "F\times\mathrm{id}_{\mathbb{X}}"'] & \mathbb{B}\times\mathbb{X}
\end{tikzcd}$$

commutes, $F\otimes_{\mathrm{Gr}}\mathrm{id}_{\mathbb{X}}$ is also a weak equivalence by 2-out-of-3. □

This allows us to prove that our model structure on DblCat is monoidal with respect to the Gray tensor product, inspired by the proof of the monoidality of Lack's model structure on 2Cat of [13, Section 7].

Notation 7.7 Given two double functors $I\colon \mathbb{A}\to\mathbb{B}$ and $J\colon \mathbb{C}\to\mathbb{D}$, we write $I\mathbin{\square} J$ for the pushout-product double functor

$$I\mathbin{\square} J\colon \mathbb{A}\otimes_{\mathrm{Gr}}\mathbb{D}\amalg_{\mathbb{A}\otimes_{\mathrm{Gr}}\mathbb{C}}\mathbb{B}\otimes_{\mathrm{Gr}}\mathbb{C}\to\mathbb{B}\otimes_{\mathrm{Gr}}\mathbb{D}.$$

Theorem 7.8 *The model structure on* DblCat *of Theorem 3.26 is monoidal with respect to the Gray tensor product $\otimes_{\mathrm{Gr}}$.*

Proof We begin by showing that whenever I and J are cofibrations, the pushout-product $I\mathbin{\square} J$ is also a cofibration; it is enough to consider the case when I and J are in the set of generating cofibrations $\mathcal{I}_w = \{I_1, I_2, I_3, I_4, I_5\}$ of Notation 3.8. Moreover, since the Gray tensor product is symmetric, if we show the result for $I \mathbin{\square} J$, then it also holds for $J \mathbin{\square} I$. Note that $I_1 \mathbin{\square} J \cong J$, which proves the cases involving I_1.

To show the cases involving I_4 or I_5, we observe the following three facts: the functors $U\boldsymbol{H}, U\boldsymbol{V} : \mathrm{DblCat} \to \mathrm{Cat}$ preserve pushouts since they are left adjoints (see Remark 4.5 of [16]); $U\boldsymbol{H}(I_4)$, $U\boldsymbol{H}(I_5)$, $U\boldsymbol{V}(I_4)$ and $U\boldsymbol{V}(I_5)$ are identities; and $U\boldsymbol{H}(\mathbb{A} \otimes_{\mathrm{Gr}} \mathbb{B})$ (resp. $U\boldsymbol{V}(\mathbb{A} \otimes_{\mathrm{Gr}} \mathbb{B})$) is completely determined by $U\boldsymbol{H}(\mathbb{A})$ and $U\boldsymbol{H}(\mathbb{B})$ (resp. $U\boldsymbol{V}(\mathbb{A})$ and $U\boldsymbol{V}(\mathbb{B})$). It then follows that $U\boldsymbol{H}(I \mathbin{\square} J)$ and $U\boldsymbol{V}(I \mathbin{\square} J)$ are isomorphisms, and thus $I \mathbin{\square} J$ is a cofibration by Theorem 3.11, if either I or J is in $\{I_4, I_5\}$.

For the remaining cases, one can check that $I_2 \mathbin{\square} I_2$ is given by the boundary inclusion $\delta(\mathbb{H}2 \otimes_{\mathrm{Gr}} \mathbb{H}2) \to \mathbb{H}2 \otimes_{\mathrm{Gr}} \mathbb{H}2$, where $\delta(\mathbb{H}2 \otimes_{\mathrm{Gr}} \mathbb{H}2)$ is obtained by removing the nonidentity squares in $\mathbb{H}2 \otimes_{\mathrm{Gr}} \mathbb{H}2$, generated by the data depicted below. Then this boundary inclusion is a cofibration by Theorem 3.11, since it is the identity on underlying horizontal and vertical categories.

Similarly, one can show that the pushout-products $I_3 \mathbin{\square} I_3$ and $I_2 \mathbin{\square} I_3$ are cofibrations, as they are given by analogously defined boundary inclusions

$$\delta(\mathbb{V}2 \otimes_{\mathrm{Gr}} \mathbb{V}2) \to \mathbb{V}2 \otimes_{\mathrm{Gr}} \mathbb{V}2 \quad \text{and} \quad \delta(\mathbb{H}2 \otimes_{\mathrm{Gr}} \mathbb{V}2) \to \mathbb{H}2 \otimes_{\mathrm{Gr}} \mathbb{V}2,$$

respectively, where the double categories $\mathbb{H}2 \otimes_{\mathrm{Gr}} \mathbb{V}2$ and $\mathbb{V}2 \otimes_{\mathrm{Gr}} \mathbb{V}2$ are generated by the data

$$\begin{tikzcd} 00 \arrow[r] \arrow[d, equal] & 10 \arrow[r] \arrow[d, phantom, "\wr\!\cong"] & 11 \arrow[d, equal] \\ 00 \arrow[r] & 01 \arrow[r] & 11 \end{tikzcd} \qquad \begin{tikzcd} 00 \arrow[r, equal] \arrow[d, "\bullet"] & 00 \arrow[d, "\bullet"] \\ 01 \arrow[r, phantom, "\cong"] \arrow[d, "\bullet"] & 10 \arrow[d, "\bullet"] \\ 11 \arrow[r, equal] & 11 \end{tikzcd} \qquad \begin{tikzcd} 00 \arrow[r] \arrow[d, "\bullet"] \arrow[dr, phantom, "\alpha"] & 10 \arrow[d, "\bullet"] \\ 01 \arrow[r] & 11 \end{tikzcd}$$

$$\mathbb{H}2 \otimes_{\mathrm{Gr}} \mathbb{H}2 \qquad\qquad \mathbb{V}2 \otimes_{\mathrm{Gr}} \mathbb{V}2 \qquad\qquad \mathbb{H}2 \otimes_{\mathrm{Gr}} \mathbb{V}2$$

It remains to show that if $I \in \mathcal{I}_w$ and $J : \mathbb{A} \to \mathbb{B}$ is a trivial cofibration, then $I \mathbin{\square} J$ is a weak equivalence. Note that I is of the form $I : \mathbb{C} \to \mathbb{D}$ with $\mathbb{C}$ cofibrant, and consider the pushout diagram

$$\begin{tikzcd} \mathbb{C} \otimes_{\mathrm{Gr}} \mathbb{A} \arrow[r, hook, "\mathrm{id}_{\mathbb{C}} \otimes_{\mathrm{Gr}} J", "\sim"'] \arrow[d, "I \otimes_{\mathrm{Gr}} \mathrm{id}_{\mathbb{A}}"'] & \mathbb{C} \otimes_{\mathrm{Gr}} \mathbb{B} \arrow[d] \arrow[ddr, bend left, "I \otimes_{\mathrm{Gr}} \mathrm{id}_{\mathbb{B}}"] & \\ \mathbb{D} \otimes_{\mathrm{Gr}} \mathbb{A} \arrow[r, hook, "K", "\sim"'] \arrow[drr, bend right, "\sim", "\mathrm{id}_{\mathbb{D}} \otimes_{\mathrm{Gr}} J"'] & \mathbb{P} \arrow[ul, phantom, "\ulcorner", very near start] \arrow[dr, dashed, "I \mathbin{\square} J"] & \\ & & \mathbb{D} \otimes_{\mathrm{Gr}} \mathbb{B} \end{tikzcd}$$

Since $\mathbb{C}$ is cofibrant and J is a cofibration, we know that $(\varnothing \to \mathbb{C}) \,\square\, J = \mathrm{id}_{\mathbb{C}} \otimes_{\mathrm{Gr}} J$ is also a cofibration by the above. Since J is a trivial cofibration by assumption, the double functor $\mathrm{id}_{\mathbb{C}} \otimes_{\mathrm{Gr}} J$ is a weak equivalence by Proposition 7.6. Then $\mathrm{id}_{\mathbb{C}} \otimes_{\mathrm{Gr}} J$ is a trivial cofibration, and therefore so is K since these are stable under pushouts. Proposition 7.6 also guarantees that $\mathrm{id}_{\mathbb{D}} \otimes_{\mathrm{Gr}} J$ is a weak equivalence, and then so is $I \,\square\, J$ by 2-out-of-3. $\square$

Remark 7.9 Recall that, by restricting the Gray tensor product $\otimes_{\mathrm{Gr}}$ in one variable along $\mathbb{H}\colon 2\mathrm{Cat} \to \mathrm{DblCat}$, we get the tensoring functor $\otimes\colon \mathrm{DblCat} \times 2\mathrm{Cat} \to \mathrm{DblCat}$ which gives an enrichment $\boldsymbol{H}[-,-]_{\mathrm{ps}}$ of DblCat over 2Cat as in [16, Proposition 7.5]. Since the functor $\mathbb{H}$ is left Quillen by Theorem 6.3, as a corollary of Theorem 7.8 we get that the model structure on DblCat of Theorem 3.26 is also 2Cat–enriched.

8 Whitehead theorem

In this section we show a Whitehead theorem for double categories, that characterizes the weak equivalences between fibrant objects (which, by Proposition 5.5, are double biequivalences) as the double functors that admit a pseudoinverse up to horizontal pseudonatural equivalence. Such a statement is reminiscent of the Whitehead theorem for 2–categories: a 2–functor $F\colon \mathcal{A} \to \mathcal{B}$ is a biequivalence if and only if there is a pseudofunctor $G\colon \mathcal{B} \to \mathcal{A}$ together with two pseudonatural equivalences $\mathrm{id}_{\mathcal{A}} \simeq GF$ and $FG \simeq \mathrm{id}_{\mathcal{B}}$.

Under the hypothesis that the double categories involved are *horizontally invariant* — defined analogously to the weakly horizontally invariant double categories with horizontal equivalences replaced by the stronger notion of horizontal isomorphisms; see [6, Theorem and Definition 4.1.7] — Grandis characterizes in [6, Theorem 4.4.5] the double functors F such that $U\boldsymbol{H}F$ and $U\boldsymbol{H}[\mathbb{V}2, F]$ are both equivalences of categories as the ones which admit a pseudoinverse up to *horizontal natural isomorphism*. In analogy, double biequivalences can be defined as the double functors such that $\boldsymbol{H}F$ and $\boldsymbol{H}[\mathbb{V}2, F]$ are biequivalences of 2–categories; see [16, Proposition 3.11]. Altogether our Whitehead theorem can be seen as a 2–categorical version of Grandis's result.

In the theorem below, whose proof is the content of this section, it is actually enough to require that the *source* be weakly horizontally invariant.

Theorem 8.1 (Whitehead theorem) *Let $\mathbb{A}$ and $\mathbb{B}$ be double categories such that $\mathbb{A}$ is weakly horizontally invariant. Then a double functor $F\colon \mathbb{A} \to \mathbb{B}$ is a weak equiva-*

lence (or equivalently, a double biequivalence) if and only if there is a pseudodouble functor $G\colon \mathbb{B} \to \mathbb{A}$ *together with horizontal pseudonatural equivalences* $\mathrm{id}_{\mathbb{A}} \simeq GF$ *and* $FG \simeq \mathrm{id}_{\mathbb{B}}$.

Remark 8.2 If $\mathbb{A}$ and $\mathbb{B}$ are cofibrant–fibrant double categories, a double functor $F\colon \mathbb{A} \to \mathbb{B}$ is a weak equivalence if and only if there is a (strict) double functor $G\colon \mathbb{B} \to \mathbb{A}$ together with horizontal pseudonatural equivalences $\mathrm{id}_{\mathbb{A}} \simeq GF$ and $FG \simeq \mathrm{id}_{\mathbb{B}}$. Indeed, the cofibrancy condition implies that the underlying horizontal and vertical categories of $\mathbb{A}$ and $\mathbb{B}$ are free, and therefore the weak inverse G can be chosen to be strict.

This retrieves a formulation of the usual Whitehead theorem for model categories — see [3, Lemma 4.24] — in our setting; such a result characterizes the weak equivalences between cofibrant–fibrant objects in a model structure as the homotopy equivalences. Indeed, the homotopies in our model structure are the horizontal pseudonatural equivalences, as we now show.

Given a weakly horizontally invariant double category $\mathbb{A}$, a path object for $\mathbb{A}$ is given by the double category $[\mathbb{H}E_{\mathrm{adj}}, \mathbb{A}]_{\mathrm{ps}}$ together with the double functors

$$\mathbb{A} \xrightarrow{W} [\mathbb{H}E_{\mathrm{adj}}, \mathbb{A}]_{\mathrm{ps}} \xrightarrow{P} \mathbb{A} \times \mathbb{A}$$

obtained by applying the functor $[-, \mathbb{A}]_{\mathrm{ps}}$ to the composite $\mathbb{1} \sqcup \mathbb{1} \to \mathbb{H}E_{\mathrm{adj}} \to \mathbb{1}$.

Since $\mathbb{1} \sqcup \mathbb{1} \to \mathbb{H}E_{\mathrm{adj}}$ is a cofibration and the model structure on DblCat is monoidal, it follows that P is a fibration in DblCat. Similarly, since $\mathbb{1} \to \mathbb{H}E_{\mathrm{adj}}$ is a trivial cofibration in DblCat, by monoidality, the induced double functor $[\mathbb{H}E_{\mathrm{adj}}, \mathbb{A}]_{\mathrm{ps}} \to \mathbb{A}$ is a trivial cofibration in DblCat. Hence, by 2-out-of-3, we get that W is a weak equivalence in DblCat, and thus $[\mathbb{H}E_{\mathrm{adj}}, \mathbb{A}]_{\mathrm{ps}}$ is a path object for $\mathbb{A}$.

Then, by definition, a homotopy in DblCat between two double functors $F, G\colon \mathbb{A} \to \mathbb{B}$ with $\mathbb{A}$ and $\mathbb{B}$ cofibrant–fibrant is a double functor $\mathbb{A} \to [\mathbb{H}E_{\mathrm{adj}}, \mathbb{B}]_{\mathrm{ps}}$, or equivalently, a double functor $\mathbb{H}E_{\mathrm{adj}} \to [\mathbb{A}, \mathbb{B}]_{\mathrm{ps}}$, whose values on the two objects of $\mathbb{H}E_{\mathrm{adj}}$ are given by F and G. By [15, Lemma A.3.3], this corresponds to a horizontal pseudonatural equivalence from F to G.

Let us now introduce what we mean by a pseudodouble functor.

Definition 8.3 A *pseudodouble functor* $G\colon \mathbb{B} \to \mathbb{A}$ consists of maps on objects, horizontal morphisms, vertical morphisms, and squares, compatible with sources and targets, which preserve

(i) horizontal compositions and identities up to coherent vertically invertible squares

$$\begin{tikzcd} GB \arrow[r, "Gb"] \arrow[d, equal] & GC \arrow[r, "Gd"] \arrow[d, phantom, "\Phi_{b,d} \wr\Vert"] & GD \arrow[d, equal] \\ GB \arrow[rr, "G(db)"'] & & GD \end{tikzcd} \qquad \begin{tikzcd} GB \arrow[r, equal] \arrow[d, equal] \arrow[dr, phantom, "\Phi_B \wr\Vert"] & GB \arrow[d, equal] \\ GB \arrow[r, "G\mathrm{id}_B"'] & GB \end{tikzcd}$$

for every object $B \in \mathbb{B}$, and every pair of composable horizontal morphisms $b\colon B \to C$ and $d\colon C \to D$ in $\mathbb{B}$;

(ii) vertical compositions and identities up to coherent horizontally invertible squares $\Psi_{v,v'}$ and Ψ_B — the transposed versions of those in (i) — for every object $B \in \mathbb{B}$, and every pair of composable vertical morphisms v and v' in $\mathbb{B}$;

(iii) horizontal and vertical compositions of squares accordingly.

For a detailed description of the coherences, the reader can see [6, Definition 3.5.1].

The pseudodouble functor G is said to be *normal* if the squares Φ_B and Ψ_B are identities for every object $B \in \mathbb{B}$.

Definition 8.4 A *horizontal pseudonatural transformation* $h\colon F \Rightarrow G$ between pseudodouble functors $F, G\colon \mathbb{A} \to \mathbb{B}$ is a pseudodouble functor $h\colon \mathbb{A} \otimes_{\mathrm{Gr}} \mathbb{H}2 \to \mathbb{B}$ which restricts to F and G under the two inclusions $\mathbb{1} \to \mathbb{H}2$. More explicitly, this consists of

(i) a horizontal morphism $h_A\colon FA \to GA$ in $\mathbb{B}$, for each object $A \in \mathbb{A}$;

(ii) a square h_u in $\mathbb{B}$

$$\begin{tikzcd} FA \arrow[r, "h_A"] \arrow[d, "Fu"'] \arrow[dr, phantom, "h_u"] & GA \arrow[d, "Gu"] \\ FA' \arrow[r, "h_{A'}"'] & GA' \end{tikzcd}$$

for each vertical morphism $u\colon A \to A'$ in $\mathbb{A}$;

(iii) a vertically invertible square h_a in $\mathbb{B}$

$$\begin{tikzcd} FA \arrow[r, "h_A"] \arrow[d, equal] & GA \arrow[r, "Ga"] \arrow[d, phantom, "h_a \wr\Vert"] & GC \arrow[d, equal] \\ FA \arrow[r, "Fa"'] & FC \arrow[r, "h_C"'] & GC \end{tikzcd}$$

for each horizontal morphism $a\colon A \to C$ in $\mathbb{A}$.

The squares in (ii) are compatible with the coherence squares of F and G for vertical compositions and identities, and the squares in (iii) are compatible with the coherence squares of F and G for horizontal compositions and identities. Together, they satisfy a pseudonaturality condition with respect to squares in $\mathbb{A}$.

A *modification* $\mu\colon (e_F \,{}^{h}_{k}\, e_G)$ between two horizontal pseudonatural transformations $h, k\colon F \Rightarrow G$ is a pseudodouble functor $\mu\colon \mathbb{A} \otimes_{\mathrm{Gr}} \mathbb{H}\Sigma\mathbb{2} \to \mathbb{B}$ which restricts to h and k under the two canonical inclusions $\mathbb{H}\mathbb{2} \to \mathbb{H}\Sigma\mathbb{2}$, where $\Sigma\mathbb{2}$ is the free 2–category on a 2–morphism. More explicitly, this consists of a square $\mu_A\colon (e_{FA} \,{}^{h_A}_{k_A}\, e_{GA})$ in $\mathbb{B}$, for each object $A \in \mathbb{A}$, satisfying horizontal and vertical coherence conditions with respect to the square components of the pseudonatural transformations h and k.

For more details about the coherence conditions, see [6, Section 3.8].

Remark 8.5 The pseudodouble functors from $\mathbb{A}$ to $\mathbb{B}$ together with the horizontal pseudonatural transformations and modifications between them form a 2–category. It can be seen as the sub–2–category of the 2–category of lax (double) functors of [6, Theorem 3.8.4] given by restriction to the pseudodouble functors.

The notion that we now introduce has also been independently considered by Grandis and Paré in [8, Section 3] under the name of *pointwise equivalences*.

Definition 8.6 Let $F, G\colon \mathbb{A} \to \mathbb{B}$ be pseudodouble functors. A *horizontal pseudonatural equivalence* $\varphi\colon F \Rightarrow G$ is an equivalence in the 2–category of pseudodouble functors $\mathbb{A} \to \mathbb{B}$, horizontal pseudonatural transformations, and modifications.

Equivalently, the horizontal pseudonatural equivalences can be described as follows; see [8, Theorem 4.4] for a proof.

Lemma 8.7 *Let $F, G\colon \mathbb{A} \to \mathbb{B}$ be pseudodouble functors. A horizontal pseudonatural transformation $\varphi\colon F \Rightarrow G$ is a horizontal pseudonatural equivalence if and only if*

(i) *the horizontal morphism $\varphi_A\colon FA \xrightarrow{\simeq} GA$ is a horizontal equivalence, for every object $A \in \mathbb{A}$, and*

(ii) *the square $\varphi_u\colon (Fu \,{}^{\varphi_A}_{\varphi_{A'}}\, Gu)$ is weakly horizontally invertible, for every vertical morphism $u\colon A \to A'$ in $\mathbb{A}$.*

We will use the term *horizontal biequivalence* to refer to the double functors which admit a pseudoinverse up to horizontal pseudonatural equivalence.

Definition 8.8 A double functor $F\colon \mathbb{A} \to \mathbb{B}$ is a *horizontal biequivalence* if there is a pseudodouble functor $G\colon \mathbb{B} \to \mathbb{A}$ together with horizontal pseudonatural equivalences $\eta\colon \mathrm{id}_{\mathbb{A}} \Rightarrow GF$ and $\epsilon\colon FG \Rightarrow \mathrm{id}_{\mathbb{B}}$.

Remark 8.9 If $F\colon \mathbb{A} \to \mathbb{B}$ is a horizontal biequivalence, there is a tuple $(G, \eta, \epsilon, \Theta, \Sigma)$ consisting of the following data:

(i) a *normal* pseudodouble functor $G\colon \mathbb{B} \to \mathbb{A}$;

(ii) a horizontal pseudonatural *adjoint* equivalence
$$(\eta\colon \mathrm{id}_{\mathbb{A}} \Rightarrow GF,\ \eta'\colon GF \Rightarrow \mathrm{id}_{\mathbb{A}},\ \lambda\colon \mathrm{id} \cong \eta'\eta,\ \kappa\colon \eta\eta' \cong \mathrm{id});$$

(iii) a horizontal pseudonatural *adjoint* equivalence
$$(\epsilon\colon FG \Rightarrow \mathrm{id}_{\mathbb{B}},\ \epsilon'\colon \mathrm{id}_{\mathbb{B}} \Rightarrow FG,\ \mu\colon \mathrm{id} \cong \epsilon'\epsilon,\ \nu\colon \epsilon\epsilon' \cong \mathrm{id});$$

(iv) two invertible modifications $\Theta\colon \mathrm{id}_F \cong \epsilon_F \circ F\eta$ and $\Sigma\colon \mathrm{id}_G \cong G\epsilon \circ \eta_G$, expressing the triangle (pseudo)identities for η and ϵ.

This follows from the fact that a pseudodouble functor is always pseudonaturally isomorphic to a normal one, and from a result by Gurski [9, Theorem 3.2] saying that a biequivalence can always be promoted to a biadjoint biequivalence, applied here to the tricategory of double categories, pseudodouble functors, horizontal pseudonatural transformations, and modifications.

Theorem 8.1 now amounts to showing that a double functor whose source is weakly horizontally invariant is a double biequivalence if and only if it is a horizontal biequivalence. However, it is always true that a horizontal biequivalence is a double biequivalence; no additional hypothesis is needed here. In order to prove this first result, we need the following lemma.

Lemma 8.10 *The data of Remark 8.9 induces an invertible modification* $\theta\colon F\eta' \cong \epsilon_F$*.*

Proof Given an object $A \in \mathbb{A}$, we define the component of θ at A to be the vertically invertible square

$$\begin{tikzcd}
FGFA \arrow[r, "F\eta'_A"] \arrow[d, equal] \arrow[dr, phantom, "\theta_A\ \cong"] & FA \arrow[d, equal] \\
FGFA \arrow[r, "\epsilon_{FA}"'] & FA
\end{tikzcd}
=
\begin{tikzcd}
FGFA \arrow[r, "F\eta'_A"] \arrow[d, equal] \arrow[dr, phantom, "e_{F\eta'_A}"] & FA \arrow[rr, equal] \arrow[d, equal] \arrow[drr, phantom, "\Theta_A\ \cong"] & & FA \arrow[d, equal] \\
FGFA \arrow[r, "F\eta'_A"'] \arrow[d, equal] \arrow[drr, phantom, "F\kappa_A\ \cong"] & FA \arrow[r, "F\eta_A"] & FGFA \arrow[r, "\epsilon_{FA}"] \arrow[d, equal] \arrow[dr, phantom, "e_{\epsilon_{FA}}"] & FA \arrow[d, equal] \\
FGFA \arrow[rr, equal] & & FGFA \arrow[r, "\epsilon_{FA}"'] & FA
\end{tikzcd}$$

The proof of horizontal and vertical coherences for θ is a standard check that stems from the constructions of the squares θ_A and from the horizontal and vertical coherences of the modifications $F\kappa\colon (F\eta)(F\eta') \cong \mathrm{id}$ and $\Theta\colon \mathrm{id} \cong \epsilon_F \circ F\eta$. □

Proposition 8.11 *If $F\colon \mathbb{A} \to \mathbb{B}$ is a horizontal biequivalence, then F is a double biequivalence.*

Proof We check that F satisfies (db1)–(db4) of Definition 2.9. Let $(F, G, \eta, \epsilon, \Theta, \Sigma)$ be the data of a horizontal adjoint biequivalence as in Remark 8.9. We first show (db1). For every object $B \in \mathbb{B}$, we want to find an object $A \in \mathbb{A}$ and a horizontal equivalence $B \xrightarrow{\simeq} FA$ in $\mathbb{B}$. Setting $A = GB$, we have that $\epsilon'_B\colon B \xrightarrow{\simeq} FGB = FA$ gives such a horizontal equivalence.

We now show (db2). Let A and C be objects in $\mathbb{A}$, and $b\colon FA \to FC$ be a horizontal morphism in $\mathbb{B}$. We want to find a horizontal morphism $a\colon A \to C$ in $\mathbb{A}$ and a vertically invertible square $(e_{FA}\ {}^{b}_{Fa}\ e_{FC})$ in $\mathbb{B}$. Let $a\colon A \to C$ be the composite

$$A \xrightarrow{\eta_A} GFA \xrightarrow{Gb} GFC \xrightarrow{\eta'_C} C.$$

We then have a vertically invertible square as desired,

$$\begin{tikzcd}
FA \arrow[rr, equal] \arrow[d, "\bullet" marking, equal] & \arrow[d, phantom, "\Theta_A\ \cong"] & FA \arrow[r, "b"] \arrow[d, "\bullet" marking, equal] \arrow[dr, phantom, "e_b"] & FC \arrow[d, "\bullet" marking, equal] \\
FA \arrow[r, "F\eta_A"] \arrow[d, "\bullet" marking, equal] \arrow[dr, phantom, "e_{F\eta_A}"] & FGFA \arrow[r, "\epsilon_{FA}"] \arrow[d, "\bullet" marking, equal] \arrow[drr, phantom, "\epsilon_b\ \cong"] & FA \arrow[r, "b"] & FC \arrow[d, "\bullet" marking, equal] \\
FA \arrow[r, "F\eta_A"] \arrow[d, "\bullet" marking, equal] \arrow[drr, phantom, "e_{(FGb)(F\eta_A)}"] & FGFA \arrow[r, "FGb"] & FGFC \arrow[r, "\epsilon_{FC}"] \arrow[d, "\bullet" marking, equal] \arrow[dr, phantom, "\theta_C^{-1}\ \cong"] & FC \arrow[d, "\bullet" marking, equal] \\
FA \arrow[r, "F\eta_A"'] & FGFA \arrow[r, "FGb"'] & FGFC \arrow[r, "F\eta'_C"'] & FC
\end{tikzcd}$$

where θ_C is the component at C of the invertible modification θ of Lemma 8.10.

We now show (db3). Let $v\colon B \twoheadrightarrow B'$ be a vertical morphism in $\mathbb{B}$. We want to find a vertical morphism $u\colon A \twoheadrightarrow A'$ in $\mathbb{A}$ and a weakly horizontally invertible square $(v \overset{\simeq}{\sim} Fu)$ in $\mathbb{B}$. Let $u\colon A \twoheadrightarrow A'$ be the vertical morphism $Gv\colon GB \twoheadrightarrow GB'$. Then ϵ'_v gives the desired weakly horizontally invertible square

$$\begin{tikzcd}
B \arrow[r, "\epsilon'_B", "\simeq"'] \arrow[d, "v"', "\bullet" marking] \arrow[dr, phantom, "\epsilon'_v\ \simeq"] & FGB \arrow[d, "FGv", "\bullet" marking] \\
B' \arrow[r, "\simeq", "\epsilon'_{B'}"'] & FGB'
\end{tikzcd}$$

Finally, we show (db4). For this, let β be a square in $\mathbb{B}$ of the form

$$\begin{tikzcd}
FA \arrow[r, "Fa"] \arrow[d, "Fu"', "\bullet" marking] \arrow[dr, phantom, "\beta"] & FC \arrow[d, "Fw", "\bullet" marking] \\
FA' \arrow[r, "Fa'"'] & FC'
\end{tikzcd}$$

We want to show that there is a unique square $\alpha\colon (u \,{}^{a}_{c}\, u')$ in $\mathbb{A}$ such that $F\alpha = \beta$. Define α to be the square given by the pasting

$$\begin{tikzcd}
A \arrow[r, "a"] \arrow[d, "u"', "\bullet" marking] \arrow[dr, phantom, "\alpha"] & C \arrow[d, "u'", "\bullet" marking] \\
A' \arrow[r, "c"'] & C'
\end{tikzcd}
\quad = \quad
\begin{tikzcd}
A \arrow[rr, equal] \arrow[d, equal, "\bullet" marking] \arrow[drr, phantom, "\lambda_A \;\cong"] & & A \arrow[r, "a"] \arrow[d, equal, "\bullet" marking] \arrow[dr, phantom, "e_a"] & C \arrow[d, equal, "\bullet" marking] \\
A \arrow[r, "\eta_A"] \arrow[d, equal, "\bullet" marking] \arrow[dr, phantom, "e_{\eta_A}"] & GFA \arrow[r, "\eta'_A"] \arrow[d, equal, "\bullet" marking] \arrow[drr, phantom, "\eta'_a \;\cong"] & A \arrow[r, "a"] & C \arrow[d, equal, "\bullet" marking] \\
A \arrow[r, "\eta_A"] \arrow[d, "u"', "\bullet" marking] \arrow[dr, phantom, "\eta_u"] & GFA \arrow[r, "GFa"] \arrow[d, "GFu", "\bullet" marking] \arrow[dr, phantom, "G\beta"] & GFC \arrow[r, "\eta'_C"] \arrow[d, "GFu'", "\bullet" marking] \arrow[dr, phantom, "\eta'_{u'}"] & C \arrow[d, "u'", "\bullet" marking] \\
A' \arrow[r, "\eta_{A'}"'] \arrow[d, equal, "\bullet" marking] \arrow[dr, phantom, "e_{\eta_{A'}}"] & GFA' \arrow[r, "GFc"'] \arrow[d, equal, "\bullet" marking] \arrow[drr, phantom, "{\eta'_c}^{-1} \;\cong"] & GFC' \arrow[r, "\eta'_{C'}"'] & C' \arrow[d, equal, "\bullet" marking] \\
A' \arrow[r, "\eta_{A'}"'] \arrow[d, equal, "\bullet" marking] \arrow[drr, phantom, "\lambda^{-1}_{A'} \;\cong"] & GFA' \arrow[r, "\eta'_{A'}"'] & A' \arrow[r, "c"'] \arrow[d, equal, "\bullet" marking] \arrow[dr, phantom, "e_c"] & C' \arrow[d, equal, "\bullet" marking] \\
A' \arrow[rr, equal] & & A' \arrow[r, "c"'] & C'
\end{tikzcd}$$

The thorough reader might check that $F\alpha = \beta$ by completing the following steps. First, transform $F\eta'_{u'}$ by using the invertible modification $\theta\colon F\eta' \cong \epsilon_F$ of Lemma 8.10. Then apply, in order: the horizontal coherence of the modification $F\nu\colon (F\eta')(F\eta) \cong \mathrm{id}$, the horizontal coherence of the modification $\Theta\colon \mathrm{id} \cong \epsilon_F \circ F\eta$, the triangle identity for (μ, ν), the compatibility of $\epsilon_F\colon FGF \Rightarrow F$ with $FG\beta$ and β, and finally the horizontal coherence of the modification $\Theta\colon \mathrm{id} \cong \epsilon_F \circ F\eta$.

Suppose now that $\alpha'\colon (u \,{}^{a}_{c}\, u')$ is another square in $\mathbb{A}$ such that $F\alpha' = \beta$. If we replace $G\beta$ with $GF\alpha'$ in the pasting diagram above, it follows from the compatibility of $\eta'\colon GF \Rightarrow \mathrm{id}_{\mathbb{A}}$ with $GF\alpha'$ and α', and the vertical coherence of the modification $\mu\colon \mathrm{id} \cong \eta'\eta$, that this pasting is also equal to α'. Therefore, we must have $\alpha = \alpha'$. This completes the proof of (db4). $\square$

It is not true in general that a double biequivalence is a horizontal biequivalence, unless we impose an additional condition on the source or on the target. In [16, Theorem 5.13] we provide a Whitehead theorem, where the target satisfies a condition related to cofibrancy in the model structure of [16]. Here, we prove that such a result holds when the source of the double biequivalence is fibrant, which completes the proof of our Whitehead theorem, Theorem 8.1.

Proposition 8.12 *Let $F\colon \mathbb{A} \to \mathbb{B}$ be a double biequivalence such that $\mathbb{A}$ is weakly horizontally invariant. Then F is a horizontal biequivalence.*

Proof We simultaneously define the pseudodouble functor $G\colon \mathbb{B} \to \mathbb{A}$ and the horizontal pseudonatural transformation $\epsilon\colon FG \Rightarrow \mathrm{id}_{\mathbb{B}}$.

G and ϵ on objects Let $B \in \mathbb{B}$ be an object. By (db1) of Definition 2.9, there is an object $A \in \mathbb{A}$ and a horizontal equivalence $b\colon FA \xrightarrow{\simeq} B$ in $\mathbb{B}$. We set $GB := A$ and $\epsilon_B := b\colon FGB \xrightarrow{\simeq} B$, and also fix horizontal equivalence data $(\epsilon_B, \epsilon'_B, \mu_B, \nu_B)$.

G and ϵ on horizontal morphisms Now let $b\colon B \to C$ be a horizontal morphism in $\mathbb{B}$. By (db2) applied to the horizontal morphism $\epsilon'_C b \epsilon_B\colon FGB \to FGC$, there is a horizontal morphism $a\colon GB \to GC$ in $\mathbb{A}$ and a vertically invertible square $\bar{\epsilon}_b$ as depicted inside the right-hand side of the pasting below. We set $Gb := a\colon GB \to GC$ and ϵ_b to be the square given by the pasting

$$
\begin{tikzcd}
FGB \arrow[r, "\epsilon_B"] \arrow[d, equal, "\bullet"{marking}] & B \arrow[r, "b"] \arrow[d, phantom, "\epsilon_b \,\cong"] & C \arrow[d, equal, "\bullet"{marking}] \\
FGB \arrow[r, "FGb"'] & FGC \arrow[r, "\epsilon_C"'] & C
\end{tikzcd}
=
\begin{tikzcd}
FGB \arrow[r, "\epsilon_B"] \arrow[d, equal, "\bullet"{marking}] & B \arrow[r, "b"] \arrow[d, phantom, "e_{b\epsilon_B}"] & C \arrow[rr, equal] \arrow[d, equal, "\bullet"{marking}] & & C \arrow[d, equal, "\bullet"{marking}] \arrow[dll, phantom, "\nu_C^{-1} \cong"] \\
FGB \arrow[r, "\epsilon_B"] \arrow[d, equal, "\bullet"{marking}] & B \arrow[r, "b"] \arrow[d, phantom, "\bar{\epsilon}_b \,\cong"] & C \arrow[r, "\epsilon'_C"] & FGC \arrow[r, "\epsilon_C"] \arrow[d, equal, "\bullet"{marking}] \arrow[dr, phantom, "e_{\epsilon_C}"] & C \arrow[d, equal, "\bullet"{marking}] \\
FGB \arrow[rrr, "Fa = FGb"'] & & & FGC \arrow[r, "\epsilon_C"'] & C
\end{tikzcd}
$$

If $b = \mathrm{id}_B$, we can choose $G\mathrm{id}_B := \mathrm{id}_{GB}$ and $\bar{\epsilon}_{\mathrm{id}_B} := \mu_B^{-1}$. Then $\epsilon_{\mathrm{id}_B} = e_{\epsilon_B}$ by the triangle identities for (μ_B, ν_B).

Horizontal coherence Given horizontal morphisms $b\colon B \to C$ and $d\colon C \to D$ in $\mathbb{B}$, we define the vertically invertible comparison square between $Gd \circ Gb$ and $G(db)$ as follows. Let us denote by $\Theta_{b,d}$ the pasting

$$
\begin{tikzcd}
FGB \arrow[rrr, "FGb"] \arrow[d, equal, "\bullet"{marking}] & & & FGC \arrow[rrr, "FGd"] \arrow[d, equal, "\bullet"{marking}] & & & FGD \arrow[d, equal, "\bullet"{marking}] \\
FGB \arrow[r, "\epsilon_B"] \arrow[d, equal, "\bullet"{marking}] \arrow[urrr, phantom, "\bar{\epsilon}_b^{-1} \cong"] & B \arrow[r, "b"] & C \arrow[r, "\epsilon'_C"] \arrow[d, equal, "\bullet"{marking}] & FGC \arrow[r, "\epsilon_C"] \arrow[urrr, phantom, "\bar{\epsilon}_d^{-1} \cong"] & C \arrow[r, "d"] \arrow[d, equal, "\bullet"{marking}] & D \arrow[r, "\epsilon'_D"] & FGD \arrow[d, equal, "\bullet"{marking}] \\
FGB \arrow[r, "\epsilon_B"'] \arrow[d, equal, "\bullet"{marking}] \arrow[ur, phantom, "e_{b\epsilon_B}"] & B \arrow[r, "b"'] & C \arrow[rr, equal] \arrow[ur, phantom, "\nu_C \cong"] & & C \arrow[r, "d"'] \arrow[ur, phantom, "e_{\epsilon'_D d}"] & D \arrow[r, "\epsilon'_D"'] & FGD \arrow[d, equal, "\bullet"{marking}] \\
FGB \arrow[rrrrrr, "FG(db)"'] \arrow[urrr, phantom, "\bar{\epsilon}_{db} \cong"] & & & & & & FGD
\end{tikzcd}
$$

Then, by (db4), there is a unique vertically invertible square $\Phi_{b,d}$ as in Definition 8.3(i) such that $F\Phi_{b,d} = \Theta_{b,d}$. In particular, one can check that with this definition of $\Phi_{b,d}$, the squares ϵ_b, ϵ_d, and ϵ_{db} satisfy the required pasting equality

$$
\begin{tikzcd}
FGB \arrow[r, "\epsilon_B"] \arrow[d, equal] \arrow[drr, phantom, "\epsilon_{db}\ \wr\!\shortparallel"] & B \arrow[r, "b"] & C \arrow[r, "d"] & D \arrow[d, equal] \\
FGB \arrow[rr, "FG(db)"'] & & FGD \arrow[r, "\epsilon_D"'] & D
\end{tikzcd}
\quad = \quad
\begin{tikzcd}
FGB \arrow[r, "\epsilon_B"] \arrow[d, equal] \arrow[dr, phantom, "\epsilon_b\ \wr\!\shortparallel"] & B \arrow[r, "b"] & C \arrow[r, "d"] \arrow[d, equal] \arrow[dr, phantom, "e_d"] & D \arrow[d, equal] \\
FGB \arrow[r, "FGb"] \arrow[d, equal] \arrow[r, phantom, shift right=4ex, ""] \arrow[dr, phantom, "e_{FGb}"] & FGC \arrow[r, "\epsilon_C"'] \arrow[d, equal] \arrow[dr, phantom, "\epsilon_d\ \wr\!\shortparallel"] & C \arrow[r, "d"'] & D \arrow[d, equal] \\
FGB \arrow[r, "FGb"'] \arrow[d, equal] \arrow[drr, phantom, "F\Phi_{b,d}\ \wr\!\shortparallel"] & FGC \arrow[r, "FGd"'] & FGD \arrow[r, "\epsilon_D"'] \arrow[d, equal] \arrow[dr, phantom, "e_{\epsilon_D}"] & D \arrow[d, equal] \\
FGB \arrow[rr, "FG(db)"'] & & FGD \arrow[r, "\epsilon_D"'] & D
\end{tikzcd}
$$

G and ϵ on vertical morphisms Now let $v\colon B \to B'$ be a vertical morphism in $\mathbb{B}$. By (db3), there is a vertical morphism $u'\colon A \to A'$ in $\mathbb{A}$ and a weakly horizontally invertible square γ_v in $\mathbb{B}$,

$$
\begin{tikzcd}
B \arrow[r, "b", "\simeq"'] \arrow[d, "v"'] \arrow[dr, phantom, "\gamma_v\ \simeq"] & FA \arrow[d, "Fu'"] \\
B' \arrow[r, "\simeq", "d"'] & FA'
\end{tikzcd}
$$

where $b\colon B \xrightarrow{\simeq} FA$ and $d\colon B' \xrightarrow{\simeq} FA'$ are horizontal equivalences. If we consider the composites of horizontal equivalences $b\epsilon_B\colon FGB \xrightarrow{\simeq} FA$ and $d\epsilon_{B'}\colon FGB' \xrightarrow{\simeq} FA'$, then by (db2) there are horizontal morphisms $a\colon GB \to A$ and $c\colon GB' \to A'$ in $\mathbb{A}$ and vertically invertible squares γ_b and γ_d:

$$
\begin{tikzcd}
FGB \arrow[r, "\epsilon_B", "\simeq"'] \arrow[d, equal] \arrow[drr, phantom, "\gamma_b\ \wr\!\shortparallel"] & B \arrow[r, "b", "\simeq"'] & FA \arrow[d, equal] \\
FGB \arrow[rr, "\simeq", "Fa"'] & & FA
\end{tikzcd}
\qquad
\begin{tikzcd}
FGB' \arrow[r, "\epsilon_{B'}", "\simeq"'] \arrow[d, equal] \arrow[drr, phantom, "\gamma_d\ \wr\!\shortparallel"] & B' \arrow[r, "d", "\simeq"'] & FA' \arrow[d, equal] \\
FGB' \arrow[rr, "\simeq", "Fc"'] & & FA'
\end{tikzcd}
$$

Since lifts of horizontal equivalences by a double biequivalence are horizontal equivalences, we have that $a\colon GB \xrightarrow{\simeq} A$ and $c\colon GB' \xrightarrow{\simeq} A'$ are horizontal equivalences in $\mathbb{A}$; thus, since $\mathbb{A}$ is weakly horizontally invariant, there is a vertical morphism $u\colon GB \to GB'$ and a weakly horizontally invertible square

$$
\begin{tikzcd}
GB \arrow[r, "a", "\simeq"'] \arrow[d, "u"'] \arrow[dr, phantom, "\alpha_v\ \simeq"] & A \arrow[d, "u'"] \\
GB' \arrow[r, "\simeq", "c"'] & A'
\end{tikzcd}
$$

We set $Gv := u\colon GB \to GB'$. To define the weakly horizontally invertible square ϵ_v, let us first fix a weak inverse γ'_v of γ_v with respect to some horizontal equivalences (b, b', λ, κ) and $(d, d', \lambda', \kappa')$. We set ϵ_v to be the square given by the pasting

$$
\begin{tikzcd}
FGB \arrow[r, "\epsilon_B"] \arrow[d, "FGv"', "\bullet" marking] \arrow[dr, phantom, "\epsilon_v"] & B \arrow[d, "v", "\bullet" marking] \\
FGB' \arrow[r, "\epsilon_{B'}"'] & B'
\end{tikzcd}
=
\begin{tikzcd}
FGB \arrow[r, "\epsilon_B"] \arrow[d, equals, "\bullet" marking] \arrow[dr, phantom, "e_{\epsilon_B}"] & B \arrow[rr, equals] \arrow[d, equals, "\bullet" marking] \arrow[drr, phantom, "\lambda \cong"] & & B \arrow[d, equals, "\bullet" marking] \\
FGB \arrow[r, "\epsilon_B"] \arrow[d, equals, "\bullet" marking] \arrow[drr, phantom, "\gamma_b \cong"] & B \arrow[r, "b"] & FA \arrow[r, "b'"] \arrow[d, equals, "\bullet" marking] \arrow[dr, phantom, "e_{b'}"] & B \arrow[d, equals, "\bullet" marking] \\
FGB \arrow[rr, "Fa"] \arrow[d, "FGv"', "\bullet" marking] \arrow[drr, phantom, "F\alpha_v \simeq"] & & FA \arrow[r, "b'"] \arrow[d, "Fu'"', "\bullet" marking] \arrow[dr, phantom, "\gamma'_v \simeq"] & B \arrow[d, "v", "\bullet" marking] \\
FGB' \arrow[rr, "Fc"'] \arrow[d, equals, "\bullet" marking] \arrow[drr, phantom, "\gamma_d^{-1} \cong"] & & FA' \arrow[r, "d'"'] \arrow[d, equals, "\bullet" marking] \arrow[dr, phantom, "e_{d'}"] & B' \arrow[d, equals, "\bullet" marking] \\
FGB' \arrow[r, "\epsilon_{B'}"'] \arrow[d, equals, "\bullet" marking] \arrow[dr, phantom, "e_{\epsilon_{B'}}"] & B' \arrow[r, "d"'] \arrow[d, equals, "\bullet" marking] \arrow[drr, phantom, "\lambda'^{-1} \cong"] & FA' \arrow[r, "d'"'] & B' \arrow[d, equals, "\bullet" marking] \\
FGB' \arrow[r, "\epsilon_{B'}"'] & B' \arrow[rr, equals] & & B'
\end{tikzcd}
$$

Note that all the squares in the pasting are weakly horizontally invertible by [15, Lemma A.2.1], and thus so is ϵ_v. We write ϵ'_v for its unique weak inverse with respect to the horizontal adjoint equivalences $(\epsilon_B, \epsilon'_B, \mu_B, \nu_B)$ and $(\epsilon_{B'}, \epsilon'_{B'}, \mu_{B'}, \nu_{B'})$, as given by Lemma 2.8.

If $v = e_B$, we can choose $Ge_B := e_{GB}$ and $\gamma_{e_B} := e_{\epsilon_B}$. Then α_{e_B} can be chosen to be the identity square at the object GB and we get $\epsilon_{e_B} = e_{\epsilon_B}$.

Vertical coherence Given vertical morphisms $v\colon B \to B'$ and $v'\colon B' \to B''$ in $\mathbb{B}$, we define the horizontally invertible comparison square between $Gv' \bullet Gv$ and $G(v'v)$ as follows. Let us denote by $\Omega_{v,v'}$ the pasting

$$
\begin{tikzcd}
FGB \arrow[rr, equals] \arrow[d, equals, "\bullet" marking] \arrow[drr, phantom, "\mu_B \cong"] & & FGB \arrow[d, equals, "\bullet" marking] \\
FGB \arrow[r, "\epsilon_B"] \arrow[d, "FGv"', "\bullet" marking] \arrow[dr, phantom, "\epsilon_v"] & B \arrow[r, "\epsilon'_B"] \arrow[d, "v", "\bullet" marking] \arrow[ddr, phantom, "\epsilon'_{v'v}"] & FGB \arrow[dd, "FG(v'v)", "\bullet" marking] \\
FGB' \arrow[r, "\epsilon_{B'}"] \arrow[d, "FGv'"', "\bullet" marking] \arrow[dr, phantom, "\epsilon_{v'}"] & B' \arrow[d, "v'", "\bullet" marking] & \\
FGB'' \arrow[r, "\epsilon_{B''}"'] \arrow[d, equals, "\bullet" marking] \arrow[drr, phantom, "\mu_{B''}^{-1} \cong"] & B'' \arrow[r, "\epsilon'_{B''}"'] & FGB'' \arrow[d, equals, "\bullet" marking] \\
FGB'' \arrow[rr, equals] & & FGB''
\end{tikzcd}
$$

Note that this square is horizontally invertible, since it is weakly horizontally invertible and its horizontal boundaries are identities. By (db4), there is a unique horizontally invertible square $\Psi_{v,v'}$ as depicted below left such that $F\Psi_{v,v'} = \Omega_{v,v'}$. In particular, one can check that, with this definition of $\Psi_{v,v'}$, the squares ϵ_v, $\epsilon_{v'}$ and $\epsilon_{v'v}$ satisfy the pasting equality below right:

$$
\begin{tikzcd}
GB \arrow[r, equal] \arrow[d, "Gv"'] & GB \arrow[dd, "G(v'v)"] \\
GB' \arrow[r, phantom, "\overset{\cong}{\Psi_{v,v'}}"] \arrow[d, "Gv'"'] & \\
GB'' \arrow[r, equal] & GB''
\end{tikzcd}
\qquad
\begin{tikzcd}
FGB \arrow[r, equal] \arrow[d, "FGv"'] & FGB \arrow[r, "\epsilon_B"] \arrow[dd, "FG(v'v)"] & B \arrow[dd, "v'v"] \\
FGB' \arrow[r, phantom, "\overset{\cong}{F\Psi_{v,v'}}"] \arrow[d, "FGv'"'] & \arrow[r, phantom, "\epsilon_{v'v}"] & \\
FGB'' \arrow[r, equal] & FGB'' \arrow[r, "\epsilon_{B''}"'] & B''
\end{tikzcd}
=
\begin{tikzcd}
FGB \arrow[r, "\epsilon_B"] \arrow[d, "FGv"'] \arrow[dr, phantom, "\epsilon_v"] & B \arrow[d, "v"] \\
FGB' \arrow[r, "\epsilon_{B'}"] \arrow[d, "FGv'"'] \arrow[dr, phantom, "\epsilon_{v'}"] & B' \arrow[d, "v'"] \\
FGB'' \arrow[r, "\epsilon_{B''}"'] & B''
\end{tikzcd}
$$

***G* on squares** Consider a square in $\mathbb{B}$

$$
\begin{tikzcd}
B \arrow[r, "b"] \arrow[d, "v"'] \arrow[dr, phantom, "\beta"] & C \arrow[d, "v'"] \\
B' \arrow[r, "d"'] & C'
\end{tikzcd}
$$

Let us denote by δ the pasting

$$
\begin{tikzcd}
FGB \arrow[rrr, "FGb"] \arrow[d, equal] \arrow[drrr, phantom, "\bar{\epsilon}_b^{-1} \cong"] & & & FGC \arrow[d, equal] \\
FGB \arrow[r, "\epsilon_B"] \arrow[d, "FGv"'] \arrow[dr, phantom, "\epsilon_v"] & B \arrow[r, "b"] \arrow[d, "v"] \arrow[dr, phantom, "\beta"] & C \arrow[r, "\epsilon'_C"] \arrow[d, "v'"] \arrow[dr, phantom, "\epsilon'_{v'}"] & FGC \arrow[d, "FGv'"] \\
FGB' \arrow[r, "\epsilon_{B'}"'] \arrow[d, equal] \arrow[drrr, phantom, "\bar{\epsilon}_d \cong"] & B' \arrow[r, "d"'] & C' \arrow[r, "\epsilon'_{C'}"'] & FGC' \arrow[d, equal] \\
FGB' \arrow[rrr, "FGd"'] & & & FGC'
\end{tikzcd}
$$

Then, by (db4), there is a unique square $\alpha\colon (Gv \,{}^{Gb}_{Gd}\, Gv')$ such that $F\alpha = \delta$. We set $G\beta := \alpha\colon (Gv \,{}^{Gb}_{Gd}\, Gv')$.

Let $b\colon B \to C$ be a horizontal morphism in $\mathbb{B}$, and $\beta = e_b\colon (e_B \,{}^{b}_{b}\, e_C)$. Then we have that $\delta = e_{FGb}$, since $\epsilon_{e_B} = e_{\epsilon_B}$ and $\epsilon'_{e_C} = e_{\epsilon_C}$, and the unique square $\alpha\colon (e_{GB} \,{}^{Gb}_{Gb}\, e_{GC})$ such that $F\alpha = e_{FGb}$ is given by e_{Gb}. Therefore, $Ge_b = e_{Gb}$.

Now let $v\colon B \twoheadrightarrow B'$ be a vertical morphism in $\mathbb{B}$, and $\beta = \mathrm{id}_v\colon (v\ {}^{\mathrm{id}_B}_{\mathrm{id}_{B'}}\ v)$. Then we have that $\delta = \mathrm{id}_{FGv}$, since $\bar{\epsilon}^{-1}_{\mathrm{id}_B} = \mu_B$ and $\bar{\epsilon}_{\mathrm{id}_{B'}} = \mu^{-1}_{B'}$, and ϵ'_v is the weak inverse of ϵ_B with respect to the horizontal adjoint equivalence data $(\epsilon_B, \epsilon'_B, \mu_B, \nu_B)$ and $(\epsilon_{B'}, \epsilon'_{B'}, \mu_{B'}, \nu_{B'})$. The unique square $\alpha\colon (Gv\ {}^{\mathrm{id}_{GB}}_{\mathrm{id}_{GB'}}\ Gv)$ such that $F\alpha = \mathrm{id}_{FGv}$ is given by id_{Gv}. Therefore, $G\mathrm{id}_v = \mathrm{id}_{Gv}$.

Naturality and adjointness of ϵ and ϵ' The assignment of G on squares is natural with the data of ϵ_B, ϵ_b and ϵ_v, and therefore the latter assemble into a horizontal pseudonatural equivalence $\epsilon\colon FG \Rightarrow \mathrm{id}_{\mathbb{B}}$. Moreover, since $(\epsilon_B, \epsilon'_B, \mu_B, \nu_B)$ are horizontal adjoint equivalences, the data of ϵ'_B, ϵ'_b and ϵ'_v also assemble into a horizontal pseudonatural equivalence $\epsilon'\colon \mathrm{id}_{\mathbb{B}} \Rightarrow FG$, where ϵ'_b is defined in a similar manner as ϵ_b was. In particular, $\epsilon\colon FG \Rightarrow \mathrm{id}_{\mathbb{B}}$ and $\epsilon'\colon \mathrm{id}_{\mathbb{B}} \Rightarrow FG$ are adjoint equivalences, where the invertible modifications are given by $\mu\colon \mathrm{id} \cong \epsilon'\epsilon$ and $\nu\colon \epsilon\epsilon' \cong \mathrm{id}$.

It remains to define the horizontal pseudonatural equivalence $\eta\colon \mathrm{id}_{\mathbb{A}} \Rightarrow GF$. For this purpose, we use the horizontal pseudonatural equivalence $\epsilon'\colon \mathrm{id}_{\mathbb{B}} \Rightarrow FG$.

η on objects Let $A \in \mathbb{A}$, and consider the horizontal equivalence $\epsilon'_{FA}\colon FA \xrightarrow{\simeq} FGFA$. By (db2), there is a horizontal morphism $a\colon A \to GFA$ and a vertically invertible square $\rho_A\colon (e_{FA}\ {}^{\epsilon'_{FA}}_{Fa}\ e_{FGFA})$. We set $\eta_A := a\colon A \to GFA$. Note that $\eta_A\colon A \xrightarrow{\simeq} GFA$ is a horizontal equivalence.

η on horizontal morphisms Let $a\colon A \to C$ be a horizontal morphism in $\mathbb{A}$. We denote by ψ_a the pasting below left. By (db4) there is a unique vertically invertible square α as below right such that $F\alpha = \psi_a$; let $\eta_a := \alpha$:

$$
\begin{tikzcd}
FA \arrow[r, "F\eta_A"] \arrow[d, equal, "\bullet" marking] \arrow[dr, phantom, "\rho_A^{-1}\ \cong"] & FGFA \arrow[r, "FGFa"] \arrow[d, equal, "\bullet" marking] \arrow[dr, phantom, "e_{FGFa}"] & FGFC \arrow[d, equal, "\bullet" marking] \\
FA \arrow[r, "\epsilon'_{FA}"'] \arrow[d, equal, "\bullet" marking] & FGFA \arrow[r, "FGFa"'] \arrow[d, phantom, "\epsilon'_{Fa}\ \cong"] & FGFC \arrow[d, equal, "\bullet" marking] \\
FA \arrow[r, "Fa"'] \arrow[d, equal, "\bullet" marking] \arrow[dr, phantom, "e_{Fa}"] & FC \arrow[r, "\epsilon'_{FC}"'] \arrow[d, equal, "\bullet" marking] \arrow[dr, phantom, "\rho_C\ \cong"] & FGFC \arrow[d, equal, "\bullet" marking] \\
FA \arrow[r, "Fa"'] & FC \arrow[r, "F\eta_C"'] & FGFC
\end{tikzcd}
\qquad
\begin{tikzcd}
A \arrow[r, "\eta_A"] \arrow[d, equal, "\bullet" marking] & GFA \arrow[r, "GFa"] \arrow[d, phantom, "\alpha\ \cong"] & GFC \arrow[d, equal, "\bullet" marking] \\
A \arrow[r, "a"'] & C \arrow[r, "\eta_C"'] & GFC
\end{tikzcd}
$$

η on vertical morphisms Let $u\colon A \to A'$ be a vertical morphism in $\mathbb{A}$. We denote by ψ_u the pasting below left:

$$\begin{tikzcd}
FA \arrow[r, "F\eta_A"] \arrow[d, equal] \arrow[dr, phantom, "\rho_A^{-1}\ \cong"] & FGFA \arrow[d, equal] \\
FA \arrow[r, "\epsilon'_{FA}"] \arrow[d, "Fu"'] \arrow[dr, phantom, "\epsilon'_{Fu}\ \simeq"] & FGFA \arrow[d, "FGFu"] \\
FA' \arrow[r, "\epsilon'_{FA'}"'] \arrow[d, equal] \arrow[dr, phantom, "\rho_{A'}\ \cong"] & FGFA' \arrow[d, equal] \\
FA' \arrow[r, "F\eta_{A'}"'] & FGFA'
\end{tikzcd}
\qquad\qquad
\begin{tikzcd}
A \arrow[r, "\eta_A"] \arrow[d, "u"'] \arrow[dr, phantom, "\gamma"] & GFA \arrow[d, "GFu"] \\
A' \arrow[r, "\eta_{A'}"'] & GFA'
\end{tikzcd}$$

Note that all the squares in ψ_u are weakly horizontally invertible by [15, Lemma A.2.1], and thus so is ψ_u. By (db4) there is a unique weakly horizontally invertible square $\gamma\colon (u\ {}^{\eta_A}_{\eta_{A'}}\ GFu)$ as above right such that $F\gamma = \psi_u$; let $\eta_u := \gamma$.

Naturality of η Since $\epsilon'\colon \mathrm{id}_{\mathbb{B}} \Rightarrow FG$ is a horizontal pseudonatural transformation, η_A, η_a, and η_u assemble into a horizontal pseudonatural transformation $\eta\colon \mathrm{id}_{\mathbb{A}} \Rightarrow GF$. Note that η is a horizontal pseudonatural equivalence, because the η_A are horizontal equivalences and the η_u are weakly horizontally invertible squares. Moreover, $\rho\colon \epsilon'_F \cong F\eta$ gives the data of an invertible modification. $\square$

References

[1] **G Böhm**, *The Gray monoidal product of double categories*, Appl. Categ. Structures 28 (2020) 477–515 MR Zbl

[2] **A Campbell**, *The folk model structure for double categories*, seminar talk (2020) Available at `http://web.science.mq.edu.au/groups/coact/seminar/cgi-bin/abstract.cgi?talkid=1616`

[3] **W G Dwyer**, **J Spaliński**, *Homotopy theories and model categories*, from "Handbook of algebraic topology" (I M James, editor), North-Holland, Amsterdam (1995) 73–126 MR Zbl

[4] **T M Fiore**, **S Paoli**, *A Thomason model structure on the category of small n–fold categories*, Algebr. Geom. Topol. 10 (2010) 1933–2008 MR Zbl

[5] **T M Fiore**, **S Paoli**, **D Pronk**, *Model structures on the category of small double categories*, Algebr. Geom. Topol. 8 (2008) 1855–1959 MR Zbl

[6] **M Grandis**, *Higher dimensional categories: from double to multiple categories*, World Sci., Hackensack, NJ (2020) MR Zbl

[7] **M Grandis**, **R Pare**, *Limits in double categories*, Cah. Topol. Géom. Différ. Catég. 40 (1999) 162–220 MR Zbl

[8] **M Grandis**, **R Paré**, *Persistent double limits*, Cah. Topol. Géom. Différ. Catég. 60 (2019) 255–297 MR Zbl

[9] **N Gurski**, *Biequivalences in tricategories*, Theory Appl. Categ. 26 (2012) 349–384 MR Zbl

[10] **M Hovey**, *Model categories*, Mathematical Surveys and Monographs 63, Amer. Math. Soc., Providence, RI (1999) MR Zbl

[11] **N Johnson**, **D Yau**, *2–Dimensional categories*, Oxford Univ. Press (2021) MR Zbl

[12] **G M Kelly**, **S Lack**, *$\mathcal{V}$–Cat is locally presentable or locally bounded if $\mathcal{V}$ is so*, Theory Appl. Categ. 8 (2001) 555–575 MR Zbl

[13] **S Lack**, *A Quillen model structure for 2–categories*, K–Theory 26 (2002) 171–205 MR Zbl

[14] **S Lack**, *A Quillen model structure for bicategories*, K–Theory 33 (2004) 185–197 MR Zbl

[15] **L Moser**, *A double $(\infty, 1)$–categorical nerve for double categories*, preprint (2020) arXiv 2007.01848

[16] **L Moser**, **M Sarazola**, **P Verdugo**, *A 2Cat-inspired model structure for double categories*, Cah. Topol. Géom. Différ. Catég. 63 (2022) 184–236 MR Zbl

[17] **E Riehl**, *A concise definition of a model category*, technical note (2009) Available at `https://emilyriehl.github.io/files/modelcat.pdf`

[18] **E Riehl**, *Categorical homotopy theory*, New Mathematical Monographs 24, Cambridge Univ. Press (2014) MR Zbl

[19] **E Riehl**, **D Verity**, *Elements of ∞–category theory*, Cambridge Studies in Advanced Mathematics 194, Cambridge Univ. Press (2022) MR Zbl

Max Planck Institute for Mathematics
Bonn, Germany

Department of Mathematics, Johns Hopkins University
Baltimore, MD, United States

Department of Mathematics and Statistics, Macquarie University
Sydney, NSW, Australia

lyne.moser@ur.de, mesarazola@jhu.edu, paula.verdugo@hdr.mq.edu.au

Received: 27 August 2020 Revised: 16 November 2021

Algebraic & Geometric Topology 23:4 (2023) 1787–1830
DOI: 10.2140/agt.2023.23.1787
Published: 14 June 2023

Residual torsion-free nilpotence, biorderability and pretzel knots

JONATHAN JOHNSON

The residual torsion-free nilpotence of the commutator subgroup of a knot group has played a key role in studying the biorderability of knot groups. A technique developed by Mayland (1975) provides a sufficient condition for the commutator subgroup of a knot group to be residually torsion-free nilpotent using work of Baumslag (1967, 1969). We apply Mayland's technique to several genus one pretzel knots and a family of pretzel knots with arbitrarily high genus. As a result, we obtain a large number of new examples of knots with biorderable knot groups. These are the first examples of biorderable knot groups for knots which are not fibered or alternating.

57K10

1 Introduction

Let J be a knot in S^3. The *knot exterior* of J is $M_J := S^3 - \nu(J)$, where $\nu(J)$ is the interior of a tubular neighborhood of J, and the *knot group* of J is $\pi_1(M_J)$. Denote the Alexander polynomial of J by Δ_J.

A group Γ is *nilpotent* if its lower central series terminates (is trivial) after finitely many steps. In other words, for some nonnegative integer n,

$$\Gamma_0 \rhd \Gamma_1 \rhd \cdots \rhd \Gamma_n = 1,$$

where $\Gamma_0 = \Gamma$ and $\Gamma_{i+1} = [\Gamma_i, \Gamma]$ for each $i = 0, \dots, n-1$. A group Γ is *residually torsion-free nilpotent* if, for every nontrivial element $x \in \Gamma$, there is a normal subgroup $N \lhd \Gamma$ such that $x \notin N$ and G/N is a torsion-free nilpotent group. We are concerned with when the commutator subgroup of a knot group is residually torsion-free nilpotent, which has applications to ribbon concordance (see Gordon [15]) and the biorderability of the knot group; see Linnell, Rhemtulla and Rolfsen [25].

Several knots are known to have groups with residually torsion-free nilpotent commutator subgroups. The commutator subgroup of fibered knot groups are finitely generated free groups, which are residually torsion-free nilpotent; see Magnus [27]. Work of Mayland and Murasugi [30] shows that the knot groups of pseudoalternating knots, whose Alexander polynomials have a prime power leading coefficient, have residually torsion-free nilpotent commutator subgroups; pseudoalternating knots are defined in Section 3. The knot groups of two-bridge knots have residually torsion-free nilpotent commutator subgroups; see Johnson [20].

There is also the following obstruction to a knot group having residually torsion-free nilpotent commutator subgroup:

Proposition 1.1 *If J is a knot in S^3 with trivial Alexander polynomial, then the commutator subgroup of $\pi_1(M_J)$ cannot be residually torsion-free nilpotent.*

Proof Let G be the commutator subgroup of $\pi_1(M_J)$. Let M^∞ be the infinite cyclic cover of M_J, the covering space of M_J corresponding to G so that $\pi_1(M^\infty) = G$; see Rolfsen [36, Chapter 7] for details. Then

$$H_1(M^\infty, \mathbb{Z}) \cong \bigoplus_{i=1}^{n} \mathbb{Z}[t, t^{-1}]/\langle a_i(t)\rangle,$$

where $a_1(t), \dots, a_n(t)$ are polynomials such that

$$\prod_{i=1}^{n} a_i(t) = \Delta_J(t).$$

Since the Alexander polynomial of J is trivial $G/[G,G] \cong H_1(M^\infty, \mathbb{Z}) = 1$, so $G = [G,G]$. It follows that every term of the lower central series of G is isomorphic to G. Suppose $N \lhd G$ is a proper normal subgroup of G. For each term of the lower central series of G/N,

$$(G/N)_i \cong G_i/N \cong G/N \neq 1,$$

so G/N cannot be nilpotent. Thus, G is not residually torsion-free nilpotent. □

Given the integers $k_1, k_2, \dots, k_n$, define $P(k_1, k_2, \dots, k_n)$ to be the *pretzel knot* represented in the diagram in Figure 1. Mayland [29] describes a technique to examine the commutator subgroup of the group of a knot bounding an unknotted minimal genus Seifert surface; see Section 2. In fact, this is the technique Mayland and Murasugi used to prove their result for pseudoalternating knots [30]. Applying Seifert's algorithm to the diagram in Figure 1 yields an unknotted minimal genus Seifert surface (see Gabai [12]) making pretzel knots ideal candidates for Mayland's technique.

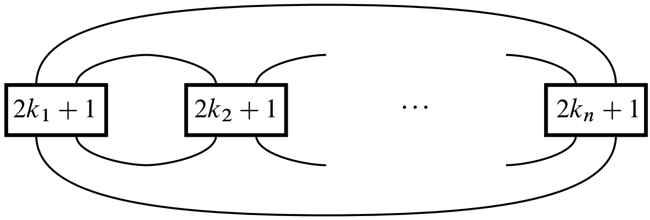


Figure 1: A pretzel knot diagram. The integers in the boxes indicate the number of right-hand half-twist when positive and left-hand half-twist when negative.

Let J be the $P(2p+1, 2q+1, 2r+1)$ pretzel knot for some integers p, q and r. J is a two-bridge knot (possibly trivial) precisely when at least one of p, q and r is equal to 0 or -1 (see Kawauchi [23, Chapter 2]) so for our purposes, we can assume that none of p, q and r are 0 or -1. Permuting the parameters $2p+1$, $2q+1$ and $2r+1$ yields the same (unoriented) knot. Also, $P(-2p-1, -2q-1, -2r-1)$ and $P(2p+1, 2q+1, 2r+1)$ are mirrors of each other. Since $\pi_1(M_J)$ is invariant of reversing orientation and mirroring, we can assume that $1 \le q \le r$.

Theorem 1.2 *Given integers p, q and r with $1 \le q \le r$ and $p \ne 0$ or -1, let J be the $P(2p+1, 2q+1, 2r+1)$ pretzel knot with Alexander polynomial Δ_J whose leading coefficient is a prime power. The commutator subgroup of $\pi_1(M_J)$ is residually torsion-free nilpotent if*

- $p \ge 1$,
- *J is $P(2p+1, 3, 2r+1)$,*
- *J is $P(-3, 2q+1, 2r+1)$ and J is not $P(-3,5,5)$, $P(-3,5,7)$, $P(-3,5,9)$, $P(-3,5,11)$ or $P(-3,7,7)$, or*
- *J is $P(-5, 2q+1, 2r+1)$ and J is not*
  - *$P(-5,7,R)$ when R is* 11, 13, 15, 17, 19, 21, 23 *or* 25,
  - *$P(-5,9,R)$ when R is* 9, 11, 13, 15 *or* 17, *or*
  - *$P(-5,11,R)$ when R is* 11 *or* 13.

Remark 1.3 Proposition 1.1 is the only known obstruction to the commutator subgroup of a genus one pretzel knot group being residually torsion-free nilpotent, so the exceptional cases in Theorem 1.2 with nontrivial Alexander polynomial remain unresolved and cannot be resolved with the technique used in this paper.

If $p \leq -2$ and $1 \leq q \leq r$, then $P(2p+1, 2q+1, 2r+1)$ is not a pseudoalternating knot; see Proposition 3.1. Therefore, all of the examples from Theorem 1.2 where $p < -1$ are new examples of knots with residually torsion-free nilpotent commutator subgroups.

In addition, we also obtain pretzel knots of arbitrarily high genus whose groups have residually torsion-free nilpotent commutator subgroups. However, we were not able to determine whether or not these knots are pseudoalternating so it is possible this result follows from Mayland and Murasugi's work.

Theorem 1.4 *If J is a $P(3, -3, \ldots, 3, -3, 2r+1)$ pretzel knot for some integer r, then the commutator subgroup of $\pi_1(M_J)$ is residually torsion-free nilpotent.*

1.1 Possible generalizations

The techniques used here have a few limitations. First, while our method can be applied to many families of genus one pretzel knots on a case by case basis, this method does not lend itself well to generalizing to all genus one pretzel knots since many of the details depend on the arithmetic properties of p, q and r. Secondly, Mayland's method requires a couple conditions (an unknotted Seifert surface satisfying the free factor property and an Alexander polynomial with prime power leading coefficient) which may not be necessary for a knot group to have residually torsion-free nilpotent commutator subgroup. Nevertheless, we make the following prediction for genus one pretzel knots.

Conjecture 1.5 If J is a genus one pretzel knot then the commutator subgroup of $\pi_1(M_J)$ is residually torsion-free nilpotent if and only if the Alexander polynomial of J is nontrivial.

1.2 Application to biorderability

A group is said to be *biorderable* if there exists a total order of the group's elements, invariant under both left and right multiplication. Chiswell, Glass and Wilson proved the following fact, using work of Linnell, Rhemtulla and Rolfsen [25], and it has been instrumental in determining the biorderability of several knot groups; see Clay, Desmarais and Naylor [8], Johnson [20] and Perron and Rolfsen [35].

Theorem 1.6 [7, Theorem B] *Let J be a knot in S^3. If $\pi_1(M_J)$ has residually torsion-free nilpotent commutator subgroup and all the roots of Δ_J are real and positive then $\pi_1(M_J)$ is biorderable.*

Furthermore, Ito obtained the following obstruction to a knot group being biorderable when the knot is rationally homologically fibered; see Section 2 for the definition of rationally homologically fibered.

Theorem 1.7 [18, Theorem 2] *Let J be a rationally homologically fibered knot. If $\pi_1(M_J)$ is biorderable then Δ_J has at least one real positive root.*

The Alexander polynomial of the pretzel knot $P(2p+1, 2q+1, 2r+1)$ has the form

$$\Delta_J(t) = Nt^2 + (1-2N)t + N,$$

where

$$N = \det\begin{pmatrix} p+q+1 & -q-1 \\ -q & q+r+1 \end{pmatrix}. \tag{1-1}$$

See Section 3 for details. Note that Δ_J has two positive real roots when $N < 0$ and two nonreal roots when $N > 0$. If $N = 0$, then $\Delta_J(t) = 1$. Therefore, we have the following proposition:

Proposition 1.8 *Let J be the $P(2p+1, 2q+1, 2r+1)$ pretzel knot, and let N be defined as in (1-1). If the commutator subgroup of $\pi_1(M_J)$ is residually torsion-free nilpotent and $N < 0$, then $\pi_1(M_J)$ is biorderable. If $N > 0$, then $\pi_1(M_J)$ is never biorderable, regardless of whether or not the commutator subgroup of $\pi_1(M_J)$ is residually torsion-free nilpotent.*

Applying Proposition 1.8 to the results in Theorem 1.2 yields the following corollary.

Corollary 1.9 *Given integers p, g and r with $1 \leq q \leq r$ and $p \neq 0$ or -1, let J be the $P(2p+1, 2q+1, 2r+1)$ pretzel knot with Alexander polynomial Δ_J.*

(1) *$\pi_1(M_J)$ is biorderable if*
- *J is $P(-3, 3, 2r+1)$,*
- *J is $P(-5, 3, 2r+1)$ and $r+4$ is a prime power, or*
- *J is $P(-5, 7, 7)$ or $P(-5, 7, 9)$.*

(2) *$\pi_1(M_J)$ is not biorderable if*
- *$p \geq 1$,*
- *J is $P(-3, 5, 2r+1)$ with $r > 3$,*
- *J is $P(-3, 2q+1, 2r+1)$ with $q \geq 2$,*
- *J is $P(-5, 7, 2r+1)$ with $r \geq 9$,*
- *J is $P(-5, 9, 2r+1)$ with $r \geq 6$, or*
- *J is $P(-5, 2q+1, 2r+1)$ with $q \geq 5$.*

We also have the following corollary to Theorem 1.4.

Corollary 1.10 *If J is the $P(3,-3,\dots,3,-3,2r+1)$ pretzel knot for some integer r, then $\pi_1(M_J)$ is biorderable.*

Details of the proof of Corollary 1.10 are provided in Section 4.

1.3 A possible connection of biorderability to branched covers

Given a knot J in S^3, let $\Sigma_n(J)$ be the *n–fold cyclic cover of S^3 branched over J*; see Rolfsen [36, Chapter 10] for the definition and construction of a cyclic branched cover. Part of the motivation for studying the biorderability of pretzel knots is to investigate the following questions.

Question 1.11 Do there exist knots with $\pi_1(M_J)$ biorderable and $\pi_1(\Sigma_n(J))$ left-orderable for some n?

Question 1.12 Does $\pi_1(M_J)$ not being biorderable imply that $\pi_1(\Sigma_n(J))$ is left-orderable for some n?

Question 1.11 is resolved here.

Theorem 1.13 *For each integer $q \geq 3$, let J_q be the $P(1-2q,2q+1,4q-3)$ pretzel knot. When $q-1$ is a prime power, $\pi_1(M_{J_q})$ is biorderable and $\pi_1(\Sigma_2(J_q))$ is left-orderable.*

Remark 1.14 Question 1.11 is still unanswered for fibered knots and alternating knots.

Question 1.12 remains unresolved as of the writing of this paper. However, some important remarks can be made about this question.

Suppose J is a pretzel knot $P(2p+1,2q+1,2r+1)$ with $1 \leq q \leq r$. When $p \geq 1$, the signature of J is nonzero which likely means that $\pi_1(\Sigma_n(J))$ is left-orderable for n sufficiently large; see Gordon [16, Corollary 1.2 and Question 1.3].

Suppose $p < -1$. By the Montesinos trick [31], the double branched cover of J is the Seifert fibered space

$$\Sigma_2(J) = M\left(0; -1, \frac{-2p-2}{-2p-1}, \frac{1}{2q+1}, \frac{1}{2r+1}\right).$$

By work of Eisenbud, Hirsch and Neumann [10], Lisca and Stipsicz [26], Jankins and Neumann [19], Naimi [34] and Boyer, Rolfsen and Wiest [4], $\Sigma_2(J)$ is left-orderable if and only if there are positive integers a and m such that the triple

$((-2p-2)/(-2p-1), 1/(2q+1), 1/(2r+1))$ is less than some permutation of the triple $(a/m, (m-a)/m, 1/m)$. This happens precisely when $1 < -p \le q$. In this case, $m = 2q$ and $a = 2q-1$. Therefore, we can state the following proposition.

Proposition 1.15 *Suppose J is the $P(2p+1, 2q+1, 2r+1)$ pretzel knot with $p < -1$ and $1 \le q \le r$. Then $\pi_1(\Sigma_2(J))$ is left-orderable if and only if $-p \le q$.*

Thus, if $p < -1$ and the double branched cover of J does not have left-orderable fundamental group, then $q < -p$ so N as defined in (1-1) is negative. Therefore, if Conjecture 1.5 is true, $\pi_1(M_J)$ would be biorderable when $q < -p$ by Proposition 1.8. In particular, if Conjecture 1.5 is true, it's not likely that any nonalternating genus one pretzel knot would be a counterexamples for Question 1.12.

There is some evidence that genus one pretzel knots with no left-orderable cyclic branched covers do exists. It is conjectured (see Boyer, Gordon and Watson [3]) that given a prime orientable closed rational homology sphere Y, $\pi_1(Y)$ is not left-orderable if and only if Y is an L-space, and Issa and Turner show that the cyclic branched covers of the $P(-3, 3, 2r+1)$ pretzel knots are all L-spaces; see [17].

Outline

In Section 2, we review how Mayland's technique [29] can be used to analyze when the commutator subgroup of a knot group is residually torsion-free nilpotent. In Section 3, we apply this technique to genus one pretzel knots and prove Theorems 1.2 and 1.13. In Section 4, we prove Theorem 1.4. Appendix A contains the proofs of some key lemmas. We also provide a chart of our results in Appendix B.

Acknowledgments

The author would like to thank Cameron Gordon for his guidance and encouragement throughout this project, Ahmad Issa and Hannah Turner for many helpful conversations, and the referee for helpful comments and critiques. This research was supported in part by NSF grant DMS-1937215.

2 Preliminaries on Mayland's technique

Mayland used a description of the commutator subgroup of a knot group to investigate when they are residual finite [29]. In this section, we show how Mayland's technique can be used to find a sufficient condition for the commutator subgroup of a knot group to be residually torsion-free nilpotent.

2.1 Mayland's technique

Let J be a knot in S^3 and suppose J bounds a minimal genus Seifert surface S such that S is *unknotted*; in other words, $\pi_1(S^3\backslash S)$ is a free group. Let $\widehat{S} = M_J \cap S$. Let G be the commutator subgroup of $\pi_1(M_J)$.

Let U be the image of a bicollar embedding $\widehat{S}\times[-1,1]\hookrightarrow M_J$ where $\widehat{S}$ is the image of $\widehat{S}\times\{0\}$, and let $M_S = M_J\backslash\widehat{S}$. Denote the images of $\widehat{S}\times(0,1]$ and $\widehat{S}\times[-1,0)$ in M_S by U^+ and U^-, respectively. Let $X=\pi_1(M_S)$, which is a free group of rank $2g$ where g is the genus of J. Consider the inclusion maps $i^+\colon U^+\to M_S$ and $i^-\colon U^-\to M_S$. Let H be the image of the induced map $i^+_*\colon\pi_1(U^+)\to\pi_1(M_S)$ and K be the image of $i^-_*\colon\pi_1(U^-)\to\pi_1(M_S)$.

For each integer n, let X_n be a copy of X, $H_n\subset X_n$ be a copy of H, and $K_n\subset X_n$ be a copy of K. The fundamental groups of U, U^+ and U^- are canonically isomorphic, and since S has minimal genus, i^+_* and i^-_* are injective. Therefore, H_n and K_{n+1} are identified with a rank $2g$ free group F. By Brown and Crowell [5, Theorem 2.1], G is an amalgamated free product of the form

$$G\cong\cdots *_F X_{-2} *_F X_{-1} *_F X_0 *_F X_1 *_F X_2 *_F \cdots. \tag{2-1}$$

Baumslag provides the following sufficient condition for a group to be residually torsion-free nilpotent when G is an ascending chain of *parafree* subgroups; see [1; 2] for a definition and discussion of parafree groups.

Proposition 2.1 [2, Proposition 2.1(i)] *Suppose G is a group which is the union of an ascending chain of subgroups*

$$G_0 < G_1 < G_2 < \cdots < G_n < \cdots < G = \bigcup_{n=1}^{\infty} G_n.$$

Suppose each G_n is parafree of the same rank. If, for each nonnegative integer n, $|G_{n+1} : G_n[G_{n+1},G_{n+1}]|$ is finite, then G is residually torsion-free nilpotent.

For each nonnegative integer m, define Z^m as follows:

$$Z^m := X_{-m} *_F X_{1-m} *_F \cdots *_F X_{m-1} *_F X_m. \tag{2-2}$$

The direct limit of the Z^m is isomorphic to G. Furthermore, since i^+_* and i^-_* are injective, the natural inclusion $Z^m\hookrightarrow Z^{m+1}$ is an embedding, so G is an ascending chain of subgroups

$$Z^0 < Z^1 < Z^2 < \cdots < Z^m < \cdots < G = \bigcup_{m=1}^{\infty} Z^m.$$

A subgroup A of a free group B is a *free factor* if $B = A * D$ for some subgroup D of B. It immediately follows that A is a free factor of B if and only if every (equivalently, at least one) free basis of A extends to a free basis of B. A theorem of Mayland provides sufficient conditions for each Z^m to be parafree.

Proposition 2.2 [29, Theorem 3.2] *If H and K are free factors of $H[X,X]$ and $K[X,X]$, respectively, and $|X : H[X,X]| = |X : K[X,X]| = p^l$ for some prime p and nonnegative integer l, then for every nonnegative m, Z^m is parafree of rank $2g$.*

The knot J is *rationally homologically fibered* if the induced map on homology, $i_h^+ : H_1(U^+;\mathbb{Q}) \to H_1(M_S;\mathbb{Q})$ (or equivalently $i_h^- : H_1(U^-;\mathbb{Q}) \to H_1(M_S;\mathbb{Q})$), is an isomorphism. Let S_+ be a Seifert matrix representing i_h^+ such that $S_- := S_+^T$ is a Seifert matrix representing i_h^-. S_+ is also a presentation matrix for the abelian group $X/H[X,X]$. Similarly, S_- is a presentation matrix for $X/K[X,X]$. Thus,

$$\frac{X}{H[X,X]} \cong \frac{X}{K[X,X]}. \tag{2-3}$$

Denote the standard form of the Alexander polynomial of J by Δ_J. For some nonnegative integer k,

$$t^k \Delta_J(t) = \det(tS_+ - S_+^T) = d_0 + d_1 t + \cdots + d_{2g} t^{2g}.$$

It is a well-known fact that $d_i = d_{2g-i}$; see [33, Chapter 6].

Proposition 2.3 *Suppose J is a knot in S^3. The following statements are equivalent*:

(a) *J is rationally homologically fibered.*

(b) *$|X : H[X,X]|$ is finite.*

(c) *$|X : K[X,X]|$ is finite.*

(d) $\deg \Delta_J = 2g$.

Proof The equivalence of (b) and (c) follows from (2-3).

Since S_+ is a presentation matrix for $X/H[X,X]$, we have that $|X : H[X,X]|$ is finite if and only if $|\det(S_+)| \neq 0$. It follows that (a) and (b) are equivalent.

Since $d_{2g} = d_0 = \det(S_+)$, we have $\deg \Delta_J = 2g$ if and only if $\det(S_+) \neq 0$, so (a) and (d) are equivalent. □

Proposition 2.4 *When J is rationally homologically fibered*,

$$|X : H[X,X]| = |X : K[X,X]| = |\Delta_J(0)|.$$

Proof When J is rationally homologically fibered

$$|X:H[X,X]| = |\det(S_+)| = |\Delta_J(0)|,$$

so the proposition follows from (2-3). □

For each nonnegative m,

$$\frac{Z^{m+1}}{Z^m[Z^{m+1},Z^{m+1}]} \cong \frac{X}{H[X,X]} \times \frac{X}{K[X,X]}.$$

So, when J is rationally homologically fibered,

$$|Z^{m+1}:Z^m[Z^{m+1},Z^{m+1}]| = |X:H[X,X]||K:H[X,X]| = \Delta_J(0)^2 \tag{2-4}$$

by Proposition 2.4.

The Seifert surface S is said to *satisfy the free factor property* if H and K are free factors of $H[X,X]$ and $K[X,X]$, respectively. Note that this property is independent of the orientation of S. A sufficient condition for the residual torsion-free nilpotence of G can be summarized as follows.

Proposition 2.5 *Suppose J is a rationally homologically fibered knot in S^3 with unknotted minimum genus Seifert surface S. If S satisfies the free factor property and $|\Delta_J(0)|$ is a prime power, then the commutator subgroup G is residually torsion-free nilpotent.*

Proof Suppose J is a rationally homologically fibered with unknotted minimum genus Seifert surface S satisfying the free factor property, and suppose $|\Delta_J(0)|$ is a prime power.

Define Z^m for each nonnegative integer m as in (2-2). By Proposition 2.4, $|X:H[X,X]|$ and $|K:H[X,X]|$ are prime powers since J is rationally homologically fibered. Thus, by Proposition 2.2, each Z^m is parafree of rank twice the genus of J.

By (2-4), $|Z^{m+1}:Z^m[Z^{m+1},Z^{m+1}]| = \Delta_J(0)^2$, so $|Z^{m+1}:Z^m[Z^{m+1},Z^{m+1}]|$ is finite. Therefore, by Proposition 2.1, G is residually torsion-free nilpotent. □

2.2 Pseudoalternating knots

A *special alternating diagram* is an alternating link diagram in which all of the crossings have the same sign. Any link with such a diagram is called a *special alternating link*. The Seifert surface described by performing Seifert's algorithm on a special alternating diagram is a *primitive flat surface*. A *generalized flat surface* is any surface which can be obtained by combining some number of primitive flat surfaces by Murasugi sums. See Gabai [11] for a definition and exposition of Murasugi sums. A link which bounds a generalized flat surface is a *pseudoalternating link*. Alternating links are pseudoalternating links. However, all torus links, many of which are not alternating, are also pseudoalternating links.

Pseudoalternating knots are rationally homologically fibered and bound surfaces satisfying the free factor condition [30, Theorem 2.5]. Therefore, the knot group of a pseudoalternating knot, whose Alexander polynomial has a prime power leading coefficient, has residually torsion-free nilpotent commutator subgroup.

3 Genus one pretzel knots

Let J be the $P(2p+1, 2q+1, 2r+1)$ pretzel knot for some integers p, q and r with $1 \le q \le r$ and $p \ne -1$ or 0. Let S be the unknotted genus one surface depicted in Figure 2, which we refer to as the *standard Seifert surface* of J. For the genus one pretzel knots which are not two-bridge knots, the standard Seifert surface is the unique Seifert surface of minimal genus, up to isotopy [13].

In this section, we analyze when S satisfies the free factor property. When $p > 0$, $P(2p+1, 2q+1, 2r+1)$ is an alternating knot, and thus $P(2p+1, 2q+1, 2r+1)$ is pseudoalternating. However, this is not true when $p \le -2$.

Proposition 3.1 *When $1 \le q \le r$ and $p \le -2$, the pretzel knot $P(2p+1, 2q+1, 2r+1)$ is not a pseudoalternating knot.*

Proof Suppose $P(2p+1, 2q+1, 2r+1)$ is pseudoalternating. When $1 \le q \le r$ and $p \le -2$, the diagram in Figure 1 has a minimal number of crossings [24, Theorem 10].

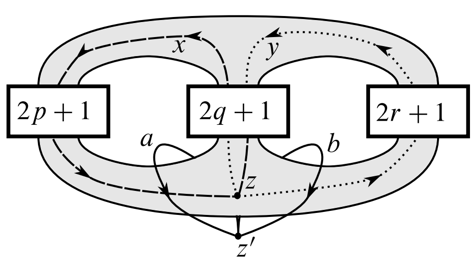


Figure 2: The Seifert surface S of $P(2p+1, 2q+1, 2r+1)$.

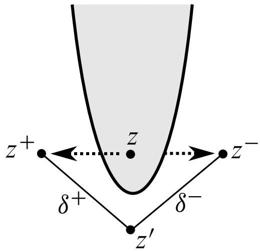


Figure 3: Isotopy of basepoints.

Since this diagram is not alternating, $P(2p+1,2q+1,2r+1)$ cannot be alternating by a theorem of Kauffman, Murasugi and Thistlethwaite [21; 22; 32; 38]. In particular, $P(2p+1,2q+1,2r+1)$ is not special alternating. Thus, $P(2p+1,2q+1,2r+1)$ must be the boundary of a surface S which is the Murasugi sum of two generalized flat surfaces, S_1 and S_2, which are not disks.

By Gabai [11], S must be a minimal genus Seifert surface, so $\chi(S)=-1$. Analyzing the effect of a Murasugi sum on the Euler characteristic yields

$$-1=\chi(S)=\chi(S_1)+\chi(S_2)-1.$$

Since S_1 and S_2 are not disks, neither S_1 nor S_2 has positive Euler characteristic. It follows that $\chi(S_1)=\chi(S_2)=0$, so S_1 and S_2 are both annuli.

The boundary of a Murasugi sum of two annuli is a double twist knot which is alternating. Thus $P(2p+1,2q+1,2r+1)$ is alternating, which is a contradiction. □

Since J is pseudoalternating when $p\geq 0$, we will only need to focus on the case when p is negative.

3.1 Mayland's technique for genus one pretzel knots

Define M_J, M_S, X, H and K as in Section 2. Here we offer a concrete description of the maps on fundamental groups i_*^+ and i_*^- for genus one pretzel knots. This is the same description used by Crowell and Trotter in [9]. Choose a basepoint z on the lower part of S, and let x and y be the classes generating $\pi_1(S,z)$ represented by the loops indicated in Figure 2. Let z^+ and z^- be push-offs of z of each side of S. Let z' be the basepoint of M_S obtained by shifting z tangentially along S through ∂S. Let δ^+ and δ^- be arcs connecting z' to z^+ and z^-, respectively; see Figure 3. Finally, let a and b be the indicated classes generating $\pi_1(M_S,z')$.

By slightly isotoping elements of $\pi_1(S,z)$ off of S, $\pi_1(U^+,z^+)$ and $\pi_1(U^-,z^-)$ are canonically identified to $\pi_1(S,z)$, which is a rank two free group F generated by x and y. The group $X := \pi_1(M_S,z')$ is a rank two free group generated by a and b. The map $i_*^+\colon F\to X$ takes a class $[\gamma]$ in $\pi_1(U^+,z^+)=F$ to the class $[\delta^+ * \gamma * (-\delta^+)]$ in $\pi_1(M_S,z')=X$. Likewise, the map $i_*^-\colon F\to X$ takes $[\gamma]$ to $[\delta^- * \gamma * (-\delta^-)]$.

With these choices, we define the elements

$$\begin{aligned} \alpha_H &:= i_*^+(x) = (b^{-1}a)^{q+1}a^p, & \alpha_K &:= i_*^-(x) = (ab^{-1})^q a^{p+1}, \\ \beta_H &:= i_*^+(y) = b^{r+1}(a^{-1}b)^q, & \beta_K &:= i_*^-(y) = b^r(ba^{-1})^{q+1}, \end{aligned} \tag{3-1}$$

so that

$$H = \langle\{\alpha_H,\beta_H\}\rangle \quad \text{and} \quad K = \langle\{\alpha_K,\beta_K\}\rangle.$$

Thus, the Seifert matrices for i_*^+ and i_*^- are

$$S_+ = \begin{pmatrix} p+q+1 & -q-1 \\ -q & q+r+1 \end{pmatrix} \quad \text{and} \quad S_- = \begin{pmatrix} p+q+1 & -q \\ -q-1 & q+r+1 \end{pmatrix}. \tag{3-2}$$

Let $N=\det S_+ = \det S_-$. Up to multiplication by a signed power of t, the Alexander polynomial of J is

$$\Delta_J(t) = Nt^2 + (1-2N)t + N.$$

When $N\neq 0$, J is rationally homologically fibered by Proposition 2.3. Simply considering the integer N can provide useful information.

Proposition 3.2 *When $N=0$, G is not residually torsion-free nilpotent.*

Proof When $N=0$ we have $\Delta_J(t)=1$, so G cannot be residually nilpotent by Proposition 1.1. □

Proposition 3.3 *If $|N|=1$, then the standard Seifert surface S does not satisfy the free factor property.*

Proof Let S be the standard Seifert surface of J, and define X, H, and K as in Section 2. Each of these are rank two free groups. Suppose S satisfies the free factor property.

When $|N|=1$ we have that $X/H[X,X]\cong X/K[X,X]\cong 1$ by Proposition 2.4, so $X=H[X,X]=K[X,X]$. Since H is a free factor of $H[X,X]$ and both are rank two free groups, $H=H[X,X]=X$. Similarly, since K is a free factor of $K[X,X]$ and both are rank two free groups, $K=X$. This implies that i_*^+ and i_*^- are isomorphisms. Thus, $\pi_1(M_J)$ is an extension of $\mathbb{Z}$ described by the short exact sequence

$$1\to X\to \pi_1(M_J)\to\mathbb{Z}\to 1.$$

The Stallings fibration theorem implies that J is a genus one fibered knot [37]. However, the only genus one fibered knots are the trefoil and the figure eight knot [6; 14], which is a contradiction since we are assuming J is not a two-bridge knot. □

In light of Proposition 2.5, to prove the commutator subgroup of $\pi_1(M_J)$ is residually torsion-free nilpotent, it is sufficient to show S satisfies the free factor property.

3.2 Outline of the procedure

In each case we use the same basic procedure, outlined below, to analyze whether or not S satisfies the free factor property.

(1) Find a presentation matrix for $X/H[X,X]$ of the form

$$\begin{pmatrix} u & v \\ 0 & w \end{pmatrix} \quad \text{or} \quad \begin{pmatrix} u & 0 \\ v & w \end{pmatrix}$$

using row operations. Note, u and w can always be made positive. Thus, $X/H[X,X]$ is isomorphic to $(\mathbb{Z}/u\mathbb{Z}) \times (\mathbb{Z}/w\mathbb{Z})$. The $\mathbb{Z}/u\mathbb{Z}$ factor is generated by the class of a, and the $\mathbb{Z}/w\mathbb{Z}$ factor is generated by the class of b.

(2) Since $X/H[X,X]$ is abelian, the set $\mathcal{C}$ is a set of coset representatives of $H[X,X]$:

$$\mathcal{C} = \{a^k b^l \mid 0 \le k < u,\ 0 \le l < w\}.$$

Given $x \in X$, denote by $\bar{x}$ the coset representative of x in $\mathcal{C}$. Define

$$x_{c,x} := cx(\overline{cx})^{-1},$$

where $c \in \mathcal{C}$ and $x \in \{a,b\}$. From this we find the following free basis for $H[X,X]$ using the Reidemeister–Schreier method:

$$\mathcal{B} = \{x_{c,x} \mid c \in \mathcal{C},\ x \in \{a,b\},\ x_{c,x} \neq 1\}.$$

See [28] for details.

(3) Use the Reidemeister–Schreier rewriting process to rewrite the generating set of H from (3-1). A word $\alpha \in H$, where $\alpha = \alpha_1^{s_1} \dots \alpha_k^{s_k}$ with $\alpha_i \in \{a,b\}$ and $s_i = \pm 1$, can be rewritten as

$$\alpha = x_{c_1,\alpha_1}^{s_1} \dots x_{c_k,\alpha_k}^{s_k},$$

where

$$c_i = \begin{cases} \overline{\alpha_1 \dots \alpha_{i-1}} & \text{when } s_i = 1, \\ \overline{\alpha_1 \dots \alpha_i} & \text{when } s_i = -1. \end{cases}$$

(4) Determine if the generating set of H can be extended to a free basis of $H[X,X]$.

(5) Repeat this procedure for K.

When the free bases of H and K can be extended to free bases of $H[X,X]$ and $K[X,X]$, respectively, S satisfies the free factor property. If the chosen basis of either H or K fails to extend, then S cannot satisfy the free factor property.

3.3 Knots whose standard Seifert surface satisfies the free factor property

Lemma 3.4 *If J is $P(-5,7,7)$ or $P(-5,7,9)$ then S satisfies the free factor property.*

Proof Suppose J is $P(-5,7,7)$. From (3-1),

$$\alpha_H = (b^{-1}a)^4a^{-3}, \quad \alpha_K = (ab^{-1})^3a^{-2},$$
$$\beta_H = b^4(a^{-1}b)^3, \quad \beta_K = b^3(ba^{-1})^4.$$

The abelian group $X/H[X,X]$ has presentation matrix

$$\begin{pmatrix} 1 & -4 \\ -3 & 7 \end{pmatrix},$$

which becomes

$$\begin{pmatrix} 1 & 1 \\ 0 & 5 \end{pmatrix}$$

after row operations.

From this we get $\mathcal{C} = \{1, b, b^2, b^3, b^4\}$ as a set of coset representatives of $H[X,X]$. We apply Reidemeister–Schreier to obtain the following free basis of $H[X,X]$:

$$\mathcal{B} = \{ab, ba, b^2ab^{-1}, b^3ab^{-2}, b^4ab^{-3}, b^5\}.$$

Label the basis elements as follows: $x_k := b^kab^{1-k}$ for $0 \le k \le 4$ and $x_5 := b^5$.

Now we can rewrite α_H and β_H in terms of $\mathcal{B}$, obtaining

$$\begin{aligned}\alpha_H &= (b^{-5})(b^4ab^{-3})(b^2ab^{-1})(ab)(b^{-5})(b^3a^{-1}b^{-4})(b^5)(b^{-1}a^{-1}) \\ &= x_5^{-1}x_4x_2x_0x_5^{-1}x_4^{-1}x_5x_0^{-1}\end{aligned}$$

and

$$\beta_H = (b^5)(b^{-1}a^{-1})(ba^{-1}b^{-2})(b^3a^{-1}b^{-4})(b^5) = x_5x_0^{-1}x_2^{-1}x_4^{-1}x_5.$$

Thus

$$\alpha_H = \beta_H^{-1}x_4^{-1}x_5x_0^{-1},$$

so

$$x_4 = x_5x_0^{-1}\alpha_H^{-1}\beta_H^{-1}$$

and

$$x_2 = \beta_H\alpha_Hx_0\beta_H^{-1}x_5x_0^{-1}.$$

Therefore, the set

$$\{\alpha_H, \beta_H, x_0, x_1, x_3, x_5\}$$

is a generating set of six elements for $H[X,X]$, and thus is a free basis. It follows that

$$H[X,X] = H * \{x_0, x_1, x_3, x_5\},$$

so H is a free factor of $H[X,X]$.

After row reductions, $X/K[X,X]$ has presentation matrix

$$\begin{pmatrix} 1 & -3 \\ 0 & 5 \end{pmatrix}.$$

From this we get a free basis of $K[X,X]$:

$$x_k := \begin{cases} b^k a b^{-3-k} & \text{for } 0 \le k \le 1, \\ b^k a b^{2-k} & \text{for } 2 \le k \le 4, \\ b^5 & \text{for } k = 5. \end{cases}$$

Rewriting α_K and β_K, we get

$$\alpha_K = (ab^{-3})(b^2 a)(b^{-5})(b^4 a b^{-2})(b a^{-1} b^{-3})(b^3 a^{-1}) = x_0 x_2 x_5^{-1} x_4 x_3^{-1} x_0^{-1}$$

and

$$\beta_K = (b^4 a^{-1} b^{-1})(b^2 a^{-1} b^{-4})(b^5)(a^{-1} b^{-2})(b^3 a^{-1}) = x_1^{-1} x_4^{-1} x_5 x_2^{-1} x_0^{-1}.$$

Thus

$$x_4 = x_5 x_2^{-1} x_0^{-1} \beta_K^{-1} x_1^{-1}$$

and

$$x_3 = x_0^{-1} \alpha_K^{-1} \beta_K^{-1} x_1^{-1}.$$

Therefore, the set

$$\{\alpha_K, \beta_K, x_0, x_1, x_2, x_5\}$$

is a free basis of $K[X,X]$ so K is a free factor of $K[X,X]$. Therefore, S satisfies the free factor property.

Suppose J is $P(-5,7,9)$. $X/H[X,X]$ has presentation matrix

$$\begin{pmatrix} 1 & -4 \\ -3 & 8 \end{pmatrix},$$

which becomes

$$\begin{pmatrix} 1 & 0 \\ 0 & 4 \end{pmatrix}$$

after row operations.

By applying Reidemeister–Schreier, we obtain the free basis $\{x_0, x_1, x_2, x_3, x_4\}$, where $x_i = b^i a b^{-i}$ for $i = 0, \dots, 3$ and $x_4 = b^4$. Then

$$\alpha_H = (b^{-1}a)^4 a^{-3} = (b^{-4})(b^3 a b^{-3})(b^2 a b^{-2})(b a b^{-1})(a^{-1})(a^{-1}) = x_4^{-1} x_3 x_2 x_1 x_0^{-2}$$

and

$$\begin{aligned}\beta_H &= b^5(a^{-1}b)^3\\ &= (b^{-4})(ba^{-1}b^{-1})(b^2a^{-1}b^{-2})(b^3a^{-1}b^{-3})(b^4)\\ &= x_4x_1^{-1}x_2^{-1}x_3^{-1}x_4.\end{aligned}$$

Thus

$$x_4 = \beta_H\alpha_H x_0^2$$

and

$$x_3 = x_4\alpha_H x_0^2 x_1^{-1}x_2^{-1}.$$

Therefore, the set

$$\{\alpha_H, \beta_H, x_0, x_1, x_2\}$$

is a free basis of $H[X,X]$, so H is a free factor of $H[X,X]$.

A similar argument shows K is a free factor of $K[X,X]$. Therefore, S satisfies the free factor property. □

Lemma 3.5 *If J is a $P(-3,3,2r+1)$ pretzel knot then S satisfies the free factor property.*

Proof From (3-1),

$$\begin{aligned}\alpha_H &= b^{-1}ab^{-1}a^{-1}, &\quad \alpha_K &= ab^{-1}a^{-1},\\ \beta_H &= b^{r+1}a^{-1}b, &\quad \beta_K &= b^{r+1}a^{-1}ba^{-1}.\end{aligned}$$

The abelian group $X/H[X,X]$ has presentation matrix

$$\begin{pmatrix}1 & 0\\ 0 & 2\end{pmatrix}$$

when r is even and

$$\begin{pmatrix}1 & -1\\ 0 & 2\end{pmatrix}$$

when r is odd.

Using $\mathcal{C} = \{1, b\}$ as a set of coset representatives, we apply Reidemeister–Schreier to obtain $\mathcal{B} = \{x_0, x_1, x_2\}$, a free basis of $H[X,X]$.

When r is even

$$x_0 = a, x_1 = bab^{-1} \quad\text{and}\quad x_2 = b^2,$$

so

$$\alpha_H = (b^{-2})(bab^{-1})(a^{-1}) = x_2^{-1}x_1x_0^{-1}$$

and

$$\beta_H = (b^{2k})(ba^{-1}b^{-1})(b^2) = x_2^k x_1^{-1}x_2,$$

where $r = 2k$.

When r is odd
$$x_0 = ab^{-1}, \quad x_1 = ba \quad \text{and} \quad x_2 = b^2,$$
so
$$\alpha_H = (b^{-2})(ba)(b^{-2})(ba^{-1}) = x_2^{-1}x_1x_2^{-1}x_0^{-1}$$
and
$$\beta_H = (b^{2k+2})(a^{-1}b^{-1})(b^2) = x_2^{k+1}x_1^{-1}x_2,$$
where $r = 2k+1$.

In either case, the set $\{\alpha_H, \beta_H, x_2\}$ is a free basis of $H[X,X]$ so H is a free factor of $H[X,X]$.

$X/K[X,X]$ has presentation matrix
$$\begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}.$$
Using $\mathcal{C} = \{1, a\}$ as a set of coset representatives, we get the free basis of $K[X,X]$, $\mathcal{B} = \{x_0, x_1, x_2\}$, where
$$x_0 = a^2, \quad x_1 = b \quad \text{and} \quad x_2 = aba^{-1}.$$
Thus,
$$\alpha_K = x_2^{-1} \quad \text{and} \quad \beta_K = x_1^{r+1}x_0^{-1}x_2.$$
The set $\{\alpha_K, \beta_K, x_1\}$ is a free basis of $K[X,X]$ so K is a free factor of $K[X,X]$. Therefore, S satisfies the free factor property. □

The proofs of the following results can be found in Appendix A.

Lemma 3.6 *If J is a $P(2p+1, 3, 2r+1)$ pretzel knot with $p < -2$ then S satisfies the free factor property.*

Lemma 3.7 *Suppose J is $P(-3, 2q+1, 2r+1)$ and one of the following conditions holds:*

(1) $q = 2$ *and* $r \geq 6$,

(2) $q = 3$ *and* $r \geq 4$,

(3) $q > 3$.

Then S satisfies the free factor property.

Lemma 3.8 *Suppose J is $P(-5, 2q+1, 2r+1)$ and one of the following conditions holds:*

(1) $q = 3$ *and* $r \geq 13$,

(2) $q = 4$ *and* $r \geq 9$,

(3) $q = 5$ and $r \geq 7$,

(4) $q > 5$.

Then S satisfies the free factor property.

3.4 Proof of Theorem 1.13

For each integer $q \geq 3$, let J_q be the pretzel knot $P(1-2q, 2q+1, 4q-3)$, so $p = -q$ and $r = 2q-2$.

Lemma 3.9 *For all $q \geq 3$, the standard Seifert surface S of J_q satisfies the free factor property.*

Proof The knot J_3 is $P(-5, 7, 9)$. Thus, for J_3, S satisfies the free factor property by Lemma 3.4.

Assume $q \geq 4$. Define X, H and K as above. After row reductions, $X/H[X, X]$ has presentation matrix

$$\begin{pmatrix} 1 & -(q+1) \\ 0 & -N \end{pmatrix},$$

where $N = -(q-1)^2$.

Let $C = -N = (q-1)^2$. Using Reidemeister–Schreier, we obtain the basis

$$\{ab^{-q-1}, bab^{-q-2}, \dots, b^{C-q-2}ab^{1-C}, b^{C-q-1}a, b^{C-q}ab^{-1}, \dots, b^{C-1}ab^{-q}, b^C\}.$$

To simplify computations we modify this basis by multiplying some of the elements by b^{-C} on the right, and obtain a free basis $\mathcal{B} = \{x_0, \dots, x_C\}$ of $H[X, X]$, where $x_k = b^k ab^{-q-1-k}$ for $k = 0, \dots, C-1$ and $x_C = b^C$.

We can rewrite α_H and β_H as

$$\begin{aligned} \alpha_H &= (b^{-1}a)^{q+1}a^{-q} \\ &= x_C^{-1}x_{C-1}x_C(x_{q-1}\cdots x_{i(q-2)-1}\cdots x_{q(q-2)-1})x_C x_{q-2}x_{q-3}^{-1}x_C^{-1} \\ &\qquad\qquad (x_{(q-3)(q+1)}^{-1}x_{(q-4)(q+1)}^{-1}\cdots x_{(q-i)(q+1)}^{-1}\cdots x_0^{-1}) \end{aligned}$$

and

$$\beta_H = b^{2q-1}(a^{-1}b)^q = x_{q-2}^{-1}x_C^{-1}(x_{q(q-2)-1}^{-1}\cdots x_{q(q-i)-1}^{-1}\cdots x_{q-1}^{-1})x_C^{-1}x_{C-1}^{-1}x_C.$$

Since $q \geq 4$, the generator x_0 appears once in the expression for α_H and does not appear in the expression for β_H. Also, since $q-2 < C-1$ and $qk-1 < C-1$ for all $k = 1, \dots, q-2$, x_{C-1} only appears once in the expression for β_H.

Thus x_{C-1} is a product of $\beta_H, x_1, \dots, x_{C-2}, x_C$ and x_0 is a product of $\alpha_H, x_1, \dots, x_C$. Therefore, the set $\{\alpha_H, \beta_H, x_1, \dots, x_{C-2}, x_C\}$ is a free basis of $H[X,X]$, so H is a free factor of $H[X,X]$.

After row reductions, $X/K[X,X]$ has presentation matrix

$$\begin{pmatrix} 1 & -q \\ 0 & C \end{pmatrix}.$$

We obtain a free basis $\mathcal{B} = \{x_0, \dots, x_C\}$ of $K[X,X]$, where $x_k = b^k a b^{-(q+k)}$ for $k = 0, \dots, C-1$ and $x_C = b^C$.

We can rewrite α_K and β_K as

$$\begin{aligned} \alpha_K &= (ab^{-1})^q a^{-q+1} \\ &= (x_0 x_{q-1} x_{2(q-1)} \cdots x_{(q-2)(q-1)}) x_C x_0 x_C^{-1} (x_{q(q-2)}^{-1} x_{q(q-3)}^{-1} \cdots x_0^{-1}) \end{aligned}$$

and

$$\beta_K = b^{2q-2}(ba^{-1})^{q+1} = x_{q-1}^{-1} x_0^{-1} x_C^{-1} (x_{(q-2)(q-1)}^{-1} x_{(q-3)(q-1)}^{-1} \cdots x_0^{-1}.$$

The generator x_q appears once in the expression for α_K and does not appear in the expression for β_K. Also, x_C only appears once in the expression for β_K. Therefore, the set $\{\alpha_K, \beta_K, x_0, \dots, x_{q-1}, x_{q+1}, \dots, x_{C-1}\}$ is a free basis of $K[X,X]$, so K is a free factor of $K[X,X]$. Hence, S satisfies the free factor property. □

Proof of Theorem 1.13 By Lemma 3.9, J_q has a Seifert surface satisfying the free factor property. The Alexander polynomial of J_q is $Nt^2 + (1-2N)t + N$ where $N = -(q-1)^2$, so J_q is rationally homologically fibered and Δ_{J_q} has two positive real roots.

When $q-1$ is a prime power, $|\Delta_{J_q}(0)| = (q-1)^2$ is also a prime power. Therefore, when $q-1$ is a prime power, $\pi_1(M_{J_q})$ has residually torsion-free nilpotent commutator subgroup by Proposition 2.5, and $\pi_1(M_{J_q})$ is biorderable by Proposition 1.8. Since $p = -q$, we have that $\Sigma_2(J_q)$ is left-orderable by Proposition 1.15 for all $q \geq 3$. □

3.5 Knots where the standard Seifert surface does not satisfy the free factor property

Lemma 3.10 *If J is $P(1-2q, 2q+1, 2q^2+1)$ or $P(1-2q, 2q+1, 2q^2-3)$ then S does not satisfy the free factor property.*

Proof If J is $P(1-2q, 2q+1, 2q^2+1)$ then $p = -q$ and $r = q^2$, and if J is $P(1-2q, 2q+1, 2q^2-3)$ then $p = -q$ and $r = q^2-2$. In both cases $|N| = 1$, so by Proposition 3.3 S does not satisfy the free factor property. □

Lemma 3.11 *Suppose J is one of*

- $P(-3,5,11)$,
- $P(-3,7,7)$,
- $P(-5,7,R)$ *for* $R = 11, 13, 21, 23$ *or* 25,
- $P(-5,9,R)$ *for* $R = 9, 13, 15$ *or* 17,
- $P(-5,11,11)$, *or*
- $P(-5,11,13)$.

Then S, the standard Seifert surface of J, does not satisfy the free factor property.

Proof If J is $P(-3,5,11)$, then $X/H[X,X]$ has presentation matrix

$$\begin{pmatrix} 1 & -1 \\ 0 & 2 \end{pmatrix}.$$

We have the free basis $\mathcal{B} = \{ab^{-1}, bab^{-2}, b^2\}$ of $H[X,X]$. Then let $x_0 = ab^{-1}$, $x_1 = bab^{-2}$ and $x_2 = b^2$, so

$$\beta_H = b^6(a^{-1}b)^2 = x_2^2 x_1^{-2} x_2.$$

Let

$$\Gamma := \frac{H[X,X]}{\langle \beta_H^{H[X,X]} \rangle} \cong \langle x_0, x_1, x_2 : x_2^3 x_1^{-2} \rangle,$$

where $\langle \beta_H^{H[X,X]} \rangle$ is the normal closure of β_H in $H[X,X]$. Suppose $\{\alpha_H, \beta_H\}$ could be extended to a basis of $H[X,X]$. Then Γ is a free group and Γ has a subgroup isomorphic to $E := \langle x_1, x_2 : x_2^3 x_1^{-2} \rangle$. The abelianization of E is $\mathbb{Z}$, but E is not abelian since x_1 and x_2 do not commute. Thus E is not free, and Γ cannot be free either, which is a contradiction.

Therefore H is not a free factor of $H[X,X]$, and S does not satisfy the free factor property.

If J is $P(-5,7,25)$, then $H[X,X]$ has a free basis $x_0 = a$, $x_1 = bab^{-1}$, $x_2 = b^2ab^{-2}$, $x_3 = b^3ab^{-3}$ and $x_4 = b^4$. Under this basis

$$\beta_H \alpha_H = b^{12} a^{-2} = x_4^3 x_0^{-2}.$$

We can extend $\{\alpha_H, \beta_H\}$ to a free basis of $H[X,X]$ if and only if $\{\alpha_H, \beta_H \alpha_H\}$ can be extended to a free basis. However, an argument similar to the previous case shows that $\beta_H \alpha_H$ cannot be extended to a basis of $H[X,X]$.

Therefore, H is not a free factor of $H[X,X]$, and S does not satisfy the free factor property.

If J is $P(-5,7,13)$ or $P(-5,7,21)$, then $H[X,X]$ has free basis $x_0=a$, $x_1=bab^{-1}$, and $x_2=b^2$. Thus, the set $\{x_0,x_1,x_2^{-1}x_1x_0\}$ is also a free basis of $H[X,X]$. Denote $x_2^{-1}x_1x_0$ by y.

Using the basis $\{x_0,x_1,y\}$,

$$\alpha_H=(b^{-1}a)^4a^{-3}=y^2x_0^{-3}.$$

An argument similar to the previous cases shows that α_H cannot be extended to a basis of $H[X,X]$. Therefore H is not a free factor of $H[X,X]$, and S does not satisfy the free factor property.

The proofs of the other cases are similar to the cases above. Here we provide the elements obstructing the free factor property.

When J is $P(-3,7,7)$,

$$\beta_H=b^4(a^{-1}b)^3=x_2x_1^{-3}x_2,$$

where $x_0=ab^{-1}$, $x_1=bab^{-2}$ and $x_2=b^2$.

When J is $P(-5,7,11)$,

$$\beta_H=b^6(a^{-1}b)^3=x_3x_2^{-3}x_3,$$

where $x_0=ab^{-1}$, $x_1=bab^{-2}$, $x_2=b^2ab^{-3}$ and $x_3=b^3$.

When J is $P(-5,7,23)$,

$$\beta_H=b^12(a^{-1}b)^3=x_3^3x_2^{-3}x_3,$$

where $x_0=ab^{-1}$, $x_1=bab^{-2}$, $x_2=b^2ab^{-3}$ and $x_3=b^3$.

When J is $P(-5,9,9)$,

$$\beta_H=b^5(a^{-1}b)^4=x_0^5(x_2^{-1}x_1x_0)^2,$$

where $x_0=b$, $x_1=aba^{-1}$ and $x_2=a^2$.

When J is $P(-5,9,13)$,

$$\beta_H=b^7(a^{-1}b)^4=x_0^7(x_2^{-1}x_1x_0)^2,$$

where $x_0=b$, $x_1=aba^{-1}$ and $x_2=a^2$.

When J is $P(-5,9,15)$,

$$\beta_K\alpha_K=b^8a^{-3}=x_4^2x_0^{-3},$$

where $x_0=a$, $x_1=bab^{-1}$, $x_2=b^2ab^{-2}$, $x_3=b^3ab^{-3}$ and $x_4=b^4$.

When J is $P(-5,9,17)$,

$$\beta_H = b^9(a^{-1}b)^4 = (x_0x_6^{-1}x_4x_2)^3(x_6^{-1}x_5x_2)^2,$$

where $x_0 = ba^2$, $x_1 = aba$, $x_2 = a^2b$, $x_3 = a^3ba^{-1}$, $x_4 = a^4ba^{-2}$, $x_5 = a^5ba^{-3}$ and $x_6 = a^6$.

When J is $P(-5,11,11)$,

$$\beta_H = b^6(a^{-1}b)^5 = x_3x_2^{-5}x_3,$$

where $x_0 = ab^{-1}$, $x_1 = bab^{-2}$, $x_2 = b^2ab^{-3}$ and $x_3 = b^3$.

When J is $P(-5,11,13)$,

$$\beta_K = b^6(a^{-1}b)^6 = (x_0x_3x_6)^3(x_0x_2x_4x_6)^2,$$

where $x_0 = ba^{-3}$, $x_1 = aba^{-4}$, $x_2 = a^2ba^{-5}$, $x_3 = a^3ba^{-6}$, $x_4 = a^4ba^{-7}$, $x_5 = a^5ba^{-8}$ and $x_6 = a^6$. □

3.6 Proof of Theorem 1.2

Lemma 3.12 *If J is $P(-3,5,7)$, $P(-5,7,17)$ or $P(-5,9,11)$ then $\pi_1(M_J)$ does not have a residually torsion-free nilpotent commutator subgroup.*

Proof For each of these knots $N = 0$, so this follows from Proposition 3.2. □

Proof of Theorem 1.2 When $p \geq 1$, S is pseudoalternating so S satisfies the free factor property [30]. Therefore, when $p \geq 1$, the knot group of $P(2p+1, 2q+1, 2r+1)$ has residually torsion-free nilpotent commutator subgroups when $|\Delta_J(0)|$ is a prime power.

The other positive results follow from applying Proposition 2.5 to Lemmas 3.4, 3.5, 3.6, 3.7, 3.8, and 3.12. □

4 Higher genus pretzel knots

In this section we prove Theorem 1.4, which presents a family of pretzel knots with arbitrarily high genus whose groups have residually torsion-free nilpotent commutator subgroups.

Let k be a positive integer, and let r be any integer. Suppose J is the $(2k+1)$–parameter pretzel knot $P(3,-3,\ldots,3,-3,2r+1)$ with genus k Seifert surface S as shown in Figure 4. Define X, H and K as in Section 2.

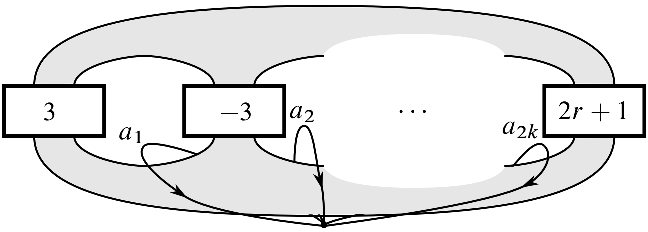


Figure 4: Seifert surface for higher genus pretzel knots.

Proof of Theorem 1.4 X is a free group of rank $2k$ with generating set $\{a_1,\dots,a_{2k}\}$ as shown in Figure 4. By choosing a suitable free basis for $\pi_1(S)$, the subgroup H has the free basis

$$\begin{aligned}
\alpha_1 &= (a_1^{-1}a_2)a_1,\\
\alpha_2 &= (a_3^{-1}a_2)^2(a_2^{-1}a_1)^2,\\
&\ \ \vdots\\
\alpha_{2i-1} &= (a_{2i-1}^{-1}a_{2i})(a_{2i-2}^{-1}a_{2i-1}),\\
\alpha_{2i} &= (a_{2i+1}^{-1}a_{2i})^2(a_{2i}^{-1}a_{2i-1})^2,\\
&\ \ \vdots\\
\alpha_{2k-1} &= (a_{2k-1}^{-1}a_{2k})(a_{2k-2}^{-1}a_{2k-1}),\\
\alpha_{2k} &= a_{2k}^{r+1}(a_{2k}^{-1}a_{2k-1})^2.
\end{aligned}$$

$X/H[X,X]$ has the presentation matrix

$$\begin{pmatrix}
0 & 1 & & & & & \\
2 & 0 & -2 & & & & \\
 & -1 & 0 & 1 & & & \\
 & & 2 & 0 & -2 & & \\
 & & & \ddots & \ddots & \ddots & \\
 & & & & -1 & 0 & 1 \\
 & & & & & 2 & r-1
\end{pmatrix}, \tag{4-1}$$

which after row operations becomes

$$\begin{pmatrix}
0 & 1 & & & & & \\
2 & 0 & 0 & & & & \\
 & 0 & 0 & 1 & & & \\
 & & 2 & 0 & 0 & & \\
 & & & \ddots & \ddots & \ddots & \\
 & & & & 0 & 0 & 1 \\
 & & & & & 2 & 0
\end{pmatrix}.$$

It follows that
$$\frac{X}{H[X,X]} \cong \bigoplus_{j=1}^{k} (\mathbb{Z}/2\mathbb{Z}),$$
where the j^{th} $\mathbb{Z}/2\mathbb{Z}$ factor is generated by the class of a_{2j} in $X/H[X,X]$, and when i is odd the class of a_i is trivial.

Define
$$a_{\boldsymbol{\sigma}} := a_1^{\sigma_1} a_3^{\sigma_2} \dots a_{2k-1}^{\sigma_k},$$
where $\boldsymbol{\sigma} = (\sigma_1, \dots, \sigma_k) \in \{0,1\}^k$. $H[X,X]$ is an index 2^k subgroup of X, so the rank of $H[X,X]$ is $2^k + 1$.

The following set is a set of coset representatives of $H[X,X]$:
$$\mathcal{C} = \{a_{\boldsymbol{\sigma}} \mid \boldsymbol{\sigma} \in \{0,1\}^k\}.$$
From $\mathcal{C}$, we find a free basis $\mathcal{B}$ of elements of the form $x_{k,\boldsymbol{\sigma}} := a_{\boldsymbol{\sigma}} a_k \overline{a_{\boldsymbol{\sigma}} a_k}^{-1}$.

We point out a few important examples of basis elements. For i odd,
$$a_i^2 = a_i a_i \overline{a_i a_i}^{-1} \in \mathcal{B}.$$
For i even,
$$a_i = 1 a_i \overline{1 a_i}^{-1} \in \mathcal{B}.$$
For i odd and j even,
$$a_i a_j a_i^{-1} = a_i a_j \overline{a_i a_j}^{-1} \in \mathcal{B}.$$
Using the basis $\mathcal{B}$ rewrite the α_i as
$$\begin{aligned}
\alpha_1 &= (a_1^{-2})(a_1 a_2 a_1^{-1})(a_1^2),\\
\alpha_2 &= (a_3^{-2})(a_3 a_2 a_3^{-1})(a_1 a_2^{-1} a_1^{-1})(a_1^2),\\
&\vdots\\
\alpha_{2i-1} &= (a_{2i-1}^{-2})(a_{2i-1} a_{2i} a_{2i-1}^{-1})(a_{2i-1} a_{2i-2}^{-1} a_{2i-1}^{-1})(a_{2i-1}^2),\\
\alpha_{2i} &= (a_{2i+1}^{-2})(a_{2i+1} a_{2i} a_{2i+1}^{-1})(a_{2i-1} a_{2i}^{-1} a_{2i-1}^{-1})(a_{2i-1}^2),\\
&\vdots\\
\alpha_{2k-1} &= (a_{2k-1}^{-2})(a_{2k-1} a_{2k} a_{2k-1}^{-1})(a_{2k-1} a_{2k-2}^{-1} a_{2k-1}^{-1})(a_{2k-1}^2),\\
\alpha_{2k} &= a_{2k}^r (a_{2k-1} a_{2k}^{-1} a_{2k-1}^{-1})(a_{2k-1}^2),
\end{aligned}$$
which can be extended to the free basis $\mathcal{B}'$ of $H[X,X]$
$$\mathcal{B}' = (\mathcal{B} - (\mathcal{B}_1 \cup \mathcal{B}_2 \cup \{a_{2k-1}^2\})) \cup \{\alpha_1, \dots, \alpha_{2k}\},$$

where

$$\mathcal{B}_1 = \{a_1a_2a_1^{-1}, a_3a_4a_3^{-1}, \dots, a_{2k-1}a_{2k}a_{2k-1}^{-1}\}$$

and

$$\mathcal{B}_2 = \{a_3a_2a_3^{-1}, a_5a_4a_5^{-1}, \dots, a_{2k-1}a_{2k-2}a_{2k-1}^{-1}\}.$$

Thus, H is a free factor of $H[X,X]$.

A similar argument shows K is a free factor of $K[X,X]$. Thus, S satisfies the free factor property.

From (4-1) we compute $|X : H[X,X]| = 2^k$, so by Proposition 2.3 J is rationally homologically fibered. Thus, S is an unknotted minimal genus Seifert surface, and J is rationally homologically fibered. It follows from Proposition 2.5 that the commutator subgroup of J is residually torsion-free nilpotent. □

Proof of Corollary 1.10 From the Seifert matrix (4-1), we compute the Alexander polynomial

$$\Delta_J(t) = (t-2)^k(2t-1)^k.$$

It follows from Theorems 1.4 and 1.6 that $\pi_1(M_J)$ is biorderable. □

Appendix A Proofs of lemmas

In this appendix, we present the proofs of Lemmas 3.6, 3.7 and 3.8. Let J be a pretzel knot $P(2p+1, 2q+1, 2r+1)$ with $1 \le q \le r$. Define the Seifert surface S and the groups $X \cong \langle a, b\rangle$, $H \cong \langle \alpha_H, \beta_H\rangle$ and $K \cong \langle \alpha_K, \beta_K\rangle$ as in Section 3.

A.1 Proof of Lemma 3.6

Lemma 3.6 *If J is a $P(2p+1, 3, 2r+1)$ pretzel knot with $p < -2$, then S satisfies the free factor property.*

Proof From (3-1),

$$\alpha_H = b^{-1}ab^{-1}a^{p+1}, \quad \alpha_K = ab^{-1}a^{p+1},$$
$$\beta_H = b^{r+1}a^{-1}b, \qquad \beta_K = b^{r+1}a^{-1}ba^{-1}.$$

The abelian group $X/H[X,X]$ has a presentation matrix

$$\begin{pmatrix} 1 & -r-2 \\ 0 & -N \end{pmatrix},$$

where $N = pr + 2p + 2r + 2 = (p+2)(r+2) - 2$, which is negative since $p \le -2$.

Let $C = -N$. Using $\mathcal{C} = \{1, b, \dots, b^C\}$ as a set of coset representatives, we apply Reidemeister–Schreier to obtain a free basis of $H[X, X]$. Modifying this basis, we get

$$\mathcal{B} = \{x_0, \dots, x_C\},$$

where $x_k := b^k a b^{-r-2-k}$ when $0 \le k \le C-1$ and $x_C := b^C$.

Using the rewriting process, we have that

$$\alpha_H = x_C^{-1} x_{C-1} (x_{C-r-4}^{-1} x_{C-2r-6}^{-1} \cdots x_{C-i(r+2)-2}^{-1} \cdots x_{r+2}^{-1} x_0^{-1})$$

and

$$\beta_H = x_C^{-1} x_{C-1}^{-1} x_C.$$

(Note that $C > r+2$ since $p < -2$, so x_{r+2} is defined.) We can extend $\{\alpha_H, \beta_H\}$ to the set $\{\alpha_H, \beta_H, x_1, \dots, x_{C-2}, x_C\}$, which is a free basis of $H[X, X]$, so H is a free factor of $H[X, X]$.

$X/K[X, X]$ has a presentation matrix

$$\begin{pmatrix} -N & 0 \\ -p-2 & 1 \end{pmatrix}.$$

Let $l = -p-2$ so $C = l(r+2)+2$. Note that l is a positive integer. We obtain a free basis of $K[X, X]$

$$\mathcal{B} = \{x_0, \dots, x_C\},$$

where $x_k := a^k b a^{l-k}$ when $0 \le k \le C-1$ and $x_C := a^C$.

Using the rewriting process,

$$\alpha_K = x_{l+1}^{-1}$$

and

$$\beta_K = x_0 x_C^{-1} x_{l(r+1)+2} x_{lr+2} x_{l(r-1)+2} \cdots x_{2l+2} x_{l+1}.$$

The set $\{\alpha_K, \beta_K, x_1, \dots, x_l, x_{l+2}, \dots, x_C\}$ is a free basis of $K[X, X]$ so K is a free factor of $K[X, X]$. Thus, S satisfies the free factor property. □

A.2 Proof of Lemma 3.7

Lemma 3.7 *Suppose J is $P(-3, 2q+1, 2r+1)$ and one of the following conditions holds:*

1. $q = 2$ *and* $r \ge 6$,
2. $q = 3$ *and* $r \ge 4$,
3. $q > 3$.

Then S satisfies the free factor property.

Proof This lemma is shown by applying the outline from Section 2 to two cases. First, we address the case when $q = 2$ and $r \geq 6$, then we show the lemma is true when $q \geq 3$, $r \geq 4$ and $q \leq r$.

Case $q = 2$ and $r \geq 6$ $X/H[X,X]$ has a presentation matrix

$$\begin{pmatrix} 1 & -3 \\ 0 & N \end{pmatrix},$$

where $N = r - 3$.

$H[X,X]$ has free basis $x_k = b^k a b^{-k-3}$ for $k = 0, \dots, N-1$, and $x_N = b^N$. Under this basis

$$\alpha_H = (b^{-1}a)^2 b^{-1} a^{-1} = x_N^{-1} x_{N-1} x_N x_1 x_0^{-1}$$

and

$$\beta_H = b^{r+1}(ba^{-1})^2 = x_N x_1^{-1} x_N x_{N-1}^{-1} x_N.$$

Since $r \geq 6$, we have $N \geq 3$, so $x_{N-1} \neq x_1$. Thus, the set $\{\alpha_H, \beta_H, x_2, \dots, x_N\}$ is a free basis of $H[X,X]$ so H is a free factor of $H[X,X]$.

$X/K[X,X]$ has a presentation matrix

$$\begin{pmatrix} 1 & -2 \\ 0 & N \end{pmatrix},$$

where $N = r - 3$.

$K[X,X]$ has free basis $x_k = b^k a b^{-k-2}$ for $k = 0, \dots, N-1$, and $x_N = b^N$. Under this basis

$$\alpha_K = (ab^{-1})^2 a^{-1} = x_0 x_1 x_0^{-1}$$

and

$$\beta_K = b^{r+1} a^{-1} (ba^{-1})^2 = x_N x_2^{-1} x_1^{-1} x_0^{-1}.$$

The set $\{\alpha_K, \beta_K, x_0, x_3, \dots, x_N\}$ is a free basis of $K[X,X]$ so K is a free factor of $K[X,X]$.

Case $q \geq 3$ and $r \geq 4$ $X/H[X,X]$ has a presentation matrix

$$\begin{pmatrix} 1 & -r \\ 0 & N \end{pmatrix},$$

where $N = qr - q - r - 1 = (q-1)(r-1) - 2$. Note that since $q \geq 3$ and $r \geq 4$, $N > r - 2 > 1$.

We then obtain a free basis $x_k = b^k a b^{-r-k}$ for $k = 0, \dots, N-1$ and $x_N = b^N$. Under this basis

$$\alpha_H = (b^{-1}a)^{q+1}a^{-3} = x_N^{-1} x_{N-1} x_N x_{r-2} x_{2r-3} \cdots x_{N-r+2} x_N x_1 x_0^{-1}$$

and

$$\beta_H = b^{r+1(a^{-1}b)^q} = x_1^{-1} x_N^{-1} x_{N-r+2}^{-1} x_{N-2r+3}^{-1} \cdots x_{r-2} x_{N-1}.$$

Since $N > r-2 > 1$, the set $\{\alpha_H, \beta_H, x_2, \dots, x_N\}$ is a free basis of $H[X,X]$ so H is a free factor of $H[X,X]$.

For K, we begin by substituting $a = a_* b_*$ and $b = b_*$ so that

$$\alpha_K = a_*^q b_*^{-1} a_*^{-1} \quad \text{and} \quad \beta_K = b_*^r a_*^{-q-1}.$$

$X/K[X,X]$ has a presentation matrix

$$\begin{pmatrix} N & 0 \\ 1-q & 1 \end{pmatrix},$$

where $N = qr - q - r - 1$.

Under the basis $x_k = a_* b_* a_*^{1-q-k}$ for $k = 0, \dots, N-1$ and $x_N = a_*^N$,

$$\alpha_K = x_1^{-1}$$

and

$$\beta_K = x_0 x_{q-1} \cdots x_{(q-1)(r-1)} x_N.$$

Similarly to H, K is a free factor of $K[X,X]$. Therefore, S satisfies the free factor property. □

A.3 Proof of Lemma 3.8

Lemma 3.8 *Suppose J is $P(-5, 2q+1, 2r+1)$ and one of the following conditions holds:*

(1) $q = 3$ and $r \geq 13$,

(2) $q = 4$ and $r \geq 9$,

(3) $q = 5$ and $r \geq 7$,

(4) $q > 5$.

Then S satisfies the free factor property.

This lemma is shown by applying the outline from Section 2 to several cases.

Lemma A.1 *If J is $P(-5,7,2r+1)$ with $r \geq 13$, then S satisfies the free factor property.*

Proof In this case, $q=3$ and $N=r-8$. $X/H[X,X]$ has a presentation matrix

$$\begin{pmatrix} 1 & -4 \\ 0 & N \end{pmatrix}.$$

We use the free basis, $x_k = b^k a b^{-4-k}$ for $k=0,\dots,N-1$ and $x_N = b^N$.

When $r=13$,

$$\alpha_H = x_5^{-1} x_4 x_5 x_2 x_5 x_0 x_5^{-1} x_4^{-1} x_0^{-1}$$

and

$$\beta_H = x_5^2 x_0^{-1} x_5^{-1} x_2^{-1} x_5^{-1} x_4^{-1} x_5,$$

so

$$\beta_H \alpha_H = x_5 x_4^{-1} x_0^{-1}.$$

The set $\{\alpha_H, \beta_H, x_1, x_3, x_4, x_5\}$ is a free basis of $H[X,X]$ so H is a free factor of $H[X,X]$.

When $r \geq 14$,

$$\alpha_H = x_N^{-1} x_{N-1} x_N x_2 x_5 x_4^{-1} x_0^{-1}$$

and

$$\beta_H = x_N x_5^{-1} x_2^{-1} x_N^{-1} x_{N-1}^{-1} x_N.$$

The set $\{\alpha_H, \beta_H, x_1, x_3, \dots, x_N\}$ is a free basis of $H[X,X]$ so H is a free factor of $H[X,X]$.

$X/K[X,X]$ has a presentation matrix

$$\begin{pmatrix} 1 & -3 \\ 0 & N \end{pmatrix}.$$

We use the free basis $x_k = b^k a b^{-3-k}$ for $k=0,\dots,N-1$ and $x_N = b^N$.

Using this basis,

$$\alpha_K = x_0 x_2 x_4 x_3^{-1} x_0^{-1}.$$

When $r=13$ or $r=14$,

$$\beta_K = x_N^2 x_{6-N}^{-1} x_N^{-1} x_4^{-1} x_2^{-1} x_0^{-1},$$

and, when $r \geq 15$,

$$\beta_K = x_N x_6^{-1} x_4^{-1} x_2^{-1} x_0^{-1}.$$

In both cases, the set $\{\alpha_K, \beta_K, x_0, x_1, x_4, \dots, x_N\}$ is a free basis of $K[X,X]$ so K is a free factor of $K[X,X]$. □

Lemma A.2 *If J is $P(-5,9,2r+1)$ with $r \geq 9$, then S satisfies the free factor property.*

Proof In this case, $q=4$ and $N=2r-10$. $X/H[X,X]$ has a presentation matrix

$$\begin{pmatrix} N & 0 \\ -2 & 1 \end{pmatrix}$$

after the substitutions $a=b_*^2a_*$ and $b=b_*$. We use the free basis $x_k=a_*^kb_*a_*^{-2-k}$ for $k=0,\dots,N-1$ and $x_N=a_*^N$.

When $r=9$,

$$\alpha_H = x_0x_3x_6x_8x_1x_2^{-1}x_8^{-1}x_7^{-1}x_5^{-1}x_2^{-1}x_0^{-1}$$

and

$$\beta_H = (x_0x_2x_4x_6x_8)^2x_0x_2x_1^{-1}x_8^{-1}x_6^{-1}x_3^{-1}x_0^{-1}.$$

The set $\{\alpha_H,\beta_H,x_0,x_1,x_2,x_4,x_6,x_7,x_8\}$ is a free basis of $H[X,X]$ so H is a free factor of $H[X,X]$.

When $r=10$,

$$\alpha_H = x_0x_3x_6x_9x_{10}x_0^{-1}x_{10}^{-1}x_7^{-1}x_5^{-1}x_2^{-1}x_0^{-1}$$

and

$$\beta_H = (x_0x_2x_4x_6x_8x_{10})^2x_0x_{10}^{-1}x_9^{-1}x_6^{-1}x_3^{-1}x_0^{-1}.$$

When $r \geq 11$,

$$\alpha_H = x_0x_3x_6x_9x_{10}^{-1}x_7^{-1}x_5^{-1}x_2^{-1}x_0^{-1}$$

and

$$\beta_H = x_0x_2\cdots x_{N-2}x_Nx_0x_2x_4x_6x_8x_{10}x_9^{-1}x_6^{-1}x_3^{-1}x_0^{-1}.$$

In both cases, the set $\{\alpha_H,\beta_H,x_0,\dots,x_6,x_8,x_{10},\dots,x_N\}$ is a free basis of $H[X,X]$.

$X/K[X,X]$ has a presentation matrix

$$\begin{pmatrix} 1 & 3-r \\ 0 & N \end{pmatrix}.$$

We use the free basis $x_k=b^kab^{3-r-k}$ for $k=0,\dots,N-1$ and $x_N=b^N$. Using this basis,

$$\alpha_K = x_0x_{r-4}x_Nx_2x_{r-2}x_{r-3}^{-1}x_0^{-1}$$

and

$$\beta_K = x_4^{-1}x_N^{-1}x_{r-2}^{-1}x_2^{-1}x_N^{-1}x_{r-4}^{-1}x_0^{-1}.$$

Since $r \geq 9$,

$$N = r - 8 + r - 2 > r - 2 > r - 3 > r - 4 > 0,$$

so the generators x_{r-2}, x_{r-3} and x_{r-4} are valid generators.

The set $\{\alpha_K, \beta_K, x_1, \ldots, x_{r-4}, x_{r-2}, \ldots, x_N\}$ is a free basis of $K[X, X]$ so K is a free factor of $K[X, X]$. □

Lemma A.3 *If J is $P(-5, 11, 2r+1)$ with $r \geq 7$, then S satisfies the free factor property.*

Proof In this case, $q = 5$ and $N = 3r - 12$. $X/H[X, X]$ has a presentation matrix

$$\begin{pmatrix} 1 & r-6 \\ 0 & N \end{pmatrix}.$$

We use the free basis $x_k = b^k a b^{r-6-k}$ for $k = 0, \ldots, N-1$ and $x_N = b^N$.

Using this basis,

$$\alpha_H = x_{2r-6}^{-1} x_N x_0^{-1}.$$

When $r = 7$,

$$\beta_H = x_9 x_0^{-1} x_2^{-1} x_4^{-1} x_6^{-1} x_8^{-1} x_9,$$

and, when $r \geq 8$,

$$\beta_H = x_{2r-5}^{-1} x_N x_2^{-1} x_{r-3}^{-1} x_{2r-8}^{-1} x_{3r-13}^{-1} x_N.$$

Note that, when $r \geq 8$,

$$N > 3r - 13 > 0,$$
$$N = r - 7 + 2r - 5 > 2r - 5 > 2r - 6 > 2r - 8 > 0,$$

and

$$N = 2r - 9 + r - 3 > r - 3 > 0,$$

so the generators x_{3r-13}, x_{2r-5}, x_{2r-6}, x_{2r-8} and x_{r-3} are valid generators.

In both cases, the set $\{\alpha_H, \beta_H, x_1, x_3, \ldots, x_N\}$ is a free basis of $H[X, X]$, so H is a free factor of $H[X, X]$.

After making the substitutions $a = b_*^2 a_*$ and $b = b_*$, $X/K[X, X]$ has a presentation matrix

$$\begin{pmatrix} N & 0 \\ 3 & 1 \end{pmatrix}.$$

We use the free basis $x_k = a_*^k b_* a_*^{3-k}$ for $k = 0, \dots, N-1$ and $x_N = b_*^N$. Using this basis,

$$\alpha_K = x_1 x_N^{-1} x_{N-1} x_{N-3} x_{N-5} x_{N-4}^{-1} x_{N-2}^{-1} x_N x_1^{-1}$$

and

$$\beta_K = x_0 x_N^{-1} x_{N-3} x_{N-6} \cdots x_3 x_0 x_N^{-1} x_{N-3} x_{N-6} x_{N-7}^{-1} x_{N-5}^{-1} x_{N-3}^{-1} x_{N-1}^{-1} x_N x_1^{-1}.$$

Since $r \geq 7$, we have $N \geq 9$, so all the generators used are valid generators.

The set $\{\alpha_K, \beta_K, x_0, \dots, x_{N-8}, x_{N-6}, x_{N-5}, x_{N-4}, x_{N-3}, x_{N-1}, x_N\}$ is a free basis of $K[X,X]$, so K is a free factor of $K[X,X]$. □

Lemma A.4 *If J is $P(-5,13,13)$ or $P(-5,13,15)$, then S satisfies the free factor property.*

Proof If J is $P(-5,13,13)$, then $p=-3$ and $q=r=6$. $X/H[X,X]$ has presentation matrix

$$\begin{pmatrix} 10 & 0 \\ -2 & 1 \end{pmatrix}.$$

We use the free basis $x_k := a^k b a^{-2-k}$ for $k = 0, \dots, 9$ and $x_{10} := a^{10}$.

Using this basis,

$$\alpha_H = x_{10}^{-1} x_8^{-1} x_7^{-1} x_6^{-1} x_5^{-1} x_4^{-1} x_3^{-1} x_2^{-1}$$

and

$$\beta_H = x_0 x_2 x_4 x_6 x_8 x_{10} x_0 x_2 x_3 x_4 x_5 x_6 x_7 x_8 x_{10},$$

so

$$\beta_H \alpha_H = x_0 x_2 x_4 x_6 x_8 x_{10} x_0.$$

The set $\{\alpha_H, \beta_H, x_0, x_1, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}\}$ is a free basis of $H[X,X]$, so H is a free factor of $H[X,X]$.

$X/K[X,X]$ has presentation matrix

$$\begin{pmatrix} 1 & 1 \\ 0 & 10 \end{pmatrix}.$$

We use the free basis $x_k := a^k b a^{1-k}$ for $k = 0, \dots, 9$ and $x_{10} := a^{10}$.

Using this basis,

$$\alpha_K = x_0 x_{10}^{-1} x_8 x_6 x_4 x_2 x_0 x_{10}^{-1} x_9^{-1} x_{10} x_0^{-1}$$

and

$$\beta_K = x_8^{-1} x_{10} x_0^{-1} x_2^{-1} x_4^{-1} x_6^{-1} x_8^{-1} x_{10} x_0^{-1}.$$

The set $\{\alpha_K, \beta_K, x_0, x_1, x_3, x_4, x_5, x_6, x_7, x_8, x_{10}\}$ is a free basis of $K[X,X]$ so K is a free factor of $K[X,X]$.

If J is $P(-5,13,15)$, then $p=-3$, $q=6$ and $r=7$. After making the substitutions $a = b_*^2 a_*$ and $b = b_*$, $X/H[X,X]$ has presentation matrix

$$\begin{pmatrix} 14 & 0 \\ 4 & 1 \end{pmatrix}.$$

We use the free basis $x_k := a_*^k b_* a_*^{4-k}$ for $k = 0, \dots, 13$ and $x_{14} := a_*^{14}$.

Using this basis,

$$\alpha_H = x_0 x_{14}^{-1} x_{11} x_8 x_5 x_2 x_{14}^{-1} x_{13} x_{14} x_0^{-1} x_3^{-1} x_7^{-1} x_{10}^{-1} x_{14} x_0^{-1}$$

and

$$\beta_H = x_0 x_{14}^{-1} x_{10} x_6 x_2 x_{14}^{-1} x_{12} x_8 x_4 x_0 x_{14}^{-1} x_{13}^{-1} x_{14} x_2^{-1} x_5^{-1} x_8^{-1} x_{11}^{-1} x_{14} x_0^{-1}.$$

The set $\{\alpha_H, \beta_H, x_0, \dots, x_5, x_8, \dots, x_{14}\}$ is a free basis of $H[X,X]$ so H is a free factor of $H[X,X]$.

$X/K[X,X]$ has presentation matrix

$$\begin{pmatrix} 1 & 2 \\ 0 & 14 \end{pmatrix}.$$

We use the free basis $x_k := b^k a b^{2-k}$ for $k = 0, \dots, 13$ and $x_{14} := b^{14}$.

Using this basis,

$$\alpha_K = x_0 x_{14}^{-1} x_{11} x_8 x_5 x_2 x_{14}^{-1} x_{13} x_{10} x_{14} x_0^{-1}$$

and

$$\beta_K = x_{10}^{-1} x_{13}^{-1} x_{14} x_2^{-1} x_5^{-1} x_8^{-1} x_{11}^{-1} x_{14} x_0^{-1},$$

so

$$\beta_H \alpha_H = x_{14} x_0^{-1}.$$

The set $\{\alpha_K, \beta_K, x_1, x_3, \dots, x_{14}\}$ is a free basis of $K[X,X]$, so K is a free factor of $K[X,X]$. □

Lemma A.5 *If J is $P(-5, 2q+1, 2r+1)$ with q even, $q \geq 6$ and $r \geq 8$, then H is a free factor of $H[X, X]$.*

Proof Let c be the integer such that $q = 2c$. $X/H[X, X]$ has a presentation matrix

$$\begin{pmatrix} 2 & -r \\ 0 & w \end{pmatrix},$$

where $w = cr - 2c - r - 1$ and $N = 2w$.

We have the set of coset representatives

$$\mathcal{C} = \{1, b, b^2, \dots, b^{w-1}, a, ab, ab^2, \dots, ab^{w-1}\}.$$

We apply Reidemeister–Schreier to find a free basis of $H[X, X]$. In the following computations we assume that the coset representative $\overline{a^2}$ of a^2 is b^r. For this to be correct, it must be true that $r < w$, which we verify here.

Since $q \geq 6$, we have $c \geq 2$, and $r \geq 8$, so

$$w = cr - 2c - r - 1 = (c-3)(r-2) + (r-7) + r > r.$$

We apply Reidemeister–Schreier to find $x_{c,x} = cx(\overline{cx})^{-1}$ for each $c \in \mathcal{C}$ and $x \in \{a, b\}$:

$$\begin{aligned}
x_{b^i,a} &= b^i a(\overline{b^i a})^{-1} &&= \begin{cases} b^i a b^{-i} a^{-1} & \text{if } 0 < i \leq w-1, \\ 1 & \text{if } i = 0, \end{cases} \\
x_{b^i,b} &= b^{i+1}(\overline{b^{i+1}})^{-1} &&= \begin{cases} 1 & \text{if } 0 \leq i < w-1, \\ b^w & \text{if } i = w-1, \end{cases} \\
x_{ab^i,a} &= ab^i a(\overline{ab^i a})^{-1} &&= \begin{cases} ab^i ab^{-i-r} & \text{if } 0 \leq i < w-r, \\ ab^i ab^{w-i-r} & \text{if } w-r \leq i \leq w-1, \end{cases} \\
x_{ab^i,b} &= ab^{i+1}(\overline{ab^{i+1}})^{-1} &&= \begin{cases} 1 & \text{if } 0 \leq i < w-1, \\ ab^w a^{-1} & \text{if } i = w-1. \end{cases}
\end{aligned}$$

The nontrivial elements $x_{c,x}$ form a basis $\{x_1, \dots, x_w, y_0, \dots, y_w\}$, where

$$x_i = \begin{cases} b^i ab^{-i} a^{-1} & \text{if } 1 \leq i \leq w-1, \\ b^w & \text{if } i = w, \end{cases}$$

and

$$y_i = \begin{cases} ab^i ab^{-i-r} & \text{if } 0 \leq i < w-r, \\ ab^i ab^{w-i-r} & \text{if } w-r \leq i < w, \\ ab^w a^{-1} & \text{if } i = w. \end{cases}$$

Using this basis,

$$\beta_H \alpha_H = y_0^{-1}$$

and

$$\beta_H = y_1^{-1} x_2^{-1} \prod_{i=1}^{c-2} (y_{\delta(i)+1}^{-1} x_{\delta(i)+2}^{-1}) y_{w-2}^{-1} x_{w-1}^{-1} x_w,$$

where

$$\delta(i) = w - i(r-2).$$

We claim that $\delta(i) \neq 0$ for all i. Since $w = (c-1)(r-2) - 3$

$$\delta(i) = w - i(r-2) = (r-2)(c-i-1) - 3,$$

so if $\delta(i) = 0$ then $(r-2)(c-i-1) = 3$. However, since $r \geq 8$, we have that $r-2$ does not divide 3.

Thus, y_1 only appears once in β_H so the set $\{\beta_H \alpha_H, \beta_H, x_1, \ldots, x_w, y_2, \ldots, y_w\}$ is a free basis of $H[X, X]$. Since $\{\beta_H \alpha_H, \beta_H\}$ is a free basis of H, H is a free factor of $H[X, X]$. □

Lemma A.6 *If J is $P(-5, 2q+1, 2r+1)$ with q odd and $q \geq 7$, then H is a free factor of $H[X, X]$.*

Proof Let c be the integer such that $q = 2c+1$. $X/H[X, X]$ has a presentation matrix

$$\begin{pmatrix} 1 & v \\ 0 & N \end{pmatrix},$$

where $v = cr - 2c - r - 2$ and $N = 2cr - 4c - r - 4 = 2v + r$.

We use the free basis $x_k = b^k a b^{v-k}$ for $k = 0, \ldots, N-1$ and $x_N = b^N$. Using this basis,

$$\beta_H \alpha_H = x_{v+r}^{-1} x_N x_0^{-1}$$

and

$$\beta_H = x_{v+r+1}^{-1} \prod_{i=0}^{2c-1} y_i,$$

where

$$y_i = \begin{cases} x_{\epsilon(i)}^{-1} & \text{if } \epsilon(i) < N - v - 1, \\ x_{\epsilon(i)}^{-1} x_N & \text{if } \epsilon(i) \geq N - v - 1, \end{cases}$$

and

$$\epsilon(i) = 2 + i(v+1) \bmod N.$$

Since $q \geq 7$, we have $c \geq 3$, and, since $r \geq 7$,

$$v = cr - 2c - r - 2 = (c-2)(r-2) + r - 6 > 1.$$

This means that

$$N = 2v + r > v + r + 1 > v + r > 0,$$

so x_{v+r} and x_{v+r+1} are valid generators.

We claim that $\epsilon(i)$ is distinct for each $i = 0, \dots, 2c-1$. Suppose that $\epsilon(i) = \epsilon(j)$ for some i and j. Then $(j-i)(v+1)$ is a multiple of N. In particular, N divides $(j-i)\gcd(N, v+1)$. Applying the Euclidean algorithm to N and $v+1$, we have

$$N = 2(v+1) + r - 2$$

and

$$v + 1 = (c-1)(r-2) - 3,$$

so

$$\gcd(N, v+1) = \gcd(r-2, 3) \leq 3.$$

The maximum value of $j - i$ is $2c - 1$. It follows that

$$N \leq 3(2c-1).$$

However, since $c \geq 3$ and $r \geq 7$,

$$N = 2cr - 4c - r - 4 = (2c-1)(r-4) + 4c - 8 \geq 3(2c-1) + 4 > 3(2c-1),$$

which is a contradiction.

Thus $x_{\epsilon(0)} = x_2$ only appears once in β_H so the set $\{\beta_H\alpha_H, \beta_H, x_1, x_3, \dots, x_N\}$ is a free basis of $H[X,X]$. Therefore, H is a free factor of $H[X,X]$. $\square$

Lemma A.7 *If J is $P(-5, 2q+1, 2r+1)$ with $q \equiv 0 \bmod 3$, $q \geq 6$ and $r \geq 8$, then K is a free factor of $K[X,X]$.*

Proof Let c be the integer such that $q = 3c$. $X/K[X,X]$ has a presentation matrix

$$\begin{pmatrix} 1 & v \\ 0 & N \end{pmatrix},$$

where $v = cr - 2c - r - 1$ and $N = 3cr - 6c - 2r - 2 = 3v + r + 1$.

We use the free basis $x_k = b^k a b^{v-k}$ for $k = 0, \dots, N-1$ and $x_N = b^N$. Using this basis,

$$\beta_K\alpha_K = x_{v+r+1}^{-1} x_{2v+r+1}^{-1} x_N x_0^{-1}$$

and

$$\beta_K = x_{v+r+1}^{-1} x_{2v+r+2}^{-1} \prod_{i=0}^{3c-2} y_i,$$

where

$$y_i = \begin{cases} x_{\zeta(i)}^{-1} & \text{if } \zeta(i) < N-v-1, \\ x_{\zeta(i)}^{-1}x_N & \text{if } \zeta(i) \geq N-v-1, \end{cases}$$

and

$$\zeta(i) = 1 + i(v+1) \bmod N.$$

Since $q \geq 6$, we have $c \geq 2$, and, since $r \geq 8$,

$$v = cr - 2c - r - 1 = (c-1)(r-5) + 3c - 6 > 1.$$

This means that

$$N = 3v + r + 1 > 2v + r + 2 > 2v + r + 1 > v + r + 1 > 0,$$

so x_{v+r+1}, x_{2v+r+1} and x_{2v+r+2} are valid generators.

Suppose that $\zeta(i) = \zeta(j)$ for some i and j. Then N divides $(j-i)\gcd(N, v+1)$. Applying the Euclidean algorithm to N and $v+1$, we have

$$N = 3(v+1) + r - 2$$

and

$$v + 1 = (c-1)(r-2) - 2,$$

so

$$N \leq 2(3c-2).$$

However, since $c \geq 2$ and $r \geq 8$,

$$N = 3cr - 6c - 2r - 2 = (3c-2)(r-4) + 6c - 10 > 2(3c-2),$$

so $\zeta(i)$ is distinct for each $i = 0, \dots, 3c-2$. Thus $x_{\zeta(0)} = x_1$ only appears once in β_K so the set $\{\beta_K\alpha_K, \beta_K, x_2, \dots, x_N\}$ is a free basis of $K[X,X]$. Therefore, K is a free factor of $K[X,X]$. □

Lemma A.8 *If J is $P(-5, 2q+1, 2r+1)$ with $q \equiv 1 \bmod 3$ and $q \geq 7$, then K is a free factor of $K[X,X]$.*

Proof Let c be the integer such that $q = 3c+1$. $X/K[X,X]$ has a presentation matrix

$$\begin{pmatrix} 1 & -v \\ 0 & N \end{pmatrix},$$

where $v = cr - 2c - 1$ and $N = 3cr - 6c - r - 4 = 3v - r - 1$.

We use the free basis $x_k = b^k a b^{-v-k}$ for $k = 0, \dots, N-1$ and $x_N = b^N$. Using this basis,

$$\beta_K \alpha_K = x_N^{-1} x_{2v}^{-1} x_v^{-1} x_0^{-1}$$

and

$$\beta_K = x_N^{-1} x_{2v}^{-1} x_{v+1}^{-1} \prod_{i=0}^{3c-1} y_i,$$

where

$$y_i = \begin{cases} x_{\eta(i)}^{-1} & \text{if } \eta(i) < N - v + 1, \\ x_{x_N^{-1}\eta(i)}^{-1} & \text{if } \eta(i) \geq N - v + 1, \end{cases}$$

and

$$\eta(i) = 2 - i(v-1) \bmod N.$$

Since $q \geq 7$, we have $c \geq 2$, and, since $r \geq 7$,

$$v = cr - 2c - 1 = (c-1)(r-8) + 4c - 8 + r + 1 > r + 1.$$

This means that

$$N = 3v - r - 1 > 2v > v + 1 > v > 0,$$

so x_v, x_{v+1} and x_{2v} are valid generators.

Suppose that $\eta(i) = \eta(j)$ for some i and j. Then N divides $(j-i)\gcd(N, v-1)$. Applying the Euclidean algorithm to N and $v-1$, we have

$$N = 3(v-1) - (r-2)$$

and

$$v - 1 = c(r-2) - 2,$$

so

$$N \leq 2(3c-1).$$

However, since $c \geq 2$ and $r \geq 7$,

$$N = 3cr - 6c - r - 4 = (3c-1)(r-4) + 6c - 8 > 2(3c-1),$$

so $\eta(i)$ is distinct for each $i = 0, \dots, 3c-2$. Thus $x_{\eta(0)} = x_2$ only appears once in β_K so the set $\{\beta_K\alpha_K, \beta_K, x_1, x_3, \dots, x_N\}$ is a free basis of $K[X, X]$. Therefore, K is a free factor of $K[X, X]$. □

Lemma A.9 *If J is $P(-5, 2q+1, 2r+1)$ with $q \equiv 2 \bmod 3$ and $q \geq 8$, then K is a free factor of $K[X, X]$.*

Proof Let c be the integer such that $q = 3c+2$. $X/K[X,X]$ has a presentation matrix
$$\begin{pmatrix} 3 & -(r+1) \\ 0 & w \end{pmatrix},$$
where $w = cr - 2c - 2$ and $N = 3w$.

We have the set of coset representatives
$$\mathcal{C} = \{1, b, b^2, \dots, b^{w-1}, a, ab, \dots, ab^{w-1}, a^2, a^2b, \dots, a^2b^{w-1}\}.$$

We apply Reidemeister–Schreier to find a free basis of $K[X,X]$.

Since $q \geq 6$ we have $c \geq 2$, and $r \geq 8$ so $r+1 < w$:
$$w = cr - 2c - 2 = (c-2)(r-2) + (r-7) + r + 1 > r + 1.$$

Thus, the coset representative, $\overline{a^3}$ is b^{r+1}.

We apply Reidemeister–Schreier to find a basis $\{x_1, \dots, x_w, y_1, \dots, y_w, z_0, \dots, z_w\}$, where
$$x_i = \begin{cases} b^i a b^{-i} a^{-1} & \text{if } 1 \leq i \leq w-1, \\ b^w & \text{if } i = w, \end{cases}$$
$$y_i = \begin{cases} ab^i ab^{-i} a^{-2} & \text{if } 1 \leq i \leq w-1, \\ ab^w a^{-1} & \text{if } i = w, \end{cases}$$
$$z_i = \begin{cases} a^2 b^i ab^{-i-r-1} & \text{if } 0 \leq i < w-r-1, \\ a^2 b^i ab^{w-i-r-1} & \text{if } w-r-1 \leq i < w, \\ a^2 b^w a^{-2} & \text{if } i = w. \end{cases}$$

Using this basis,
$$\beta_K \alpha_K = z_0^{-1}$$
and
$$\beta_K = z_0^{-1} y_1^{-1} x_2^{-1} \prod_{i=1}^{c-1} (z_{\delta(i)}^{-1} y_{\delta(i)+1}^{-1} x_{\delta(i)+2}^{-1}) z_{w-2}^{-1} y_{w-1}^{-1} y_w,$$
where
$$\delta(i) = w - i(r-2).$$

Since $w = c(r-2) - 2$,
$$\delta(i) - 1 = w - i(r-2) - 1 = (r-2)(c-i) - 3,$$
so if $\delta(i) = 1$ then $(r-2)(c-i) = 3$. However, $r-2$ does not divide 3 since $r \geq 7$, so $\delta(i)$ is never 1, so y_1 only appears once in β_K.

Therefore, the set $\{\beta_K\alpha_K, \beta_K, x_1, \dots, x_w, y_2, \dots, y_w, z_1, \dots, z_w\}$ is a free basis of $K[X,X]$. Since $\{\beta_K\alpha_K, \beta_K\}$ is a free basis of K, K is a free factor of $K[X,X]$. $\square$

Appendix B Chart of results

Table 1 summarizes the results we've found for the pretzel knots $P(-3, Q, R)$ and $P(-5, Q, R)$ where $Q = 2q + 1$ and $R = 2r + 1$. The shapes around the cells in each chart indicate whether or not the knot's standard Seifert surface S satisfies the

values of Q \ values of R	3	5	7	9	11	13	15
3	−2	−2	−2	−2	−2	−2	−2
5		$\boxed{-1}$	0	$\boxed{1}$	$\boxed{2}$	3	4
7			$\boxed{2}$	4	6	8	10
9				7	10	13	16
11					14	18	22

values of Q \ values of R	3	5	7	9	11	13	15	17	19	21	23	25	27
3	−5	−6	−7	−8	−9	−10	−11	−12	−13	−14	−15	−16	−17
5		−6	−6	−6	−6	−6	−6	−6	−6	−6	−6	−6	−6
7			−5	−4	$\boxed{-3}$	$\boxed{-2}$	$\boxed{-1}$	0	$\boxed{1}$	$\boxed{2}$	$\boxed{3}$	$\boxed{4}$	5
9				$\boxed{-2}$	0	$\boxed{2}$	$\boxed{4}$	$\boxed{6}$	8	10	12	14	16
11					$\boxed{3}$	$\boxed{6}$	9	12	15	18	21	24	27
13						10	14	18	22	26	30	34	38
15							19	24	29	34	39	44	49

Table 1: The results for some $P(-3, Q, R)$ (top) and $P(-5, Q, R)$ (bottom) pretzel knots where $Q = 2q + 1$ and $R = 2r + 1$. The integer in each cell is the value of N. Each cell is in a circle if the knot's standard Seifert surface satisfies the free factor property, and in a square if the knot's standard Seifert surface does not satisfy the free factor property.

free factor property. Cells of knots with trivial Alexander polynomial have no shapes. The integer in each cell is the value of $N = \det(S_+) = \det(S_-)$, which is also the leading coefficient of the Alexander polynomial. If a pretzel knot's cell is contained in a circle and N is a prime power, then the knot group has residually torsion-free nilpotent commutator subgroup. If in addition $N < 0$, then the knot group is biorderable.

References

[1] **G Baumslag**, *Groups with the same lower central sequence as a relatively free group, I: The groups*, Trans. Amer. Math. Soc. 129 (1967) 308–321 MR Zbl

[2] **G Baumslag**, *Groups with the same lower central sequence as a relatively free group, II: Properties*, Trans. Amer. Math. Soc. 142 (1969) 507–538 MR Zbl

[3] **S Boyer**, **C M Gordon**, **L Watson**, *On L-spaces and left-orderable fundamental groups*, Math. Ann. 356 (2013) 1213–1245 MR Zbl

[4] **S Boyer**, **D Rolfsen**, **B Wiest**, *Orderable 3–manifold groups*, Ann. Inst. Fourier (Grenoble) 55 (2005) 243–288 MR Zbl

[5] **E M Brown**, **R H Crowell**, *The augmentation subgroup of a link*, J. Math. Mech. 15 (1966) 1065–1074 MR Zbl

[6] **G Burde**, **H Zieschang**, *Neuwirthsche Knoten und Flächenabbildungen*, Abh. Math. Sem. Univ. Hamburg 31 (1967) 239–246 MR Zbl

[7] **I M Chiswell**, **A M W Glass**, **J S Wilson**, *Residual nilpotence and ordering in one-relator groups and knot groups*, Math. Proc. Cambridge Philos. Soc. 158 (2015) 275–288 MR Zbl

[8] **A Clay**, **C Desmarais**, **P Naylor**, *Testing bi-orderability of knot groups*, Canad. Math. Bull. 59 (2016) 472–482 MR Zbl

[9] **R H Crowell**, **H F Trotter**, *A class of pretzel knots*, Duke Math. J. 30 (1963) 373–377 MR Zbl

[10] **D Eisenbud**, **U Hirsch**, **W Neumann**, *Transverse foliations of Seifert bundles and self-homeomorphism of the circle*, Comment. Math. Helv. 56 (1981) 638–660 MR Zbl

[11] **D Gabai**, *The Murasugi sum is a natural geometric operation, II*, from "Combinatorial methods in topology and algebraic geometry" (J R Harper, R Mandelbaum, editors), Contemp. Math. 44, Amer. Math. Soc., Providence, RI (1985) 93–100 MR Zbl

[12] **D Gabai**, *Genera of the arborescent links*, Mem. Amer. Math. Soc. 339, Amer. Math. Soc., Providence, RI (1986) MR Zbl

[13] **H Goda**, **M Ishiwata**, *A classification of Seifert surfaces for some pretzel links*, Kobe J. Math. 23 (2006) 11–28 MR Zbl

[14] **F González-Acuña**, *Dehn's construction on knots*, Bol. Soc. Mat. Mexicana 15 (1970) 58–79 MR Zbl

[15] **C M Gordon**, *Ribbon concordance of knots in the 3–sphere*, Math. Ann. 257 (1981) 157–170 MR Zbl

[16] **C M Gordon**, *Riley's conjecture on* $\mathrm{SL}(2,\mathbb{R})$ *representations of 2–bridge knots*, J. Knot Theory Ramifications 26 (2017) art. id. 1740003 MR Zbl

[17] **A Issa**, **H Turner**, *Links all of whose cyclic branched covers are L-spaces*, preprint (2020) arXiv 2008.06127

[18] **T Ito**, *Alexander polynomial obstruction of bi-orderability for rationally homologically fibered knot groups*, New York J. Math. 23 (2017) 497–503 MR Zbl

[19] **M Jankins**, **W D Neumann**, *Rotation numbers of products of circle homeomorphisms*, Math. Ann. 271 (1985) 381–400 MR Zbl

[20] **J Johnson**, *Residual torsion-free nilpotence, bi-orderability and two-bridge links*, Canad. J. Math. (online publication January 2023)

[21] **L H Kauffman**, *State models and the Jones polynomial*, Topology 26 (1987) 395–407 MR Zbl

[22] **L H Kauffman**, *New invariants in the theory of knots*, from "On the geometry of differentiable manifolds", Astérisque 163-164, Soc. Math. France, Paris (1988) 137–219 MR Zbl

[23] **A Kawauchi**, *A survey of knot theory*, Birkhäuser, Basel (1996) MR Zbl

[24] **W B R Lickorish**, **M B Thistlethwaite**, *Some links with nontrivial polynomials and their crossing-numbers*, Comment. Math. Helv. 63 (1988) 527–539 MR Zbl

[25] **P A Linnell**, **A H Rhemtulla**, **D P O Rolfsen**, *Invariant group orderings and Galois conjugates*, J. Algebra 319 (2008) 4891–4898 MR Zbl

[26] **P Lisca**, **A I Stipsicz**, *On the existence of tight contact structures on Seifert fibered 3–manifolds*, Duke Math. J. 148 (2009) 175–209 MR Zbl

[27] **W Magnus**, *Beziehungen zwischen Gruppen und Idealen in einem speziellen Ring*, Math. Ann. 111 (1935) 259–280 MR Zbl

[28] **W Magnus**, **A Karrass**, **D Solitar**, *Combinatorial group theory: presentations of groups in terms of generators and relations*, Interscience, New York (1966) MR Zbl

[29] **E J Mayland, Jr**, *The residual finiteness of the classical knot groups*, Canadian J. Math. 27 (1975) 1092–1099 MR Zbl

[30] **E J Mayland, Jr**, **K Murasugi**, *On a structural property of the groups of alternating links*, Canadian J. Math. 28 (1976) 568–588 MR Zbl

[31] **J M Montesinos**, *Surgery on links and double branched covers of* S^3, from "Knots, groups, and 3–manifolds (papers dedicated to the memory of R H Fox)" (L P Neuwirth, editor), Ann. of Math. Studies 84, Princeton Univ. Press (1975) 227–259 MR Zbl

[32] **K Murasugi**, *Jones polynomials and classical conjectures in knot theory*, Topology 26 (1987) 187–194 MR Zbl

[33] **K Murasugi**, *Knot theory and its applications*, Birkhäuser, Boston, MA (1996) MR Zbl

[34] **R Naimi**, *Foliations transverse to fibers of Seifert manifolds*, Comment. Math. Helv. 69 (1994) 155–162 MR Zbl

[35] **B Perron**, **D Rolfsen**, *On orderability of fibred knot groups*, Math. Proc. Cambridge Philos. Soc. 135 (2003) 147–153 MR Zbl

[36] **D Rolfsen**, *Knots and links*, Mathematics Lecture Series 7, Publish or Perish, Berkeley, CA (1976) MR Zbl

[37] **J Stallings**, *On fibering certain* 3*–manifolds*, from "Topology of 3–manifolds and related topics", Prentice-Hall, Englewood Cliffs, NJ (1962) 95–100 MR Zbl

[38] **M B Thistlethwaite**, *A spanning tree expansion of the Jones polynomial*, Topology 26 (1987) 297–309 MR Zbl

Department of Mathematics, University of Texas at Austin
Austin, TX, United States

jonjohnson@utexas.edu

Received: 13 September 2020 Revised: 3 November 2021

Algebraic & Geometric Topology **23**:4 (2023) 1831–1848
DOI: 10.2140/agt.2023.23.1831
Published: 14 June 2023

Maximal knotless graphs

LINDSAY EAKINS
THOMAS FLEMING
THOMAS MATTMAN

A graph is maximal knotless if it is edge maximal for the property of knotless embedding in $\mathbb{R}^3$. We show that such a graph has at least $\frac{7}{4}|V|$ edges, and construct an infinite family of maximal knotless graphs with $|E| < \frac{5}{2}|V|$. With the exception of $|E| = 22$, we show that for any $|E| \geq 20$ there exists a maximal knotless graph of size $|E|$. We classify the maximal knotless graphs through nine vertices and 20 edges. We determine which of these maxnik graphs are the clique sum of smaller graphs and construct an infinite family of maxnik graphs that are not clique sums.

05C10; 57K10, 57M15

1 Introduction

A graph G is *maximal planar* if it is edge maximal for the property of being a planar graph. That is, G is either a planar complete graph, or else adding any missing edge to G results in a nonplanar graph. Maximal planar graphs are triangulations and are characterized by the number of edges: a planar graph with $|V| \geq 3$ is maximal planar if and only if $|E| = 3|V| - 6$.

Naturally, planarity is not the only property of graphs that can be studied with respect to edge maximality. A graph is *intrinsically linked* if every embedding of the graph in $\mathbb{R}^3$ contains a nonsplit link. Some early results on *maximal linkless* (or *maxnil*) graphs — those that are edge maximal for the property of not being intrinsically linked — include a family of maximal linkless graphs with $3|V| - 3$ edges (see Jørgensen [10]), and the fact that the graph $Q(13, 3)$ is a splitter for intrinsic linking, a property that implies it is maximal linkless; see Maharry [13]. Recently there have been several new results including families of maxnil graphs with $3|V| - 3$ edges (rediscovering Jørgensen's examples, see Dehkordi and Farr [4]), with $\frac{14}{5}|V|$ edges (see Aires [1]) and with $\frac{25}{12}|V|$

edges (see Naimi, A Pavelescu, and E Pavelescu [16]). Lower bounds for the number of edges required for a maxnil graph have been established [1], and methods for creating new maxnil graphs via clique sum have been developed [16].

We extend this work with what appears to be the first study of *maximal knotless* graphs. A graph is *intrinsically knotted* (*IK*) if every embedding in $\mathbb{R}^3$ includes a nontrivially knotted cycle, and a graph is *not IK* or *nIK* if it has a knotless embedding, that is, an embedding in which every cycle is a trivial knot. We will call a graph that is edge maximal for the nIK property maximal knotless or *maxnik*.

In Section 2, we establish a connection between maximal 2–apex graphs and maxnik graphs, specifically that a 2–apex graph is maxnik if and only if it is maximally 2–apex. This connection is instrumental in allowing the identification of all maxnik graphs with nine or fewer vertices, and with 20 or fewer edges. We remark that there is an analogous connection between maximal apex graphs and maxnil graphs that may be of independent interest.

We consider clique sums of maxnik graphs in Section 3, and are able to establish similar, if weaker, results to those of [16]. Most importantly, we show that the edge sum of two maxnik graphs G_1 and G_2 on an edge e is maxnik if e is nontriangular (ie not part of a 3–cycle) in at least one G_i. Similarly, we provide conditions that ensure that the clique sum over K_3 of two maxnik graphs is again maxnik. These results are used in Section 4 to construct new maxnik graphs from those found in Section 2.

We then turn to studying general properties of maxnik graphs in Section 4. We establish a lower bound for the number of edges in a maxnik graph of $\frac{7}{4}|V|$, and construct an infinite family of maxnik graphs with fewer than $\frac{5}{2}|V|$ edges. A maximal planar graph has $|E| = 3|V| - 6$, and maximal k–apex graphs also have a fixed number of edges depending on $|V|$. In contrast, the number of edges in maxnil and maxnik graphs can vary. We show that, except for $|E| = 22$, given any $|E| \geq 20$, there exists a maxnik graph of size $|E|$.

We will call a maxnik graph *composite* if it is the clique sum of two smaller graphs. Otherwise we say it is *prime*. These terms are analogous to knots, where a knot is composite if is the connected sum of two nontrivial knots, and prime otherwise. The infinite families of maxnik graphs constructed in Section 4 are all composite, as they are clique sums of smaller maxnik graphs. In Section 5, we classify the maxnik graphs found in Section 2 and construct an infinite family of prime maxnik graphs.

2 Classification through order nine and size 20

For a graph G, let $\delta(G)$ denote the minimal degree, the smallest degree among the vertices of G. Similarly, $\Delta(G)$ is the maximal degree.

Theorem 2.1 *A maxnik graph is* 2–*connected. If* $|V| \geq 3$, *then* $\delta(G) \geq 2$. *If* $|V| \geq 7$, *then* $20 \leq |E| \leq 5|V| - 15$.

Proof Suppose G is maxnik. If G is not connected, then, in a knotless embedding, add an edge e to connect two components. This is a knotless embedding of $G + e$, contradicting G being maximal knotless.

Suppose G has connectivity one with cut vertex v. Label the two components of $G \setminus v$ as A and B. Let a be a neighbor of v in A and b be a neighbor of v in B. These must exist as G is connected. We will argue that $G + ab$ is also nIK, a contradiction.

Let A' denote the subgraph induced by A and the vertex v and similarly define B'. Since G is maxnik, the subgraphs A' and B' are both nIK. Embed A' and B' so that they are knotless and disjoint with a plane separating them. Isotope the edges va and vb to lie on the plane so that the two copies of the vertex v are identified. In this way, we obtain an embedding of G with A' on one side of the separating plane and B' on the other side of the plane so that v, a, and b are the only vertices on the separating plane and va and vb the only edges. Next add edge ab so that the triangle abv bounds a disk.

Any cycle contained in A (or in B) is an unknot. Any cycle c that uses vertices from both A and B must use at least two vertices in the triangle abv. Since abv bounds a disk, this means the cycle c is a connected sum of a cycle in A and a cycle in B. Since those are unknots, c must be as well. This shows $G + ab$ is nIK, contradicting G being maxnik. So a maxnik graph cannot have connectivity one and must be 2–connected.

Suppose G is a maxnik graph with $|V| \geq 3$. Since G is connected, $\delta(G) > 0$. If $v \in V(G)$ has degree one, let u be the neighbor of v. Since G is connected and $|V| \geq 3$, u must have another neighbor $w \neq v$. In a knotless embedding of G, we can introduce the edge vw that closely follows the path vuw. This gives a knotless embedding of $G + vw$, contradicting the maximality of G.

Suppose G is maxnik with $|V| \geq 7$. The lower bound on size is a consequence of the observation [9; 14] that an IK graph has at least 21 edges. The upper bound follows, as a graph with $|E| \geq 5|V| - 14$ has a K_7 minor and is therefore IK [3; 12]. □

In Theorem 4.3 below, we construct an infinite family of maxnik graphs, each with $\delta(G) = 2$.

We say that a graph is *apex* if it is planar or it becomes planar on deletion of a single vertex (the apex). Similarly, a graph is 2–*apex* if it is apex or becomes apex on deletion of a single vertex and a *maximal* 2–*apex* graph is one that is edge maximal for the 2–apex property.

Theorem 2.2 *A* 2*–apex graph is maxnik if and only if it is maximal* 2*–apex.*

Proof Let G be 2–apex. If G is not maximal 2–apex, then there is an edge e such that $G+e$ is 2–apex, hence nIK [2; 17]. This shows that G is not maxnik. Conversely, if G is maximal 2–apex there are two cases, depending on $|V|$. If $n = |V| < 7$, then K_n is 2–apex, so $G = K_n$. But, K_n is also nIK and therefore maxnik. If $|V| \geq 7$, then $|E| = 5|V| - 15$. Since G is 2–apex, it is nIK. Adding any edge e, we have $G+e$ with $5|V| - 14$ edges. It follows that G has a K_7 minor and is IK [3; 14]. This shows that G is maxnik. □

A similar result, with essentially the same proof, holds for maxnil.

Theorem 2.3 *An apex graph is maxnil if and only if it is maximal apex.*

Theorem 2.4 *For $|V| = n \leq 6$, K_n is the only maxnik graph. The only maxnik graphs for $n = 7$ and* 8 *are the three* 2*–apex graphs derived from triangulations on five and six vertices.*

Proof In [14, Proposition 1.4] it's shown that every nIK graph of order eight or less is 2–apex. So, the maxnik graphs are the maximal 2–apex graphs. For $n \leq 6$, all graphs are 2–apex, so K_n is the only maximal knotless graph. For $n = 7$, the maximal 2–apex graph is K_7^-, formed by adding two vertices to the unique graph with a planar triangulation on five vertices, K_5^-. The two maximal planar graphs on eight vertices are formed by adding two vertices to the two triangulations on six vertices, the octahedron and a graph whose complement is a 3–path. We will call these graphs $K_8 - 3$ disjoint edges and $K_8 - P_3$. □

Let E_9 (called N_9 in [8]) be the nIK nine vertex graph in the Heawood family. Figure 2 in Section 4 below shows a knotless [14] embedding of E_9.

Theorem 2.5 *The graph E_9 is maxnik.*

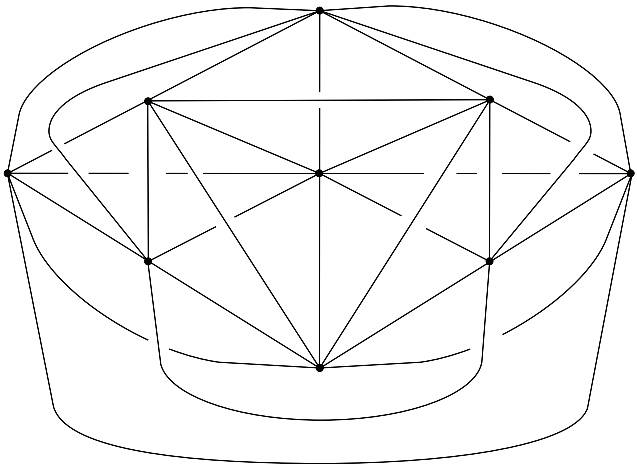


Figure 1: A knotless embedding of $G_{9,29}$.

Proof That E_9 is nIK is established in [14]. Up to symmetry, there are two types of edges that may be added. One type yields the graph $E_9 + e$, shown to be IK (in fact minor minimal IK or MMIK) in [7]. The other possible addition yields a graph that has as a subgraph F_9 in the Heawood family. Kohara and Suzuki [11] established that F_9 is MMIK. □

Theorem 2.6 *There are seven maxnik graphs of order nine.*

Proof The seven graphs are the five maximal 2–apex graphs with 30 edges, E_9, and the graph $G_{9,29}$, shown in Figure 1. Note that $G_{9,29}$ is the complement of $K_1 \sqcup K_2 \sqcup C_6$. Theorems 2.2 and 2.5 show that six of these seven graphs are maxnik. To see that $G_{9,29}$ is as well, note that the embedding shown in Figure 1, due to Ramin Naimi (personal communication, 2011), is knotless. Up to symmetry, there are two ways to add an edge to the graph. In either case, the new graph has a K_7 minor and is IK.

It remains to argue that no other graphs of order nine are maxnik. We know that order nine graphs with size 21 or less are either IK, the graph E_9, or else 2–apex; see [14, Propositions 1.6 and 1.7]. Using Theorem 2.2, this completes the argument for graphs with $|E| \leq 21$. Suppose G is maxnik of order nine with $|E| \geq 22$. By Theorem 2.1, we can assume $|E| \leq 30$. If G is 2–apex, by Theorem 2.2, it is one of the five maximal 2–apex graphs. So, we can assume G is not 2–apex. The minor minimal not 2–apex (MMN2A) graphs through order nine are classified in [15]. With a few exceptions these graphs are also MMIK. If G has an IK minor (including an MMIK minor) it is IK and not maxnik. So, we can assume G has as a minor a graph that is MMN2A, but not MMIK. There are three such graphs. One is E_9, the other two, G_{26} and G_{27}, have 26 and 27 edges. In Theorem 2.5, we showed that E_9 is maxnik. The other two are subgraphs of $G_{9,29}$. To complete the proof, we observe that any order nine graph that contains G_{26} is either a subgraph of $G_{9,29}$ or else IK and similarly for G_{27}. In fact, for those that are IK, we can verify this by finding an MMIK minor, either in the K_7 or $K_{3,3,1,1}$ family, or else the graph $G_{9,28}$ described in [7]. □

Theorem 2.7 *The only maxnik graph of size* 20 *is* K_7^-. *There are seven maxnik graphs with at most* 20 *edges.*

Proof Work above establishes this through order nine. The seven maxnik graphs with at most 20 edges are the seven on seven or fewer vertices. Suppose G of order ten or more and size 20 is maxnik. By [14, Theorem 2.1], G is 2–apex and therefore maximal 2–apex. But this means $|E| = 5|V| - 15 \geq 35$, a contradiction. □

Remark 2.8 A computer search suggests that E_9 is the only maxnik graph of size 21. The search makes use of the 92 known MMIK graphs of size 22; see [5].

3 Clique sums of maxnik graphs

Clique sums of maxnil graphs were studied in [16], and we will show similar, if weaker, versions in the case of maxnik graphs. These results are used in Section 4. A *clique* in a graph is a complete subgraph. When graphs G and H both contain the same clique K_n, we can form a new graph $G \cup_{K_n} H$, called the *clique sum*, from the disjoint union by identifying the vertices in the two copies of K_n.

Lemma 3.1 *For* $t \leq 2$, *the clique sum over* K_t *of nIK graphs is nIK.*

Proof Let G_1 and G_2 be nIK graphs, and let $\Gamma(G)$ denote the set of all cycles in G. Let G be the clique sum of G_i over a clique of size t. Let f_i be an embedding of G_i that contains no nontrivial knot.

Suppose $t = 1$. We may extend the f_i to an embedding of G by embedding $f_1(G_1)$ in 3–space with $z > 0$, and $f_2(G_2)$ with $z < 0$. $G = G_1 \cup_v G_2$, so by isotoping vertex v from each G_i to the plane $z = 0$ and identifying them there, we have an embedding $f(G)$. A closed cycle in G must be contained in a single G_i, and hence given $c \in \Gamma(G)$, then $c \in \Gamma(G_i)$ for some i. As the embeddings $f_i(G_i)$ contain no nontrivial knot, c must be the unknot, and hence G is nIK.

Suppose $t = 2$. We may extend the f_i to an embedding of G by embedding $f_1(G_1)$ in 3–space with $z > 0$, and $f_2(G_2)$ with $z < 0$. $G = G_1 \cup_e G_2$, so by shrinking the edge e in each G_i and then isotoping them to the plane $z = 0$ and identifying them there, we have an embedding $f(G)$. A closed cycle $c \in \Gamma(G)$ must either be an element of $\Gamma(G_i)$, or $c = c_1 \# c_2$, with $c_i \in \Gamma(G_i)$. As the embeddings $f_i(G_i)$ contain no nontrivial knot, in the first case c is the unknot, and in the second it is the connected sum of unknots and hence unknotted. Thus, G is nIK. □

For $H_1, H_2, \dots, H_k$ subgraphs of graph G, let $\langle H_1, H_2, \dots, H_k \rangle_G$ denote the subgraph induced by the vertices of the subgraphs.

Lemma 3.2 *Let G be a maxnik graph with a vertex cut set $S = \{x, y\}$, and let $G_1, G_2, \dots, G_r$ denote the connected components of $G \setminus S$. Then $xy \in E(G)$ and $\langle G_i, S \rangle_G$ is maxnik for all $1 \le i \le r$.*

Proof As G is 2–connected by Theorem 2.1, each of x and y has at least one neighbor in each G_i. Suppose $xy \notin G$. Form $G' = G + xy$ and let $G'_i = \langle G_i, S \rangle_{G'}$. For each i, edge xy is in G'_i. But G'_i is a minor of G, as there exists G_j with $i \ne j$ since S is separating, and there exists a path from x to y in G_j as G_j is connected. Thus in $\langle G_i, G_j, S \rangle_G$, we may contract G_j to x to obtain a graph isomorphic to G'_i. Thus, G'_i is nIK. So, by Lemma 3.1, $G' = G'_1 \cup_{xy} G'_2 \cup_{xy} \cdots \cup_{xy} G'_r$ is nIK. This contradicts the fact that G is maxnik, and hence $xy \in E(G)$.

Suppose that one or more of the G_i are not maxnik. Then add edges as needed to each G_i to form graphs H_i that are maxnik. Then the graph $H = H_1 \cup_{xy} H_2 \cup_{xy} \cdots \cup_{xy} H_r$ is nIK by Lemma 3.1 and contains G as a subgraph. As G is maxnik, $G = H$ and hence $G_i = H_i$ for all i, so every G_i is maxnik as well. □

We say that an edge in a graph is *triangular* if it is part of a triangle or 3–cycle. Similarly, the edge is *nontriangular* if it is part of no 3–cycle in the graph.

Lemma 3.3 *Let G_1 and G_2 be maxnik graphs. Pick an edge in each G_i and label it e. Then $G = G_1 \cup_e G_2$ is maxnik if e is nontriangular in at least one G_i.*

Proof Suppose that e is nontriangular in G_1 and has endpoints x and y. Add an edge ab to the graph G. The graph G is nIK by Lemma 3.1. If both $a, b \in G_i$ for some i, then $G + ab$ is IK, as the G_i are each maxnik. Thus, we may assume that $a \in G_1$ and $b \in G_2$. The edge e is nontriangular in G_1, so vertex a is not adjacent to both endpoints of e. We may assume that a is not adjacent to x. As G_2 is connected, we construct a minor G' of $G + ab$ by contracting the whole of G_2 to vertex x. Note that as $b \in G_2$, we have the edge ax in G', and in fact $G' = G_1 + ax$. As G_1 is maxnik, G' is IK and so is $G + ab$. Thus, G is maxnik. □

Lemma 3.4 *For $i = 1, 2$, let G_i be maxnik, containing a 3–cycle C_i, and admitting a knotless embedding such that C_i bounds a disk whose interior is disjoint from the graph. Then the clique sum G over K_3 formed by identifying C_1 and C_2 is nIK. Moreover, G is maxnik if C_i is not part of a K_4 in at least one G_i.*

Proof Let f_i be the knotless embedding of G_i. Embed the $f_i(G_i)$ so that they are separated by a plane. We may then extend this to an embedding $f(G)$ by isotoping the C_i to the separating plane and identifying them there.

Let $\Gamma(G)$ denote the set of all cycles in G. As the cycles C_i bound a disk in $f(G)$, if a closed cycle $c \in \Gamma(G)$ is not contained in one of the $f_i(G_i)$, then $c = c_1 \# c_2$, with $c_i \in \Gamma(G_i)$. As the embeddings $f_i(G_i)$ contain no nontrivial knot, in the first case c is the unknot, and in the second, it is the connected sum of unknots and hence unknotted. Thus, G is nIK.

Suppose C_1 is not contained in a 4–clique in G_1. We will show $G + ab$ is IK, and hence G is maxnik. As the G_i are maxnik, we may assume that $a \in G_1$ and $b \in G_2$, as otherwise $G + ab$ is IK. As C_1 is not contained in a 4–clique in G_1, there exists a vertex x in C_1 that is not adjacent to a. As G_2 is connected, there is a path from b to x. Contract G_2 to x. This graph contains $G_1 + ax$ as a minor, and hence is IK, as G_1 is maxnik and does not contain edge ax. Thus, G is maxnik. □

4 Bounds on maximal knotless graphs

We now consider maximal knotless graphs in general and establish bounds on the possible number of edges, and the maximal and minimal degrees. We first show a lemma that will be useful for establishing a lower bound. A similar result holds for maximal linkless graphs as well.

Lemma 4.1 *Suppose G is maxnik and contains a vertex v of degree three. Then all neighbors of v are adjacent to each other.*

Proof Label the neighbors of v as x_1, x_2, and x_3. Let $E_v = \{x_1x_2, x_1x_3, x_2x_3\}$ and $E = E(G)$. Delete the edges in $E \cap E_v$ to form $G_Y = G \setminus (E \cap E_v)$. Then add back all the edges of E_v to form $G' = G_Y + E_v$. We will show $G = G'$.

As G is maxnik, G_Y has an embedding f with no nontrivial knot. We may extend f to an embedding of G' by embedding each edge x_ix_j so that the 3–cycle x_ivx_j bounds a disk.

Let $\Gamma(G)$ denote the set of all cycles in the graph G. Suppose c is a cycle in $\Gamma(G')$. If c does not contain one or more edges x_ix_j, then $c \in \Gamma(G_Y)$, and hence is a trivial cycle in $f(G')$. Suppose that c does contain one or more edges x_ix_j. There are three possibilities: c is a 3–cycle x_ivx_j and bounds a disk, c includes a path of the form $x_ix_jvx_k$ with $\{i,j,k\} = \{1,2,3\}$, or c does not include the vertex v. In the first case c is trivial as it bounds a disk. If c does not contain v, then, since the cycles x_ivx_j bound disks, c is isotopic to $c' \in \Gamma(G_Y)$ and hence trivial. Similarly, if c includes a path $x_ix_jvx_k$, using the disk x_ivx_j we can isotope the path to x_ivx_k to make c isotopic to $c' \in \Gamma(G_Y)$ and hence trivial.

Thus, G' has an embedding with no nontrivial knot. As G is maxnik, G cannot be a proper subgraph of G', and hence $G = G'$. □

Theorem 4.2 *If G is maxnik with $|V| \geq 5$, then $|E| \geq \frac{7}{4}|V|$.*

Proof By Theorem 2.4, K_5 is the only maxnik graph with order five and it satisfies the conclusion of the theorem.

Suppose H has the least number of vertices among counterexamples to the theorem. We will consider a vertex v of minimal degree in H. If $\deg(v) \geq 4$, then H has $|E| \geq 2|V|$ and hence is not a counterexample, so $\deg(v) \leq 3$. By Theorem 2.1, $\deg(v) \geq 2$, so we need only consider v of degree two or three.

Suppose $\deg(v) = 2$. We will argue that $H' = H \setminus v$ is also maxnik with $|E'| < \frac{7}{4}|V'|$, contradicting our assumption that H was a minimal counterexample. Let $N(v) = \{w, x\}$ and note that $wx \in E(H)$. Otherwise, in an unknotted embedding of H, we could add the edge wx so that the 3–cycle vwx bounds a disk. This will not introduce a knot into the embedding and contradicts the maximality of H.

As a subgraph of H, H' is nIK. Suppose it is not maxnik because there is an edge ab such that $H'+ab$ remains nIK. In a knotless embedding of $H'+ab$, we can add the vertex v and its two edges so the 3–cycle vwx bounds a disk. This will not introduce a knot into the embedding and shows that $H+ab$ is also nIK, contradicting the maximality of H. Thus, no such graph H with a vertex of degree two can exist.

So we may assume that $\deg(v)=3$. Here we cannot apply the techniques of [1], as $Y\nabla$ moves do not preserve intrinsic knotting [6]. However, Lemma 4.1 allows us to show the average degree of H is actually at least 3.5, and hence H is not a counterexample.

Divide the vertices of H into three sets: $A=\{\text{vertices of degree } 3\}$, $B=\{\text{vertices of degree} > 3 \text{ that are neighbors of vertices in } A\}$, and $C=\{\text{all other vertices of } H\}$. Form the graph $H'=H\setminus C$. All vertices in C have degree four or greater, so it suffices to show that the vertices in each connected component of H' have average degree 3.5 or higher.

A vertex a_{i1} of degree three has three neighbors, label them b_{i1}, b_{i2}, and a_{i2}, where a_{i2} is a neighbor of minimal degree. By Lemma 4.1, the neighbors of a_{i1} are mutually adjacent. If $\deg(a_{i2})=3$, we continue. If not, delete all edges incident on a_{i2} except those between a_{i2} and $\{a_{i1},b_{i1},b_{i2}\}$. This creates a subgraph of H' with strictly fewer edges; we will abuse notation and continue to call it H'. Vertex a_{i2} now has degree three in H', and we move it to set A.

If a_{i2} had degree greater than three in H, then, since it has the minimal degree among the neighbors of a_{i1}, $\deg(b_{ij})\geq 4$ and $b_{i1},b_{i2}\in B$. If $\deg(a_{i2})=3$ in H, vertices a_{ij} are adjacent only to each other and the b_{ij}. If either of the b_{ij} have degree three in H, then H can be disconnected by deleting the other b_{ij}. This is a contradiction as H is maxnik and must be 2–connected by Theorem 2.1. Thus, the b_{ij} are in B.

Consider the connected component of v in H', call it H'_1. We will calculate the total degree of the vertices in H'_1 and divide by the number of vertices. Suppose there are n vertices from set A and m vertices from set B in H'_1 for a total of $n+m$ vertices. Each vertex from set A has degree three, so the contribution to total degree from set A is $3n$. Each vertex in A is adjacent to exactly two of the b_{ij}, so the total degree contribution for set B is at least $2n$ from edges to set A. Further, H'_1 is connected. As a_{ij} is only adjacent to $b_{i'j'}$ if $i=i'$, there must be at least $m-1$ edges between the b_{ij}, which adds $2(m-1)$ to the total degree. This gives an average degree of $(5n+2m-2)/(n+m)$ in H'_1. However, H is 2–connected by Theorem 2.1, so there must be at least two edges from H'_1 to its complement in H. So within H, these

$\lvert V\rvert$	1	2	3	4	5	6	7	8	9
$\min(\lvert E\rvert/\lvert V\rvert)$	0	$\frac{1}{2}$	1	$\frac{3}{2}$	2	$\frac{5}{2}$	$\frac{20}{7}$	$\frac{25}{8}$	$\frac{21}{9}$

Table 1: The least ratios of size to order for maxnik graphs through order nine.

vertices must have average degree greater than or equal to $(5n+2m)/(n+m)$. Note that $2 \le m \le n$, and $(5n+2m)/(n+m)$ attains its minimum at $m=n$. The minimum is $\frac{7}{2}$, and hence H must have $|E| \ge \frac{7}{4}|V|$. □

Theorem 4.3 *There exist maxnik graphs with $|E| < \frac{5}{2}|V|$ edges for arbitrarily large $|V|$.*

Proof Let e be an edge of E_9 connecting a degree four vertex to one of degree five. Edge e is nontriangular and there are five other edges symmetric to it. Using Lemma 3.3, take k copies of E_9 glued along edge e. The resulting graph has $7k+2$ vertices and $20k+1$ edges. Gluing on five K_3 graphs in each E_9 on the other nontriangular edges gives an additional $5k$ vertices and $10k$ edges. So, for each $k \ge 1$, we have a graph G with $n = 12k+2$ vertices and $m = 30k+1$ edges. Then $m = 30(n-2)\frac{1}{12}+1 = \frac{5}{2}n-5$. □

These two theorems suggest the following question. In Table 1 we give the least ratios through order nine.

Question 4.4 What is the minimal number of edges for a maxnik graph of n vertices?

For maximal planar graphs, $|E| = 3|V|-6$. Similarly, maximal k–apex graphs have a fixed number of edges depending on $|V|$. In contrast, as with maximal linkless graphs, the number of edges in a maxnik graph can vary. In fact, with the exception of $|E| = 22$, for any $|E| \ge 20$, there exists a maxnik graph of that size.

Theorem 4.5 *Let $n \ge 20$ and $n \ne 22$. Then there exists a maxnik graph with $|E| = n$.*

Proof The graph K_7^- is maxnik of size 20 by Theorem 2.4. The graph E_9 has a knotless embedding where the 3–cycle abc bounds a disk [14], shown in Figure 2. As no vertex in E_9 is adjacent to all three of these vertices, we may use Lemma 3.4 to construct a maxnik graph of size 24 by taking a clique sum over K_3 of E_9 and K_4. So, we may assume $n \ge 21$ and $n \notin \{22, 24\}$.

The graph E_9 has size 21 and six nontriangular edges. Let G_i denote the maxnik graph obtained from i copies of E_9 by gluing along nontriangular edges.

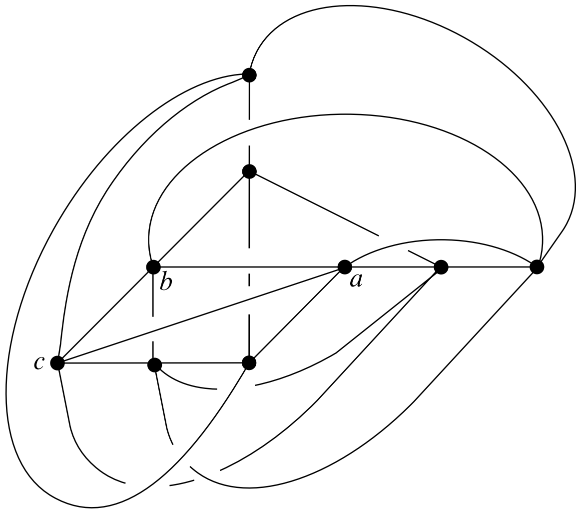


Figure 2: A knotless embedding of E_9.

Note that $|E(G_{i+1})|-|E(G_i)|=20$, and that G_i contains at least six nontriangular edges for any i. We will now work by induction. Suppose that maxnik graphs exist for size $n<|E(G_i)|$ and for size $|E(G_i)|+1$ and size $|E(G_i)|+3$. Then it suffices to show that there exist maxnik graphs of size $|E(G_i)|+k$ for $4\le k\le 19$ and $k\in\{0,2,21,23\}$.

Clearly a maxnik graph of size $|E(G_i)|+0$ exists, as G_i is maxnik. We may form a new maxnik graph from G_i by gluing a copy of K_m (for $3\le m\le 6$) along a nontriangular edge of G_i by Lemma 3.3. As G_i has at least six nontriangular edges, we can glue on up to six such graphs, each adding $\binom{m}{2}-1$ edges. Thus, to prove the result we need only to be able to form the desired values of k using six or fewer addends from the set $\{2,5,9,14\}$. This is clearly possible.

In the base case $i=1$, we have a maxnik graph of size $|E(G_1)|=21$, and we excluded graphs of size 22 and 24 ($|E(G_1)|+1$ and $|E(G_1)|+3$) above. Thus we may form maxnik graphs of size $|E(G_1)|+k$ for the k of interest as before. □

Remark 4.6 A computer search shows there are no size 22 maxnik graphs. Our strategy is based on the classification through size 22 of the obstructions to 2–apex in [15]. Let's call such graphs MMN2A (minor minimal not 2–apex). All but eight of the graphs in the classification are MMIK. Two exceptions are 4–regular of order 11, the other six are in the Heawood family.

A maximal 2–apex graph has $5n-15$ edges where n is the number of vertices. By Theorem 2.2 a maxnik graph G of size 22 is not 2–apex and therefore has an MMN2A minor. Since G is nIK, it must have one of the eight exceptions as a minor. Using a computer, we verified that no size 22 expansion of any of these eight graphs is maxnik.

Theorem 4.5 implies that there are maxnik graphs of nearly every size. Note that there are maxnik graphs of any order, as there exist maximal 2–apex graphs of any order and by Theorem 2.2 these graphs are maxnik.

We have considered the minimal number and the possible number of edges in a maxnik graph. We now consider other aspects of maxnik graphs' structure, in particular, the maximal and minimal degree. Since $\Delta(G)=|V|-1$ for maximal 2–apex graphs, there are maxnik graphs with arbitrarily large $\Delta(G)$.

Proposition 4.7 *The complete graph K_3 is the only maxnik graph with maximal degree two.*

Proof Suppose G is maxnik with $\Delta(G)=2$. Then $|G|\geq 3$ and, by Theorem 2.1, $\delta(G)=2$ and G is connected. So G is a cycle. Now, a cycle is planar, hence 2–apex, and by Theorem 2.2 G is maximal 2–apex. However, a cycle is not maximal 2–apex unless it is K_3. □

Note that Lemma 4.1 has the following two immediate corollaries:

Corollary 4.8 *If a graph G is maxnik and has $\Delta(G)=3$, then G is 3–regular.*

Corollary 4.9 *If a graph G is maxnik and 3–regular, then $G=K_4$.*

These results motivate the following question:

Question 4.10 Do there exist regular maxnik graphs other than K_n with $n<7$?

A maximal 2–apex graph will have $\Delta(G)=|V|-1$ and $\delta(G)\leq 7$, so if there is such a regular maxnik graph with $|V|\geq 7$ it is not 2–apex. However, through order nine, our two examples of maxnik non-2–apex graphs are both close to regular, having $\Delta(G)-\delta(G)\leq 2$. This suggests the answer to our question is likely yes.

For $\delta(G)$, Theorem 2.1 gives a lower bound of two that is realized by the infinite family of Theorem 4.3. On the other hand, by starting with a planar triangulation of minimum

$\lvert V\rvert$	1	2	3	4	5	6	7	8	9
$\delta(G)$	0	1	2	3	4	5	5	5 or 6	4 to 7
$\Delta(G)$	0	1	2	3	4	5	6	7	5 to 8

Table 2: Maximal and minimal degrees of maxnik graphs through order nine.

degree five, we can construct graphs with $\delta(G) = 7$ that are maximal 2–apex, and hence maxnik. At the same time, since a graph with $|E| \geq 5|V| - 14$ has a K_7 minor and is IK [3; 12], a maxnik graph must have $\delta(G) \leq 9$. It seems likely that there are examples that realize this upper bound on $\delta(G)$. Table 2 records the range of degrees for maxnik graphs through order nine.

5 Prime and composite maxnik graphs

We will call a graph *composite* if it is the clique sum of two graphs. Otherwise it is *prime*. These terms are analogous to knots, where a knot is composite if is the connected sum of two nontrivial knots, and prime otherwise. In this section, we classify the maxnik graphs described earlier in this paper as prime and composite. We remark that it may be of interest to study other instances of prime graphs, for example, prime maximal planar or prime maxnil.

The infinite families of maxnik graphs constructed in Section 4 are all composite, as they are clique sums of smaller maxnik graphs.

Note that K_n is prime, so all maxnik graphs of order six or less are prime.

Proposition 5.1 *The following maxnik graphs are composite*: K_7^-, $K_8 - P_3$, *and four of the five maximal* 2*–apex graphs on nine vertices*, *specifically big-*Y, *long-*Y, *hat and house.*

Proof The graph K_7^- is formed from two copies of K_6 summed over a 5–clique.

The graph $K_8 - P_3$ is formed from K_7^- clique sum K_6 over a 5–clique, where the 5–clique contains exactly one endpoint of the missing edge.

Big-Y is formed from $K_8 - P_3$ clique sum K_6 over a 5–clique, where the 5–clique contains both of the terminal vertices of the 3–path.

Long-Y is formed from K_8- 3 disjoint edges clique sum K_6 over a 5–clique.

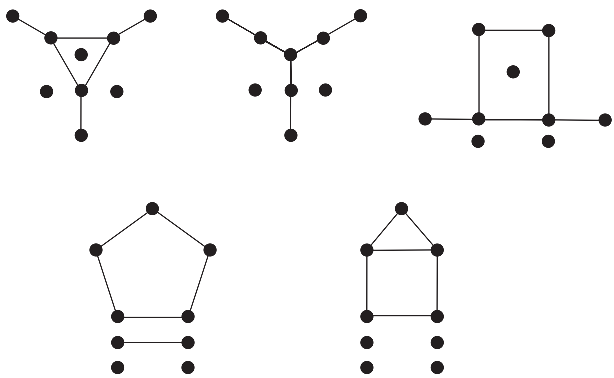


Figure 3: Complements of the maximal 2–apex graphs of order nine. Top row, left to right: big-Y, long-Y, and hat. Bottom row: pentagon–bar and house.

Hat is formed from $K_8 - P_3$ clique sum K_6 over a 5–clique, where the 5–clique contains one terminal vertex and one (nonadjacent) interior vertex of the 3–path.

House is formed from $K_8 - P_3$ clique sum K_6 over a 5–clique, where the 5–clique contains one interior vertex of the 3–path. □

Lemma 5.2 *If G^c is of the form $K_2 \amalg H$, then either G is prime, or G is the clique sum of two copies of K_n over an $n-1$ clique.*

Proof Call the two vertices of the K_2 in G^c v_1 and v_2. Suppose that G is a clique sum of G_1 and G_2 over a clique C. We cannot have both v_1 and v_2 in C, as edge v_1v_2 is in G^c. Without loss of generality, we may assume that v_1 is in $G_1 \setminus C$. So, in G^c, v_1 must be adjacent to every vertex of $G_2 \setminus C$. Thus $G_2 \setminus C$ is v_2. As the only neighbor of v_1 in G^c is v_2, v_1 is adjacent to every vertex in C. Similarly for v_2. Thus if G is composite, it is the clique sum of K_n and K_n over an $n-1$ clique. □

Corollary 5.3 *The following maxnik graphs are prime: pentagon–bar, $G_{9,29}$ and $K_8 - 3$ disjoint edges.*

Proof Each of these graphs has a complement of the form $K_2 \amalg H$. As these graphs are not of the form $K_n -$ a single edge, they are prime by Lemma 5.2. □

Note that if G is a clique sum over a t–clique, it is not $(t+1)$–connected.

Proposition 5.4 *The maxnik graph E_9 is prime.*

Proof The largest clique in E_9 is a 3–clique, but E_9 is 4–connected and hence must be prime. □

Lemma 5.5 *If $G = H * K_2$, and G is 2–apex, then G is prime maxnik if and only if H is prime maximal planar.*

Proof As G is 2–apex, it is maxnik if and only if it is maximal 2–apex, and G is maximal 2–apex if and only if H is maximal planar.

If H is composite, then H is the clique sum of H_1 and H_2 over a t–clique. So G is the clique sum of $H_1 * K_2$ and $H_2 * K_2$ over a $(t+2)$–clique, and hence G is composite.

As G is maxnik, it must be 2–connected. Hence if G is composite, it must be G_1 clique sum G_2 over a t–clique C, with $t \geq 2$. Label two of the vertices in C as v_1 and v_2. Then H is the clique sum of $G_1 \setminus \{v_1, v_2\}$ and $G_2 \setminus \{v_1, v_2\}$ over $C \setminus \{v_1, v_2\}$, and thus composite. □

Corollary 5.6 *There exist prime maxnik graphs of arbitrarily large size, and of any order ≥ 8.*

Proof The octahedron graph is maximal planar and 4–connected. The largest clique it contains is a 3–clique, so it is prime. New triangulations formed by repeated subdivision of a single edge are 4–connected and maximal planar, but have no 4–clique, hence are prime as well. Thus all of these graphs give prime maxnik examples when joined with K_2. □

We remark that the construction of this family of graphs is similar to the maxnil families with $3n-3$ edges due to Jørgensen [10] and $3n-5$ edges due to Naimi, Pavelescu, and Pavelescu [16].

References

[1] **M Aires**, *On the number of edges in maximally linkless graphs*, J. Graph Theory 98 (2021) 383–388 MR

[2] **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 MR Zbl

[3] **J Campbell**, **T W Mattman**, **R Ottman**, **J Pyzer**, **M Rodrigues**, **S Williams**, *Intrinsic knotting and linking of almost complete graphs*, Kobe J. Math. 25 (2008) 39–58 MR Zbl

[4] **H R Dehkordi**, **G Farr**, *Non-separating planar graphs*, Electron. J. Combin. 28 (2021) art. id. 1.11 MR Zbl

[5] **E Flapan**, **T W Mattman**, **B Mellor**, **R Naimi**, **R Nikkuni**, *Recent developments in spatial graph theory*, from "Knots, links, spatial graphs, and algebraic invariants" (E Flapan, A Henrich, A Kaestner, S Nelson, editors), Contemp. Math. 689, Amer. Math. Soc., Providence, RI (2017) 81–102 MR Zbl

[6] **E Flapan**, **R Naimi**, *The Y-triangle move does not preserve intrinsic knottedness*, Osaka J. Math. 45 (2008) 107–111 MR Zbl

[7] **N Goldberg**, **T W Mattman**, **R Naimi**, *Many, many more intrinsically knotted graphs*, Algebr. Geom. Topol. 14 (2014) 1801–1823 MR Zbl

[8] **R Hanaki**, **R Nikkuni**, **K Taniyama**, **A Yamazaki**, *On intrinsically knotted or completely 3–linked graphs*, Pacific J. Math. 252 (2011) 407–425 MR Zbl

[9] **B Johnson**, **M E Kidwell**, **T S Michael**, *Intrinsically knotted graphs have at least 21 edges*, J. Knot Theory Ramifications 19 (2010) 1423–1429 MR Zbl

[10] **L K Jørgensen**, *Some maximal graphs that are not contractible to K_6*, art. id. R 89-28, Institut for Elektroniske Systemer, Aalborg Universitet (1989)

[11] **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 MR Zbl

[12] **W Mader**, *Homomorphiesätze für Graphen*, Math. Ann. 178 (1968) 154–168 MR Zbl

[13] **J Maharry**, *A splitter for graphs with no Petersen family minor*, J. Combin. Theory Ser. B 72 (1998) 136–139 MR Zbl

[14] **T W Mattman**, *Graphs of 20 edges are 2–apex, hence unknotted*, Algebr. Geom. Topol. 11 (2011) 691–718 MR Zbl

[15] **T W Mattman**, **M Pierce**, *The K_{n+5} and $K_{3^2,1^n}$ families and obstructions to n–apex*, from "Knots, links, spatial graphs, and algebraic invariants" (E Flapan, A Henrich, A Kaestner, S Nelson, editors), Contemp. Math. 689, Amer. Math. Soc., Providence, RI (2017) 137–158 MR Zbl

[16] **R Naimi**, **A Pavelescu**, **E Pavelescu**, *New bounds on maximal linkless graphs*, preprint (2020) arXiv 2007.10522 To appear in Algebr. Geom. Topol.

[17] **M Ozawa**, **Y Tsutsumi**, *Primitive spatial graphs and graph minors*, Rev. Mat. Complut. 20 (2007) 391–406 MR Zbl

Department of Mathematics and Statistics, California State University, Chico
Chico, CA, United States

New York, NY, United States

Department of Mathematics and Statistics, California State University at Chico
Chico, CA, United States

lpepper@mail.csuchico.edu, thomasrfleming@gmail.com,
tmattman@csuchico.edu

Received: 12 January 2021 Revised: 26 August 2021

Algebraic & Geometric Topology 23:4 (2023) 1849–1889
DOI: 10.2140/agt.2023.23.1849
Published: 14 June 2023

Distinguishing Legendrian knots with trivial orientation-preserving symmetry group

IVAN DYNNIKOV
VLADIMIR SHASTIN

Recent work of I Dynnikov and M Prasolov proposes a new method of comparing Legendrian knots. In general, to apply the method requires a lot of technical work. In particular, one needs to search all rectangular diagrams of surfaces realizing certain dividing configurations. We show that in the case when the orientation-preserving symmetry group of the knot is trivial, this exhaustive search is not needed, which simplifies the procedure considerably. This allows one to distinguish Legendrian knots in certain cases when the computation of the known algebraic invariants is infeasible or not informative. In particular, we disprove that when $A \subset \mathbb{R}^3$ is an annulus tangent to the standard contact structure along ∂A, then the two components of ∂A are always equivalent Legendrian knots. A candidate counterexample was proposed recently by Dynnikov and Prasolov, but the proof of the fact that the two components of ∂A are not Legendrian equivalent was not given. Now this work is accomplished. It is also shown here that the problem of comparing two Legendrian knots having the same topological type is algorithmically solvable provided that the orientation-preserving symmetry group of these knots is trivial.

57K10, 57K33

Introduction

Deciding whether or knot two Legendrian knots in $\mathbb{S}^3$ having the same classical invariants (see definitions below) are Legendrian isotopic is not an easy task in general. There are two major tools used for classification of Legendrian knots of a fixed topological type: Legendrian knot invariants having algebraic nature (see Chekanov [3], Eliashberg [13], Fuchs [20], Ng [35; 36], Ozsváth, Szabó and Thurston [37] and Pushkar' and Chekanov [38]), and Giroux's convex surfaces endowed with the characteristic foliation (see Eliashberg and Fraser [14; 15], Etnyre and Honda [16], Etnyre, LaFountain and Tosun [17], Etnyre, Ng and Vértesi [18] and Etnyre and Vértesi [19]).

The Legendrian knot atlas by W Chongchitmate and L Ng [4] summarizes the classification results for Legendrian knots having arc index at most 9. As one can see from [4] there are still many gaps in the classification even for knots with a small arc index/crossing number. Namely, there are many pairs of Legendrian knot types which are conjectured to be distinct, but are not distinguished by means of the existing methods.

The works [9; 10] by Dynnikov and Prasolov propose a new combinatorial technique for dealing with Giroux's convex surfaces. This includes a combinatorial presentation of convex surfaces in $\mathbb{S}^3$ and a method that allows one, in certain cases, to decide whether or not a convex surface with a prescribed topological structure of the dividing set exists.

The method of [9; 10] is useful for distinguishing Legendrian knots, but it requires, in each individual case, a substantial amount of technical work and a smart choice of a Giroux convex surface whose boundary contains one of the knots under examination.

In the present paper we show that there is a way to make this smart choice in the case when the examined knots have no topological (orientation-preserving) symmetries, so that the remaining technical work described in [10] becomes unnecessary, as the result is known in advance. This makes the procedure completely algorithmic and allows us, in particular, to distinguish two specific Legendrian knots for which computation of the known algebraic invariants is infeasible due to the large complexity of the knot presentations.

The two knots in question are of interest due to the fact that they cobound an annulus embedded in $\mathbb{S}^3$ and have zero relative Thurston–Bennequin and rotation invariants. They were proposed in [9] as a candidate counterexample to the claim of Gospodinov [27] that the two boundary components of such an annulus must be Legendrian isotopic.

The main technical result of this paper was announced by us in [11] without complete proof. In Dynnikov [7] the method of this paper is used to show that one can compare transverse link types in a similar fashion provided that the orientation-preserving symmetry group of the links is trivial. In a forthcoming paper by Dynnikov and Prasolov it will be shown how to drop the no-symmetry assumption and to produce algorithms for comparing Legendrian and transverse link types in the general case.

The paper is organized as follows. In Section 1 we recall the definition of a Legendrian knot, and introduce the basic notation. In Section 2 we discuss annuli with Legendrian boundary whose components have zero relative Thurston–Bennequin number. In Section 3 we define the orientation-preserving symmetry group of a knot and introduce some $\mathbb{S}^3$–related notation. In Section 4 we recall the definition of a rectangular diagram of a knot and discuss the relation of rectangular diagrams to Legendrian knots. Section 5 discusses rectangular diagrams of surfaces. Here we describe the smart choice of a surface mentioned above (Lemma 5.4). In Section 6 we prove the triviality of the orientation-preserving symmetry group of the concrete knots that are discussed in the paper (modulo the proof of hyperbolicity of the two complicated knots cobounding an annulus but not Legendrian equivalent, which is postponed till Section 7). In Section 7 we prove a number of statements about the nonequivalence of the considered Legendrian knots.

Acknowledgement The work is supported by the Russian Science Foundation under grant 19-11-00151.

1 Legendrian knots

All general statements about knots in this paper can be extended to many-component links. To simplify the exposition, we omit the corresponding formulations, which are pretty obvious but sometimes slightly more complicated.

All knots in this paper are assumed to be oriented. The knot obtained from a knot K by reversing the orientation is denoted by $-K$.

Definition 1.1 Let ξ be a *contact structure* in the three-space $\mathbb{R}^3$, that is, a smooth 2–plane distribution that locally has the form $\ker\alpha$, where α is a differential 1–form such that $\alpha\wedge d\alpha$ does not vanish. A smooth curve γ in $\mathbb{R}^3$ is called *ξ–Legendrian* if it is tangent to ξ at every point $p\in\gamma$.

A ξ–Legendrian knot is a knot in $\mathbb{R}^3$ which is a ξ–Legendrian curve. Two ξ–Legendrian knots K and K' are said to be *equivalent* if there is a diffeomorphism $\varphi\colon\mathbb{R}^3\to\mathbb{R}^3$ preserving ξ such that $\varphi(K)=K'$ — this is equivalent to saying that there is an isotopy from K to K' through Legendrian knots.

The contact structure $\xi_+=\ker(x\,dy+dz)$, where x,y,z are the coordinates in $\mathbb{R}^3$, will be referred to as *the standard contact structure*. If $\xi=\xi_+$ we often abbreviate "ξ–Legendrian" to "Legendrian".

In this paper we also deal with the contact structure

$$\xi_-=\ker(x\,dy-dz),$$

which is a mirror image of ξ_+.

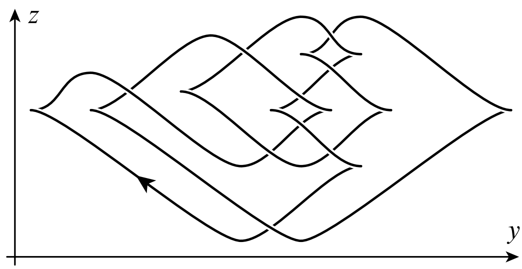


Figure 1: Front projection of a Legendrian knot.

We denote by $r_-, r_| \colon \mathbb{R}^3 \to \mathbb{R}^3$ the orthogonal reflections in the xy– and xz–planes, respectively,

$$r_-(x,y,z) = (x,y,-z), \quad r_|(x,y,z) = (x,-y,z).$$

Clearly, if K is a ξ_+–Legendrian knot, then $r_-(K)$ and $r_|(K)$ are ξ_-–Legendrian knots, and vice versa. It is also clear that the contact structures ξ_+ and ξ_- are invariant under the transformation $r_- \circ r_| \colon (x,y,z) \mapsto (x,-y,-z)$ (however, if the contact structures are endowed with an orientation, then the latter is flipped).

It is well known that a Legendrian knot in $\mathbb{R}^3$ is uniquely recovered from its *front projection*, which is defined as the projection to the yz–plane along the x–axis, provided that this projection is generic — a projection is *generic* if it has only finitely many cusps and only double self-intersections, which are also required to be disjoint from cusps. Note that a front projection always has cusps, since the tangent line to the projection cannot be parallel to the z–axis. Note also that at every double point of the projection, the arc having smaller slope dz/dy is overpassing.

An example of a generic front projection is shown in Figure 1.

There are two well-known integer invariants of Legendrian knots called the Thurston–Bennequin number and the rotation number. We recall their definitions.

Definition 1.2 *The Thurston–Bennequin number* $\mathrm{tb}(K)$ of a Legendrian knot K having generic front projection is defined as

$$\mathrm{tb}(K) = w(K) - \tfrac{1}{2}c(K),$$

where $w(K)$ is the writhe of the projection (that is, the algebraic number of double points), and $c(K)$ is the total number of cusps of the projection.

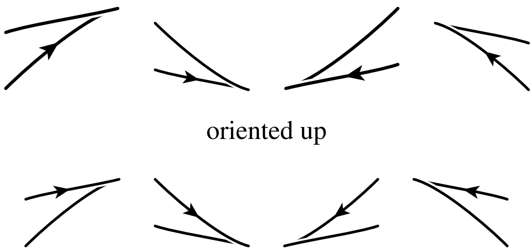


Figure 2: Cusps oriented up and down.

Definition 1.3 A cusp of a front projection is said to be *oriented up* if the outgoing arc appears above the incoming one, and *oriented down* otherwise; see Figure 2.

The rotation number $r(K)$ of a Legendrian knot K having generic front projection is defined as

$$r(K) = \tfrac{1}{2}(c_{\mathrm{down}}(K) - c_{\mathrm{up}}(K)),$$

where $c_{\mathrm{down}}(K)$ (resp. $c_{\mathrm{up}}(K)$) is the number of cusps of the front projection of K oriented down (resp. up).

For instance, if K is the Legendrian knot shown in Figure 1, then $\mathrm{tb}(K) = -10$ and $r(K) = 1$.

The topological meaning of tb and r is as follows. Let v be a normal vector field to ξ. Then $\mathrm{tb}(K)$ is the linking number $\mathrm{lk}(K, K')$, where K' is obtained from K by a small shift along v. The rotation number $r(K)$ is equal to the degree of the map $K \to \mathbb{S}^1$ defined in a local parametrization $(x(t), y(t), z(t))$ of K by

$$(x, y, z) \mapsto \frac{(\dot{x}, \dot{y})}{\sqrt{\dot{x}^2 + \dot{y}^2}}.$$

If K is a Legendrian knot, then by *the classical invariants of* K one means the topological type of K together with $\mathrm{tb}(K)$ and $r(K)$.

Sometimes the classical invariants determine the equivalence class of a Legendrian knot completely, in which case the knot is said to be Legendrian simple. This occurs, for instance, when K is an unknot [14; 15], a figure eight knot, or a torus knot [16]. But many examples of Legendrian nonsimple knots are known.

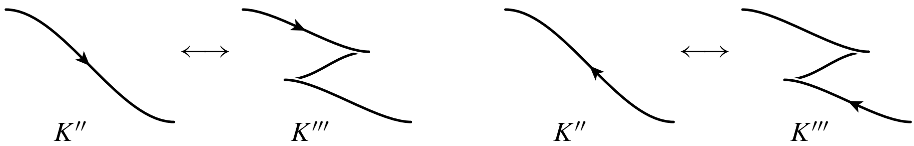


Figure 3: Stabilizations and destabilizations of Legendrian knots: positive, left, and negative, right.

Definition 1.4 Let K and K' be Legendrian knots. We say that K' is obtained from K by a *positive stabilization* (resp. *negative stabilization*), and K is obtained from K' by a *positive destabilization* (resp. *negative destabilization*), if there are Legendrian knots K'' and K''' equivalent to K and K', resp. such that the front projection of K''' is obtained from the front projection of K'' by a local modification shown in Figure 3, left (resp. right).

A positive (resp. negative) stabilization shifts the (tb, r) pair of the Legendrian knot by $(-1, 1)$ (resp. by $(-1, -1)$), so stabilizations and destabilizations always change the equivalence class of a Legendrian knot. If K is a Legendrian knot we denote by $S_+(K)$ (resp. $S_-(K)$) the result of a positive (resp. negative) stabilization applied to K.

One can see that the equivalence class of the Legendrian knot $S_+(K)$ is well defined. If $\mathscr{L}$ is an equivalence class of Legendrian knots, then by $S_+(\mathscr{L})$ (resp. $S_-(\mathscr{L})$) we denote the class $\{S_+(K) : K \in \mathscr{L}\}$ (resp. $\{S_-(K) : K \in \mathscr{L}\}$).

Remark 1.5 In the case of links having more than one component, the result of a stabilization, viewed up to Legendrian equivalence, depends on which component of the link the modification shown in Figure 3 is applied to, so the notation should be refined accordingly.

As shown in [21], any two Legendrian knots that have the same topological type can be obtained from one another by a sequence of stabilizations and destabilizations.

Definition 1.6 If K is a ξ_+–Legendrian or ξ_-–Legendrian knot then the image of K under the transformation $r_- \circ r_{|}$ is called *the Legendrian mirror of* K and denoted by $\mu(K)$.

Note that in terms of the respective front projections Legendrian mirroring is just a rotation by π around the origin. It preserves the Thurston–Bennequin number of the knot and reverses the sign of its rotation number. Thus, if K is a Legendrian knot with $r(K) = 0$, then K and $\mu(K)$ have the same classical invariants. However, it happens pretty often in this case that $\mu(K)$ and K are not equivalent Legendrian knots; see examples in Section 7.

Similarly, if K is a Legendrian knot whose topological type is invertible, then $-\mu(K)$ and K have the same classical invariants, but may not be equivalent Legendrian knots.

Definition 1.7 If K is a ξ_-–Legendrian knot, the Thurston–Bennequin and rotation numbers of K, as well as positive and negative stabilizations, are defined by using the mirror image $r_|(K)$ as follows:

$$\operatorname{tb}(K) = \operatorname{tb}(r_|(K)), \qquad r(K) = r(r_|(K)),$$
$$S_+(K) = r_|\big(S_+(r_|(K))\big), \quad S_-(K) = r_|\big(S_-(r_|(K))\big).$$

2 Annuli

Definition 2.1 Let K be a Legendrian knot, and let F be an oriented compact surface embedded in $\mathbb{R}^3$ such that $K \subset \partial F$ and the orientation of K agrees with the induced orientation of ∂F. Let also v be a normal vector field to ξ_+. The *Thurston–Bennequin number of K relative to F*, denoted by $\operatorname{tb}(K; F)$, is the intersection index of F with a knot obtained from K by a small shift along v.

If F is an arbitrary compact surface embedded in $\mathbb{R}^3$ such that $K \subset \partial F$, then $\operatorname{tb}(K; F)$ is defined as $\operatorname{tb}(K; F')$, where F' is the appropriately oriented intersection of a small tubular neighborhood U of K with F (the shift of K along v should then be chosen so small that the shifted knot does not escape from U).

Let K be a Legendrian knot, and let $F \subset \mathbb{R}^3$ be a compact surface such that $K \subset \partial F$. It is elementary to see that the following three conditions are equivalent:

(i) $\operatorname{tb}(K; F) = 0$.

(ii) F is isotopic relative to K to a surface F' such that F' is tangent to ξ_+ along K.

(iii) F is isotopic relative to K to a surface F' such that F' is transverse to ξ_+ along K.

In 3–dimensional contact topology, Giroux's convex surfaces play a fundamental role [24; 25; 26]. Especially important are convex annuli with Legendrian boundary and relative Thurston–Bennequin numbers of both boundary component equal to zero, since, vaguely speaking, any closed convex surface, viewed up to isotopy in the class of convex surfaces, can be built up from such annuli by gluing along a Legendrian graph.

Let $A \subset \mathbb{R}^3$ be an annulus with boundary consisting of two Legendrian knots K_1 and K_2 such that $\mathrm{tb}(K_1; A) = \mathrm{tb}(K_2; A) = 0$ and $\partial A = K_1 \cup (-K_2)$. Then the knots K_1 and K_2 have the same classical invariants, and it is natural to ask whether they must always be equivalent as Legendrian knots.

A quick look at this problem reveals no obvious reason why K_1 and K_2 must be equivalent, but constructing a counterexample appears to be tricky.

Theorem 8.1 of [27], which is given without a complete proof, implies that K_1 and K_2 are always equivalent Legendrian knots even in a more general situation in which $\mathbb{R}^3$ is replaced by an arbitrary tight contact 3–manifold.

However, counterexamples to this more general claim appeared earlier in a work of P Ghiggini [23] (without a special emphasis on the phenomenon), the simplest of which is as follows. Endow the three-dimensional torus $\mathbb{T}^3 = (\mathbb{R}/\mathbb{Z})^3$ with the contact structure $\xi = \ker(\sin(2\pi z)\,dx + \cos(2\pi z)\,dy)$, and take the annulus $(\mathbb{R}/\mathbb{Z}) \times \{0\} \times \left[0; \frac{1}{2}\right]$ for A. This annulus is clearly tangent to ξ along ∂A, but the boundary components are not Legendrian isotopic according to [23, Proposition 7.1]. (The fact that $(\mathbb{T}^3, \xi)$ is a tight contact manifold was established earlier by E Giroux.)

In this example, and in similar ones from [23], any connected component of ∂A can be taken to the other by a contactomorphism of $(\mathbb{T}^3, \xi)$. So, it is important here that the group of contactomorphisms of $(\mathbb{T}^3, \xi)$ is disconnected, which is not the case for the standard contact structure on $\mathbb{R}^3$. Another feature of this example is that the boundary components of A are not nullhomologous.

The following statement shows that the assertion of [27, Theorem 8.1] is false in the case of $\mathbb{R}^3$, too.

Theorem 2.2 *There is an oriented annulus $A \subset \mathbb{R}^3$ with boundary $\partial A = K_1 \cup (-K_2)$ such that K_1 and K_2 are nonequivalent Legendrian knots having zero Thurston–Bennequin number relative to A.*

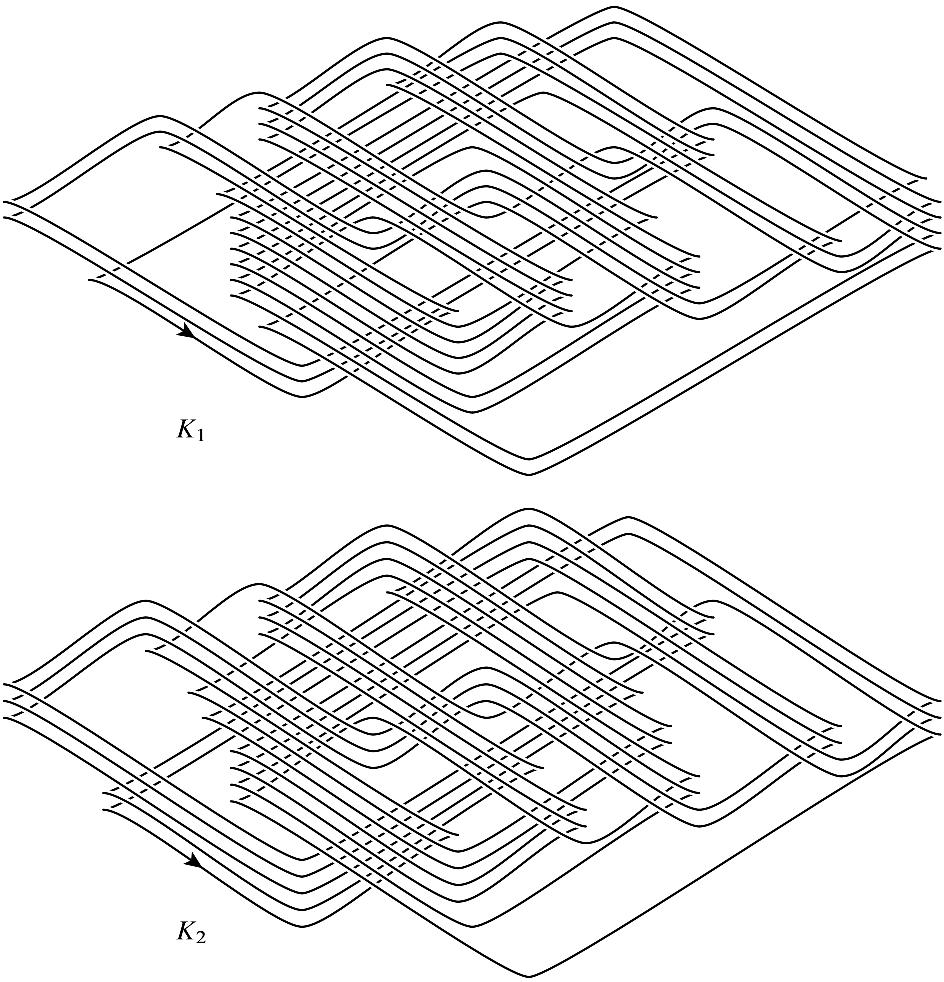


Figure 4: Nonequivalent Legendrian knots K_1 and K_2 cobounding an annulus A such that $\mathrm{tb}(K_1; A) = \mathrm{tb}(K_2; A) = 0$ and $\partial A = K_1 \cup (-K_2)$.

The proof is by producing an explicit example, and the example we use here is proposed by Dynnikov and Prasolov in [9]. Front projections of the Legendrian knots from this example are shown in Figure 4. It is shown in [9] that they cobound an embedded annulus such that

$$\mathrm{tb}(K_1; A) = \mathrm{tb}(K_2; A) = 0,$$

and it has remained unproved that K_1 and K_2 are not Legendrian equivalent.

The proof of Theorem 2.2 is given in Section 7.

3 $\mathbb{S}^3$ settings: the orientation-preserving symmetry group

By $\mathbb{S}^3$ we denote the unit 3–sphere in $\mathbb{R}^4$, which we identify with the group SU(2) in the standard way. We use the parametrization

$$(\theta,\varphi,\tau)\mapsto\begin{pmatrix}\cos\big(\tfrac12\pi\tau\big)e^{\mathbb{i}\varphi} & \sin\big(\tfrac12\pi\tau\big)e^{\mathbb{i}\theta}\\ -\sin\big(\tfrac12\pi\tau\big)e^{-\mathbb{i}\theta} & \cos\big(\tfrac12\pi\tau\big)e^{-\mathbb{i}\varphi}\end{pmatrix},$$

where $(\theta,\varphi,\tau)\in(\mathbb{R}/(2\pi\mathbb{Z}))\times(\mathbb{R}/(2\pi\mathbb{Z}))\times[0;1]$. The coordinate system (θ,φ,τ) can also be viewed as the one coming from the join construction $\mathbb{S}^3\cong\mathbb{S}^1*\mathbb{S}^1$, with θ the coordinate on $\mathbb{S}^1_{\tau=1}$, and φ on $\mathbb{S}^1_{\tau=0}$. Let α_+ be the right-invariant 1–form on $\mathbb{S}^3\cong\mathrm{SU}(2)$ given by

$$\alpha_+(X)=\tfrac12\operatorname{tr}\left(X^{-1}\begin{pmatrix}\mathbb{i} & 0\\ 0 & -\mathbb{i}\end{pmatrix}dX\right)=\sin^2\big(\tfrac12\pi\tau\big)\,d\theta+\cos^2\big(\tfrac12\pi\tau\big)\,d\varphi.$$

It is known [22] that, for any point $p\in\mathbb{S}^3$, there is a diffeomorphism ϕ from $\mathbb{R}^3$ to $\mathbb{S}^3\setminus\{p\}$ that takes the contact structure ξ_+ to the one defined by α_+, that is, to $\ker\alpha_+$. For this reason, the latter is denoted by ξ_+, too. Two Legendrian knots in $\mathbb{R}^3$ are equivalent if and only if so are their images under ϕ in $\mathbb{S}^3$. We will switch between the $\mathbb{R}^3$ and $\mathbb{S}^3$ settings depending on which is more suitable in the current context. The $\mathbb{R}^3$ settings are usually more visual, but sometimes are not appropriate. In particular, the definition of the knot symmetry group given below requires the $\mathbb{S}^3$ settings.

Definition 3.1 Let K be a smooth knot in $\mathbb{S}^3$. Denote by $\mathrm{Diff}^*(\mathbb{S}^3;K)$ the group of diffeomorphisms of $\mathbb{S}^3$ preserving the orientation of $\mathbb{S}^3$ and the orientation of K, and by $\mathrm{Diff}^*_0(\mathbb{S}^3;K)$ the connected component of this group containing the identity. The group $\mathrm{Diff}^*(\mathbb{S}^3;K)/\mathrm{Diff}^*_0(\mathbb{S}^3;K)$ is called *the orientation-preserving symmetry group of K* and denoted by $\mathrm{Sym}^*(K)$.

Clearly the group $\mathrm{Sym}^*(K)$ depends only on the topological type of K. In this paper we are dealing with knots K for which $\mathrm{Sym}^*(K)$ is a trivial group.

In the $\mathbb{S}^3$ settings, we also define the mirror image ξ_- of ξ_+ as

$$\xi_-=\ker\big(\sin^2\big(\tfrac12\pi\tau\big)\,d\theta-\cos^2\big(\tfrac12\pi\tau\big)\,d\varphi\big).$$

4 Rectangular diagrams of knots

We denote by $\mathbb{T}^2$ the two-dimensional torus $\mathbb{S}^1\times\mathbb{S}^1$, and by θ and φ the angular coordinates on the first and the second $\mathbb{S}^1$ factor, respectively.

Definition 4.1 *An oriented rectangular diagram of a link* is a finite subset $R \subset \mathbb{T}^2$ with an assignment $+$ or $-$ to every point in R such that every meridian $\{\theta\} \times \mathbb{S}^1$ and every longitude $\mathbb{S}^1 \times \{\varphi\}$ contains either no point or exactly two points from R, and in the latter case one of the points is assigned $+$ and the other $-$. The points in R are called *vertices* of R, and the pairs $\{u, v\} \subset R$ satisfying $\theta(u) = \theta(v)$ (resp. $\varphi(u) = \varphi(v)$) are called *vertical edges* (resp. *horizontal edges*) of R.

A rectangular diagram of a link is defined similarly, without assignment $+$ or $-$ to vertices.

An (oriented) rectangular diagram R of a link is called *an* (*oriented*) *rectangular diagram of a knot* if it is *connected* in the sense that, for any two vertices $v, v' \in R$, there exists a sequence $v_0 = v,\ v_1,\ v_2, \dots,\ v_k = v'$ of vertices of R such that any pair v_{i-1}, v_i of successive elements in it is an edge of R.

From the combinatorial point of view, oriented rectangular diagrams of links are the same thing as grid diagrams [32] viewed up to cyclic permutations of rows and columns. They are also nearly the same thing as arc-presentations; see [6].

Convention In this paper we mostly work with oriented knots and knot diagrams. For brevity, unless a rectangular diagram is explicitly specified as unoriented, it is assumed to be oriented.

With every rectangular diagram of a knot R one associates a knot, denoted by $\widehat{R}$, in $\mathbb{S}^3$ as follows. For a vertex $v \in R$, denote by $\hat{v}$ the image of the arc $v \times [0;1]$ in $\mathbb{S}^3 \cong \mathbb{S}^1 * \mathbb{S}^1 = (\mathbb{T}^2 \times [0;1])/\sim$ oriented from 0 to 1 if v is assigned $+$, and from 1 to 0 otherwise. The knot $\widehat{R}$ is by definition $\bigcup_{v \in V} \hat{v}$.

To get a planar diagram of a knot in $\mathbb{R}^3$ equivalent to $\widehat{R}$, one can proceed as follows. Cut the torus $\mathbb{T}^2$ along a meridian and a longitude not passing through a vertex of R to get a square. For every edge $\{u, v\}$ of R join u and v by a straight line segment, and let vertical segments overpass horizontal ones at every crossing point. Vertical edges are oriented from $+$ to $-$, and the horizontal ones from $-$ to $+$; see Figure 5. One can show (see [6]) that the obtained planar diagram represents a knot equivalent to $\widehat{R}$.

For two distinct points $x, y \in \mathbb{S}^1$, we denote by $[x; y]$ the arc of $\mathbb{S}^1$ such that, with respect to the standard orientation of $\mathbb{S}^1$, it has the starting point at x, and the endpoint at y.

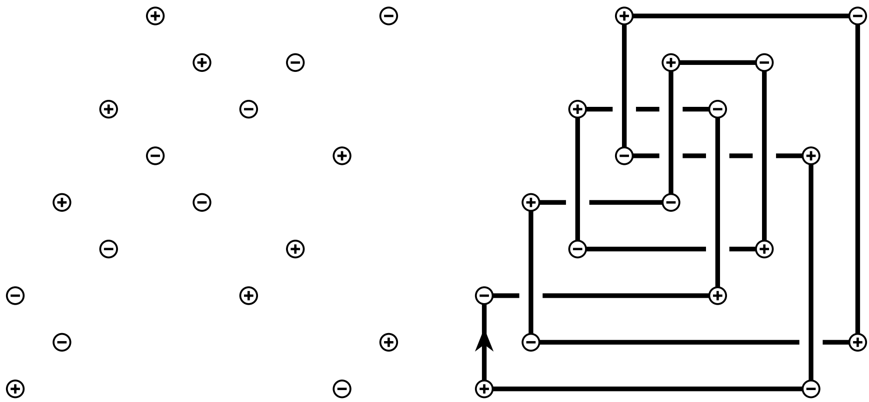


Figure 5: A rectangular diagram of a knot and the corresponding planar diagram.

Definition 4.2 Let R_1 and R_2 be rectangular diagrams of a knot such that, for some $\theta_1, \theta_2, \varphi_1, \varphi_2 \in \mathbb{S}^1$, we have:

(1) $\theta_1 \neq \theta_2$, $\varphi_1 \neq \varphi_2$.

(2) The symmetric difference $R_1 \triangle R_2$ is $\{\theta_1, \theta_2\} \times \{\varphi_1, \varphi_2\}$.

(3) $R_1 \triangle R_2$ contains an edge of one of the diagrams R_1, R_2.

(4) None of R_1 and R_2 is a subset of the other.

(5) The intersection of the rectangle $[\theta_1; \theta_2] \times [\varphi_1; \varphi_2]$ with $R_1 \cup R_2$ consists of its vertices, that is, $\{\theta_1, \theta_2\} \times \{\varphi_1, \varphi_2\}$.

(6) Each $v \in R_1 \cap R_2$ is assigned the same sign in R_1 as in R_2.

Then we say that the passage $R_1 \mapsto R_2$ is *an elementary move*.

An elementary move $R_1 \mapsto R_2$ is called

- *an exchange move* if $|R_1| = |R_2|$,
- *a stabilization move* if $|R_2| = |R_1| + 2$, and
- *a destabilization move* if $|R_2| = |R_1| - 2$,

where $|R|$ denotes the number of vertices of R.

We distinguish two *types* and four *oriented types* of stabilizations and destabilizations as follows.

Definition 4.3 Let $R_1 \mapsto R_2$ be a stabilization, and let θ_1, θ_2 and φ_1, φ_2 be as in Definition 4.2. Denote by V the set of vertices of the rectangle $[\theta_1; \theta_2] \times [\varphi_1; \varphi_2]$. We say that the stabilization $R_1 \mapsto R_2$ and the destabilization $R_2 \mapsto R_1$ are of *type* I (resp. of *type* II) if $R_1 \cap V \in \{(\theta_1, \varphi_1), (\theta_2, \varphi_2)\}$ (resp. $R_1 \cap V \in \{(\theta_1, \varphi_2), (\theta_2, \varphi_1)\}$).

Let $\varphi_0 \in \{\varphi_1, \varphi_2\}$ be such that $\{\theta_1, \theta_2\} \times \{\varphi_0\} \subset R_2$. The stabilization $R_1 \mapsto R_2$ and the destabilization $R_2 \mapsto R_1$ are of *oriented type* $\vec{\mathrm{I}}$ (resp. of *oriented type* $\vec{\mathrm{II}}$) if they are of type I (resp. of type II) and (θ_2, φ_0) is a positive vertex of R_2. The stabilization $R_1 \mapsto R_2$ and the destabilization $R_2 \mapsto R_1$ are of *oriented type* $\overleftarrow{\mathrm{I}}$ (resp. of *oriented type* $\overleftarrow{\mathrm{II}}$) if they are of type I (resp. of type II) and (θ_2, φ_0) is a negative vertex of R_2.

Our notation for stabilization types follows [8]. The correspondence with the notation of [37] is as follows:

notation of [8]	$\vec{\mathrm{I}}$	$\overleftarrow{\mathrm{I}}$	$\vec{\mathrm{II}}$	$\overleftarrow{\mathrm{II}}$
notation of [37]	X:*NE*, O:*SW*	X:*SW*, O:*NE*	X:*SE*, O:*NW*	X:*NW*, O:*SE*

With every rectangular diagram of a knot R we associate an equivalence class $\mathscr{L}_+(R)$ of ξ_+–Legendrian knots and an equivalence class $\mathscr{L}_-(R)$ of ξ_-–Legendrian knots as follows. The front projection of a representative of $\mathscr{L}_+(R)$ (resp. of $\mathscr{L}_-(R)$) is obtained from R in the following three steps:

(1) Produce a conventional planar diagram from R as described above.

(2) Rotate it counterclockwise (resp. clockwise) by any angle between 0 and $\pi/2$.

(3) Smooth out.

See Figure 6 for an example.

Theorem 4.4 [37] *Let R_1 and R_2 be rectangular diagrams of a knot. The classes $\mathscr{L}_+(R_1)$ and $\mathscr{L}_+(R_2)$ (resp. $\mathscr{L}_-(R_1)$ and $\mathscr{L}_-(R_2)$) coincide if and only if the diagrams R_1 and R_2 are related by a finite sequence of elementary moves in which all stabilizations and destabilizations are of type I (resp. of type II).*

Moreover, if $R_1 \mapsto R_2$ is a stabilization of oriented type T, then

$$\mathscr{L}_-(R_2) = \begin{cases} S_-(\mathscr{L}_-(R_1)) & \text{if } T = \overleftarrow{\mathrm{I}}, \\ S_+(\mathscr{L}_-(R_1)) & \text{if } T = \vec{\mathrm{I}}, \end{cases}$$

$$\mathscr{L}_+(R_2) = \begin{cases} S_+(\mathscr{L}_+(R_1)) & \text{if } T = \overleftarrow{\mathrm{II}}, \\ S_-(\mathscr{L}_+(R_1)) & \text{if } T = \vec{\mathrm{II}}. \end{cases}$$

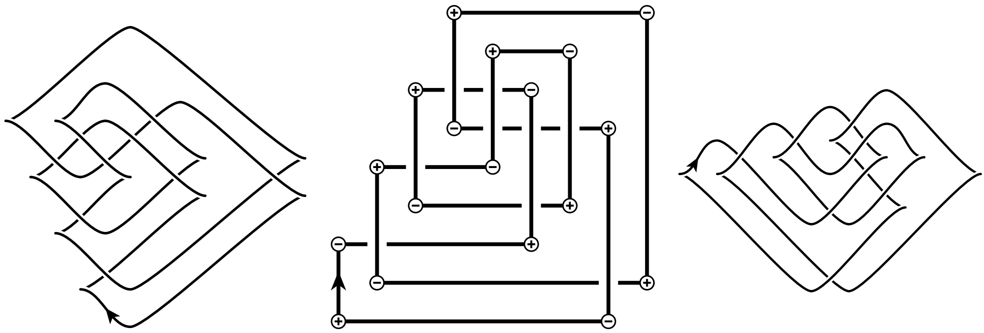


Figure 6: Legendrian knots associated with a rectangular diagram of a knot. Left: a representative of $\mathscr{L}_+(R)$. Center: R. Right: a representative of $\mathscr{L}_-(R)$.

The following is the key result of the present work.

Theorem 4.5 *Let K be a knot with trivial orientation-preserving symmetry group, and let R_1 and R_2 be rectangular diagrams of knots isotopic to K. Then the following two conditions are equivalent*:

(i) *We have $\mathscr{L}_+(R_1) = \mathscr{L}_+(R_2)$ and $\mathscr{L}_-(R_1) = \mathscr{L}_-(R_2)$.*

(ii) *The diagram R_2 can be obtained from R_1 by a sequence of exchange moves.*

The proof is given in the next section.

5 Rectangular diagrams of surfaces

Here we recall some definitions from [9; 10].

By a *rectangle* we mean a subset $r \subset \mathbb{T}^2$ of the form $[\theta_1;\theta_2]\times[\varphi_1;\varphi_2]$. Two rectangles r_1 and r_2 are said to be *compatible* if their intersection satisfies one of the following:

(1) $r_1 \cap r_2$ is empty.

(2) $r_1 \cap r_2$ is a subset of vertices of r_1 (equivalently, of r_2).

(3) $r_1 \cap r_2$ is a rectangle disjoint from the vertices of both rectangles r_1 and r_2.

Definition 5.1 *A rectangular diagram of a surface* is a collection $\Pi = \{r_1, \dots, r_k\}$ of pairwise compatible rectangles in $\mathbb{T}^2$ such that every meridian $\{\theta\}\times\mathbb{S}^1$ and every longitude $\mathbb{S}^1\times\{\varphi\}$ of the torus contains at most two free vertices, where by *a free vertex* we mean a point that is a vertex of exactly one rectangle in Π.

The set of all free vertices of Π is called *the boundary of* Π and denoted by $\partial\Pi$.

One can see that the boundary of a rectangular diagram of a surface is an unoriented rectangular diagram of a link. In particular, for any rectangle r, the boundary of $\{r\}$ is the set of vertices of r, and $\widehat{\partial\{r\}}$ is an unknot.

To every rectangular diagram of a surface Π we associate a C^1–smooth surface $\widehat{\Pi} \subset \mathbb{S}^3$ with piecewise smooth boundary, as we now describe.

By *the torus projection* we mean the map $\mathfrak{t}\colon \mathbb{S}^3 \setminus \left(\mathbb{S}^1_{\tau=1} \cup \mathbb{S}^1_{\tau=0}\right) \to \mathbb{T}^2$ defined by $(\theta, \varphi, \tau) \mapsto (\theta, \varphi)$. With every rectangle $r \subset \mathbb{T}^2$ one can associate a disc $\widehat{r} \subset \mathbb{S}^3$ having the form of a curved quadrilateral contained in $\overline{\mathfrak{t}^{-1}(r)}$ and spanning the loop $\widehat{\partial\{r\}}$ so that the following conditions hold:

(1) For each rectangle r, the restriction of $\mathfrak{t}$ to the interior of $\widehat{r}$ is a one-to-one map onto the interior of r.

(2) If r_1 and r_2 are compatible rectangles, then the interiors of $\widehat{r}_1$ and $\widehat{r}_2$ are disjoint.

(3) If $r = [\theta_1;\theta_2] \times [\varphi_1;\varphi_2]$, then $\widehat{r}$ is tangent to ξ_+ along the sides $\widehat{(\theta_1,\varphi_2)}$ and $\widehat{(\theta_2,\varphi_1)}$, and to ξ_- along the sides $\widehat{(\theta_1,\varphi_1)}$ and $\widehat{(\theta_2,\varphi_2)}$.

An explicit way to define the discs $\widehat{r}$, which are referred to as *tiles*, is given in [9, Section 2.3].

The surface $\widehat{\Pi}$ associated with a rectangular diagram of a surface Π is then defined as

$$\widehat{\Pi} = \bigcup_{r \in \Pi} \widehat{r}.$$

One can show that we have $\partial\widehat{\Pi} = \widehat{\partial\Pi}$ and, for any connected component R of $\partial\Pi$, the relative Thurston–Bennequin number $\mathrm{tb}_+(\widehat{R};\widehat{\Pi})$ (resp. $\mathrm{tb}_-(\widehat{R};\widehat{\Pi})$) equals minus half the number of vertices of R which are bottom-right or top-left (resp. bottom-left or top-right) vertices of some rectangles of Π.

On every rectangular diagram of a surface Π we introduce two binary relations, $\mathbin{\cdot^{\cdot}}$ and $\mathbin{^{\cdot}\cdot}$, which keep the information about which vertices are shared between two rectangles from Π. Namely, given $r_1, r_2 \in \Pi$, then $r_1 \mathbin{\cdot^{\cdot}} r_2$ means that r_1 and r_2 have the form

$$r_1 = [\theta_1;\theta_2] \times [\varphi_1;\varphi_2] \quad \text{and} \quad r_2 = [\theta_2;\theta_3] \times [\varphi_2;\varphi_3],$$

and $r_1 \mathbin{^{\cdot}\cdot} r_2$ means that r_1 and r_2 have the form

$$r_1 = [\theta_1;\theta_2] \times [\varphi_2;\varphi_3] \quad \text{and} \quad r_2 = [\theta_2;\theta_3] \times [\varphi_1;\varphi_2].$$

Proposition 5.2 *Let R_1 and R_2 be rectangular diagrams of a knot such that the knots $\widehat{R}_1$ and $\widehat{R}_2$ are topologically equivalent and have trivial orientation-preserving symmetry group. Suppose that $\mathscr{L}_+(R_1) = \mathscr{L}_+(R_2)$ and $\mathscr{L}_-(R_1) = \mathscr{L}_-(R_2)$.*

Then, for any rectangular diagram of a surface $\Pi = \{r_1, \dots, r_m\}$ such that $R_1 \subset \partial\Pi$, there exists a rectangular diagram of a surface $\Pi' = \{r'_1, \dots, r'_m\}$ and a rectangular diagram of a knot R'_2 such that

1. *R_2 and R'_2 are related by a sequence of exchange moves;*
2. *there exists an orientation-preserving self-homeomorphism of $\mathbb{S}^3$ that takes $\widehat{R}_1$ to $\widehat{R}'_2$ and $\widehat{r}_i$ to $\widehat{r}'_i$ for $i = 1, \dots, m$;*
3. *$r_i \mathrel{.^{\cdot}} r_j \iff r'_i \mathrel{.^{\cdot}} r'_j$ and $r_i \mathrel{^{\cdot}.} r_j \iff r'_i \mathrel{^{\cdot}.} r'_j$.*

Proof This statement is a consequence of the results of [10, Section 2], namely, of Theorems 2.1 and 2.2, as we will now see. The reader is referred to [10, Section 2] for the terminology that we use here.

Denote by $D = (\delta_+, \delta_-)$ a canonical dividing configuration of $\widehat{\Pi}$. By hypothesis we have $\mathrm{tb}_+(R_1) = \mathrm{tb}_+(R_2)$ and $\mathrm{tb}_-(R_1) = \mathrm{tb}_-(R_2)$, which implies that $\widehat{\Pi}$ is both $+$–compatible and $-$–compatible with R_2. By [10, Theorem 2.1], there exist a proper $+$–realization (Π_+, ϕ_+) of δ_+ and a proper $-$–realization (Π_-, ϕ_-) of δ_- at R_2.

Since the orientation-preserving symmetry group of $\widehat{R}_2$ is trivial, there is an isotopy from ϕ_+ to ϕ_- preserving $\widehat{R}_2$. One can clearly find a $-$–realization (Π_-, ϕ'_-) at R_2 of an abstract dividing set equivalent to δ_- such that there be an isotopy from ϕ_+ to ϕ'_- that fixes $\widehat{R}_2$ pointwise.

By [10, Theorem 2.2] this implies the existence of a proper realization (Π', ϕ) of D and a rectangular diagram of a knot R'_2 obtained from R_2 by a sequence of exchange moves, and such that $\phi(\widehat{R}_1) = \widehat{R}'_2$, which is just a reformulation of the assertion of Proposition 5.2. □

Definition 5.3 Two rectangular diagrams of a surface (or of a knot) are said to be *combinatorially equivalent* if one can be taken to the other by a homeomorphism $\mathbb{T}^2 \to \mathbb{T}^2 \cong \mathbb{S}^1 \times \mathbb{S}^1$ of the form $f \times g$, where f and g are orientation-preserving homeomorphisms of the circle $\mathbb{S}^1$.

Let Π be a rectangular diagram of a surface. The relations $\mathrel{.^{\cdot}}$ and $\mathrel{^{\cdot}.}$ on Π defined above constitute what is called in [10] the (equivalence class of a) *dividing code* of Π. In other words, two diagrams Π_1 and Π_2 have equivalent dividing codes if there is a bijection $\Pi_1 \to \Pi_2$ that preserves the relations $\mathrel{\cdot^{\textstyle\cdot}}$ and $\mathrel{^{\textstyle\cdot}\cdot}$. In general, this does not imply that the diagrams Π_1 and Π_2 are combinatorially equivalent; see [10, Figure 2.2] for an example.

Lemma 5.4 *For any rectangular diagram of a link R, there exists a rectangular diagram of a surface Π such that*

(1) $R \subset \partial\Pi$,

(2) *whenever a rectangular diagram of a surface Π' has the same dividing code as Π has, the diagrams Π and Π' are combinatorially equivalent.*

Proof For simplicity we assume that R is connected. In the case of a many-component link the proof is essentially the same, but a cosmetic change of notation is needed.

Let

$$(\theta_1, \varphi_1),\ (\theta_1, \varphi_2),\ (\theta_2, \varphi_2), \dots, (\theta_{n-1}, \varphi_n),\ (\theta_n, \varphi_n),\ (\theta_n, \varphi_1)$$

be the vertices of R. We put $\theta_0 = \theta_n$ and $\varphi_0 = \varphi_n$.

Pick an $\varepsilon > 0$ not larger than the length of any of the intervals $[\theta_i; \theta_j]$ and $[\varphi_i; \varphi_j]$ with $i \neq j$. For $i \in \{1, 2, \dots, n\}$ and $j \in \{0, 1, 2, 3, 4, 5\}$ write

$$\theta_{i,j} = \theta_i + \tfrac{1}{6} j\varepsilon \quad \text{and} \quad \varphi_{i,j} = \varphi_i + \tfrac{1}{6} j\varepsilon.$$

The sought-for diagram Π is constructed in the following four steps, illustrated in Figure 7.

Step 1 Put

$$\Pi_1 = \{[\theta_{i,0}; \theta_{i,3}] \times [\varphi_{i,3}; \varphi_{i+1,0}], [\theta_{i,3}; \theta_{i+1,0}] \times [\varphi_{i+1,0}; \varphi_{i+1,3}]\}_{i=0,1,\dots,n}.$$

Step 2 A rectangular diagram of a surface is uniquely defined by the union of its rectangles. Define Π_2 so that

$$\bigcup_{r \in \Pi_2} r = \overline{\bigcup_{r \in \Pi_1} r \setminus \bigcup_{r, r' \in \Pi_1;\, r \neq r'} (r \cap r')}.$$

Step 3 Define Π_3 by

$$\bigcup_{r \in \Pi_3} r = \overline{\bigcup_{r \in \Pi_2} r \mathbin{\triangle} \bigcup_{i=1}^{n} \bigl(([\theta_{i,1}; \theta_{i,2}] \cup [\theta_{i,4}; \theta_{i,5}]) \times \mathbb{S}^1}\bigr).$$

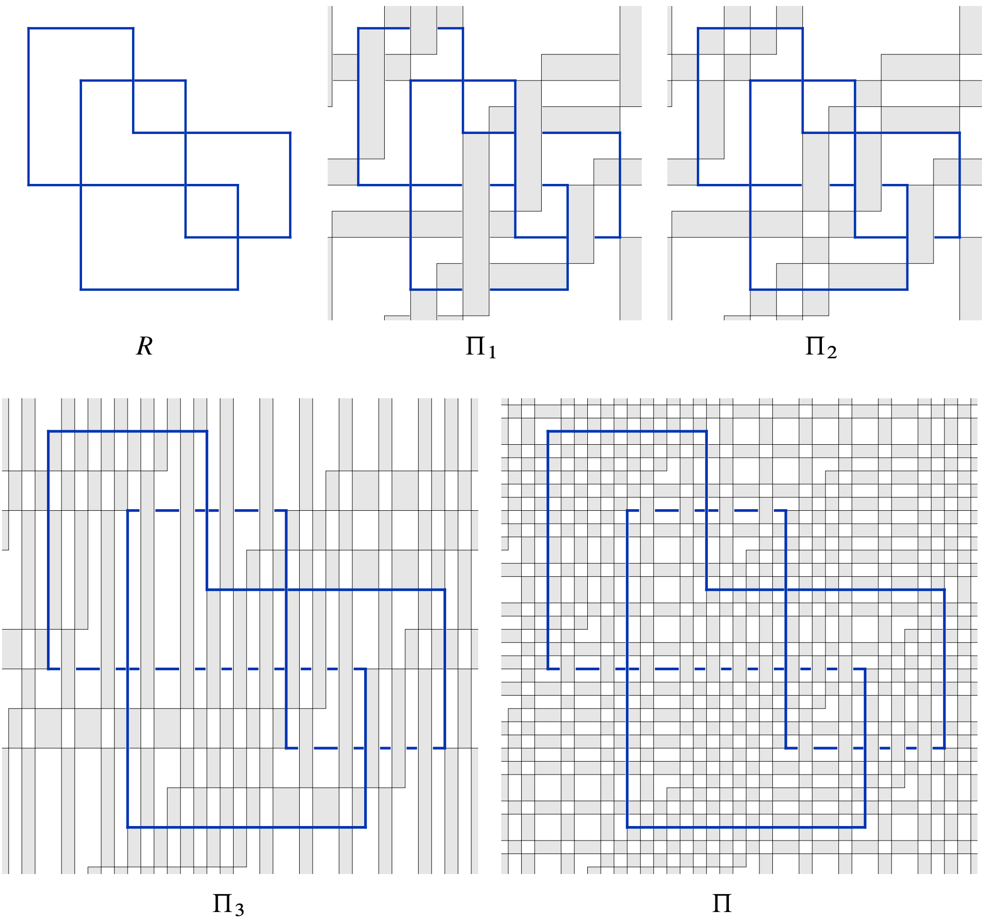


Figure 7: Constructing the diagram Π in the proof of Lemma 5.4.

Step 4 Finally, Π is defined by

$$\bigcup_{r\in\Pi} r = \overline{\bigcup_{r\in\Pi_3} r \triangle \bigcup_{i=1}^{n}\big(\mathbb{S}^1\times([\varphi_{i,1};\varphi_{i,2}]\cup[\varphi_{i,4};\varphi_{i,5}])\big)}.$$

One can see that $R\subset\partial\Pi_1=\partial\Pi_2=\partial\Pi_3=\partial\Pi$. We claim that the combinatorial type of Π is uniquely recovered from the dividing code of Π.

Indeed, suppose we have forgotten the values of $\theta_{i,j}$ and $\varphi_{i,j}$, and keep only the information about which pairs $(\theta_{i,j},\varphi_{i',j'})$ are vertices of which rectangles in Π (this information is extracted from the dividing code).

For any $i \in \{1, 2, \dots, n\}$ and $j \in \{1, 2, 4, 5\}$, the point $(\theta_{i,j}, \varphi_{1,1})$ is a vertex of some rectangle in Π. Hence the cyclic order on $\{\theta_{i,j}\}_{i\in\{1,2,\dots,n\};\, j\in\{1,2,4,5\}} \subset \mathbb{S}^1$ is prescribed by the dividing code.

For each $i \in \{1, 2, \dots, n\}$, let i^- denote the unique element of $\{1, 2, \dots, n\}$ such that $(\theta_{i^-}; \theta_i) \times \mathbb{S}^1$ does not contain vertices of R. One can see that for any $i \in \{1, 2, \dots, n\}$ there exist j and j' in $\{1, 2, \dots, n\}$ such that $(\theta_{i^-,5}, \varphi_{j,1})$, $(\theta_{i,0}, \varphi_{j,1})$, $(\theta_{i,1}, \varphi_{j,1})$, $(\theta_{i,2}, \varphi_{j',1})$, $(\theta_{i,3}, \varphi_{j',1})$, $(\theta_{i,4}, \varphi_{j',1})$ are vertices of some rectangles in Π. This prescribes the cyclic order on $\{\theta_{i^-,5}, \theta_{i,0}, \theta_{i,1}\}$ and $\{\theta_{i,2}, \theta_{i,3}, \theta_{i,4}\}$ for any i. Therefore, the cyclic order on $\{\theta_{i,j}\}_{i\in\{1,2,\dots,n\};\, j\in\{0,1,2,3,4,5\}}$ is completely determined by the dividing code.

Similarly, the cyclic order on $\{\varphi_{i,j}\}_{i\in\{1,2,\dots,n\};\, j\in\{0,1,2,3,4,5\}}$ is completely determined by the dividing code, and hence so is the combinatorial type of Π. □

Proof of Theorem 4.5 By Lemma 5.4 we can find a rectangular diagram of a surface Π such that $R_1 \subset \partial\Pi$ and the combinatorial type of Π is determined by the dividing code of Π. We pick such a Π and apply Proposition 5.2. Since the combinatorial type of Π is determined by the dividing code of Π, we may strengthen the assertion of Proposition 5.2 in this case by claiming additionally that $\Pi' = \Pi$ and $R_2' = R_1$, which implies the assertion of the theorem. □

6 Triviality of the orientation-preserving symmetry groups of some knots

We use Rolfsen's knot notation [39]. Knots with crossing number $\leqslant 10$ are well-studied (see [29; 30]), and the existing results about them imply the following.

Proposition 6.1 *The orientation-preserving symmetry group of each of the knots 9_{42}, 9_{43}, 9_{44}, 9_{45}, 10_{128} and 10_{160} is trivial.*

The concrete sources for this statement are as follows. All knots listed in Proposition 6.1 are known to be invertible (this can be seen from their pictures in [39]), so the assertion is equivalent to saying that the symmetry group of each of the knots is $\mathbb{Z}_2$.

The knots 9_{42}, 9_{43}, 9_{44}, 9_{45} and 10_{128} are Montesinos knots (introduced in [33]):

$$9_{42} = K\left(\tfrac{2}{5}, \tfrac{1}{3}, -\tfrac{1}{2}\right), \quad 9_{43} = K\left(\tfrac{3}{5}, \tfrac{1}{3}, -\tfrac{1}{2}\right),$$
$$9_{44} = K\left(\tfrac{2}{5}, \tfrac{2}{3}, -\tfrac{1}{2}\right), \quad 9_{45} = K\left(\tfrac{3}{5}, \tfrac{2}{3}, -\tfrac{1}{2}\right),$$
$$10_{128} = K\left(\tfrac{3}{7}, \tfrac{1}{3}, -\tfrac{1}{2}\right).$$

The knots 9_{42}, 9_{43}, 9_{44}, 9_{45} are elliptic Montesinos knots, for which the symmetry group was computed by M Sakuma [40]. The symmetry group of the knot 10_{128} was computed by M Boileau and B Zimmermann [1]. Both works are based on the technique which is due to F Bonahon and L Siebenmann [2].

The fact that the knot 10_{160} is not periodic was established by U Lüdicke [31], and that it is not freely periodic was shown by R Hartley [28].

Proposition 6.2 *The orientation-preserving symmetry group of the* (*topologically equivalent*) *knots K_1 and K_2 in Figure 4 is trivial.*

Proof We use the classical methods of the above-mentioned works with some technical improvements needed for reducing the amount of computation. "A direct check" below refers to a computation that requires only a few minutes of a modern computer's processor time and standard well-known algorithms.

The first direct check is to see that the Alexander polynomial of K_1 and K_2 is

$$\Delta(t)=t^{20}-t^{19}+t^{18}-3t^{17}+3t^{16}-5t^{15}+10t^{14}-5t^{13}+6t^{12}-14t^{11} \\ +15t^{10}-14t^{9}+6t^{8}-5t^{7}+10t^{6}-5t^{5}+3t^{4}-3t^{3}+t^{2}-t+1. \tag{1}$$

According to Murasugi [34], if a knot has period p with p prime, then the Alexander polynomial of this knot reduced modulo p is either the p^{th} power of a polynomial with coefficients in $\mathbb{Z}_p$ or has a factor of the form $(1+t+\cdots+t^d)^{p-1}$, where $d\geqslant 1$. It is a direct check that neither of these occurs in the case of the polynomial (1) for prime $p\leqslant 19$, and for $p>19$ the corresponding verification is trivial.

By Hartley [28], to prove that our knot has not a free period equal to p it suffices to ensure that $\Delta(t^p)$ does not have a self-reciprocal factor of degree $\deg\Delta(t)=20$. For prime $p<100$, it can be checked directly that $\Delta(t^p)$ is irreducible.

Suppose, for some prime $p>100$, we have a factorization $\Delta(t^p)=f(t)\cdot g(t)$ with self-reciprocal $f(t),g(t)\in\mathbb{Z}[t]$ such that $\deg f=20$. Since $\Delta(0)=1$ we may assume that $f(0)=1$ without loss of generality. For a self-reciprocal polynomial $q(t)$ of even degree, we denote by $\tilde{q}(t)$ the Laurent polynomial $t^{-(\deg q)/2}q(t)$.

For any $\alpha\in\{1,e^{\pi\mathring{\imath}/3},\mathring{\imath},e^{2\pi\mathring{\imath}/3},-1\}$, we have

(i) $\alpha^p\in\{\alpha,\overline{\alpha}\}$,

(ii) $\widetilde{\Delta}(\alpha)=\widetilde{\Delta}(\overline{\alpha})$ and $\widetilde{f}(\alpha)=\widetilde{f}(\overline{\alpha})$,

(iii) $\Delta(\alpha), f(\alpha), g(\alpha)\in\mathbb{Z}$.

For $a = (a_1, a_2, a_3, a_4, a_5)$, let $\ell_a(t) \in \mathbb{R}[t]$ be a self-reciprocal polynomial of even degree not exceeding 8 such that $\widetilde{\ell}_a(t)$ takes the values a_1, a_2, a_3, a_4, a_5 at the points $t = 1, e^{\pi \mathring{\imath}/3}, \mathring{\imath}, e^{2\pi \mathring{\imath}/3}, -1$, respectively. This polynomial is clearly unique.

Now let $a \in \mathbb{Z}^5$ be the list of values of $\widetilde{f}$ at the points $1, e^{\pi \mathring{\imath}/3}, \mathring{\imath}, e^{2\pi \mathring{\imath}/3}, -1$. Then the polynomial $t^{10}(\widetilde{f}(t) - \widetilde{\ell}(t))$ is divisible by $(t^6-1)(t^2+1)$. Since this polynomial is also self-reciprocal, it is actually divisible by $(t^6-1)(t^2+1)(t-1)$. Thus, we have

$$(2)\quad f(t) = t^{10}\,\widetilde{\ell}_a(t) + (t^6-1)(t^2+1)(t-1) \times \big(t^{11} + b_1 t^{10} + b_2 t^9 + b_3 t^8 + b_4 t^7 + b_5 t^6 + b_5 t^5 + b_4 t^4 + b_3 t^3 + b_2 t^2 + b_1 t + 1\big).$$

Since $\widetilde{\ell}_a$ may have nonzero coefficients only in front of t^k with $k \in [-4; 4]$, we see that $f(t) \in \mathbb{Z}[t]$ implies $b_i \in \mathbb{Z}$ for $i = 1, \dots, 5$, and $\ell_a(t) \in \mathbb{Z}[t]$.

One easily finds that the values of $\widetilde{\Delta}(t)$ at the points $t = 1, e^{\pi \mathring{\imath}/3}, \mathring{\imath}, e^{2\pi \mathring{\imath}/3}, -1$ are $1, -7, 17, 13, 113$, respectively. Therefore, a_1, a_2, a_3, a_4, a_5 must be divisors of $1, -7, 13, 13, 113$, respectively. Together with the condition $\ell_a(t) \in \mathbb{Z}[t]$ this leaves us only the 32 options for a listed in Table 1.

It is another direct check that all roots of Δ are located inside the circle $\{z \in \mathbb{C} : |z| < \frac{3}{2}\}$. Therefore, the roots of f are contained in the circle $\{z \in \mathbb{C} : |z| < \left(\frac{3}{2}\right)^{1/p}\}$.

For $k \in \mathbb{N}$, denote by p_k the k^{th} Newton sum of f, that is, the sum of the k^{th} powers of the roots. They must be integers, and their absolute values are estimated by

$$(3)\qquad |p_k| < 20 \cdot \left(\tfrac{3}{2}\right)^{k/p}.$$

Since $p > 100$, this implies, in particular, that

$$(4)\qquad |p_k| \leqslant 20 \quad \text{for } k = 1, 2, 3, 4, 5.$$

Let c_k, $k = 1, 2, \dots, 19$, denote the coefficients of f, so

$$f = 1 + c_1 t + c_2 t^2 + \dots + c_{19} t^{19} + t^{20} \quad \text{with } c_i = c_{20-i}.$$

The first (equivalently, the last) five of them are related with p_i by Newton's identities

$$\begin{aligned}
-p_1 &= c_1,\\
-p_2 &= c_1 p_1 + 2c_2,\\
-p_3 &= c_1 p_2 + c_2 p_1 + 3c_3,\\
-p_4 &= c_1 p_3 + c_2 p_2 + c_3 p_1 + 4c_4,\\
-p_5 &= c_1 p_4 + c_2 p_3 + c_3 p_2 + c_4 p_1 + 5c_5.
\end{aligned}$$

a	$l_a(t)$
$\pm(1,1,1,1,1)$	± 1
$\pm(1,1,1,13,1)$	$\pm(-2t^8+2t^7-2t^5+5t^4-2t^3+2t-2)$
$\pm(1,1,17,1,1)$	$\pm(4t^8-4t^6+t^4-4t^2+4)$
$\pm(1,1,17,13,1)$	$\pm(2t^8+2t^7-4t^6-2t^5+5t^4-2t^3-4t^2+2t+2)$
$\pm(1,-1,1,1,113)$	$\pm(5t^8-9t^7+14t^6-19t^5+19t^4-19t^3+14t^2-9t+5)$
$\pm(1,-1,1,13,113)$	$\pm(3t^8-7t^7+14t^6-21t^5+23t^4-21t^3+14t^2-7t+3)$
$\pm(1,-1,17,1,113)$	$\pm(9t^8-9t^7+10t^6-19t^5+19t^4-19t^3+10t^2-9t+9)$
$\pm(1,-1,17,13,113)$	$\pm(7t^8-7t^7+10t^6-21t^5+23t^4-21t^3+10t^2-7t+7)$
$\pm(1,7,1,1,1)$	$\pm(-t^8-t^7+t^5+3t^4+t^3-t-1)$
$\pm(1,7,1,13,1)$	$\pm(-3t^8+t^7-t^5+7t^4-t^3+t-3)$
$\pm(1,7,17,1,1)$	$\pm(3t^8-t^7-4t^6+t^5+3t^4+t^3-4t^2-t+3)$
$\pm(1,7,17,13,1)$	$\pm(t^8+t^7-4t^6-t^5+7t^4-t^3-4t^2+t+1)$
$\pm(1,-7,1,1,113)$	$\pm(6t^8-8t^7+14t^6-20t^5+17t^4-20t^3+14t^2-8t+6)$
$\pm(1,-7,1,13,113)$	$\pm(4t^8-6t^7+14t^6-22t^5+21t^4-22t^3+14t^2-6t+4)$
$\pm(1,-7,17,1,113)$	$\pm(10t^8-8t^7+10t^6-20t^5+17t^4-20t^3+10t^2-8t+10)$
$\pm(1,-7,17,13,113)$	$\pm(8t^8-6t^7+10t^6-22t^5+21t^4-22t^3+10t^2-6t+8)$

Table 1

This Diophantine system has exactly 971 865 solutions satisfying (4), which can be searched (another direct check). The coefficients b_1, b_2, b_3, b_4, b_5 in (2) can obviously be expressed through c_1, c_2, c_3, c_4, c_5. Thus, we get only $32 \cdot 971865 = 31\,099\,680$ possible candidates for f, and it is the last direct check that the k^{th} Newton sum of each of the obtained polynomials violates (3) for some $k \leqslant 31$ with any $p > 100$; a contradiction.

We have thus established that the orientation-preserving symmetry group of the knots K_1 and K_2 has no finite-order elements. It remains to ensure that these knots are not satellite knots; that is, they are hyperbolic. A way to verify this is explained in the appendix. □

Proposition 6.2 is also directly confirmed by the SnapPy program [5]. For the reader's convenience, we provide here a Dowker–Thistlethwaite code of the diagram of K_1 shown in Figure 4 (the numeration of the crossings starts from the arrowhead):

$-462, -346, -76, -218, 156, 472, 356, 66, 208, 126, -324, 444, 132, 202,$
$60, 362, 180, -478, -338, -284, -452, 246, 302, 188, -460, 400, -296, -492,$
$-450, -286, -230, -88, -172, -122, -418, 352, 468, 160, -276, 220, 154, 474,$
$334, 384, 412, 502, -442, -24, 134, 200, 58, 40, 146, -366, -184, -222,$

$-80, 8, 314, 264, 380, 416, 506, -168, -92, -234, -290, -446, 196, 54,$
$456, -488, -46, -282, -340, -480, -430, -272, -116, -424, -372, -256, 508, 20,$
$504, 414, 382, 266, -84, -226, 36, 150, 278, 344, -396, -458, -52, -294,$
$-494, -448, -288, -232, -90, -170, -124, 68, 354, 470, 158, -274, -428, 152,$
$476, 336, 386, 410, 500, 94, -22, 136, 198, 56, 42, -392, -140, 192,$
$50, 2, 320, -350, -72, -214, -12, -332, 176, 312, 114, 426, -482, -342,$
$106, 304, 388, 408, 498, -96, -238, 402, 250, -142, -394, -104, -364, -182,$
$-224, -82, -10, -216, -74, -348, -464, 166, 128, 206, 64, 358, -268, -434,$
$-32, 306, 108, 368, -484, 162, 466, -376, -420, -120, -174, -86, -228, 38,$
$148, 280, -186, 4, 318, 260, -70, -212, -14, -330, -436, -30, -102, -244,$
$454, 144, 486, -252, 194, -138, -242, -100, -28, -438, -328, -16, -210, 378,$
$262, 316, 6, -78, 112, 310, 178, 360, 62, 204, 130, -236, 292, 496,$
$406, 390, -44, -490, -298, 398, 254, 164, -258, -374, -422, -118, -270, -432,$
$-34, 308, 110, 370, -48, 190, 300, 248, 404, -240, -98, -26, -440, -326,$
$-18, -322.$

7 Applications

Theorem 7.1 *There exists an algorithm that decides in finite time whether or not two given Legendrian knots, L_1 and L_2, say, are equivalent, provided that they are topologically equivalent and have trivial orientation-preserving symmetry group.*

Proof It is understood that L_1 and L_2 are presented in a combinatorial way that allows one to recover actual curves in $\mathbb{R}^3$. Whichever presentation is chosen, it can always be converted into rectangular diagrams. So, we assume that we are given two rectangular diagrams of a knot, R_1 and R_2, say, such that $\mathscr{L}_+(R_1) \ni L_1$ and $\mathscr{L}_+(R_2) \ni L_2$.

By [8, Theorem 7] there exists a rectangular diagram of a knot R_3 such that $\mathscr{L}_+(R_3) = \mathscr{L}_+(R_1)$ and $\mathscr{L}_-(R_3) = \mathscr{L}_-(R_2)$. By Theorem 4.4 this is equivalent to saying that there exists a sequence of elementary moves transforming R_1 to R_3 (resp. R_3 to R_2) including only exchange moves and type I (resp. type II) stabilizations and destabilizations. Therefore, such an R_3 can be found by an exhaustive search of sequences of elementary moves starting at R_1 in which all type I stabilizations and destabilizations occur before all type II ones. Indeed, the combinatorial types of such sequences are enumerable.The search terminates once a sequence with the above properties arriving at R_2 is encountered. By [8, Theorem 7] this must eventually happen.

Once R_3 is found we check whether or not it is related to R_2 by a sequence of exchange moves. The latter can produce only finitely many combinatorial types of diagrams from the given one, so this process is finite. According to Theorem 4.5, the diagrams R_2 and R_3 are related by a sequence of exchange moves if and only if $\mathscr{L}_+(R_2) = \mathscr{L}_+(R_3)$, which is equivalent to $\mathscr{L}_+(R_1) = \mathscr{L}_+(R_2)$. □

Now we use Theorem 4.5 to establish some facts that are left in [4] as conjectures. These involve knots with trivial orientation-preserving symmetry group, which are listed in Proposition 6.1 above.

For a rectangular diagram of a knot R, the set of all rectangular diagrams obtained from R by a sequence of exchange moves is called *the exchange class of R*.

In what follows we use the following notation system. ξ_+–Legendrian classes of knots having topological type m_n are denoted by m_n^{k+}, $k = 1, 2, \dots$, or simply m_n^+ if we need to consider only one Legendrian class and its images under μ and orientation reversal. Similarly, for ξ_-–Legendrian classes we use notation of the form m_n^{k-} or m_n^-, and for exchange classes $m_n^{k\mathrm{R}}$ or m_n^{R}.

The $\xi_\pm$–Legendrian classes and exchange classes of interest to us are defined by specifying a representative. In order to help the reader to see the correspondence with the notation of [4], we define the ξ_-–Legendrian classes via their mirror images, which are ξ_+–Legendrian classes.

We use the same notation for natural operations on (exchange classes of) rectangular diagrams as for Legendrian knots: $-$ for orientation reversal, $r_|$ and r_- for the horizontal and the vertical flip, respectively, and μ for $r_| \circ r_-$. One can see that if X is an exchange class, then $\mathscr{L}_\pm(-X) = -\mathscr{L}_\pm(X)$, $\mathscr{L}_\pm(\mu(X)) = \mu(\mathscr{L}_\pm(X))$ and $\mathscr{L}_\pm(r_|(X)) = r_|(\mathscr{L}_\mp(X))$.

Proposition 7.2 *For the classes 9_{42}^+ and 9_{42}^-, whose representatives are shown in Figure 8, we have $9_{42}^+ = -9_{42}^+ \neq \mu(9_{42}^+)$ and $9_{42}^- \neq -9_{42}^- = \mu(9_{42}^-)$.*

Proof We use the exchange class 9_{42}^{R} of the diagram shown in Figure 8 on the right. Black vertices are positive, and white ones are negative.

It is an easy check that the diagram representing the class 9_{42}^{R} in Figure 8 admits no nontrivial (that is, changing the combinatorial type) exchange move, and its combinatorial type changes under reversing the orientation and under its composition with the rotation μ. We conclude from this that

$$9_{42}^{\mathrm{R}} \neq -9_{42}^{\mathrm{R}} \quad \text{and} \quad -9_{42}^{\mathrm{R}} \neq \mu(9_{42}^{\mathrm{R}}). \tag{5}$$

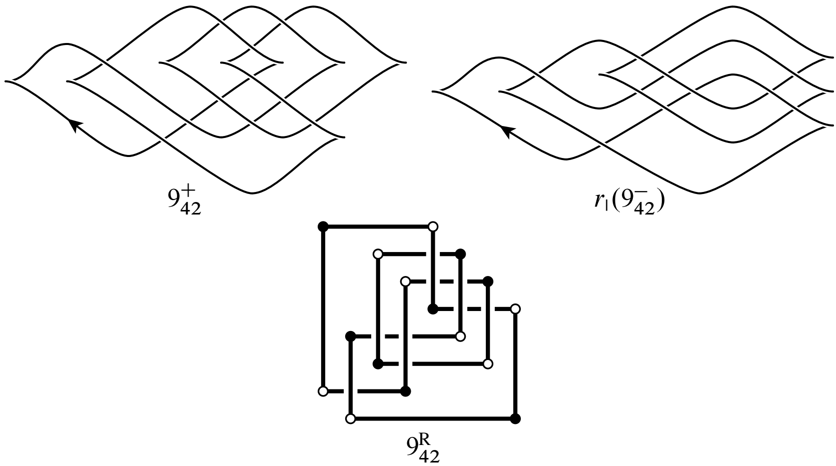
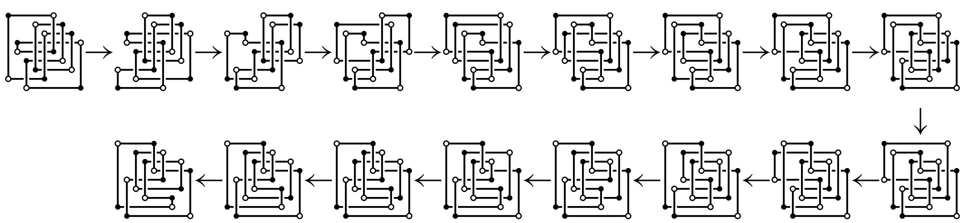
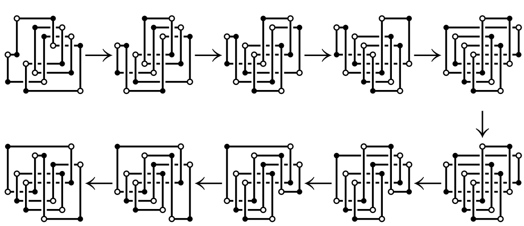


Figure 8: Legendrian knots in Proposition 7.2 and an exchange class representing both.

Now we verify directly that $S_{\vec{\mathrm{I}}}(9^{\mathrm{R}}_{42}) = S_{\vec{\mathrm{I}}}(-9^{\mathrm{R}}_{42})$:



And $S_{\vec{\mathrm{II}}}(-9^{\mathrm{R}}_{42}) = S_{\vec{\mathrm{II}}}(\mu(9^{\mathrm{R}}_{42}))$:



By Theorem 4.4 this implies

$$\mathscr{L}_{+}(9^{\mathrm{R}}_{42}) = \mathscr{L}_{+}(-9^{\mathrm{R}}_{42}) \quad \text{and} \quad \mathscr{L}_{-}(-9^{\mathrm{R}}_{42}) = \mathscr{L}_{-}(\mu(9^{\mathrm{R}}_{42})). \tag{6}$$

From Proposition 6.1 and Theorem 4.5 we conclude that

$$(5) \text{ and } (6) \quad \Longrightarrow \quad \mathscr{L}_{-}(9^{\mathrm{R}}_{42}) \neq \mathscr{L}_{-}(-9^{\mathrm{R}}_{42}) \text{ and } \mathscr{L}_{+}(-9^{\mathrm{R}}_{42}) \neq \mathscr{L}_{+}(\mu(9^{\mathrm{R}}_{42})).$$

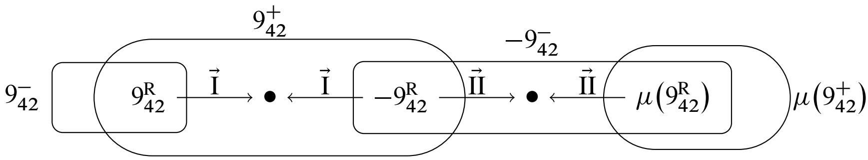
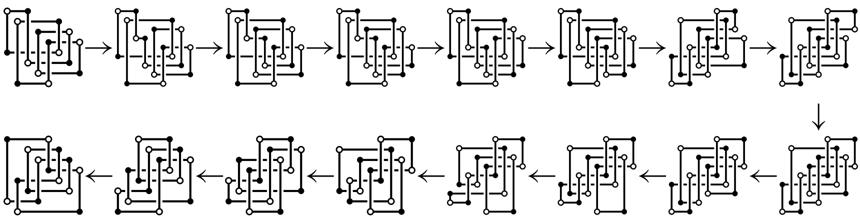
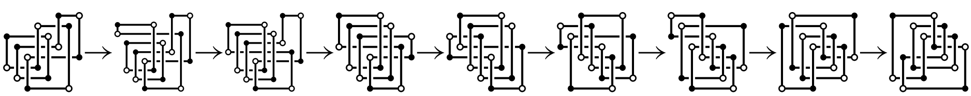


Figure 9: Proof of Proposition 7.2.

The front projections in Figure 8 are obtained as described in Section 4 (see Figure 6) from rectangular diagrams that can be easily guessed from the pictures. We find using Theorem 4.4 that $9^+_{42} = \mathscr{L}_+(9^{\mathrm{R}}_{42})$:



And $9^-_{42} = \mathscr{L}_-(9^{\mathrm{R}}_{42})$:



This completes the proof. □

The proof of Proposition 7.2 is summarized in Figure 9. In what follows we present the proofs by similar schemes, omitting the verbal description. For a routine check of all equalities and inequalities of exchange classes used in the proofs, the reader is referred to [12].

In the proofs of Propositions 7.4, 7.6 and 7.8, we may also silently use symmetries: an inequality $X \neq Y$, where X and Y are some Legendrian or exchange classes, is equivalent to either of $-X \neq -Y$ and $\mu(X) \neq \mu(Y)$. Another use of symmetries is as follows. If X and Y are Legendrian classes such that $X = -X$ and $Y \neq -Y$ (similarly for μ or $-\mu$ in place of $-$), then we immediately know that $X \notin \{Y, -Y, \mu(Y), -\mu(Y)\}$.

Proposition 7.3 *For the $\xi_\pm$–Legendrian classes whose representatives are shown in Figure 10, we have $9^+_{43} \neq -9^+_{43}$ and $9^-_{43} \neq -\mu(9^-_{43})$.*

The proof is presented in Figure 11.

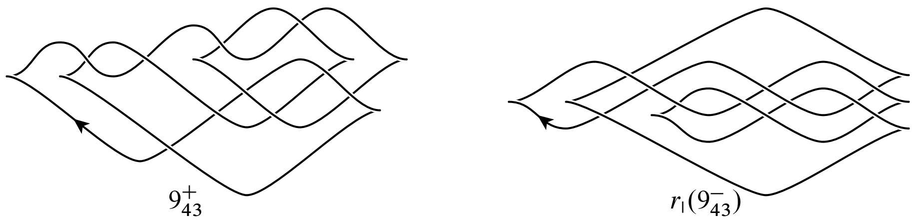


Figure 10: Legendrian knots in Proposition 7.3.

Proposition 7.4 *For the $\xi_\pm$–Legendrian classes whose representatives are shown in Figure 12, the following statements hold*:

(i) *The ξ_+–Legendrian classes 9_{44}^{1+}, 9_{44}^{2+}, 9_{44}^{3+}, $-\mu(9_{44}^{1+})$, $-\mu(9_{44}^{2+})$ and $-\mu(9_{44}^{3+})$ are pairwise distinct.*

(ii) *For $k \in \{1,2,3,4\}$ the ξ_+–Legendrian classes $S_+^k(9_{44}^{1+})$, $S_+^k(9_{44}^{2+})$ and $S_+^k(9_{44}^{3+})$ are pairwise distinct.*

(iii) *The ξ_-–Legendrian classes 9_{44}^{-} and -9_{44}^{-} are distinct.*

Proof Representatives of the exchange classes involved in the proof are shown in Figure 13. It is established in [4] that $9_{44}^{1+}, -\mu(9_{44}^{1+}) \notin \{9_{44}^{2+}, 9_{44}^{3+}\}$ and $S_+^k(9_{44}^{1+}) \notin \{S_+^k(9_{44}^{2+}), S_+^k(9_{44}^{3+})\}$ for any $k \in \mathbb{N}$. The proof of the remaining claims is presented in Figure 14 (where some of the known facts are also reproved). □

Remark 7.5 It is conjectured in [4] that the ξ_+–Legendrian classes $S_+^k(9_{44}^{2+})$ and $S_+^k(9_{44}^{3+})$ are distinct for any $k \in \mathbb{N}$, not only $k \leqslant 4$. The method of this paper allows us,

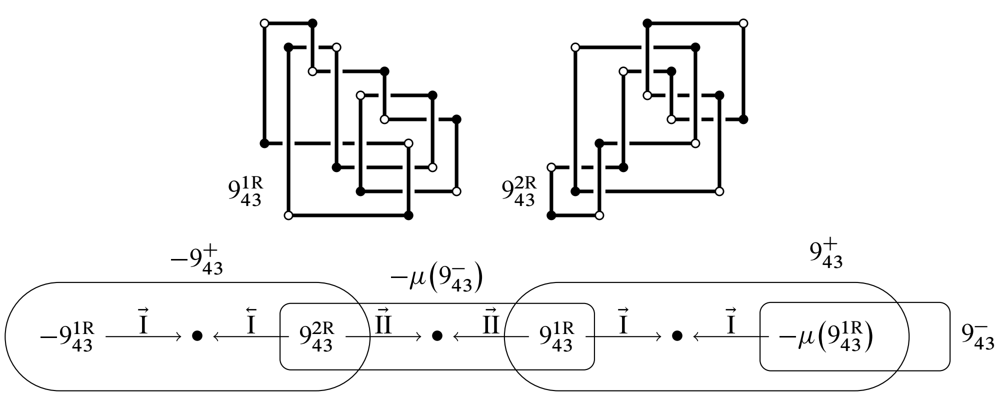


Figure 11: Proof of Proposition 7.3.

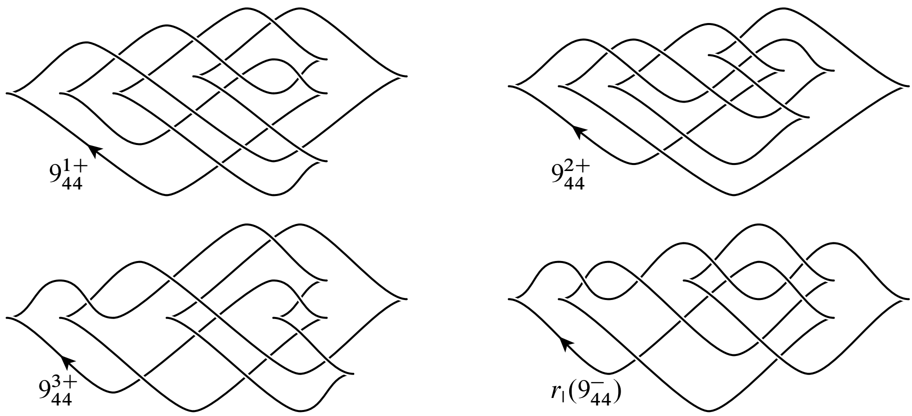


Figure 12: The knots in Proposition 7.4.

in principle, to test the claim for any fixed k, and this has been done by the authors for $k \leqslant 4$. (For larger k, the simple — and far from being optimized — exhaustive search, which we used to test diagrams for exchange-equivalence, takes too much time.)

Proving the claim for all k is equivalent to distinguishing certain transverse knots. The present technique has been upgraded in [7] to an algorithmic solution of this problem (in the case of knots with trivial orientation-preserving symmetry group). This reduces the task of verifying the inequality $S_+^k(9_{44}^{2+}) \neq S_+^k(9_{44}^{3+})$ for all $k \in \mathbb{N}$ to a finite exhaustive search, which is still to be done.

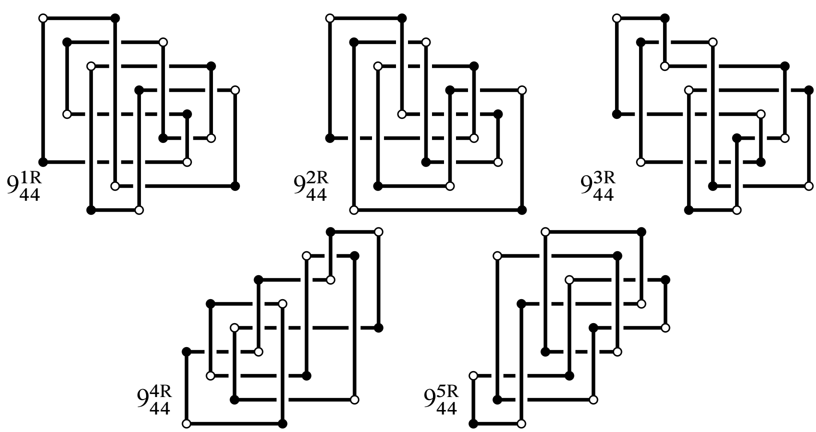


Figure 13: Exchange classes used in the proof of Proposition 7.4.

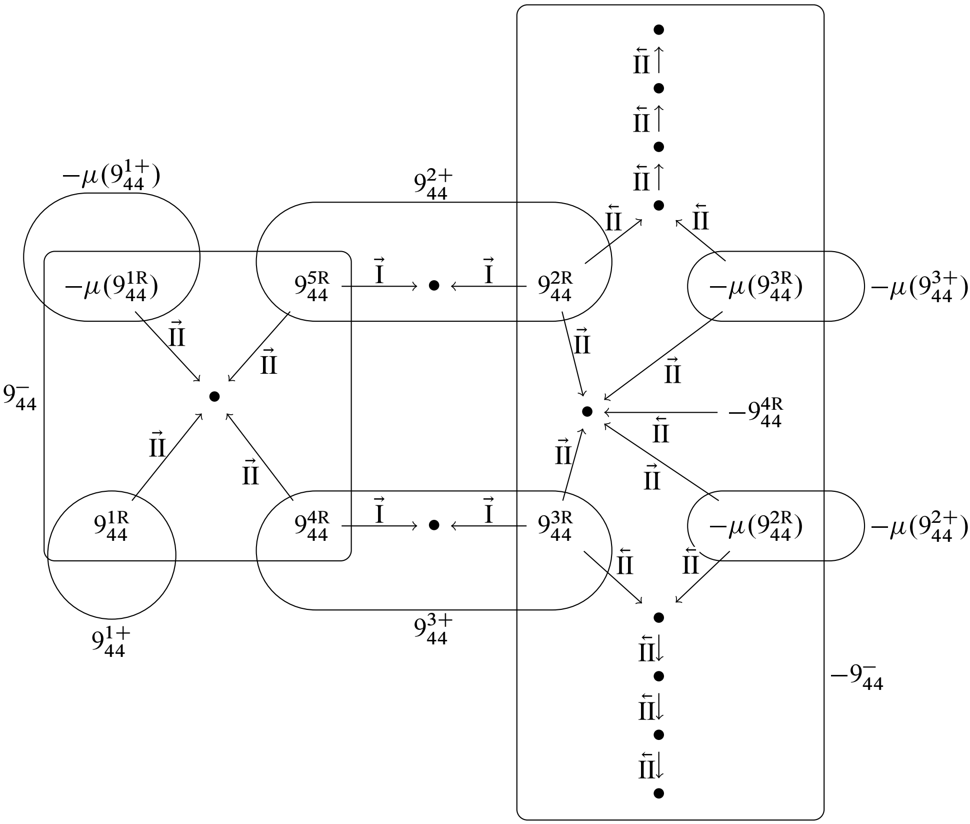


Figure 14: Proof of Proposition 7.4.

A similar remark applies to part (iii) of Proposition 7.6 and parts (ii) of Propositions 7.7 and 7.8. In the last two cases the required exhaustive search appears to be trivial, so the question for the knots 10_{128} and 10_{160} is settled in [7] completely.

Proposition 7.6 *For the $\xi_\pm$–Legendrian classes whose representatives are shown in Figure 15 the following statements hold*:

(i) $9_{45}^{1+}, 9_{45}^{2+}, 9_{45}^{3+}, -\mu(9_{45}^{1+})$ *and* $-\mu(9_{45}^{3+})$ *are pairwise distinct.*

(ii) $9_{45}^{1-}, -9_{45}^{1-}, \mu(9_{45}^{1-}), -\mu(9_{45}^{1-}), 9_{45}^{2-}$ *and* $\mu(9_{45}^{2-})$ *are pairwise distinct.*

(iii) *For $k \in \{1, 2, 3\}$ the ξ_-–Legendrian classes $S_+^k(9_{45}^{2-})$ and $S_+^k(-\mu(9_{45}^{2-}))$ are distinct.*

Proof Representatives of the exchange classes involved in the proof are shown in Figure 16. It is established in [4] that $9_{45}^{2+} = -\mu(9_{45}^{2+})$. So, to prove part (i) of the

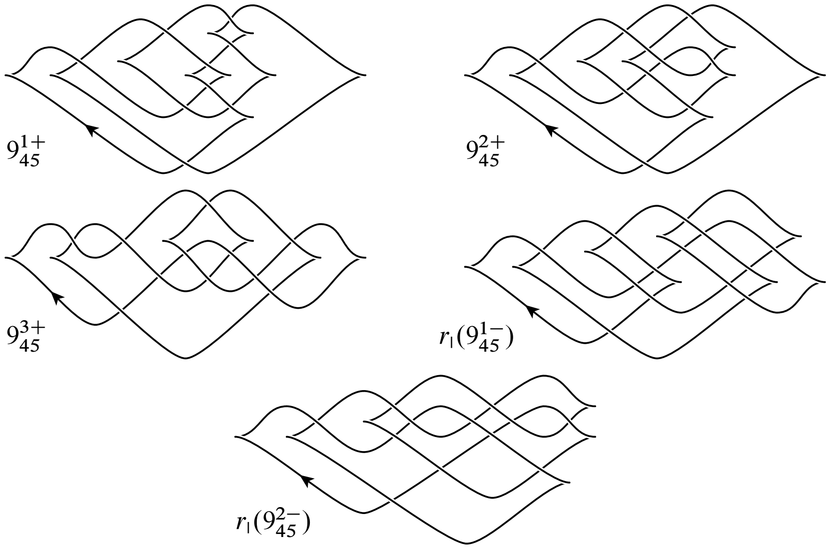


Figure 15: The knots in Proposition 7.6.

proposition, it suffices to show that 9^{1+}_{45}, 9^{3+}_{45}, $-\mu(9^{1+}_{45})$, and $-\mu(9^{3+}_{45})$ are pairwise distinct. The proof of this and of part (iii) is presented in Figure 17.

It is established in [4] that $9^{2-}_{45} = -9^{2-}_{45} \notin \{9^{1-}_{45}, -9^{1-}_{45}, \mu(9^{1-}_{45}), -\mu(9^{1-}_{45}), \mu(9^{2-}_{45})\}$, so it remains to show that 9^{1-}_{45}, -9^{1-}_{45}, $\mu(9^{1-}_{45})$ and $-\mu(9^{1-}_{45})$ are pairwise distinct. To this end, it suffices to show that some three of these four classes are pairwise distinct. This is done in Figure 18. □

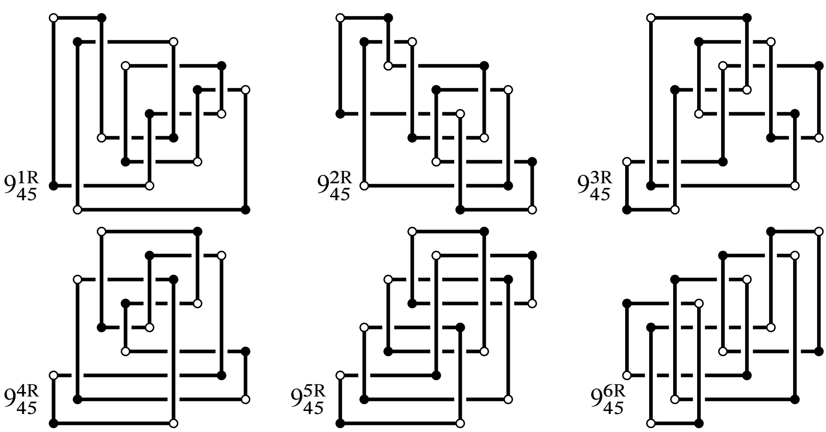


Figure 16: Exchange classes used in the proof of Proposition 7.6.

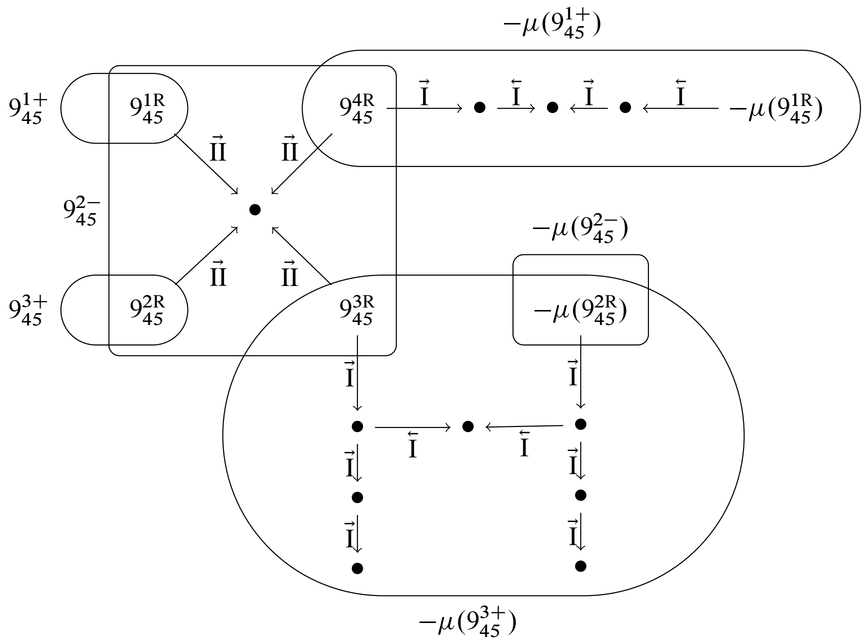


Figure 17: Proof of parts (i) and (iii) of Proposition 7.6.

Proposition 7.7 *For the Legendrian classes whose representatives are shown in Figure 19, the following statements hold*:

(i) *The Legendrian classes* 10_{128}^{1+}, 10_{128}^{2+}, $-\mu(10_{128}^{1+})$ *and* $-\mu(10_{128}^{2+})$ *are pairwise distinct.*

(ii) *For any* $k \in \{1,2,3,4\}$, *the Legendrian classes* $S_-^k(10_{128}^{1+}) = S_-^k(10_{128}^{2+})$ *and* $S_-^k(-\mu(10_{128}^{1+})) = S_-^k(-\mu(10_{128}^{2+}))$ *are distinct.*

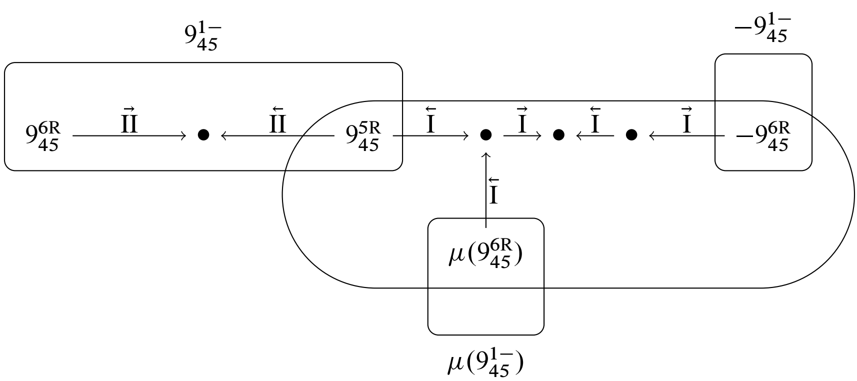


Figure 18: Proof of part (ii) of Proposition 7.6.

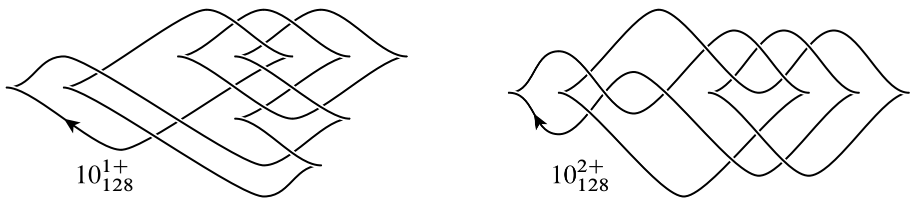


Figure 19: Knots in Proposition 7.7.

Proof The proof is presented in Figure 20. □

Proposition 7.8 *For the Legendrian classes whose representatives are shown in Figure 21, the following statements hold*:

(i) *The Legendrian classes* $10^{1+}_{160} = -10^{1+}_{160}$, $\mu(10^{1+}_{160})$, 10^{2+}_{160}, -10^{2+}_{160}, $\mu(10^{2+}_{160})$ *and* $-\mu(10^{2+}_{160})$ *are pairwise distinct.*

(ii) *For any* $k \in \{1, 2, 3, 4\}$, *the Legendrian classes* $S^k_-(10^{2+}_{160})$ *and* $S^k_-(-10^{2+}_{160})$ *are distinct.*

Proof The proof is presented in Figure 22. (The ξ_-–Legendrian class 10^{-}_{160} can be guessed from the scheme. We don't provide a picture as this class is not involved in any of our statements.) □

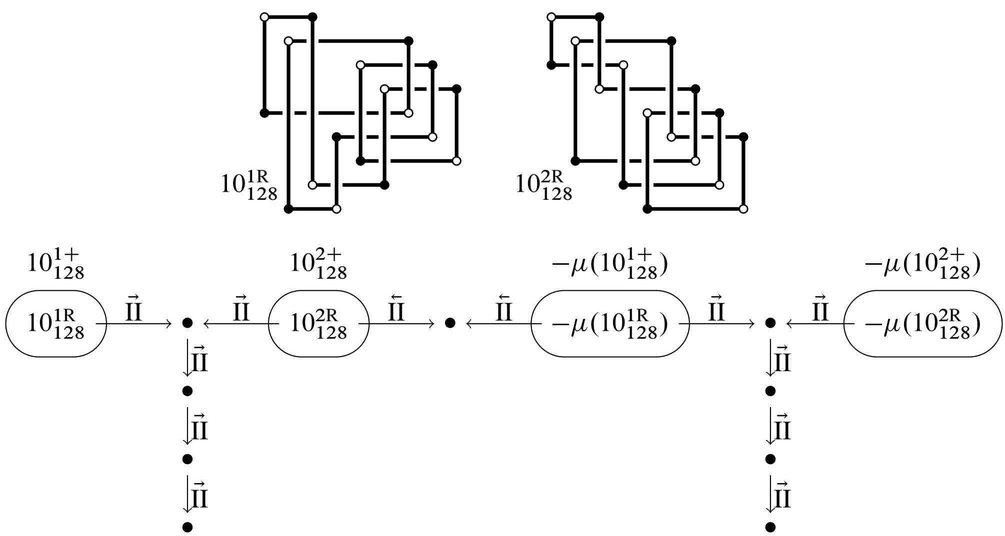


Figure 20: Proof of Proposition 7.7.

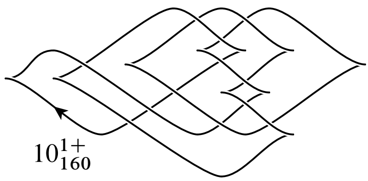


Figure 21: Knots in Proposition 7.8.

Remark 7.9 The fact that $10^{1+}_{160} \notin \{10^{2+}_{160}, -10^{2+}_{160}, \mu(10^{2+}_{160}), -\mu(10^{2+}_{160})\}$ and that $10^{1+}_{160} = -10^{1+}_{160}$ is established already in [4].

Proof of Theorem 2.2 The front projections of K_1 and K_2 shown in Figure 4 are produced from two rectangular diagrams R_1 and R_2, respectively, via the procedure described in Section 4 and illustrated in Figure 6. Thus, $K_i \in \mathscr{L}_+(R_i)$ for $i = 1, 2$.

Now we recall the origin of R_1 and R_2. Shown in [9, Figure 35] is a rectangular diagram Π of a surface such that

(i) the associated surface $\widehat{\Pi}$ is an annulus,

(ii) the relative Thurston–Bennequin numbers $\operatorname{tb}(\widehat{R}_i; \widehat{\Pi})$, $i = 1, 2$, vanish,

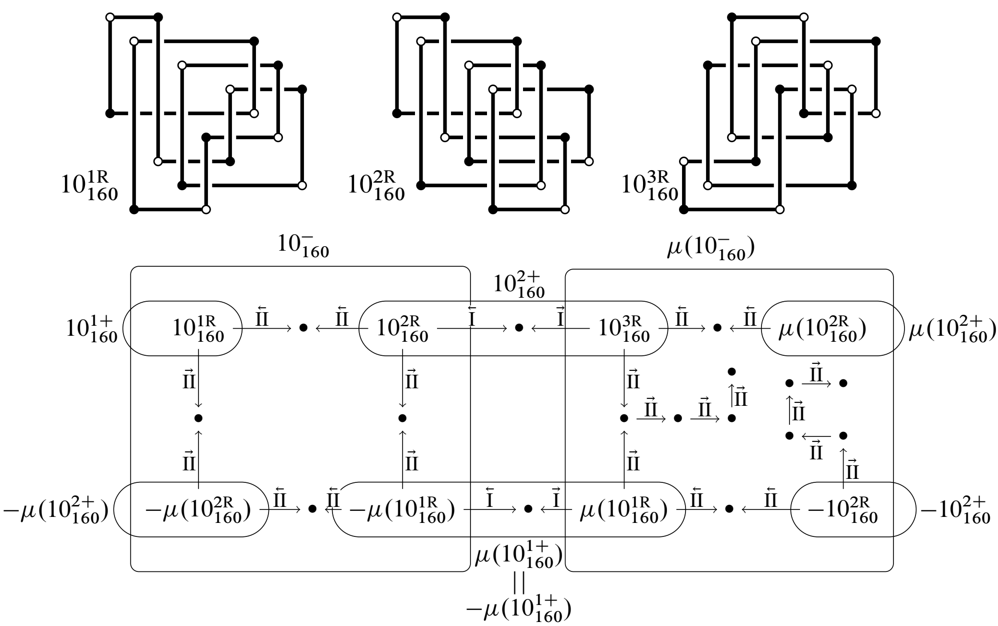


Figure 22: Proof of Proposition 7.8.

(iii) $\widehat{\Pi}$ can be endowed with an orientation so that $\partial\widehat{\Pi} = \widehat{R}_1 \cup (-\widehat{R}_2)$,

(iv) Π has the form $\{r_i\}_{i=1,2,\dots,74}$, where, for each $i = 1, \dots, 74$, the intersection $r_{i-1} \cap r_i$ is the bottom left vertex of r_i (we put $r_0 = r_{74}$).

The last condition in this list means that there are $\theta_0, \theta_1, \dots, \theta_{74} = \theta_0 \in \mathbb{S}^1$ and $\varphi_0, \varphi_1, \dots, \varphi_{74} = \varphi_0 \in \mathbb{S}^1$ such that $r_i = [\theta_{i-1}; \theta_i] \times [\varphi_{i-1}; \varphi_i]$ and $R_1 \cup R_2 = \{(\theta_{i-1}, \varphi_i), (\theta_i, \varphi_{i-1})\}_{i=1,\dots,74}$. Moreover, the signs of the vertices $(\theta_{i-1}, \varphi_i)$ and $(\theta_i, \varphi_{i-1})$ in $R_1 \cup R_2$ are opposite.

We now show that a sequence of elementary moves including a type II stabilization, exchange moves, and a type II destabilization transforms $R_1 \cup R_2$ to a rectangular diagram of a link in which the connected components become combinatorially equivalent. To this end, pick an $\varepsilon > 0$ smaller than one half of the length of any interval $[\theta_i; \theta_j]$ and $[\varphi_i; \varphi_j]$ with $i \neq j$, and make the following replacements in $R_1 \cup R_2$:

$$\begin{aligned}
(\theta_1, \varphi_0) &\rightsquigarrow (\theta_0 - \varepsilon, \varphi_0), (\theta_0 - \varepsilon, \varphi_1 - \varepsilon), (\theta_1, \varphi_1 - \varepsilon) && \text{(type II stabilization)},\\
(\theta_1, \varphi_1 - \varepsilon), (\theta_1, \varphi_2) &\rightsquigarrow (\theta_2 - \varepsilon, \varphi_1 - \varepsilon), (\theta_2 - \varepsilon, \varphi_2) && \text{(exchange)},\\
(\theta_2 - \varepsilon, \varphi_2), (\theta_3, \varphi_2) &\rightsquigarrow (\theta_2 - \varepsilon, \varphi_3 - \varepsilon), (\theta_3, \varphi_3 - \varepsilon) && \text{(exchange)},\\
(\theta_3, \varphi_3 - \varepsilon), (\theta_3, \varphi_4) &\rightsquigarrow (\theta_4 - \varepsilon, \varphi_3 - \varepsilon), (\theta_4 - \varepsilon, \varphi_4) && \text{(exchange)},\\
&\ \vdots\\
(\theta_{72} - \varepsilon, \varphi_{72}), (\theta_{73}, \varphi_{72}) &\rightsquigarrow (\theta_{72} - \varepsilon, \varphi_{73} - \varepsilon), (\theta_{73}, \varphi_{73} - \varepsilon) && \text{(exchange)},\\
(\theta_{73}, \varphi_{73} - \varepsilon), (\theta_{73}, \varphi_0), (\theta_0 + \varepsilon, \varphi_0) &\rightsquigarrow (\theta_0 - \varepsilon, \varphi_{73} - \varepsilon) && \text{(type II destabilization)}.
\end{aligned}$$

This sequence of moves is illustrated in Figure 23.

This proves that $\mathscr{L}_-(R_1) = \mathscr{L}_-(R_2)$. The diagrams R_1 and R_2 are not combinatorially equivalent and do not admit any nontrivial exchange move. The knots represented by R_1 and R_2 have trivial orientation-preserving symmetry group by Proposition 6.2. Therefore, by Theorem 4.5, $\mathscr{L}_+(R_1) \neq \mathscr{L}_+(R_2)$. □

Appendix K_1 and K_2 are not satellite knots

Here we explain how to verify, with very little computation, that the complement of K_1 (and K_2) contains no incompressible nonboundary-parallel torus. To do so we use a method that can be viewed as a modification of Haken's method of normal surfaces, which allows one, in general, to find all incompressible surfaces of minimal genus. Haken's algorithm in general has very high computational complexity, which makes

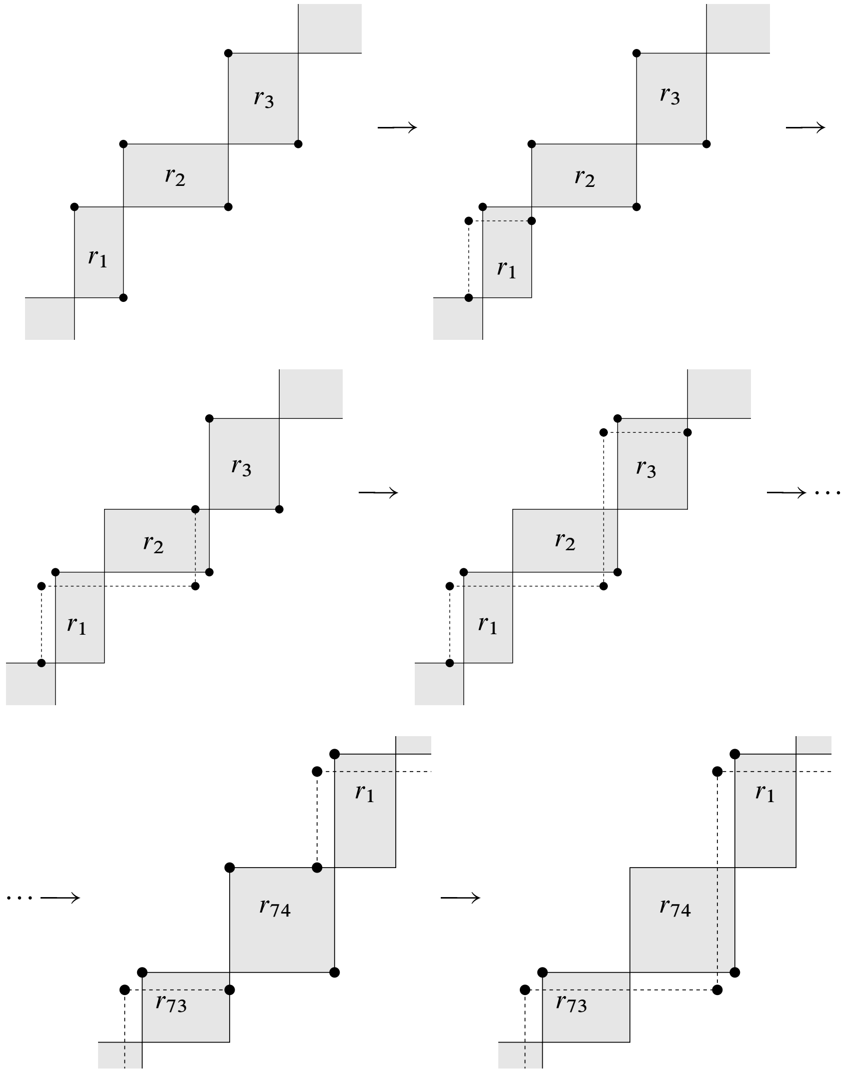


Figure 23: Transforming one of R_1 and R_2 to the other by elementary moves.

it infeasible to implement in most cases. However, in certain cases, including our particular one, a modified version of Haken's method can be efficiently used to search all incompressible surfaces of *nonnegative Euler characteristic*.

First we describe the general idea for the reader well familiar with the difficulties in using Haken's method in practice. Haken's normal surfaces are encoded by certain *normal coordinates* $x_1, \dots, x_N$, which take integer values. To determine a normal surface they must satisfy a bunch of conditions that are naturally partitioned into the following three groups:

(i) *Nonnegativity conditions*, which are the inequalities $x_i \geqslant 0$, $i = 1, \dots, N$.

(ii) *Matching conditions*, which are linear equations with integer coefficients.

(iii) *Compatibility conditions*, which are equations of the form $x_i x_j = 0$ for some set of pairs (i, j).

The Euler characteristic of a normal surface F can be expressed as a linear combination of the normal coordinates of F in numerous ways, and some of these expressions have only nonpositive coefficients. If we are looking for normal surfaces of nonnegative Euler characteristic, for any such expression $\sum_i a_i x_i$ with nonpositive coefficients a_i, we may add the inequality $\sum_i a_i x_i \geqslant 0$ to the system. Together with the nonnegativity conditions, this implies $x_i = 0$ whenever $a_i < 0$. This reduces the number of variables in the system, and chances are that, after the reduction, the space of solutions of the system of matching equations alone has very small dimension.

Now we turn to our concrete case. The idea explained above will be realized in quite different terms. The reduction of variables will occur in Lemma A.2.

The rectangular diagrams from which the Legendrian knots K_1 and K_2 shown in Figure 4 are produced have 37 edges of each direction. For this reason we rescale the coordinates θ, φ on $\mathbb{T}^2$ so that they take values in $\mathbb{R}/(37 \cdot \mathbb{Z})$, and the vertices of the diagrams will form a subset of $\mathbb{Z}_{37} \times \mathbb{Z}_{37}$.

We will work with the knot K_1. The corresponding rectangular diagram of a knot, which we denote by R, has the following list of vertices:

$(0,13)$, $(0,28)$, $(1,14)$, $(1,35)$, $(2,15)$, $(2,36)$, $(3,0)$, $(3,19)$, $(4,1)$, $(4,22)$, $(5,6)$, $(5,23)$, $(6,7)$, $(6,24)$, $(7,9)$, $(7,25)$, $(8,10)$, $(8,26)$, $(9,11)$, $(9,27)$, $(10,12)$, $(10,29)$, $(11,13)$, $(11,34)$, $(12,20)$, $(12,35)$, $(13,21)$, $(13,36)$, $(14,8)$, $(14,22)$, $(15,9)$, $(15,31)$, $(16,10)$, $(16,32)$, $(17,11)$, $(17,33)$, $(18,18)$, $(18,34)$, $(19,4)$, $(19,19)$, $(20,5)$, $(20,20)$, $(21,6)$, $(21,21)$, $(22,7)$, $(22,23)$, $(23,8)$, $(23,30)$, $(24,12)$, $(24,31)$, $(25,16)$, $(25,32)$, $(26,17)$, $(26,33)$, $(27,2)$, $(27,18)$, $(28,3)$, $(28,24)$, $(29,4)$, $(29,28)$, $(30,14)$, $(30,29)$, $(31,15)$, $(31,30)$, $(32,0)$, $(32,16)$, $(33,1)$, $(33,17)$, $(34,2)$, $(34,25)$, $(35,3)$, $(35,26)$, $(36,5)$, $(36,27)$.

According to [9, Theorem 1], any incompressible torus in the complement of $\widehat{R}$ is isotopic to a surface of the form $\widehat{\Pi}$, where Π is a rectangular diagram of a surface. Let such a diagram Π be chosen so that the number of rectangles in Π is as minimal as possible (which is equivalent to requesting that $\widehat{\Pi}$ has minimal possible number of intersections with $\mathbb{S}^1_{\tau=0} \cup \mathbb{S}^1_{\tau=1}$). We fix it from now on.

To any rectangle $r = [\theta';\theta''] \times [\varphi';\varphi'']$ with $\{\theta',\theta'',\varphi',\varphi''\} \cap \mathbb{Z}_{37} = \varnothing$, we associate a *type*, which is a 4–tuple $(i,j,k,l) \in (\mathbb{Z}_{37})^4$ defined by the conditions

$$\theta' \in (i;i+1), \quad \varphi' \in (j;j+1), \quad \theta'' \in (k;k+1), \quad \varphi'' \in (l;l+1).$$

Since $\partial\widehat{\Pi} = \varnothing$, we have $\{\theta',\theta'',\varphi',\varphi''\} \cap \mathbb{Z}_{37} = \varnothing$, so every rectangle in Π has a type.

Recall from [10] that by *an occupied level* of Π we mean any meridian $m_{\theta_0} = \{\theta_0\} \times \mathbb{S}^1$ or any longitude $\ell_{\varphi_0} = \mathbb{S}^1 \times \{\varphi_0\}$ that contains a vertex of some rectangle in Π.

Lemma A.1 *There are no rectangles in* Π *of type* (i,j,k,l) *with* $i = k$ *or* $j = l$*.*

Proof Let $r = [\theta';\theta''] \times [\varphi';\varphi'']$ be a rectangle of Π such that the annulus $(\theta';\theta'') \times \mathbb{S}^1$ contains no occupied level of Π. Then the interval $(\theta';\theta'')$ contains at least one point from $\mathbb{Z}_{37}$ since otherwise the number of intersections of $\widehat{\Pi}$ with $\mathbb{S}^1_{\tau=1}$ could be reduced by an isotopy.

This implies that for *any* rectangle $r = [\theta';\theta''] \times [\varphi';\varphi'']$ of Π, the intersection $(\theta';\theta'') \cap \mathbb{Z}_{37}$ is nonempty. Indeed, if there is an occupied level of Π contained in $(\theta';\theta'') \times \mathbb{S}^1$, then there is a rectangle $[\theta''';\theta''''] \times [\varphi''';\varphi'''']$ in Π with $[\theta''';\theta''''] \subset (\theta';\theta'')$. By taking the narrowest such rectangle we will have that $(\theta''';\theta'''') \times \mathbb{S}^1$ contains no occupied level of Π, and hence $(\theta''';\theta'''')$ has a nonempty intersection with $\mathbb{Z}_{37}$. Similarly, $(\varphi';\varphi'') \cap \mathbb{Z}_{37} \neq \varnothing$ for any rectangle of Π.

Now let (i,j,k,l) be the type of some rectangle $r = [\theta';\theta''] \times [\varphi';\varphi''] \in \Pi$. The equality $i = k$ would mean that

$$(\theta';\theta'') \subset (i;i+1) \quad \text{or} \quad (\theta'';\theta') \subset (i;i+1).$$

The former case is impossible as we have just seen. In the latter case, we must have $(\varphi';\varphi'') \subset (j;j+1)$ as otherwise r would contain a vertex of R. Therefore, this case also does not occur, and we have $i \neq k$.

The inequality $j \neq l$ is established similarly. □

The type (i,j,k,l) of a rectangle r is said to be *admissible* if $r \cap R = \varnothing$. It is said to be *maximal* if it is admissible, and none of the types $(i-1,j,k,l)$, $(i,j-1,k,l)$, $(i,j,k+1,l)$ and $(i,j,k,l+1)$ is admissible.

Lemma A.2 *The type of any rectangle in* Π *is maximal.*

Proof Here we will use the fact that the diagram R is *rigid*, which means that it admits no nontrivial exchange move. In other words, for any two neighboring edges

$\{(i, j_1), (i, j_2)\}$, $\{(i+1, j_3), (i+1, j_4)\}$ or $\{(j_1, i), (j_2, i)\}$, $\{(j_3, i+1), (j_4, i+1)\}$ of R, exactly one of j_3, j_4 lies in $(j_1; j_2)$, and the other lies in $(j_2; j_1)$.

Let $\{(i, j_1), (i, j_2)\}$ and $\{(i+1, j_3), (i+1, j_4)\}$ be two neighboring vertical edges of R, and let m_{θ_0} with $\theta_0 \in (i; i+1)$ be an occupied level of Π. Since the surface $\widehat{\Pi}$ is closed, the whole meridian m_{θ_0} is covered by the vertical sides of rectangles in Π. Therefore, there are rectangles $r_1, r_2, \dots, r_{2p} \in \Pi$ of the form

$$r_{2k-1} = [\theta_{2k-1}; \theta_0] \times [\varphi_{2k-1}; \varphi_{2k}], \quad r_{2k} = [\theta_0; \theta_{2k}] \times [\varphi_{2k}; \varphi_{2k+1}],$$

where $k = 1, \dots, p$ and $\varphi_{2p+1} = \varphi_1$.

We claim that each interval $[\varphi_k; \varphi_{k+1}]$, $k = 1, \dots, 2p$, contains at most one of j_1, j_2, j_3, j_4. Indeed, let k be odd. Then r_k has the form $[\theta_k; \theta_0] \times [\varphi_k; \varphi_{k+1}]$. Since it is disjoint from $R \supset \{(i, j_1), (i, j_2)\}$, we must have either $[\varphi_k; \varphi_{k+1}] \subset (j_1; j_2)$ or $[\varphi_k; \varphi_{k+1}] \subset (j_2; j_1)$. Due to rigidity of R, each of the intervals $(j_1; j_2)$ and $(j_2; j_1)$ contains exactly one of j_3, j_4, hence the claim. In the case when k is even, the proof is similar, with the roles of $\{j_1, j_2\}$ and $\{j_3, j_4\}$ exchanged.

Thus, p is at least 2. We now claim that p is exactly 2. Indeed, the number of tiles of $\widehat{\Pi}$ attached to the vertex corresponding to m_{θ_0} is equal to $2p$, and we have just seen that $2p \geqslant 4$. The same applies similarly to any other vertex of the tiling. Since every tile is a 4–gon and the surface $\widehat{\Pi}$ is a torus, every vertex of the tiling must be adjacent to *exactly* four tiles.

The equality $p = 2$ implies that every interval $(\varphi_k; \varphi_{k+1})$, $k = 1, 2, 3, 4$, contains exactly one of j_1, j_2, j_3, j_4, which means that the rectangles

$$\begin{aligned} &[\theta_1; \theta_0 + 1] \times [\varphi_1; \varphi_2], \quad [\theta_0 - 1; \theta_2] \times [\varphi_2; \varphi_3], \\ &[\theta_3; \theta_0 + 1] \times [\varphi_3; \varphi_4], \quad [\theta_0 - 1; \theta_4] \times [\varphi_4; \varphi_1], \end{aligned}$$

are not of an admissible type. In other words, whenever Π contains a rectangle of type (i, j, k, l) (resp. of type (k, l, i, j)), the type $(i-1, j, k, l)$ (resp. $(k, l, i+1, j)$) is not admissible. Since $i \in \mathbb{Z}_{37}$ was chosen arbitrarily, we can put it another way: whenever Π contains a rectangle of type (i, j, k, l), the types $(i-1, j, k, l)$ and $(i, j, k+1, l)$ are not admissible.

Similar reasoning applied to a horizontal occupied level ℓ_{φ_0} of Π instead of m_{θ_0} shows that whenever Π contains a rectangle of type (i, j, k, l), the types $(i, j-1, k, l)$ and $(i, j, k, l+1)$ are not admissible. Therefore, every rectangle in Π is of a maximal type. $\square$

A simple exhaustive search shows that there are exactly 623 maximal types of rectangles for R. For every maximal type (i,j,k,l), we denote by $x_{i,j,k,l}$ the number of rectangles of type (i,j,k,l) in Π. From the fact that every vertex of a rectangle in Π is shared by exactly two rectangles, which are disjoint otherwise, we get the *matching conditions*

$$\sum_{k,l\in\mathbb{Z}_{37}} x_{i,j,k,l} = \sum_{k,l\in\mathbb{Z}_{37}} x_{k,l,i,j}, \quad (i,j)\in(\mathbb{Z}_{37})^2, \tag{7}$$

where we put $x_{i,j,k,l}=0$ unless (i,j,k,l) is a maximal type. For a complete list of maximal types and matching conditions the reader is referred to [12].

It is now a direct check that the system (7) is of rank 621, and thus has two-dimensional solution space. It is another direct check that only one solution in this space, up to positive scale, satisfies the nonnegativity conditions $x_{i,j,k,l}\geqslant 0$. Therefore, there exists at most one isotopy class of incompressible tori in the complement of K_1, which implies that every incompressible torus is boundary-parallel.

References

[1] **M Boileau**, **B Zimmermann**, *Symmetries of nonelliptic Montesinos links*, Math. Ann. 277 (1987) 563–584 MR Zbl

[2] **F Bonahon**, **L C Siebenmann**, *New geometric splittings of classical knots and the classification and symmetries of arborescent knots*, unpublished manuscript (2016) Available at `https://dornsife.usc.edu/assets/sites/1191/docs/Preprints/BonSieb.pdf`

[3] **Y Chekanov**, *Differential algebra of Legendrian links*, Invent. Math. 150 (2002) 441–483 MR Zbl

[4] **W Chongchitmate**, **L Ng**, *An atlas of Legendrian knots*, Exp. Math. 22 (2013) 26–37 MR Zbl

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

[6] **I A Dynnikov**, *Arc-presentations of links: monotonic simplification*, Fund. Math. 190 (2006) 29–76 MR Zbl

[7] **I A Dynnikov**, *Transverse-Legendrian links*, Sib. Èlektron. Mat. Izv. 16 (2019) 1960–1980 MR Zbl

[8] **I A Dynnikov**, **M V Prasolov**, *Bypasses for rectangular diagrams: a proof of the Jones conjecture and related questions*, Trans. Moscow Math. Soc. (2013) 97–144 MR Zbl In Russian; translated in Trans. Moscow Math. Soc. (2013) 97–144

[9] **I A Dynnikov**, **M V Prasolov**, *Rectangular diagrams of surfaces: representability*, Mat. Sb. 208 (2017) 55–108 MR Zbl In Russian; translated in Sb. Math. 208 (2017) 791–841

[10] **I Dynnikov**, **M Prasolov**, *Rectangular diagrams of surfaces: distinguishing Legendrian knots*, J. Topol. 14 (2021) 701–860 MR Zbl

[11] **I A Dynnikov**, **V A Shastin**, *On the equivalence of Legendrian knots*, Uspekhi Mat. Nauk 73 (2018) 195–196 MR Zbl In Russian; translated in Russian Math. Surveys 73 (2018) 1125–1127

[12] **I Dynnikov**, **V Shastin**, *Distinguishing Legendrian knots with trivial orientation-preserving symmetry group*, preprint (2021) arXiv 1810.06460v3

[13] **Y Eliashberg**, *Invariants in contact topology*, from "Proceedings of the International Congress of Mathematicians, II" (G Fischer, U Rehmann, editors), Deutsche Math. Vereinigung, Berlin (1998) 327–338 MR Zbl

[14] **Y Eliashberg**, **M Fraser**, *Classification of topologically trivial Legendrian knots*, from "Geometry, topology, and dynamics" (F Lalonde, editor), CRM Proc. Lecture Notes 15, Amer. Math. Soc., Providence, RI (1998) 17–51 MR Zbl

[15] **Y Eliashberg**, **M Fraser**, *Topologically trivial Legendrian knots*, J. Symplectic Geom. 7 (2009) 77–127 MR Zbl

[16] **J B Etnyre**, **K Honda**, *Knots and contact geometry, I: Torus knots and the figure eight knot*, J. Symplectic Geom. 1 (2001) 63–120 MR Zbl

[17] **J B Etnyre**, **D J LaFountain**, **B Tosun**, *Legendrian and transverse cables of positive torus knots*, Geom. Topol. 16 (2012) 1639–1689 MR Zbl

[18] **J B Etnyre**, **L L Ng**, **V Vértesi**, *Legendrian and transverse twist knots*, J. Eur. Math. Soc. 15 (2013) 969–995 MR Zbl

[19] **J Etnyre**, **V Vértesi**, *Legendrian satellites*, Int. Math. Res. Not. 2018 (2018) 7241–7304 MR Zbl

[20] **D Fuchs**, *Chekanov–Eliashberg invariant of Legendrian knots: existence of augmentations*, J. Geom. Phys. 47 (2003) 43–65 MR Zbl

[21] **D Fuchs**, **S Tabachnikov**, *Invariants of Legendrian and transverse knots in the standard contact space*, Topology 36 (1997) 1025–1053 MR Zbl

[22] **H Geiges**, *An introduction to contact topology*, Cambridge Studies in Advanced Mathematics 109, Cambridge Univ. Press (2008) MR Zbl

[23] **P Ghiggini**, *Linear Legendrian curves in T^3*, Math. Proc. Cambridge Philos. Soc. 140 (2006) 451–473 MR Zbl

[24] **E Giroux**, *Convexité en topologie de contact*, Comment. Math. Helv. 66 (1991) 637–677 MR Zbl

[25] **E Giroux**, *Structures de contact en dimension trois et bifurcations des feuilletages de surfaces*, Invent. Math. 141 (2000) 615–689 MR Zbl

[26] **E Giroux**, *Structures de contact sur les variétés fibrées en cercles audessus d'une surface*, Comment. Math. Helv. 76 (2001) 218–262 MR Zbl

[27] **G D Gospodinov**, *Relative knot invariants: properties and applications*, preprint (2009) arXiv 0909.4326

[28] **R Hartley**, *Knots with free period*, Canadian J. Math. 33 (1981) 91–102 MR Zbl

[29] **S R Henry**, **J R Weeks**, *Symmetry groups of hyperbolic knots and links*, J. Knot Theory Ramifications 1 (1992) 185–201 MR Zbl

[30] **K Kodama**, **M Sakuma**, *Symmetry groups of prime knots up to* 10 *crossings*, from "Knots 90" (A Kawauchi, editor), de Gruyter, Berlin (1992) 323–340 MR Zbl

[31] **U Lüdicke**, *Zyklische Knoten*, Arch. Math. (Basel) 32 (1979) 588–599 MR Zbl

[32] **C Manolescu**, **P Ozsváth**, **S Sarkar**, *A combinatorial description of knot Floer homology*, Ann. of Math. 169 (2009) 633–660 MR Zbl

[33] **J M Montesinos**, *Variedades de Seifert que son recubridores ciclicos ramificados de dos hojas*, Bol. Soc. Mat. Mexicana 18 (1973) 1–32 MR Zbl

[34] **K Murasugi**, *On periodic knots*, Comment. Math. Helv. 46 (1971) 162–174 MR Zbl

[35] **L L Ng**, *Computable Legendrian invariants*, Topology 42 (2003) 55–82 MR Zbl

[36] **L Ng**, *Combinatorial knot contact homology and transverse knots*, Adv. Math. 227 (2011) 2189–2219 MR Zbl

[37] **P Ozsváth**, **Z Szabó**, **D Thurston**, *Legendrian knots, transverse knots and combinatorial Floer homology*, Geom. Topol. 12 (2008) 941–980 MR Zbl

[38] **P E Pushkar'**, **Y V Chekanov**, *Combinatorics of fronts of Legendrian links, and Arnol'd's 4–conjectures*, Uspekhi Mat. Nauk 60 (2005) 99–154 MR Zbl In Russian; translated in Russian Math. Surveys 60 (2005) 95–149

[39] **D Rolfsen**, *Knots and links*, Mathematics Lecture Series 7, Publish or Perish, Berkeley, CA (1976) MR Zbl

[40] **M Sakuma**, *The geometries of spherical Montesinos links*, Kobe J. Math. 7 (1990) 167–190 MR Zbl

ID: *V A Steklov Mathematical Institute, Russian Academy of Science*
Moscow, Russia

VS: *Department of Mechanics and Mathematics, Moscow State University*
Moscow, Russia

ID, VS: *St Petersburg State University*
Saint Petersburg, Russia

dynnikov@mech.math.msu.su, vashast@gmail.com

Received: 27 March 2021 Revised: 30 July 2021

Algebraic & Geometric Topology **23**:4 (2023) 1891–1934
DOI: 10.2140/agt.2023.23.1891
Published: 14 June 2023

A quantum invariant of links in $T^2 \times I$ with volume conjecture behavior

JOE BONINGER

We define a polynomial invariant J_n^T of links in the thickened torus. We call J_n^T the n^{th} toroidal colored Jones polynomial, and show it satisfies many properties of the original colored Jones polynomial. Most significantly, J_n^T exhibits volume conjecture behavior. We prove the volume conjecture for the two-by-two square weave, and provide computational evidence for other links. We also give two equivalent constructions of J_n^T, one as a generalized operator invariant we call a pseudo-operator invariant, and another using the Kauffman bracket skein module of the torus. Finally, we show J_n^T produces invariants of biperiodic and virtual links. To our knowledge, J_n^T gives the first example of volume conjecture behavior in a virtual (nonclassical) link.

57K14, 81R50

1 Introduction

A growing body of evidence supports the idea that the asymptotic growth rate of quantum invariants of links and 3–manifolds encodes geometric information. This hypothesis was initiated by the well-known volume conjecture of Kashaev, Murakami and Murakami.

Conjecture 1.1 [12; 24] *For a knot $K \subset S^3$, let $J_n(K; e^{2\pi i/n})$ be the n^{th} colored Jones polynomial of K evaluated at $e^{2\pi i/n}$. Then*

$$\lim_{n\to\infty} \frac{2\pi}{n} \log |J_n(K; e^{2\pi i/n})| = \mathrm{Vol}(S^3 \setminus K).$$

Here Vol indicates *simplicial volume*, which we define to be the sum of the hyperbolic volumes of the hyperbolic pieces in the Jaco–Shalen–Johannson decomposition of $S^3 \setminus K$; see Soma [29]. We say a quantum invariant exhibits *volume conjecture behavior* if theoretical or computational evidence supports a conjectured limit as above.

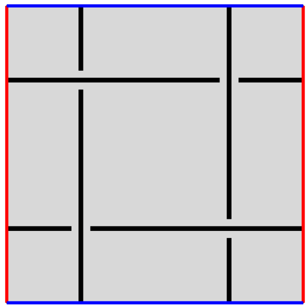


Figure 1: The two-by-two square weave (opposite sides of the diagram are identified).

Conjecture 1.1 has been generalized to other 3–manifolds in several ways. In [6], Costantino extended the colored Jones polynomial to links in $\#_k(S^2 \times S^1)$ using Turaev's theory of shadows and proved the volume conjecture for an infinite family of hyperbolic links. More recently, Chen and Yang [5] discovered volume conjecture behavior exhibited by the Witten–Reshetikhin–Turaev and Turaev–Viro invariants of 3–manifolds, two quantum invariants closely related to the colored Jones polynomial. These conjectures have been verified in many cases; see Detcherry, Kalfagianni and Yang [7] and Ohtsuki [26].

We define a polynomial invariant J_n^T, for $n \in \mathbb{N}$, of oriented links in the thickened torus, $T^2 \times I$. We call J_n^T the n^{th} *toroidal colored Jones polynomial*, and show it satisfies many properties of the colored Jones polynomial for links in S^3. For example, we give one construction of J_n^T using the theory of operator invariants, and another using the Kauffman bracket skein module of $T^2 \times I$. Significantly, J_n^T is the first example of volume conjecture behavior in the Kauffman bracket skein module of a manifold other than S^3.

We state the volume conjecture for J_n^T precisely as follows.

Conjecture 1.2 *For any link $L \subset T^2 \times I$ such that $(T^2 \times I) \setminus L$ is hyperbolic,*

$$\lim_{n\to\infty} \frac{2\pi}{n} \log |J_n^T(L; e^{2\pi i/n})| = \mathrm{Vol}((T^2 \times I) \setminus L).$$

Here the simplicial volume Vol is simply the hyperbolic volume of $(T^2 \times I) \setminus L$. We prove Conjecture 1.2 for the two-by-two square weave $W \subset T^2 \times I$ shown in Figure 1.

Theorem 6.2 *We have*

$$\lim_{n\to\infty} \frac{2\pi}{n} \log |J_n^T(W; e^{2\pi i/n})| = 4v_{\mathrm{oct}} = \mathrm{Vol}((T^2 \times I) \setminus W),$$

where $v_{\mathrm{oct}} \approx 3.6638$ is the volume of the regular ideal hyperbolic octahedron.

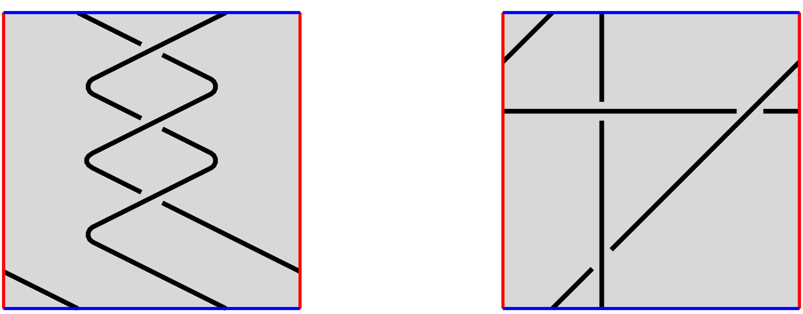
link	$(2\pi/n)\cdot\log\lvert J^T_n(L;e^{2\pi i/n})\rvert$ at $n=$ 10	20	30	50	75	100	Vol
2.1	5.4685	5.5004	5.4843	5.4548	5.4309	5.4215	5.3335
3.2	7.5047	7.6976	7.7393	7.7566	7.7564	7.7528	7.7069
3.5	5.9817	6.2649	6.3345	6.3733	6.3836	6.3852	6.3545
3.7	9.0885	9.3732	9.4523	9.5017	9.5182	9.5231	9.5034
B	7.1834	7.3637	7.3903	7.3953	7.3891	7.3825	$2v_{\text{oct}} \approx 7.3278$
ℓ	9.5569	9.9321	10.0405	10.1130	10.1411	10.1519	$10v_{\text{tet}} \approx 10.149$

Table 1: Computational evidence for Conjecture 1.2.

In addition to Theorem 6.2, the computations in Table 1 support our volume conjecture. Each row gives the normalized logarithm of the modulus of the toroidal colored Jones polynomial of a certain link, at the relevant root of unity, for different values of n. The first four rows are genus one virtual knots in Green's table [10] — each of these corresponds to a knot in $T^2 \times I$ (see Kuperberg [20]) with volume computed by Adams, Eisenberg, Greenberg, Kapoor, Liang, O'Connor, Pacheco-Tallaj and Wang in [1]. In the fifth and sixth rows, B and ℓ refer respectively to the virtual 2–braid and triaxial weave shown in Figure 2. (The geometry of ℓ is discussed by Champanerkar, Kofman and Purcell in [4].) Finally, $v_{\text{tet}} \approx 1.0149$ is the volume of the regular ideal hyperbolic tetrahedron.

Pseudo-operator invariant

Volume conjecture behavior is not the only interesting feature of J^T_n. Like the original colored Jones polynomial, J^T_n is defined using the theory of operator invariants and



The virtual 2–braid B. The triaxial weave ℓ.

Figure 2

the quantum group $\mathscr{A} = U_q(\mathrm{sl}(2,\mathbb{C}))$, the quantized universal enveloping algebra of $\mathrm{sl}(2,\mathbb{C})$ specialized to a root of unity q. Briefly, given a link $L \subset T^2 \times I$ with diagram $D \subset T^2$, we use the flat geometry of T^2 to label certain points of D as critical points. We then assign $\mathscr{A}$–linear operators to each critical point of D and use these local assignments to compute J_n^T as a state sum. This is similar to the construction of the colored Jones polynomial of links in S^3, with a key conceptual difference: in S^3, the local assignments of $\mathscr{A}$–linear operators to critical points (crossings and local extrema) extend to a global assignment of a single $\mathscr{A}$–linear operator to the entire link. With J_n^T no global assignment is possible, and for this reason we refer to J_n^T as a *pseudo-operator invariant*. The theory of pseudo-operator invariants, which generalizes the theory of operator invariants, may have applications beyond the invariant J_n^T. In Section 3, we develop this theory in detail and in the process construct another invariant $\hat{J}^T_{\boldsymbol{n},q}$ of framed, unoriented links in $T^2 \times I$. The invariant $\hat{J}^T_{\boldsymbol{n},q}$ is analogous to the invariant $J_{L,\boldsymbol{n}}$ of Kirby and Melvin [17], where $\boldsymbol{n} = (n_1, \dots, n_k)$ is a multi-integer indicating an integer n_i assigned to each component of L.

Skein module invariant

We also consider an SU(2) toroidal colored Jones polynomial obtained by specializing to the quantum group $\mathrm{SU}(2)_q$. We show that if $C \subset T^2$ is a contractible, simple closed curve, the level two SU(2) invariant $\hat{J}^T_{\boldsymbol{2},q} = \hat{J}^T_{(2,\dots,2),q}$ satisfies

$$\hat{J}^T_{\boldsymbol{2},q}(C) = -q^{1/2} - q^{-1/2}.$$

If $C \subset T^2$ is a simple closed curve which is not contractible,

$$\hat{J}^T_{\boldsymbol{2},q}(C) = 2.$$

Additionally, we prove $\hat{J}^T_{\boldsymbol{2},q}$ satisfies the Kauffman bracket skein relation. These observations motivate the following definition and theorem, which characterize $\hat{J}^T_{\boldsymbol{2},q}$ skein-theoretically.

Definition 5.2 Define a Kauffman-type bracket $\langle * \rangle_\tau \in \mathbb{Z}[A^{\pm 1}, z]$ on link diagrams in T^2 (and framed links in $T^2 \times I$) by the relations:

(a) $\langle \varnothing \rangle_\tau = 1$.

(b) Let $C \subset T^2$ be a simple closed curve disjoint from a diagram $D \subset T^2$.

  (i) If C is contractible, $\langle C \sqcup D \rangle_\tau = (-A^2 - A^{-2})\langle D \rangle_\tau$.

  (ii) If C is not contractible, $\langle C \sqcup D \rangle_\tau = z\langle D \rangle_\tau$.

(c) $\langle \times \rangle_\tau = A\langle)(\rangle_\tau + A^{-1}\langle \asymp \rangle_\tau$.

Here A and z are indeterminates.

Theorem 5.3 *For any framed link $L \subset T^2 \times I$,*

$$\hat{J}^T_{\boldsymbol{2},q}(L) = \langle L \rangle_\tau |_{A^4=q,z=2}.$$

As a corollary, for any oriented, unframed link $L \subset T^2 \times I$ with diagram $D \subset T^2$,

$$J^T_2(L;q) = [-A^{-3w(D)} \langle D \rangle_\tau]|_{A^4=q,z=2}.$$

This gives a skein-theoretic construction of the toroidal Jones polynomial generalizing that of the usual Jones polynomial. In fact, our Theorem 5.4 and Corollary 5.6 prove much stronger statements defining $\hat{J}^T_{\boldsymbol{n},q}$ and J^T_n skein-theoretically for all $\boldsymbol{n}$ and n; to accomplish this we use the Kauffman bracket skein module of the thickened torus.

Why $z = 2$?

Relations (a), (b) and (c) in Definition 5.2 are identical to the relations defining the standard Kauffman bracket [14], with the additional stipulation in (b)(ii) that essential, simple closed curves can be removed from a diagram by multiplying by z. (A somewhat similar bracket is defined by Krushkal in [19].) To obtain Theorem 5.3, and for a geometrically motivated theory, it is necessary to fix $z = 2$. Indeed, only when $z = 2$ do we obtain an R–matrix, allowing us to do calculations as in Table 1. Proposition 3.8 below shows *any* pseudo-operator invariant takes the value 2 on essential, simple closed curves in T^2, if those curves have been colored by a 2–dimensional representation of a quantum group. In the appendix we examine this property further using rotation number and Lin and Wang's definition of the usual Jones polynomial [21] — see Proposition A.1 and the following discussion.

Comparison with J_n

For any link L in $T^2 \times I$, there exists a link $\hat{L} \subset S^3$ such that $(T^2 \times I) \setminus L$ and $S^3 \setminus \hat{L}$ are homeomorphic: $\hat{L}$ has a Hopf sublink H whose components are the cores of the tori which make up $S^3 \setminus (T^2 \times I)$; see Figure 3. We show $J^T_n(L)$ and $J_n(\hat{L})$ are fundamentally distinct invariants.

A key difference between the two is that J^T_n is unchanged by orientation-preserving homeomorphisms of the torus:

Proposition 4.5 *If the link diagram $D' \subset T^2$ is obtained from a diagram $D \subset T^2$ by an orientation-preserving homeomorphism of T^2, then for the corresponding links $L', L \subset T^2 \times I$, we have $J^T_n(L';q) = J^T_n(L;q)$ for all n.*

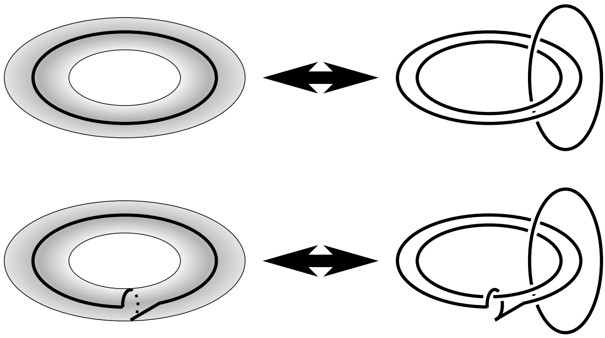


Figure 3: Nonisotopic links with homeomorphic complements in $T^2 \times I$ and S^3.

Using this proposition, we can construct infinite families of nonisotopic links in $T^2 \times I$ with identical toroidal colored Jones polynomials, whose corresponding links in S^3 all have distinct colored Jones polynomials. See Figure 3 for a simple example, where the links on the left in $T^2 \times I$ have the same toroidal colored Jones polynomials, but the corresponding links on the right in S^3 have different colored Jones polynomials.

While this makes J_n^T a less sensitive invariant than J_n, it also makes J_n^T applicable in a wider range of contexts. In Section 8, for example, we show J_n^T gives invariants of virtual links and biperiodic links. To our knowledge, J_n^T is the first invariant of virtual links to exhibit volume conjecture behavior in a nonclassical setting.

Finally, while $J_n^T(L;q)$ and $J_n(\hat{L};q)$ are different invariants, there is an important special case when the toroidal colored Jones polynomial and usual colored Jones polynomial completely determine each other (see Figure 12):

Theorem 7.3 *Let L' be a link in S^3, and consider an inclusion of L' in an embedded 2–sphere in $T^2 \times I$. Let $K \subset T^2 \times I$ be a knot projecting to an essential, simple closed curve in $T^2 \times \{0\}$, and let L be a connect sum $L = L' \# K$. Then*

$$J_n^T(L;q) = n \cdot J_n(L';q)$$

for all n.

An immediate corollary of Theorem 7.3 is that, for L and L' as in the theorem,

$$\lim_{n\to\infty} \frac{2\pi}{n} \log |J_n^T(L; e^{2\pi i/n})| = \lim_{n\to\infty} \frac{2\pi}{n} \log |J_n(L'; e^{2\pi i/n})|.$$

In Section 7, we use this fact to prove that a suitable generalization of our volume conjecture, Conjecture 1.2, implies the original volume conjecture, Conjecture 1.1 — see

Conjecture 7.1 and Corollary 7.5 below. It is not clear whether the reverse implication is true.

Outline

In Section 2, we review Kauffman bracket skein modules and operator invariants. In Section 3 we define a general pseudo-operator invariant Φ of framed links in $T^2 \times I$, and in Section 4 we specialize Φ to $U_q(\mathrm{sl}(2,\mathbb{C}))$ to obtain J_n^T and $\hat{J}_{\boldsymbol{n},q}^T$. In Section 5 we define these invariants skein-theoretically. In Section 6 we prove Theorem 6.2, and in Section 7 we discuss generalizations of Conjecture 1.2. In particular, we consider the case of nonhyperbolic links in $T^2 \times I$ and show that a generalization of Conjecture 1.2 implies the original volume conjecture. In Section 8 we discuss J_n^T as an invariant of biperiodic and virtual links. Finally, in the appendix, we study the behavior of J_2^T through the lens of Lin and Wang's formulation of the Jones polynomial [21].

Acknowledgements

We thank Ilya Kofman for his help and guidance with this project, and Hitoshi Murakami, Adam Sikora and Abhijit Champanerkar for helpful comments.

2 Background

2.1 Kauffman bracket skein modules

For a 3–manifold M and indeterminate A, let $\mathscr{L}(M)$ be the free $\mathbb{Z}[A^{\pm 1}]$–module generated by regular isotopy classes of framed links in M. The *Kauffman bracket skein module* of M [27; 30], $\mathscr{S}(M)$, is the quotient of $\mathscr{L}(M)$ by the submodule generated by the relations

(i) $\bigcirc \sqcup L = (-A^2 - A^{-2})L$,

(ii) $\times = A\,)(\, + A^{-1} \asymp$.

The links in each expression above are identical except in a ball where they look as shown, and all diagrams are assumed to have blackboard framing. Each link $L \subset M$ is represented in $\mathscr{S}(M)$ by $\langle L\rangle$, called the *Kauffman bracket* of L. If $M = \Sigma \times I$, Σ an orientable surface, we also denote the skein module of M by $\mathscr{S}(\Sigma)$. In this case, gluing two copies of $\Sigma \times I$ together along a boundary component gives $\mathscr{S}(\Sigma)$ the structure of a $\mathbb{Z}[A^{\pm 1}]$–algebra.

As an algebra, the skein module $\mathcal{S}(\mathcal{A})$ of the thickened annulus $\mathcal{A}\times I$ is generated by a copy of its core with framing parallel to $\mathcal{A}\times\{0\}$. Sending this core to z gives an algebra isomorphism $\mathcal{S}(\mathcal{A})\cong\mathbb{Z}[A^{\pm1}][z]$, so that the set $\{1,z,z^2,z^3,\dots\}$ is a basis of $\mathcal{S}(\mathcal{A})$ as a $\mathbb{Z}[A^{\pm1}]$–module. An alternate basis for $\mathcal{S}(\mathcal{A})$ is given by the Chebyshev polynomials $S_j(z)$, $j\geq 0$, defined recursively by

$$S_0(z)=1,\quad S_1(z)=z,\quad S_{j+1}(z)=zS_j-S_{j-1}. \tag{1}$$

If L is a link in M with k components, we can construct a multilinear map

$$\langle\cdots\rangle_L\colon\mathcal{S}(\mathcal{A})^{\otimes k}\to\mathcal{S}(M), \tag{2}$$

called the *Kauffman multibracket*, as follows. For $z^{i_j}\in\mathbb{Z}[A^{\pm1}][z]\cong\mathcal{S}(\mathcal{A})$, $i_j\geq 0$, let $L^{i_1,\dots,i_k}$ be the framed link in M obtained by cabling the j^{th} component of L by i_j parallel copies of itself. Define

$$\langle z^{i_1},\dots,z^{i_k}\rangle_L=\langle L^{i_1,\dots,i_k}\rangle$$

and extend $\mathbb{Z}[A^{\pm1}]$–multilinearly to all of $\mathcal{S}(\mathcal{A})$.

Sending the empty link to 1 gives an isomorphism from $\mathcal{S}(S^3)$ to $\mathbb{Z}[A^{\pm1}]$. Thus, for a link $L\in S^3$ with k components, the Kauffman multibracket is a map

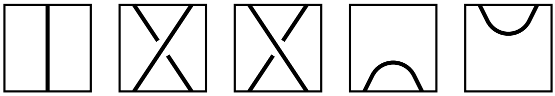
$$\langle\cdots\rangle_L\colon\mathcal{S}(\mathcal{A})^{\otimes k}\to\mathbb{Z}[A^{\pm1}].$$

Let L be an oriented, unframed link in S^3 with k components and D a diagram for L with writhe $w(D)$. The n^{th} *colored Jones polynomial* of L, $J_n(L;q)$, is defined by

$$J_n(L;q)=\left[\frac{((-1)^{n-1}A^{n^2-1})^{-w(D)}}{-A^2-A^{-2}}\langle S_{n-1}(z),\dots,S_{n-1}(z)\rangle_D\right]\Bigg|_{A^4=q}. \tag{3}$$

In Section 5 we study the Kauffman bracket skein module $\mathcal{S}(T^2)$ of the thickened torus $T^2\times I$ and its associated Kauffman multibracket. $\mathcal{S}(T^2)$ is generated as an algebra by isotopy classes of simple closed curves in T^2, which are in bijection with the set of tuples $(a,b)\in\mathbb{Z}^2$ such that either $a=b=0$, or a and b are coprime, modulo the relation $(a,b)\sim(-a,-b)$. We think of (a,b) as the curve homotopic to a times a meridian plus b times a longitude, and write $(a,b)^m$ to indicate m parallel copies of such a curve. Additionally, to avoid ambiguity, we denote the image of a link $L\subset T^2\times I$ in $\mathcal{S}(T^2)$ by $\langle L\rangle_T$ and use $\langle\cdots\rangle_{T,L}$ to mean the multibracket map determined by L,

$$\langle\cdots\rangle_{T,L}\colon\mathcal{S}(\mathcal{A})^{\otimes k}\to\mathcal{S}(T^2). \tag{4}$$



I $\quad R \quad L \quad \cap \quad \cup$

Figure 4: Elementary diagrams.

2.2 Tangle operators

An alternate definition of the colored Jones polynomial comes from the theory of *tangle operators*. The exposition here follows [17, Section 3].

Recall that a *tangle* T is a 1–manifold properly embedded (up to isotopy) in the unit cube $I^3 \subset \mathbb{R}^3$ with $\partial T \subset \{\frac{1}{2}\} \times I \times \partial I$, and define $\partial_- T = T \cap (I^2 \times \{0\})$ and $\partial_+ T = T \cap (I^2 \times \{1\})$. Choosing a regular projection onto $\{0\} \times I^2$ gives a *tangle diagram* of T.

For two tangles S and T, denote by $S \otimes T$ the tangle formed by placing S and T side by side so the boundary $I \times \{1\} \times I$ of S equals the boundary $I \times \{0\} \times I$ of T. Similarly, by $S \circ T$ we mean the tangle formed by stacking S and T vertically so $\partial_+ T = \partial_- S$; this operation can be performed only if $|\partial_+ T| = |\partial_- S|$. With these operations, the set of all tangle diagrams is generated by the five *elementary diagrams* I, R, L, $\cap$ (called a *cap*), and $\cup$ (called a *cup*) shown in Figure 4. Below, we assume tangles are equipped with orientations and framings.

Fix a quasitriangular Hopf algebra $(\mathscr{A}, \check{R})$ with R–matrix $\check{R} = \sum \alpha_i \otimes \beta_i \in \mathscr{A} \otimes \mathscr{A}$ and define a *V–coloring* of a tangle T (or one of its diagrams) to be an assignment of an $\mathscr{A}$–module to each component of T. This induces a coloring of ∂T as follows: If C is a component of color V, we assign V to each endpoint of C where C is oriented downward and the dual module V^* to each endpoint where C is oriented upward. Tensoring from left to right gives *boundary $\mathscr{A}$–modules* $T_\pm$ assigned to $\partial_\pm T$ with the empty tensor product defined to be $\mathbb{C}$.

Suppose $\mathscr{A}$ contains a unit μ with the properties

(i) $\mu \alpha \mu^{-1} = S^2(\alpha)$ for all $\alpha \in \mathscr{A}$, where S is the antipode of $\mathscr{A}$,

(ii) $\sum \alpha_i \mu^{-1} \beta_i = \sum \beta_i \mu \alpha_i$.

We call such a unit a *good unit* of $\mathscr{A}$. In this case, by the following fundamental result, any tangle T gives an $\mathscr{A}$–linear map $T_- \to T_+$.

Theorem 2.1 [17; 28] *There exist unique $\mathscr{A}$–linear operators*

$$\mathcal{F}_T = \mathcal{F}_T^{\mathscr{A},\check{R},\mu} : T_- \to T_+$$

assigned to each colored framed tangle T which satisfy

$$\mathcal{F}_{T \circ T'} = \mathcal{F}_T \circ \mathcal{F}_{T'}, \quad \mathcal{F}_{T \otimes T'} = \mathcal{F}_T \otimes \mathcal{F}_{T'},$$

and for the tangles given by the elementary diagrams with blackboard framing,

$$\begin{aligned} \mathcal{F}_I &= \mathrm{Id}, \\ \mathcal{F}_R = R \quad &\text{and} \quad \mathcal{F}_L = R^{-1}, \\ \mathcal{F}_{\curvearrowright} = E \quad &\text{and} \quad \mathcal{F}_{\curvearrowleft} = E_\mu, \\ \mathcal{F}_{\circlearrowleft} = N \quad &\text{and} \quad \mathcal{F}_{\circlearrowright} = N_{\mu^{-1}}, \end{aligned}$$

where $R = \tau \circ \check{R}$, τ the transposition map $\alpha \otimes \beta \mapsto \beta \otimes \alpha$. Additionally $E(f \otimes x) = f(x)$, $E_\mu(x \otimes f) = f(\mu x)$, $N(1) = \sum e_i \otimes e^i$ and $N_{\mu^{-1}}(1) = \sum e^i \otimes (\mu^{-1} e_i)$ for any basis e_i.

The map R is also called an R–matrix, and the map $\mathcal{F}_T$ is called the *operator invariant* of T.

Remark 2.2 The quasitriangular Hopf algebra in this construction can be replaced more generally with a *ribbon category* [32].

We set $\mathscr{A} = \mathscr{A}_q = \mathcal{U}_q(\mathrm{sl}(2,\mathbb{C}))$, the quantized universal enveloping algebra of $\mathrm{sl}(2,\mathbb{C})$ specialized to $q = e^{2\pi i/r}$ (see [13]), and fix a certain good unit μ. We also limit tangle colorings to a distinguished set of $\mathscr{A}$–modules $\{V^1, \dots, V^r\}$, V^n coming from the unique n–dimensional irreducible representation of $\mathrm{sl}(2,\mathbb{C})$ [28]. If $L \subset I^2$ is a k–component link and $\boldsymbol{n} = (n_1, \dots, n_k)$ a multi-integer, let $L(\boldsymbol{n})$ be the $\mathscr{A}$–colored link L with j^{th} component colored V^{n_j}. With this setup, the colored Jones polynomial is defined to be

$$J_n(L;q) = (q^{(n^2-1)/4})^{-w(D)} \frac{\{1\}}{\{n\}} \mathcal{F}^{\mathscr{A}_q}_{L(n,\dots,n)} = (q^{(n^2-1)/4})^{-w(D)} \frac{1}{[n]} \mathcal{F}^{\mathscr{A}_q}_{L(n,\dots,n)}, \tag{5}$$

where $q = e^{2\pi i/r}$, $r \in \mathbb{N}$, and $w(D)$ is the writhe of the tangle diagram of L. The terms $\{m\}$ and $[m]$ are the quantum integers defined by

$$\{m\} = \{m\}_q = q^{m/2} - q^{-m/2}, \quad [m] = [m]_q = \frac{\{m\}}{\{1\}}.$$

The boundary $\mathscr{A}$–modules of any colored link $L(\boldsymbol{n})$ are both $\mathbb{C}$, so $\mathcal{F}^{\mathscr{A}_q}_{L(\boldsymbol{n})}$ is a linear map from $\mathbb{C}$ to $\mathbb{C}$ — a scalar. This scalar is a Laurent polynomial in q.

In fact, the invariant J_n as defined in (5) is not strictly equal to J_n as defined in (3) — for example, the two definitions differ by a sign on the two-component unlink. Achieving precise equality requires a normalization of (5) equivalent to specializing $\mathcal{F}$ to the quantum group $\mathrm{SU}(2)_q$ rather than $U_q(\mathrm{sl}(2, \mathbb{C}))$ [17; 22]. For this reason, we refer to the invariant (3) as the $\mathrm{SU}(2)$ *colored Jones polynomial*.

In the following section we generalize the theory of tangle operators to links in $T^2 \times I$, leading in Section 4 to the definition of the toroidal colored Jones polynomial.

3 Pseudo-operator invariants

For a link L in the thickened torus $T^2 \times I$, we take a regular projection to $T^2 \times \{0\}$ to obtain a link diagram $D \subset T^2$. Let $\pi : \mathbb{R}^2 \to T^2$ be a smooth, orientation-preserving covering map with fundamental domain the unit square $I^2 \subset \mathbb{R}^2$ and deck transformations generated by horizontal and vertical unit shifts of $\mathbb{R}^2$. Let $\widetilde{D} = \pi^{-1}(D)$. Define $p \in D$ to be a *local extremum* of D, so that a small neighborhood of p is a cap or cup, if p has a lift $\tilde{p}$ which is a local extremum of $\widetilde{D}$ with respect to the height function $(x, y) \mapsto y$ on $\mathbb{R}^2$.

Definition 3.1 A point $p \in D$ is a *critical point* of D if it is a local extremum or crossing point. A *torus diagram* $D \subset T^2$ is a regular projection of a smooth link $L \subset T^2 \times I$ onto $T^2 \times \{0\}$ such that critical points are isolated.

Below, all diagrams in T^2 are assumed to be torus diagrams.

Fix a quasitriangular Hopf algebra $(\mathcal{A}, R)$ and good unit $\mu \in \mathcal{A}$. As in Section 2.2, a *V–coloring* (or simply *coloring*) of D is an assignment of an $\mathcal{A}$–module to each link component.

We now define an invariant $\Phi_\pi = \Phi_\pi^{\mathcal{A}, R, \mu}$ of oriented V–colored link diagrams in T^2 with framing parallel to $T^2 \times \{0\}$. Let $D \subset T^2$ be such a diagram and P the set of critical points of D. For each $p \in P$, there exists a small rectangular neighborhood $T(p)$ of p and a local section ψ of π, $\psi : T(p) \to \mathbb{R}^2$, giving $T(p)$ the structure of an oriented, blackboard-framed, elementary tangle diagram. In this way Theorem 2.1 assigns an $\mathcal{A}$–linear operator $\mathcal{F}_{T(p)}$ to each $T(p)$, for $p \in P$, with boundary $\mathcal{A}$–modules $T(p)_\pm$. We cannot generally extend these local assignments to a global assignment of an $\mathcal{A}$–linear operator to D, as Theorem 2.1 does for tangle diagrams in I^2. However, the local assignments of operators to each critical point still allow us to give D a value in $\mathbb{C}$ using the *state sum* formulation of the theory, as explained below.

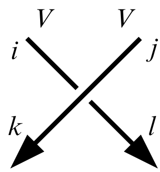


Figure 5: A V–colored tangle $T(p)$ near a crossing point p, labeled by a state σ.

For each $\mathcal{A}$–module V, fix a basis B_V of V as a $\mathbb{C}$–vector space. Removing the set of critical points P from D breaks it into components, each colored by some V and each oriented upward or downward when lifted to $\mathbb{R}^2$. A *state* σ is an assignment of a label $\sigma(S)$ to each component S of $D \setminus P$ as follows: If S is colored by the module V and oriented downward, $\sigma(S)$ is an element of B_V. If S is oriented upward, $\sigma(S)$ is an element of the dual basis B_{V^*}.

A state σ determines a *weight* $\omega_p(\sigma)$ of each critical point. For each $p \in P$, taking tensor products of the labels $\sigma(S)$ of the strands above and below p gives basis elements $\sigma(p)_\pm$ of the modules $T(p)_\pm$. Define the weight $\omega_p(\sigma) \in \mathbb{C}$ to be the coefficient of $\sigma(p)_+$ in $\mathcal{F}_{T(p)}(\sigma(p)_-)$, and define the weight of the state σ by

$$\omega(\sigma) = \prod_{p \in P} \omega_p(\sigma), \tag{6}$$

where the empty product (if D contains no critical points) is defined to be 1. Finally, set

$$\Phi_\pi(D) = \sum_\sigma \omega(\sigma), \tag{7}$$

where the sum is over all states of D.

For an example computation of the weight of a critical point, let p be the crossing point with neighborhood $T(p)$ shown in Figure 5. Viewing $T(p)$ as a tangle diagram, both tangle components of $T(p)$ are colored by the same $\mathcal{A}$–module V, so

$$T(p)_- = T(p)_+ = V \otimes V.$$

Additionally, since $T(p)$ is a positive crossing, $\mathcal{F}_{T(p)} = R$, viewed as a map from $V \otimes V$ to itself. In the given state σ, the diagram components of $T(p) \setminus p$ are assigned basis elements $e_i, e_j, e_k, e_l \in V$ as shown, where $\{e_0, e_1, \dots, e_{n-1}\}$ is a basis for V. We have $\sigma(p)_- = e_k \otimes e_l$, $\sigma(p)_+ = e_i \otimes e_j$, and if R satisfies

$$R(e_k \otimes e_l) = R_{kl}^{0,0}(e_0 \otimes e_0) + \cdots + R_{kl}^{ij}(e_i \otimes e_j) + \cdots + R_{kl}^{n-1,n-1}(e_{n-1} \otimes e_{n-1}),$$

where the $R_{kl}^{st} \in \mathbb{C}$ are scalars, $0 \le s, t \le n-1$, then $\omega_p(\sigma) = R_{kl}^{ij}$.

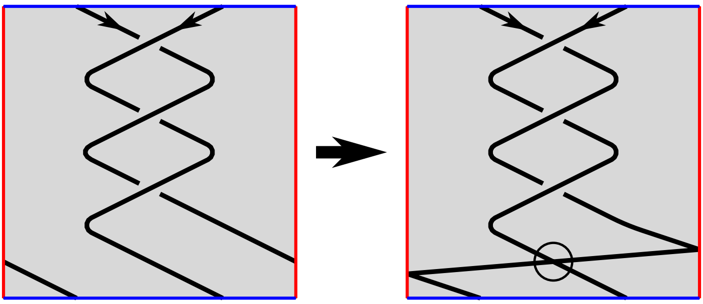


Figure 6: Constructing a virtual tangle diagram from a torus diagram.

Lemma 3.2 *The value $\Phi_\pi(D)$ does not depend on the choice of bases of the colors V.*

Proof To prove the lemma we give an alternate construction of Φ_π. Recall I^2 is a fundamental domain for the covering map $\pi\colon \mathbb{R}^2 \to T^2$; adjusting D if necessary, we assume no critical points of $\widetilde{D} = \pi^{-1}(D)$ occur in ∂I^2. Shift $\widetilde{D} \cap (\{1\} \times I)$, the points of $\widetilde{D}$ intersecting the right side of I^2, slightly upward by an isotopy of I^2 which is the identity on $\{0\} \times I$ and does not change the set of critical points of $\widetilde{D}$. Because the left and right sides of I^2 are identified in the torus, each point $q \in \widetilde{D} \cap (\{0\} \times I)$ has a corresponding point $q' \in \widetilde{D} \cap (\{1\} \times I)$, slightly higher than q as a result of the isotopy. As a final step of the construction, connect each q and q' by a curve $c_q\colon I \to I^2$ which satisfies $c_q(0) = q$, $c_q(1) = q'$, and is monotonically increasing in height. This produces a *virtual tangle diagram* $D' = (I^2 \cap \widetilde{D}) \cup_{q \in D \cap (\{0\} \times I)} c_q(I)$ whose classical (ie nonvirtual) critical points are the same as the critical points of D and whose virtual crossings are any point where some $c_q(I)$, for $q \in \widetilde{D} \cap (\{0\} \times I)$, intersects another point of D'. See Figure 6 for an example, where the virtual crossing on the right is circled. We assume virtual crossings are isolated from other critical points.

The coloring of D induces a coloring of D' in an obvious way. As before, let P be the set of (classical and virtual) critical points of D' with $T(p)$ a small rectangular neighborhood of $p \in P$. If $p \in P$ is a classical critical point, the functor $\mathcal{F}$ of Theorem 2.1 associates an $\mathcal{A}$–linear operator $\mathcal{F}_{T(p)}$ to $T(p)$ which agrees with the operator assigned to $T(p)$ in the construction of Φ_π. If $p \in P$ is a virtual crossing, define $\mathcal{F}_{T(p)}$ to be the transposition map $\tau : \alpha \otimes \beta \mapsto \beta \otimes \alpha$ as in [15]. (This is a $\mathbb{C}$–linear map but not generally an $\mathcal{A}$–linear one.) Because $\partial_- D'$ and $\partial_+ D'$ are identified in the torus, $D'_- = D'_+ = V$ for some $\mathcal{A}$–module V. Thus, extending the local operator assignments $\mathcal{F}_{T(p)}$, for $p \in P$, as in Theorem 2.1 associates D' with a $\mathbb{C}$–linear map $\phi\colon V \to V$. Define $\Phi'_\pi(D') = \mathrm{Tr}(\phi)$; we claim $\Phi'_\pi(D') = \Phi_\pi(D)$.

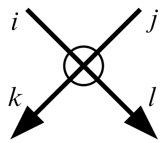


Figure 7: A state assignment near a virtual crossing.

Computing $\Phi'_\pi(D')$ as a state sum, as in [15; 17; 25], shows the two invariants agree. Fix a basis for each $\mathscr{A}$–module. As in the construction of Φ_π, a state is an assignment of basis elements to components of $D' \setminus P$ and the weight of a state is the product of the weights of the critical points. Taking the trace of ϕ ensures identified strands of D'_- and D'_+ are assigned the same basis element in any state with nonzero weight. If the strands near a virtual crossing p are assigned basis elements e_i, e_j, e_k and e_l, as in Figure 7, the weight of p is $\delta_i^l \delta_j^k$, where δ is the Kronecker delta. This ensures the identified strands on either side of p have the same state, in which case p has weight 1.

If we compute $\Phi_\pi(D)$ using the same bases, we have

$$\Phi_\pi(D) = \sum_\sigma \omega(\sigma) = \mathrm{Tr}(\phi) = \Phi'_\pi(D').$$

This shows the definition of $\Phi'_\pi(D')$ does not depend on the choice of curves c_q, for $q \in D \cap (\{0\} \times I)$. Since Φ'_π does not depend on a choice of basis, neither does Φ_π. □

We write Φ rather than Φ_π in the next definition because we will ultimately show Φ does not depend on the choice of covering map π. Before proving this, however, we give the main result of the section.

Definition 3.3 Let $L \subset T^2 \times I$ be a framed, oriented, V–colored link and D a diagram for L with framing parallel to T^2. Define the *pseudo-operator invariant* of L, depending on $(\mathscr{A}, R)$ and μ, by

$$\Phi(L) = \Phi^{\mathscr{A},R,\mu}(L) = \Phi_\pi^{\mathscr{A},R,\mu}(D).$$

Theorem 3.4 *Φ is an invariant of framed, oriented, V–colored links in $T^2 \times I$. That is, if D_1 and D_2 are two diagrams of a framed, oriented, V–colored link $L \subset T^2 \times I$ with each having framing parallel to $T^2 \times \{0\}$, then $\Phi(D_1) = \Phi(D_2)$.*

Proof Consider the lift $\widetilde{D}_i$ of D_i to $\mathbb{R}^2$ for $i = 1, 2$. By construction, $\widetilde{D}_i$ is the diagram of a biperiodic link $\tilde{L} \subset \mathbb{R}^2 \times I$ such that the critical points of $\widetilde{D}_i$ are lifts of critical points of D_i. Let $f_t(x)\colon I \times (T^2 \times I) \to T^2 \times I$ be an ambient isotopy carrying D_1 to D_2, such that $f_0 \equiv \mathrm{Id}$ and $f_1(D_1) = D_2$. Then f_t lifts to a biperiodic isotopy $\tilde{f}_t$ of

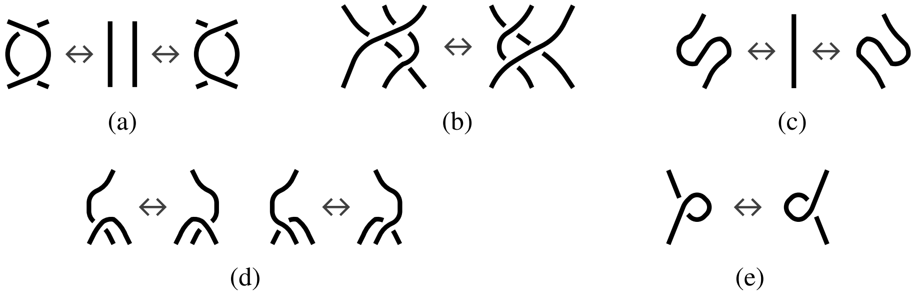


Figure 8: Local tangle moves.

$\mathbb{R}^2 \times I$ taking $\widetilde{D}_1$ to $\widetilde{D}_2$. Because $\widetilde{D}_1$ and $\widetilde{D}_2$ are locally blackboard-framed tangle diagrams, a well-known theorem [8; 28] asserts that $\tilde{f}_t$ decomposes into a sequence $\tilde{g}_t$ of diagram-preserving isotopies and the moves shown in Figure 8 (with all possible orientations). We assume the isotopies and moves are biperiodic, ie applied to each lifted copy of a region of D_1 simultaneously.

Because $\tilde{g}_t$ is biperiodic, it descends to a sequence g_t of the same moves on T^2 carrying D_1 to D_2. Hence it suffices to check invariance of Φ under each local move, which follows from properties of $\mathscr{A}$, R and μ. For example, the equation $R^{-1} \circ R = \mathrm{Id} = R \circ R^{-1}$ implies invariance under move (a). Move (b) follows from the fact that R satisfies the Yang–Baxter equation [31], and moves (c)–(e) also follow from properties of R and μ — see [17, Theorem 3.6] for details. □

Remark 3.5 The construction of Φ_π given in the proof of Lemma 3.2 is similar to Kauffman's quantum invariant for virtual links [15], in that virtual crossings are associated with the transposition map τ. However, the two invariants have significant differences. We can think of the virtual diagram D' in the proof of Lemma 3.2 as the diagram of a tangle on a cylinder $S^1 \times I$: the original diagram D sits on the "front" of the cylinder, while the added curves c_q circle around the "back". This is one difference between our invariant and Kauffman's — the use of a cylinder to create the virtual diagram rather than a torus. Another difference is that Kauffman's invariant is defined in the context of *rotational virtual knot theory* (see [16]) — it is not invariant under virtual Reidemeister I-moves. We achieve invariance under virtual I–moves by placing all classical critical points on the front of the cylinder, where the orientation of the cylinder matches the orientation of the virtual diagram. If a critical point were moved to the back of the cylinder, that point's orientation on the cylinder would not match its orientation in the virtual diagram, and the two local operator assignments in the two constructions of Φ_π would disagree.

We use the phrase "pseudo-operator invariant" because, as remarked above, a torus diagram $D \subset T^2$ cannot generally be associated with an $\mathscr{A}$–linear operator using our construction. It is interesting that the local assignments of $\mathscr{A}$–linear operators to critical points of D still allow us to define $\Phi(D)$, which seems to encode geometric information about the link L. For an example of computing Φ with a specific $\mathscr{A}$, see Section 6.

The fact that $\Phi(D) = \Phi_\pi(D)$ does not depend on π follows from the proposition below, which shows Φ_π is invariant under orientation-preserving homeomorphisms of T^2.

Proposition 3.6 *Let $D, D' \subset T^2$ be oriented, V–colored link diagrams with blackboard framing. If f is an orientation-preserving homeomorphism of T^2 satisfying $f(D) = D'$, then $\Phi_\pi(D) = \Phi_\pi(D')$.*

Proof Since Φ_π is an isotopy invariant, it suffices to prove the theorem for a set of homeomorphisms generating the mapping class group $\mathrm{Mod}(T^2)$. To this end, we consider two Dehn twists, about two curves in T^2 which lift via π to horizontal and vertical lines in $\mathbb{R}^2$. Let $l \subset T^2$ be a simple closed curve lifting to a vertical line in $\mathbb{R}^2$ such that l contains no critical points of D, and choose a bicollar neighborhood $N(l)$ of l satisfying that

(i) no critical points of D occur within $N(l)$,

(ii) each connected component of $D \cap N(l)$ intersects l only once, transversely.

Now suppose $f : T^2 \to T^2$ is an upward twist (from left to right) about l which is the identity outside of $N(l)$. See Figure 9 for an example. Let c be a component of $D \cap N(l)$ — then c is a curve which increases or decreases monotonically as it travels across $N(l)$ from left to right. If c is increasing, $f(c)$ is also monotonically increasing and contains no critical points. If c is decreasing, $f(c)$ contains a minimum to the left of l and a maximum to the right of l and no critical points other than these; see Figure 9. The cases of twisting downward and twisting about horizontal lines are similar. Finally, because f is injective, the crossing points of D' are the same as those of D.

It follows that the only critical points of $f(D) = D'$ which do not occur in D are max–min pairs formed as above. When $\Phi_\pi(D')$ is computed as a state sum, the weights of these max–min pairs cancel as in identity (c) of Figure 8. We conclude $\Phi_\pi(D) = \Phi_\pi(D')$. □

Corollary 3.7 *The value $\Phi_\pi(D)$ does not depend on the choice of covering map $\pi : \mathbb{R}^2 \to T^2$.*

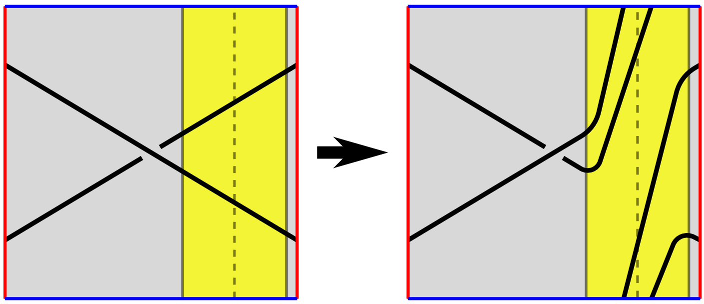


Figure 9: Twisting a diagram.

Proof Let $\pi_1, \pi_2\colon \mathbb{R}^2 \to T^2$ be two smooth, orientation-preserving covering maps with fundamental domain I^2. The uniqueness property of covering spaces gives an orientation-preserving homeomorphism $\tilde{f}\colon \mathbb{R}^2 \to \mathbb{R}^2$ which satisfies $\pi_2 \circ \tilde{f} = \pi_1$, and since π_1 and π_2 have the same fundamental domain and deck transformations, $\tilde{f}$ descends to an orientation-preserving homeomorphism f of T^2 satisfying $f \circ \pi_1 = \pi_2$. By Proposition 3.6, and noting the value $\Phi_{\pi_i}(D)$ is completely determined by the lift ${\pi_i}^{-1}(D)$,

$$\Phi_{\pi_1}(D) = \Phi_{\pi_1}(f(D)) = \Phi_{f\circ\pi_1}(D) = \Phi_{\pi_2}(D). \qquad \square$$

We conclude the section with a general property of pseudo-operator invariants. Though the proposition is a simple observation, it motivates the skein theory to come in Section 5.

Proposition 3.8 *Let $K \subset T^2 \times I$ be a knot projecting to an essential, simple closed curve in $T^2 \times I$ with framing parallel to $T^2 \times \{0\}$. If K is colored by an n–dimensional $\mathcal{A}$–module V,*

$$\Phi(K) = n.$$

Proof Applying Proposition 3.6, we may assume without loss of generality that K lifts to a vertical line in $\mathbb{R}^2$. Then the proof of Lemma 3.2 shows $\Phi(K)$ is the trace of the identity map of V, so $\Phi(K) = \dim(V) = n$. $\square$

4 Quantum invariants for sl$(2, \mathbb{C})$ and the toroidal colored Jones polynomial

4.1 An invariant of framed, unoriented links in $T^2 \times I$

As with the colored Jones polynomial, we now specialize to $\mathcal{A} = \mathcal{A}_q = \mathcal{U}_q(\mathrm{sl}(2, \mathbb{C}))$, for $q = e^{2\pi i/r}$, and limit $\mathcal{A}$–modules to the set $\{V^1, \dots, V^r\}$ as in Section 2.2.

For a link $L \subset T^2 \times I$ with k components and multi-integer $\boldsymbol{n} = (n_1, \dots, n_k)$, let $L(\boldsymbol{n}) = L(n_1, \dots, n_k)$ be the V–colored link L with j^{th} component colored by V^{n_j}. By $1 \le \boldsymbol{n} \le r$ in the definition below we mean that $1 \le n_j \le r$ for all n_j.

Definition 4.1 Given a framed, unoriented link $L \subset T^2 \times I$ with k components, fix an orientation of each component. For $q = e^{2\pi i/r}$, $1 \le \boldsymbol{n} \le r$, define $\hat{J}^T_{\boldsymbol{n},q}(L)$ by

$$\hat{J}^T_{\boldsymbol{n},q}(L) = \Phi^{\mathscr{A}_q, R, \mu}(L(\boldsymbol{n})).$$

Theorem 4.2 *$\hat{J}^T_{\boldsymbol{n},q}(L)$ is an invariant of framed, unoriented links in $T^2 \times I$. That is, $\hat{J}^T_{\boldsymbol{n},q}(L)$ does not depend on the orientation chosen for each component of L.*

Proof Let $D, D' \subset T^2$ be oriented diagrams of $L(\boldsymbol{n})$ with framing parallel to $T^2 \times \{0\}$ and D' obtained from D by changing the orientation of a link component C. It suffices to show

$$\Phi^{\mathscr{A}_q, R, \mu}(D) = \Phi^{\mathscr{A}_q, R, \mu}(D').$$

Suppose C is colored by V^n and let C' be the corresponding component of D', also colored by V^n. Let $p \in C$ be a critical point, p' the same point of C', and $T(p)$, $T(p')$ small rectangular neighborhoods of each. Then each copy of V^n coming from C in $T(p)_\pm$ corresponds to a copy of $(V^n)^*$ in $T(p')_\pm$ and vice versa. V^n is self-dual as an $\mathscr{A}$–module via a canonical isomorphism $\varphi\colon (V^n)^* \to V^n$, and we use φ to identify the modules $T(p)_\pm$ and $T(p')_\pm$.

We apply Lemma 3.18 and Remark 3.26 of [17], which state that if n is odd, $\mathcal{F}_{T(p)} = \mathcal{F}_{T(p')}$ as maps from $T(p)_-$ to $T(p)_+$ for all p. Thus $\Phi(D) = \Phi(D')$ if n is odd. Suppose n is even. If p is a crossing then $\mathcal{F}_{T(p)} = \mathcal{F}_{T(p')}$, and if p is an extreme point then $\mathcal{F}_{T(p)} = -\mathcal{F}_{T(p')}$. Self-duality of V^n induces a bijection between the states of D and the states of D', and it follows that if σ is a state of D and σ' the corresponding state of D', $\omega(\sigma) = (-1)^j \omega(\sigma')$, where j is the total number of extreme points of C. Since $C \subset T^2$ is a closed curve, j is even and $\omega(\sigma) = \omega(\sigma')$. We conclude $\Phi(D) = \Phi(D')$ if n is even. □

The invariant $\hat{J}^T_{\boldsymbol{n},q}$ should be thought of as a toroidal analogue of the invariant $J_{L,\boldsymbol{k}}$ of [17]. One might also be reminded of the Kauffman bracket skein module, another invariant of framed, unoriented links — this comparison will be made precise in the next section. Like the invariant $J_{L,\boldsymbol{k}}$ of [17] or the Kauffman bracket skein module of S^3, $\hat{J}^T_{\boldsymbol{n},q}$ can be normalized to obtain an invariant of oriented, unframed links in

$T^2 \times I$ analogous to the colored Jones polynomial.

4.2 The toroidal colored Jones polynomial

To create an invariant of unframed links, we use the fact that as an endomorphism of $V^n \otimes V^n$, $0 \le n \le r$, the R–matrix R satisfies [17; 30]

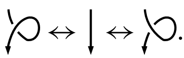
$$(\mathrm{Id} \otimes E_\mu)(R^{\pm 1} \otimes \mathrm{Id})(\mathrm{Id} \otimes N) = q^{\pm(n^2-1)/4}\,\mathrm{Id},$$

where $N\colon \mathbb{C} \to V^n \otimes (V^n)^*$ and $E_\mu\colon V^n \otimes (V^n)^* \to \mathbb{C}$ are the maps given in Theorem 2.1. Pictorially, this is equivalent to the two equations

(8) $\hat{J}^T_{\boldsymbol{n},q}($$) = q^{(n^2-1)/4} \cdot \hat{J}^T_{\boldsymbol{n},q}($$)$, $\hat{J}^T_{\boldsymbol{n},q}($$) = q^{-(n^2-1)/4} \cdot \hat{J}^T_{\boldsymbol{n},q}($$)$,

where the diagrams represent oriented, unframed links and the component shown is colored by V^n.

If $L \subset T^2 \times I$ is oriented and unframed, any two diagrams of L are related by a sequence of the moves in Figure 8 and additions or removals of curls,

(9) 

This fact, combined with (8), motivates Definition 4.3. Similar to above, given a diagram $D \subset T^2$ of a k–component link and multi-integer $\boldsymbol{n} = (n_1, \dots, n_k)$, let $D(\boldsymbol{n})$ indicate D with j^{th} component colored by V^{n_j}. Define $\hat{J}^T_{\boldsymbol{n},q}(D) := \Phi^{\mathscr{A}_q, R, \mu}(D(\boldsymbol{n}))$.

Definition 4.3 Let $L \subset T^2 \times I$ be an oriented, unframed link with k components, $D \subset T^2$ a diagram of L, and $C_1, \dots, C_k$ the link components of D. Define $J^T_{\boldsymbol{n},q}(L)$ by

$$J^T_{\boldsymbol{n},q}(L) = q^{\alpha/4}\,\hat{J}^T_{\boldsymbol{n},q}(D),$$

where $\alpha = -\sum_{j=1}^k w(C_j) \cdot ({n_j}^2 - 1)$ and $w(C_j)$ is the writhe of C_j, ie the sum of the signs of its self-crossings.

It follows from (8) and the proof of Theorem 3.4 that $J^T_{\boldsymbol{n},q}(L)$ is an invariant of oriented, unframed links in $T^2 \times I$.

If all components of L are given the same color, ie if $\boldsymbol{n} = (n, \dots, n)$ for some $n \in \mathbb{N}$, we can define a similar invariant which agrees with $J^T_{\boldsymbol{n},q}$ if L is a knot. This next definition is our analogue of the colored Jones polynomial.

Definition 4.4 For an oriented, unframed link $L \subset T^2 \times I$ with diagram D and $n \in \mathbb{N}$, define the n^{th} *toroidal colored Jones polynomial* $J_n^T(L;q)$ of L by

$$J_n^T(L;q) = (q^{(n^2-1)/4})^{-w(D)} \hat{J}^T_{(n,\dots,n),q}(D),$$

where $w(D)$ is the writhe of D.

Compare Definition 4.4 with (5) — the definitions are analogous except for a factor of $1/[n]$. This factor may be included in the definition of J_n because J_n is always divisible by $[n]$ as a Laurent polynomial in q; this follows from the $\mathcal{A}$–linearity of the operator in Theorem 2.1. Because there is no guarantee of global $\mathcal{A}$–linearity in the construction of J_n^T, we cannot divide by $[n]$. In particular, if $q = e^{2\pi i/n}$,

$$[n] = \frac{e^{\pi i} - e^{-\pi i}}{\{1\}} = 0. \tag{10}$$

Additionally, the root of unity q may be replaced by an indeterminate in the definition of $J_n^T(L;q)$ without affecting calculations — see, for example, [25]. This justifies thinking of $J_n^T(L;q)$ as a Laurent polynomial in q and accounts for our slight change in notation. We prefer J_n^T to $J^T_{\boldsymbol{n},q}$ for simplicity in calculations, and don't make use of $J^T_{\boldsymbol{n},q}$ outside of this section. We also remark that, because Dehn twists do not affect the signs of crossings in a diagram, Proposition 3.6 extends to J_n^T and $J^T_{\boldsymbol{n},q}$:

Proposition 4.5 *Let $L, L' \subset T^2 \times I$ be oriented links with respective diagrams $D, D' \subset T^2$. If f is an orientation-preserving homeomorphism of T^2 satisfying $f(D) = D'$, then $J_n^T(L;q) = J_n^T(L';q)$ and $J^T_{\boldsymbol{n},q}(L) = J^T_{\boldsymbol{n},q}(L')$ for all n and $\boldsymbol{n}$.*

As a final result of the subsection, we give a cabling formula for $\hat{J}^T_{\boldsymbol{n},q}$ analogous to the cabling formula for the invariant $J_{L,\boldsymbol{k}}$; see [17, Theorem 4.15]. For a framed, unoriented link $L \subset T^2 \times I$, this expresses the value $\hat{J}^T_{\boldsymbol{n},q}(L)$ in terms of $\hat{J}^T_{\boldsymbol{2},q} = \hat{J}^T_{(2,\dots,2),q}$ evaluated on certain cablings of L. This will allow us to develop $\hat{J}^T_{\boldsymbol{n},q}$ from a skein-theoretic viewpoint in the next section.

Let L be a k–component link and $\boldsymbol{n} = (n_1, \dots, n_k)$ a multi-integer. As in Section 2.1, denote by $L^{\boldsymbol{n}}$ the cabling of L which replaces the j^{th} component of L by n_j parallel pushoffs of itself, oriented compatibly if the link is oriented, with associated diagram $D^{\boldsymbol{n}}$. Below, the sum is over all $\boldsymbol{i} = (i_1, \dots, i_k)$ with $1 \le i_j \le \frac{1}{2}(n_j - 1)$, and

$$(-1)^{\boldsymbol{i}} \binom{\boldsymbol{n}-1-\boldsymbol{i}}{\boldsymbol{i}} = \prod_{j=1}^{k} (-1)^{i_j} \binom{n_j-1-i_j}{i_j}.$$

Theorem 4.6 (cabling formula) *Let $L \subset (T^2 \times I)$ be a framed, unoriented link, $\boldsymbol{n}$ a coloring of L, and D a torus diagram for L with framing parallel to $T^2 \times \{0\}$. Then*

$$\hat{J}^T_{\boldsymbol{n},q}(L) = \sum_{\boldsymbol{i}=0}^{(\boldsymbol{n}-1)/2} (-1)^{\boldsymbol{i}} \binom{\boldsymbol{n}-1-\boldsymbol{i}}{\boldsymbol{i}} \hat{J}^T_{\boldsymbol{2},q}(L^{\boldsymbol{n}-1-2\boldsymbol{i}})$$
$$= \sum_{\boldsymbol{i}=0}^{(\boldsymbol{n}-1)/2} (-1)^{\boldsymbol{i}} \binom{\boldsymbol{n}-1-\boldsymbol{i}}{\boldsymbol{i}} q^{3w(D^{\boldsymbol{n}-1-2\boldsymbol{i}})/4} J^T_2(L^{\boldsymbol{n}-1-2\boldsymbol{i}}; q).$$

We sketch the proof, following closely the proof of Theorem 4.15 in [17]. We first require a lemma — cf [17, Lemma 3.10] — which gives useful properties of pseudo-operator invariants.

Lemma 4.7 *Let $(\mathscr{A}, R)$ be a quasitriangular Hopf algebra with good unit μ. Let $D \subset T^2 \times I$ be a colored torus diagram and C a link component of D colored by V.*

(a) *If $V = X \oplus Y$, or more generally V is an extension of Y by X (ie there is a short exact sequence $0 \to X \to V \to Y \to 0$ of $\mathscr{A}$–modules), then*
$$\Phi(D) = \Phi(D_X) + \Phi(D_Y),$$
where D_Z denotes the torus diagram obtained by changing the color of C to Z.

(b) *If $V = X \otimes Y$, then*
$$\Phi(D) = \Phi(D_{XY}),$$
where D_{XY} is the diagram obtained by replacing C by two parallel pushoffs of itself (using the framing) colored by X and Y, respectively.

Proof To prove (a), fix bases B_V, B_X and B_Y so that $B_X \subset B_V$ (viewing X as a subspace of V) and B_Y is the projection of $\widetilde{B}_Y = B_V - B_X$. We call state labels from B_X or B_{X^*} *X–labels*, whereas those from $\widetilde{B}_Y$ or $\widetilde{B}_{Y^*}$ are *Y–labels*.

If σ is a state of D with nonzero weight, then the corresponding labels on the arcs of C must be either all X–labels (written $\sigma|C \subset X$) or all Y–labels (written $\sigma|C \subset Y$). This follows from the $\mathscr{A}$–invariance of $X \subset V$ (and dually of $Y^* \subset V^*$) — if the component of C on one side of a critical point has an X–label and the component on the other side has a Y–label, the weight of the critical point in that state will be zero. From this, we see

$$\Phi(D) = \sum \omega(\sigma) = \sum_{\sigma|C\subset X} \omega(\sigma) + \sum_{\sigma|C\subset Y} \omega(\sigma) = \Phi(D_X) + \Phi(D_Y),$$

as desired.

Statement (b) is a fundamental property of operator invariants — see [32, Chapter 1] or [17, Lemma 3.10]. As with statement (a), we extend to the case of pseudo-operator invariants by considering each critical point individually. □

Proof of Theorem 4.6 It is a classical result — see [17, Corollary 2.15] — that, for $0 \le n < r$, the equality

$$V^{n+1} = \sum_{j=0}^{n/2} (-1)^j \binom{n-j}{j} (V^2)^{n-2j} \tag{11}$$

holds in the representation ring of $\mathscr{A}_q$, where $q = e^{2\pi i/r}$ and the sum is over all j with $0 \le 2j \le n$. Here V^n refers to the $\mathscr{A}_q$–module V^n, while $(V^n)^j$ indicates the j^{th} tensor product of V^n with itself. The proof of (11) uses the fact that the modules V^n satisfy the recurrence relation $V^{n+1} = V^2 V^n - V^{n-1}$, the same recurrence relation defining the Chebyshev polynomials in (1).

The first equality of Theorem 4.6 now follows from combining Lemma 4.7 and (11), and the second equality comes from Definition 4.4. □

Remark 4.8 Toroidal analogues of other link invariants can be constructed by considering quantum groups other than $U_q(\mathrm{sl}(2,\mathbb{C}))$. For example, letting $\mathscr{A} = U_q(\mathrm{sl}(m,\mathbb{C}))$ for general m gives a toroidal analogue of the specialization $P_L(q^m, q - q^{-1})$, where P_L is the two-variable HOMFLY polynomial of a link L. Letting $\mathscr{A} = U_q G$, where $G = \mathrm{so}(m)$ or $\mathrm{sp}(2m)$, leads to a toroidal analogue of a certain specialization of the two-variable Kauffman polynomial, depending on the choice of G; see [28, Section 6.1]. In Section 5, we'll construct a toroidal analogue of the SU(2) colored Jones polynomial by setting $\mathscr{A} = \mathrm{SU}(2)_q$.

5 The toroidal colored Jones polynomial and skein theory

In this section only, we consider the invariant defined by specializing Φ to the quantum group $\mathrm{SU}(2)_q$ rather than $U_q(\mathrm{sl}(2,\mathbb{C}))$. This constitutes a certain normalization of the $\mathrm{sl}(2,\mathbb{C})$ invariant and we denote the SU(2) version by the same notation, $\hat{J}^T_{\boldsymbol{n},q}$. This is consistent with literature on the colored Jones polynomial and the operators involved are discussed, for example, in [18; 22].

The goal of the section is to develop the SU(2) toroidal colored Jones polynomial skein-theoretically. This begins with an observation about the level two framed invariant $\hat{J}_{\mathbf{2},q}$.

Lemma 5.1 *The level two* SU(2) *invariant $\hat{J}^T_{\mathbf{2},q}$ has the following properties*:

(a) $\hat{J}^T_{\mathbf{2},q}(\varnothing) = 1$.

(b) *Let $C \subset T^2$ be a simple closed curve disjoint from a diagram $D \subset T^2$.*

  (i) *If C is contractible, $\hat{J}^T_{\mathbf{2},q}(C \sqcup D) = (-q^{1/2} - q^{-1/2})\hat{J}^T_{\mathbf{2},q}(D) = -[2]\hat{J}^T_{\mathbf{2},q}(D)$.*

  (ii) *If C is not contractible, $\hat{J}^T_{\mathbf{2},q}(C \sqcup D) = 2\hat{J}^T_{\mathbf{2},q}(D)$.*

(c) $\hat{J}^T_{\mathbf{2},q}(\times) = q^{1/4}\hat{J}^T_{\mathbf{2},q}()\,() + q^{-1/4}\hat{J}^T_{\mathbf{2},q}(\asymp)$.

Proof Properties (a), (b)(i) and (c) are identical to the relations defining the usual Kauffman bracket. Because they are local properties — (b)(i) is local in the sense that we can assume C exists in a coordinate neighborhood of T^2 — the proofs are the same as for the usual SU(2) Jones polynomial; see [22, Theorem 4.1]. Each property reduces to an algebraic statement about the quantum group $\mathrm{SU}(2)_q$.

To prove property (b)(ii) holds for $\hat{J}^T_{\mathbf{2},q}$, suppose $C \subset T^2$ is a simple, closed essential curve. Then $\hat{J}^T_{2,q}(C) = 2$ by Proposition 3.8. The general statement follows from the multiplicativity of $\hat{J}_{n,q}$ on disjoint diagrams; that is,

$$\hat{J}^T_{\mathbf{2},q}(U \sqcup D) = \hat{J}^T_{\mathbf{2},q}(U) \cdot \hat{J}^T_{\mathbf{2},q}(D) = 2\hat{J}^T_{\mathbf{2},q}(D). \qquad \square$$

Lemma 5.1 leads to the following definition and theorem:

Definition 5.2 Define a Kauffman-type bracket $\langle * \rangle_\tau \in \mathbb{Z}[A^{\pm 1}, z]$ on link diagrams in T^2 (and framed links in $T^2 \times I$) by the relations:

(a) $\langle \varnothing \rangle_\tau = 1$.

(b) Let $C \subset T^2$ be a simple closed curve disjoint from a diagram $D \subset T^2$.

  (i) If C is contractible, $\langle C \sqcup D \rangle_\tau = (-A^2 - A^{-2})\langle D \rangle_\tau$.

  (ii) If C is not contractible, $\langle C \sqcup D \rangle_\tau = z \cdot \langle D \rangle_\tau$.

(c) $\langle \times \rangle_\tau = A\langle)\,(\rangle_\tau + A^{-1}\langle \asymp \rangle_\tau$.

Theorem 5.3 *For any framed link $L \subset T^2 \times I$,*

$$\hat{J}^T_{\mathbf{2},q}(L) = \langle L \rangle_\tau|_{A^4=q, z=2}.$$

To extend Theorem 5.3 to all values of $\boldsymbol{n}$ for $\hat{J}_{\boldsymbol{n},q}$, we consider the skein module of the thickened torus, $\mathcal{S}(T^2)$. As Section 2.1 discusses, a basis for $\mathcal{S}(T^2)$ as a

$\mathbb{Z}[A^{\pm 1}]$–module is given by positive powers of the tuples (a,b) such that either $a = b = 0$ or a and b are coprime. Let $p_2\colon \mathcal{S}(T^2) \to \mathbb{Z}[A^{\pm 1}]$ be the $\mathbb{Z}[A^{\pm 1}]$–linear map defined by

$$p_2((a,b)^m) = \begin{cases} 1 & \text{if } a = b = 0, \\ 2^m & \text{otherwise.} \end{cases}$$

Then it's clear that, for any framed link $L \subset T^2 \times I$,

$$\langle L \rangle_\tau |_{z=2} = p_2(\langle L \rangle_T),$$

where $\langle L \rangle_T$ is the class of L in $\mathcal{S}(T^2)$.

We can now state the full result using $\mathcal{S}(T^2)$.

Theorem 5.4 *For any oriented, unframed link $L \subset T^2 \times I$ with diagram $D \subset T^2$,*

$$\hat{J}^T_{\boldsymbol{n},q}(L) = p_2(\langle S_{n_1-1}(z), \dots, S_{n_k-1}(z)\rangle_{T,D})|_{A^4=q},$$

where $\hat{J}^T_{\boldsymbol{n},q}$ is the SU(2) *invariant and $\langle * \rangle_{T,D}\colon \mathcal{S}(\mathcal{A})^{\otimes k} \to \mathcal{S}(T^2)$ is the multibracket of (4).*

Proof A closed formula for the n^{th} Chebyshev polynomial, as defined in (1), is given by

$$S_n(z) = \sum_{j=0}^{n/2} (-1)^j \binom{n-j}{j} z^{n-2j} \tag{12}$$

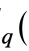
where the sum is over all integers j with $0 \le 2j \le n$. Let $L \subset T^2 \times I$ be a k–component link and $\boldsymbol{n} = (n_1, \dots, n_k)$ a multi-integer. Applying the above formula and the multilinearity of p_2 and the Kauffman multibracket, we have

$$\begin{aligned} p_2(\langle S_{n_1}(z), \dots, S_{n_k}(z)\rangle_{T,D}) &= \sum_{\boldsymbol{j}=0}^{\boldsymbol{n}/2} (-1)^{\boldsymbol{j}} \binom{\boldsymbol{n}-\boldsymbol{j}}{\boldsymbol{j}} p_2(\langle L^{\boldsymbol{n}-2\boldsymbol{j}}\rangle_T) \\ &= \sum_{\boldsymbol{j}=0}^{\boldsymbol{n}/2} (-1)^{\boldsymbol{j}} \binom{\boldsymbol{n}-\boldsymbol{j}}{\boldsymbol{j}} \hat{J}^T_{\boldsymbol{2},q}(L^{\boldsymbol{n}-2\boldsymbol{j}}) \\ &= \hat{J}^T_{\boldsymbol{n}-1,q}(L). \end{aligned}$$

The second equality is Theorem 5.3. The third comes from Theorem 4.6, which applies in the SU(2) theory since the same relation $V^{j+1} = V^2 V^j - V^{j-1}$ holds in the representation ring. □

Having constructed the SU(2) $\hat{J}_{\boldsymbol{n},q}$ skein-theoretically, we can define a skein-theoretic toroidal colored Jones polynomial. The SU(2) version of $\hat{J}^T_{\boldsymbol{n},q}$ satisfies [23]

$\hat{J}^T_{\boldsymbol{n},q}($  $) = (-1)^{n-1} q^{(n^2-1)/4} \cdot \hat{J}^T_{\boldsymbol{n},q}($  $),$

$\hat{J}^T_{\boldsymbol{n},q}($  $) = (-1)^{n-1} q^{-(n^2-1)/4} \cdot \hat{J}^T_{\boldsymbol{n},q}($  $),$

where the strand shown in the diagram is colored by V^n. Subsequently:

Definition 5.5 The SU(2) *toroidal colored Jones polynomial* J_n^T of an oriented, unframed link $L \subset T^2 \times I$ with diagram D is defined by

$$J_n^T(L;q) = ((-1)^{n-1} q^{(n^2-1)/4})^{-w(D)} \hat{J}^T_{(n,\dots,n),q}(D).$$

We immediately have:

Corollary 5.6 *The* SU(2) *toroidal colored Jones polynomial is skein-theoretically defined by*

$$J_n^T(L;q) = \big[((-1)^{n-1} A^{n^2-1})^{-w(D)} p_2(\langle S_{n-1}(z), \dots, S_{n-1}(z)\rangle_D)\big]\big|_{A^4=q}.$$

Compare the right side of Theorem 5.4 with (3) — the missing factor of $1/(-A^2-A^{-2})$ is analogous to the missing $1/[n]$ factor in the $U_q(\mathrm{sl}(2,\mathbb{C}))$ case.

The skein theoretic definitions of Theorem 5.4 and Corollary 5.6 let us extend the SU(2) invariants $\hat{J}^T_{\boldsymbol{n},q}$ and J_n^T to links in orientable manifolds other than $T^2 \times I$, using the bracket $\langle * \rangle_\tau$ of Definition 5.2 (with $z = 2$) as a generalized Kauffman bracket. In S^3 the bracket $\langle * \rangle_\tau$ coincides with the usual Kauffman bracket, and thus the invariant $\hat{J}^T_{\boldsymbol{n},q}$ defined in S^3 is exactly the invariant J_L of [22]. The SU(2) toroidal colored Jones polynomial, defined skein-theoretically in S^3, satisfies

$$J_n^T(L;q) = (-A^2 - A^{-2})|_{A^4=q} \cdot J_n(L;q),$$

where J_n is the SU(2) colored Jones polynomial of (3).

6 The volume conjecture for the two-by-two square weave

We now prove the volume conjecture for the *two-by-two square weave* $W \subset T^2 \times I$, the link shown in Figure 1. More generally, a diagram for the $2k$*-by-*$2l$ *square weave* $W_{2k,2l}$, for $k, l \in \mathbb{N}$, is made by tiling W to form a rectangular grid with $2k$ rows and $2l$ columns of crossings. We consider only even dimensions to ensure the diagram is alternating on the torus.

The complement of $W_{2k,2l}$ in $T^2 \times I$ is geometrically simple — Champanerkar, Kofman and Purcell [3] describe a complete hyperbolic structure for $(T^2 \times I) \setminus W_{2k,2l}$ consisting of $4kl$ regular ideal hyperbolic octahedra, one for each crossing. Thus,

$$\mathrm{Vol}(W_{2k,2l}) = 4kl \cdot v_{\mathrm{oct}}. \tag{13}$$

Separately, we have the following proposition.

Proposition 6.1 *Let $q = e^{2\pi i/n}$. If $L \subset T^2 \times I$ is an oriented link which has a diagram $D \subset T^2$ with c crossings,*

$$\lim_{n\to\infty} \frac{2\pi}{n} \log |J_n^T(L;q)| \le c \cdot v_{\mathrm{oct}}.$$

Proof The result follows from work of Garoufalidis and Lê [9, Corollary 8.10 and Theorem 1.13]. Recall the definition of J_n^T as a state sum (Definition 4.4 and equations (6) and (7)),

$$J_n^T(L;q) = (q^{(n^2-1)/4})^{-w(D)} \sum_{\sigma} \prod_{p\in P} \omega_p(\sigma),$$

where P is the set of critical points of D and σ is a state of D. Since q is a root of unity, $|J_n^T| = \left|\sum_\sigma \prod_{p\in P} \omega_p(\sigma)\right|$.

Let m be the number of connected components of $D \setminus P$; then D has n^m states. If σ' is a state with maximum modulus weight, then

$$\begin{aligned} \log |J_n^T| &= \log \left| \sum_\sigma \prod_{p\in P} \omega_p(\sigma) \right| \\ &\le \log \left| n^m \prod_{p\in P} \omega_p(\sigma') \right| = m \log n + \sum_{p\in P} \log |\omega_p(\sigma')|. \end{aligned} \tag{14}$$

Fix the standard basis for $V^n \cong \mathbb{C}^n$. In this basis, if $p \in P$ is an extremum then $|\omega_p(\sigma')| = 1$. If p is a crossing then $\omega_p(\sigma')$ is an element of the R–matrix of $\mathcal{A}_q$ in the given basis, and Garoufalidis and Lê proved

$$\frac{2\pi}{n} \lim_{n\to\infty} \log |\omega_p(\sigma')| \le v_{\mathrm{oct}}. \tag{15}$$

Considering $\lim_{n\to\infty} \frac{2\pi}{n} \log |J_n^T|$, the summands $\log |\omega_p(\sigma')|$ with p a crossing are the only summands of (14) which don't vanish asymptotically. Thus, applying (15),

$$\lim_{n\to\infty} \frac{2\pi}{n} \log |J_n^T| \le c \cdot v_{\mathrm{oct}}. \qquad \square$$

Equation (13) and Proposition 6.1 together give

$$\lim_{n\to\infty} \frac{2\pi}{n} \log |J_n^T(W_{2k,2l}; e^{2\pi i/n})| \le 4kl \cdot v_{\mathrm{oct}} = \mathrm{Vol}((T^2 \times I) \setminus W_{2k,2l}). \tag{16}$$

In other words, for the rectangular weave $W_{2k,2l}$, the asymptotic growth of the toroidal colored Jones polynomial is bounded above by the volume of the complement. This makes the $2k$-by-$2l$ rectangular weave a natural object of study for our volume conjecture, Conjecture 1.2 — to prove the conjecture for a link in this family, we need only show the upper bound in (16) is achieved.

Before proving Theorem 6.2, we give a formula for $J_n^T(W;q) = J_n^T(W_{2,2};q)$. As an isomorphism of $V^n \otimes V^n$ the R–matrix R of $\mathcal{A}_q$ is defined by weights $R_{kl}^{ij} \in \mathbb{C}$,

$$R(e_k \otimes e_l) = \sum_{i,j=0}^{n-1} R_{kl}^{ij} e_i \otimes e_j,$$

where $\{e_0, \dots, e_{n-1}\}$ is a preferred basis of V^n. We have

$$R_{kl}^{ij} = \sum_{m=0}^{\min(n-1-i,j)} \delta_{i+m}^{l} \delta_{j-m}^{k} \cdot q^{\alpha} \cdot \frac{\{l\}!\{n-1-k\}!}{\{i\}!\{m\}!\{n-1-j\}!}, \tag{17}$$

where

$$\alpha = \bigl(i - \tfrac{1}{2}(n-1)\bigr)\bigl(j - \tfrac{1}{2}(n-1)\bigr) - \tfrac{1}{2}m(i-j) - \tfrac{1}{4}m(m+1),$$

δ is the Kronecker delta, and $\{m\}! = \{m\}\{m-1\}\cdots\{2\}\{1\}$. Similarly,

$$R^{-1} \colon V^n \otimes V^n \to V^n \otimes V^n$$

is defined by the scalars

$$(R^{-1})_{kl}^{ij} = \sum_{m=0}^{\min(n-1-i,j)} \delta_{i-m}^{l} \delta_{j+m}^{k} \cdot (-1)^m \cdot q^{\beta} \cdot \frac{\{k\}!\{n-1-l\}!}{\{j\}!\{m\}!\{n-1-i\}!}, \tag{18}$$

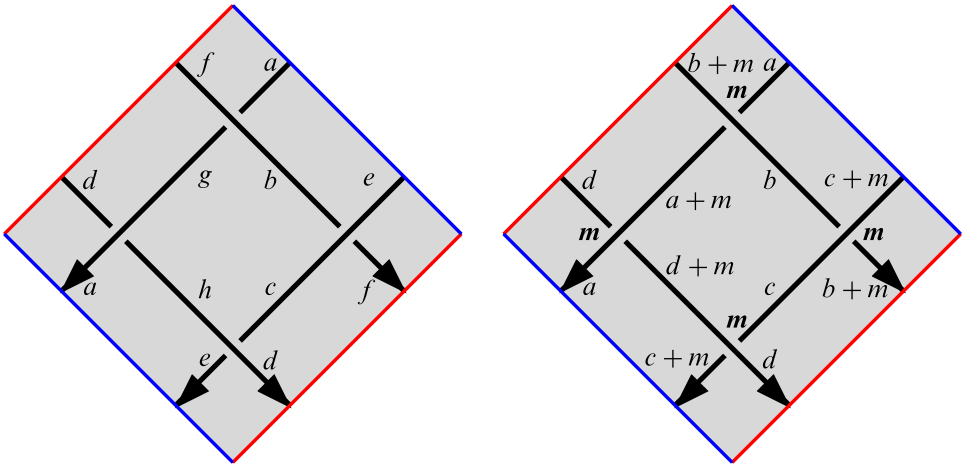
where

$$\beta = -\bigl(i - \tfrac{1}{2}(n-1)\bigr)\bigl(j - \tfrac{1}{2}(n-1)\bigr) - \tfrac{1}{2}m(i-j) + \tfrac{1}{4}m(m+1).$$

Let D be the diagram for W in Figure 10, left, with components labeled by basis elements as shown — implicitly we've chosen a covering map where crossings are oriented downward and there are no maxima or minima. We have

$$J_n^T(W;q) = \sum_{a,b,\dots,g,h=0}^{n-1} (R^{-1})_{gb}^{fa} R_{cf}^{be} R_{ah}^{dg} (R^{-1})_{ed}^{hc}.$$

The index m in (17) and (18) is sometimes thought of as the "label" of the associated crossing [9], with the corresponding summand its "weight". In this way a state of D becomes a labeling of both strands and crossings with integers between 0 and $n-1$,



Each crossing is labeled m.

Figure 10: States of D.

and the Kronecker deltas in (17) and (18) imply we need only consider states whose crossing labels are as shown in Figure 11.

Assigning labels to strands and crossings of D according to these rules gives the diagram in Figure 10, right — the structure of W forces each crossing to have the same label in any nonzero state. We obtain the formula

$$\begin{aligned}
&J_n^T(W;q)\\
&\quad=\sum_{m=0}^{n-1}\sum_{a,b,c,d=0}^{n-m-1}(R^{-1})^{b+m,a}_{a+m,b}R^{b,c+m}_{c,b+m}R^{d,a+m}_{a,d+m}(R^{-1})^{d+m,c}_{c+m,d}\\
&\quad=\sum_{m=0}^{n-1}\sum_{a,b,c,d=0}^{n-m-1}q^{(a-c)(d-b)}\frac{\{a+m\}!\{n-1-a\}!}{\{a\}!\{m\}!\{n-1-a-m\}!}\cdot\frac{\{b+m\}!\{n-1-b\}!}{\{b\}!\{m\}!\{n-1-b-m\}!}\\
&\qquad\qquad\cdot\frac{\{c+m\}!\{n-1-c\}!}{\{c\}!\{m\}!\{n-1-c-m\}!}\cdot\frac{\{d+m\}!\{n-1-d\}!}{\{d\}!\{m\}!\{n-1-d-m\}!}.
\end{aligned}$$

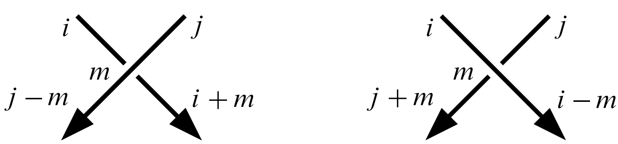


Figure 11: States of crossings with nonzero weights.

For the remainder of the section, fix $q = e^{2\pi i/n}$. This allows us to apply the identity [9, equation 38]

$$\{k\}! = (\sqrt{-1})^{n-1} \frac{n}{\{n-1-k\}!},$$

where $k \in \mathbb{N}$ and $0 \le k \le n-1$, and the formula above becomes

$$J_n^T(W;q) = \sum_{m=0}^{n-1} \sum_{a,b,c,d=0}^{n-m-1} q^{(a-c)(d-b)} \frac{1}{(\{m\}!)^4} \cdot \left(\frac{\{a+m\}!\{b+m\}!\{c+m\}!\{d+m\}!}{\{a\}!\{b\}!\{c\}!\{d\}!} \right)^2. \tag{19}$$

Theorem 6.2 $\displaystyle\lim_{n\to\infty} \frac{2\pi}{n} \log |J_n^T(W; e^{2\pi i/n})| = 4v_{\mathrm{oct}} = \mathrm{Vol}((T^2 \times I) \setminus W).$

Proof By (16), we need only show

$$\lim_{n\to\infty} \frac{2\pi}{n} \log |J_n^T(W; e^{2\pi i/n})| \ge 4v_{\mathrm{oct}} = \mathrm{Vol}((T^2 \times I) \setminus W). \tag{20}$$

Let $\eta(a,b,c,d) = q^{(a-c)(d-b)}$ and

$$\rho(a,b,c,d,m) = \frac{1}{(\{m\}!)^4} \cdot \left(\frac{\{a+m\}!\{b+m\}!\{c+m\}!\{d+m\}!}{\{a\}!\{b\}!\{c\}!\{d\}!} \right)^2.$$

Then (19) becomes

$$J_n^T(W;q) = \sum_{m=0}^{n-1} \sum_{a,b,c,d=0}^{n-m-1} \eta(a,b,c,d) \cdot \rho(a,b,c,d,m). \tag{21}$$

For $k \in \mathbb{Z}$ with $0 \le k \le n-1$, $\{k\} = 2i \sin(k\pi/n) = ix$ where x is a nonnegative real number. Thus $(\{m\}!)^4$ is a nonnegative real number for all values of m. Furthermore, since

$$\begin{aligned} \frac{\{a+m\}!\{b+m\}!\{c+m\}!\{d+m\}!}{\{a\}!\{b\}!\{c\}!\{d\}!} &= \prod_{k=1}^{m} \{a+k\}\{b+k\}\{c+k\}\{d+k\} \\ &= 16 \prod_{k=1}^{m} \sin\frac{\pi(a+k)}{n} \sin\frac{\pi(b+k)}{n} \sin\frac{\pi(c+k)}{n} \sin\frac{\pi(d+k)}{n}, \end{aligned}$$

$\rho(a,b,c,d,m)$ is a nonnegative real number for all relevant values of a, b, c, d, and m.

If $a=c$ then $\eta(a,b,c,d)=1$ and $\eta(a,b,c,d)\cdot\rho(a,b,c,d,m)$ is a real number. If $a\neq c$, since $\rho(a,b,c,d,m)=\rho(a,c,b,d,m)$ and $\eta(a,b,c,d)=\eta(a,c,b,d)^{-1}$,

$$\begin{aligned}\eta(a,b,c,d)\rho(a,b,c,d,m)+\eta(a,c,b,d)\rho(a,c,b,d,m)&\\ &=(\eta(a,b,c,d)+\eta(a,b,c,d)^{-1})\cdot\rho(a,b,c,d,m)\\ &=2\cos\Big(\frac{2\pi(a-c)(d-b)}{n}\Big)\cdot\rho(a,b,c,d,m)\in\mathbb{R}.\end{aligned}$$

Pairing up the summands of (21) this way we see $J_n^T(W;q)$ is a real number; in fact

$$\begin{aligned}J_n^T(W;q)&=\mathrm{Re}(J_n^T(W;q))\\ &=\sum_{k=0}^{n-1}\sum_{a,b,c,d=0}^{n-k-1}\cos\Big(\frac{2\pi(a-c)(d-b)}{n}\Big)\cdot|\rho(a,b,c,d,m)|.\end{aligned}\tag{22}$$

Using the identity $\cos(\alpha-\beta)=\cos(\alpha)\cos(\beta)+\sin(\alpha)\sin(\beta)$, we rewrite (22) as

$$\begin{aligned}J_n^T(W;q)&\\ &=\sum_{m=0}^{n-1}\sum_{a,b,c,d=0}^{n-m-1}\cos\Big(\frac{2\pi(a-c)(d-b)}{n}\Big)\bigg|\frac{\{a+m\}!\{b+m\}!\{c+m\}!\{d+m\}!}{\{a\}!\{b\}!\{c\}!\{d\}!}\bigg|^2\bigg|\frac{1}{\{m\}!}\bigg|^4\\ &=\sum_{m=0}^{n-1}\sum_{b,d=0}^{n-m-1}\bigg|\frac{1}{\{m\}!}\bigg|^4\bigg|\frac{\{b+m\}!\{d+m\}!}{\{b\}!\{d\}!}\bigg|^2\cdot\nu(b,d,m),\end{aligned}\tag{23}$$

where

$$\begin{aligned}\nu(b,d,m)&=\sum_{a,c=0}^{n-m-1}\cos\Big(\frac{2\pi a(d-b)}{n}-\frac{2\pi c(d-b)}{n}\Big)\bigg|\frac{\{a+m\}!\{c+m\}!}{\{a\}!\{m\}!}\bigg|^2\\ &=\sum_{a,c=0}^{n-m-1}\cos\Big(\frac{2\pi a(d-b)}{n}\Big)\cos\Big(\frac{2\pi c(d-b)}{n}\Big)\bigg|\frac{\{a+m\}!\{c+m\}!}{\{a\}!\{m\}!}\bigg|^2\\ &\qquad+\sum_{a,c=0}^{n-m-1}\sin\Big(\frac{2\pi a(d-b)}{n}\Big)\sin\Big(\frac{2\pi c(d-b)}{n}\Big)\bigg|\frac{\{a+m\}!\{c+m\}!}{\{a\}!\{m\}!}\bigg|^2\\ &=\bigg(\sum_{a=0}^{n-m-1}\cos\Big(\frac{2\pi a(d-b)}{n}\Big)\bigg|\frac{\{a+m\}!}{\{a\}!}\bigg|^2\bigg)^2+\bigg(\sum_{a=0}^{n-m-1}\sin\Big(\frac{2\pi a(d-b)}{n}\Big)\bigg|\frac{\{a+m\}!}{\{a\}!}\bigg|^2\bigg)^2.\end{aligned}$$

In particular, $\nu(b,d,m)$ is a nonnegative real number for all values of b,d, and m.

Because each summand of (23) is nonnegative and real, $J_n^T(W;q)$ is bounded below for all n by the summand of (23) with $m=\lfloor n/2\rfloor$ and $b=d=\lfloor n/4\rfloor$. Additionally,

$\nu(\lfloor n/4 \rfloor, \lfloor n/4 \rfloor, \lfloor n/2 \rfloor)$ is bounded below by the summand of the equation above with $a = \lfloor n/4 \rfloor$. We have

$$J_n^T(W;q) \geq \left|\frac{1}{\{\lfloor n/2 \rfloor\}!}\right|^4 \left|\frac{\{\lfloor n/4 \rfloor + \lfloor n/2 \rfloor\}!\{\lfloor n/4 \rfloor + \lfloor n/2 \rfloor\}!}{\{\lfloor n/4 \rfloor\}!\{\lfloor n/4 \rfloor\}!}\right|^2 \nu\left(\left\lfloor \frac{n}{4} \right\rfloor, \left\lfloor \frac{n}{4} \right\rfloor, \left\lfloor \frac{n}{2} \right\rfloor\right)$$
$$\geq \left|\frac{1}{\{\lfloor n/2 \rfloor\}!}\right|^4 \left|\frac{\{\lfloor n/4 \rfloor + \lfloor n/2 \rfloor\}!}{\{\lfloor n/4 \rfloor\}!}\right|^8$$
$$\geq \min\left(\left|\frac{1}{\{\lfloor n/2 \rfloor\}!}\right|^4 \left|\frac{\{\lfloor 3n/4 \rfloor\}!}{\{\lfloor n/4 \rfloor\}!}\right|^8, \left|\frac{1}{\{\lfloor n/2 \rfloor\}!}\right|^4 \left|\frac{\{\lfloor 3n/4 - 1 \rfloor\}!}{\{\lfloor n/4 \rfloor\}!}\right|^8\right). \tag{24}$$

Garoufalidis and Lê [9] proved that, for $\alpha \in (0, n)$,

$$\log |\{\lfloor \alpha \rfloor\}!| = -\frac{n}{\pi} \Lambda\left(\pi \frac{\alpha}{n}\right) + O(\log n).$$

Here Λ is the Lobachevsky function $\Lambda(z) = -\int_0^z \log |2 \sin \zeta| \, d\zeta$ and $O(\log n)$ is an expression bounded by $C \log n$ for a constant C independent of n. Applying this to (24) gives

$$\lim_{n \to \infty} \frac{2\pi}{n} \log |J_n^T(W;q)|$$
$$\geq \lim_{n \to \infty} 8\Lambda\left(\frac{\pi}{2}\right) + 16\Lambda\left(\frac{\pi}{4}\right) - 16 \cdot \min\left(\Lambda\left(\frac{3\pi}{4}\right), \Lambda\left(\frac{(3n-4)\pi}{4n}\right)\right) + \frac{O \log(n)}{n}$$
$$= 4\left(2\Lambda\left(\frac{\pi}{2}\right) + 4\Lambda\left(\frac{\pi}{4}\right) - 4\Lambda\left(\frac{3\pi}{4}\right)\right) = 4v_{\mathrm{oct}},$$

proving the theorem. □

7 Generalizing the volume conjecture for links in $T^2 \times I$

7.1 Simplicial volume

With the original volume conjecture, Conjecture 1.1, in mind, we generalize our Conjecture 1.2 to links which may not be hyperbolic.

Conjecture 7.1 *For any link $L \subset T^2 \times I$ such that $(T^2 \times I) \setminus L$ is irreducible,*

$$\lim_{n \to \infty} \frac{2\pi}{n} \log |J_n^T(L; e^{2\pi i/n})| = \mathrm{Vol}((T^2 \times I) \setminus L), \tag{25}$$

where $n > 0$ runs over all odd integers.

As in Conjecture 1.1, Vol refers to *simplicial volume* — the sum of the volumes of the hyperbolic pieces in the JSJ decomposition of $(T^2 \times I) \setminus L$. By *irreducible*, we mean that every smooth embedded 2–sphere in $(T^2 \times I) \setminus L$ bounds a 3–ball. The irreducibility condition and the restriction to odd n are, in fact, necessary if one wishes to generalize the original volume conjecture, Conjecture 1.1, from knots to links. For a link $L \subset S^3$, $S^3 \setminus L$ being irreducible is equivalent to L not being a split link, a class of links for which the colored Jones polynomial is known to vanish [24]. Separately, van der Veen has constructed a class of nonsplit links called *Whitehead chains* for which the colored Jones polynomial vanishes at even values of n [33]. Before discussing the necessity of these two conditions in Conjecture 7.1, we give some positive results.

Call $L' \subset S^3$ a *VC–verified link* if Conjecture 1.1 is known to hold for L'. That is, L' is VC–verified if

$$\lim_{n\to\infty} \frac{2\pi}{n} \log |J_n(L'; e^{2\pi i/n})| = \mathrm{Vol}(S^3 \setminus L'),$$

where the limit runs over odd $n > 0$. VC–verified links include the figure eight knot, the Borromean rings, and others — see [25, Chapter 3] for a somewhat recent, comprehensive list.

Theorem 7.2 *Let $L' \subset S^3$ be a VC–verified link, and consider an inclusion of L' in an embedded 2–sphere in $T^2 \times I$. Let $K \subset T^2 \times I$ be a knot projecting to an essential, simple closed curve in $T^2 \times \{0\}$, and let L be a connect sum $L = L' \# K$. Then Conjecture 7.1 holds for L.*

See Figure 12 for an example where L' is the figure eight knot and K is a meridian. The main ingredient in the proof of Theorem 7.2 is the following relationship between J_n^T and J_n.

Theorem 7.3 *Let L' be a link in S^3, and consider an inclusion of L' in an embedded 2–sphere in $T^2 \times I$. Let $K \subset T^2 \times I$ be a knot projecting to an essential, simple closed curve in $T^2 \times \{0\}$, and let L be a connect sum $L = L' \# K$. Then*

$$J_n^T(L; q) = n \cdot J_n(L'; q).$$

Proof Using Proposition 4.5, we assume K is a meridian. Then we can choose a diagram $D \subset T^2$ for L and a lift, $\widetilde{D}$, of D to $\mathbb{R}^2$ such that $D' = \widetilde{D} \cap I^2$ is a diagram of L' as a $(1,1)$–tangle. (See Figure 12, where L' is the figure eight knot.) Coloring D' by V^n, Theorem 2.1 associates an $\mathcal{A}$–linear map $\phi\colon V^n \to V^n$ to D'. The irreducibility of V^n implies ϕ is a scalar multiple of the identity, and, after accounting for writhe, this scalar is $J_n(L'; q)$ — see [17, Lemma 3.9] and [9; 25].

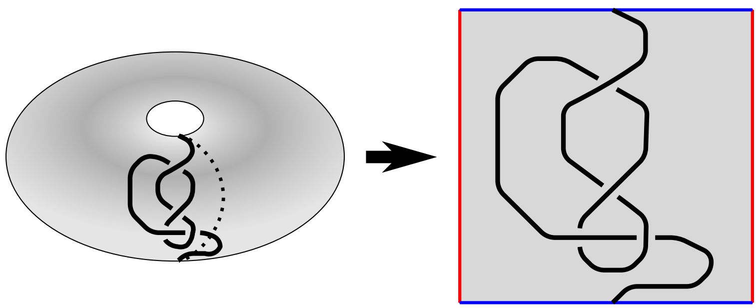


Figure 12: Composing the figure eight knot with a meridian in $T^2 \times I$.

On the other hand, by the proof of Lemma 3.2, $J_n^T(L,q)$ is the trace of ϕ (corrected for writhe). We conclude

$$J_n^T(L;q) = n \cdot J_n(L';q). \qquad \square$$

Theorem 7.2 follows.

Proof of Theorem 7.2 By Theorem 7.3, $J_n^T(L;q) = n \cdot J_n(L';q)$. Therefore

$$\lim_{n\to\infty} \frac{2\pi}{n} \log |J_n^T(L;e^{2\pi i/n})| = \lim_{n\to\infty} \frac{2\pi}{n} \log |J_n(L';e^{2\pi i/n})|. \tag{26}$$

The complement $(T^2 \times I) \setminus K$ is homeomorphic to $S^3 \setminus H'$, where H' is the link shown in Figure 13, left. This implies $(T^2 \times I) \setminus L$ is homeomorphic to $S^3 \setminus (L' \# H')$, where the composition is formed as in Figure 13, right. By [29],

$$\mathrm{Vol}((T^2 \times I) \setminus L) = \mathrm{Vol}(S^3 \setminus (L' \# H')) = \mathrm{Vol}(S^3 \setminus L') + \mathrm{Vol}(S^3 \setminus H') = \mathrm{Vol}(S^3 \setminus L'),$$

and combining this with (26) gives

$$\begin{aligned} \lim_{n\to\infty} \frac{2\pi}{n} \log |J_n^T(L;e^{2\pi i/n})| &= \lim_{n\to\infty} \frac{2\pi}{n} \log |J_n(L';e^{2\pi i/n})| \\ &= \mathrm{Vol}(S^3 \setminus L') = \mathrm{Vol}((T^2 \times I) \setminus L). \qquad \square \end{aligned}$$

Figure 13: The links H' and $L' \# H'$ in S^3, from the proof of Theorem 7.2, with L' the figure eight knot.

We've shown any positive result for the original Conjecture 1.1 gives a positive result for Conjecture 7.1. The proof also shows why restricting to odd n is necessary — if we let $L' \subset S^3$ be a Whitehead chain, as in [33], the link $L' \# K \subset T^2 \times I$ (defined as above) will satisfy Conjecture 7.1 but the toroidal colored Jones polynomial will vanish for even n.

Using the nice behavior of simplicial volume and the toroidal colored Jones polynomial under split unions of links, we can push the result of Theorem 7.2 further. We define a *split union* $L = L_1 \sqcup L_2$ of links $L_1, L_2 \subset T^2 \times I$ to be a union such that L admits a torus diagram which is a disjoint union of diagrams of L_1 and L_2. Additionally, define a *torus link* to be a link in $T^2 \times I$ with a diagram consisting of a set of disjoint, simple closed curves in T^2.

Corollary 7.4 *Let $L'_1, L'_2, \dots, L'_m \subset S^3$ be VC–verified links, and define K as in Theorem 7.2. Let $L_i = L'_i \# K$ for $i = 1, \dots, m$. Then Conjecture 7.1 holds for the split union $L = L_1 \sqcup L_2 \sqcup \dots \sqcup L_m$. In particular, Conjecture 7.1 holds for all torus links with no nullhomotopic components.*

Proof The result follows just as in Theorem 7.2 after checking that

$$J_n^T(L) = J_n^T(L_1) \cdot J_n^T(L_2) \cdots J_n^T(L_m)$$

and

$$\mathrm{Vol}(L) = \mathrm{Vol}(L_1) + \mathrm{Vol}(L_2) + \dots + \mathrm{Vol}(L_m).$$

To prove the second statement, let L'_i be the unknot for all i — every torus link with no nullhomotopic components can be obtained this way. Alternatively, one could use Proposition 3.8 and a direct computation. The result also holds for torus links with nullhomotopic components (see Proposition 7.6 below), but the complement of such a link in $T^2 \times I$ is not irreducible. □

If we view links in the thickened torus as generalizations of $(1,1)$–tangles, as the proof of Theorem 7.3 suggests, and think of the colored Jones polynomial as an invariant of $(1,1)$–tangles, the toroidal colored Jones polynomial becomes a generalization of the colored Jones polynomial rather than a toroidal analogue. This view is supported by Corollary 7.5 below, which shows Conjecture 7.1 implies Conjecture 1.1.

Corollary 7.5 *For knots, Conjecture 7.1 implies Conjecture 1.1.*

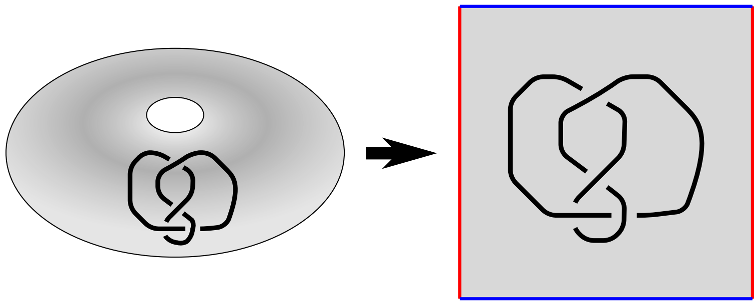


Figure 14: A nullhomotopic inclusion of the figure eight knot in $T^2 \times I$.

Proof Given a knot $K' \subset S^3$, let $L = K' \# K$ with K defined as above. Then, assuming Conjecture 7.1,

$$\lim_{n\to\infty} \frac{2\pi}{n} \log |J_n(K'; e^{2\pi i/n})| = \lim_{n\to\infty} \frac{2\pi}{n} \log |J_n^T(L; e^{2\pi i/n})| = \mathrm{Vol}((T^2 \times I) \setminus L) = \mathrm{Vol}(S^3 \setminus K'). \qquad \square$$

In this sense, Conjecture 7.1 generalizes Conjecture 1.1. It is interesting to note that Conjecture 1.1 does not seem to imply Conjecture 7.1.

As we noted earlier, just as Conjecture 1.1 fails for split links [24], Conjecture 7.1 fails for links in $T^2 \times I$ which have one or more nullhomotopic split components. By a *nullhomotopic split component* of a link $L \subset T^2 \times I$, we mean a sublink $L' \subset L$ such that L' is contained in an embedded 2–sphere in $(T^2 \times I) \setminus L$, and no proper sublink of L' is contained in such a sphere. This is implied by the following:

Proposition 7.6 *If $L \subset T^2 \times I$ is a link with a nullhomotopic split component, $J_n^T(L; e^{2\pi i/n}) = 0$ for all n. In particular, if L is nullhomotopic, $J_n^T(L; e^{2\pi i/n}) = 0$ for all n.*

Proof Let $L \subset T^2 \times I$ be a link with nullhomotopic split component L_1, and let $L_2 = L \setminus L_1$. Then $J_n^T(L; q) = J_n^T(L_1; q) \cdot J_n^T(L_2; q)$, so it suffices to show $J_n^T(L_1, e^{2\pi i/n}) = 0$.

Since L_1 is nullhomotopic, it has a torus diagram D which lifts to a diagram $\widetilde{D} \subset \mathbb{R}^2$ such that $\widetilde{D} \cap I^2$ is a diagram for L_1 as a link in S^3. See Figure 14, where L_1 is the figure eight knot. A direct computation shows $J_n^T(L_1; q) = [n] J_n(L_1; q)$, and $[n] = 0$ when $q = e^{2\pi i/n}$. $\square$

Remark 7.7 In [34], van der Veen noted that Conjecture 1.1 can be changed to account for split links by choosing a different normalization of the colored Jones polynomial. Essentially, each split component adds a factor of $[n]$ to J_n — if a link L has s split components, we can divide by $[n]^s$ to obtain a nonzero value at the root $q = e^{2\pi i/n}$.

Analogously, if $L \subset T^2 \times I$ has s *nullhomotopic* split components, we can ask whether

$$\lim_{n\to\infty} \frac{2\pi}{n} \log \left| \frac{1}{[n]^s} J_n^T(L; e^{2\pi i/n}) \right| = \mathrm{Vol}((T^2 \times I) \setminus L).$$

Replacing (25) in Conjecture 7.1 with the above equation, we can remove the hypothesis that $(T^2 \times I) \setminus L$ be irreducible.

7.2 Higher-genus surfaces

Taking a different direction, one could attempt to generalize Conjecture 1.2 to links in thickened surfaces of genus greater than one. As we noted earlier, while there is no obvious way to define pseudo-operator invariants for links in these surfaces, Corollary 5.6 lets us define the SU(2) toroidal colored Jones polynomial skein-theoretically in any orientable manifold.

As defined, volume conjecture behavior is unlikely to occur in thickened surfaces of genus greater than one. To see why, let $\Sigma \times I$ be such a thickened surface containing a link L. Since $\Sigma \times I$ has boundary components which are not spheres or tori, there is not a unique way to assign a complete hyperbolic structure to the complement of L. One way to resolve this ambiguity, as in [1], is to choose the hyperbolic structure on $(\Sigma \times I) \setminus L$ which has totally geodesic boundary. If such a structure exists, $(\Sigma \times I) \setminus L$ is called *tg–hyperbolic* and it has a finite *tg–hyperbolic volume*.

Proposition 6.1 says that, in the case of a link L in the thickened torus with crossing number c,

$$\lim_{n\to\infty} \frac{2\pi}{n} \log |J_n^T(L; e^{2\pi i/n})| \le c \cdot v_{\mathrm{oct}}. \tag{27}$$

A similar bound exists for links in S^3 — see [9, Theorem 1.13] — and we conjecture that (27) holds for J_n^T for links in any genus thickened surface. In surfaces with genus greater than one, however, there are many links whose tg–hyperbolic volume exceeds this bound. Consider, for example, the virtual link 3.1 of [10] viewed as a link in the thickened orientable surface of genus two: its crossing number is three and its tg–hyperbolic volume is $\approx 18.75 > 3v_{\mathrm{oct}}$ [1]. Thus, volume convergence as defined above is not possible if (27) holds for genus two surfaces and we choose the tg–hyperbolic structure on the complement of L.

This does not mean *no* volume conjecture can exist for links in higher genus surfaces — just that any such conjecture would need to look different from Conjecture 1.1 and Conjecture 1.2. It may be interesting to examine what kind of relationship can exist between the SU(2) toroidal colored Jones polynomial of a link in a higher-genus surface and its tg–hyperbolic volume.

8 The toroidal colored Jones polynomial as an invariant of biperiodic and virtual links

Beyond its volume conjecture behavior, the toroidal colored Jones polynomial may be useful as an invariant of biperiodic and virtual links. A *biperiodic link* is a properly embedded 1–manifold $\tilde{L} \subset \mathbb{R}^2 \times I$, such that $\tilde{L}$ is invariant under translations by a 2–dimensional lattice Λ and $L = \tilde{L}/\Lambda$ is a link in $T^2 \times I$; see [4]. We call Λ *maximal* if it is not properly contained in another invariant lattice for $\tilde{L}$, in which case the resulting link $L \subset T^2$ is a *minimal representative* of $\tilde{L}$. For a given biperiodic link $\tilde{L}$, there are many possible choices of minimal representative. However, if $L_1, L_2 \subset T^2 \times I$ are two minimal representatives of $\tilde{L} \subset \mathbb{R}^2 \times I$ with respective diagrams $D_1, D_2 \subset T^2$, then there exists an orientation-preserving homeomorphism f of T^2 such that $f(D_1) = D_2$; see [11, Proposition 2.1]. Hence, Proposition 4.5 gives the following:

Theorem 8.1 *If $\tilde{L} \subset \mathbb{R}^2 \times I$ is a biperiodic link and $L \subset T^2 \times I$ is a minimal representative of $\tilde{L}$, define $J_n^T(\tilde{L}) = J_n^T(L)$. Then J_n^T is an invariant of biperiodic links in $\mathbb{R}^2 \times I$.*

Another nonclassical type of link, *virtual links*, are an area of extensive study — see [15] for an introduction. By [2; 20], any virtual link L' is represented uniquely by a link L in a minimal-genus thickened surface, up to an orientation-preserving homeomorphism of the surface. The $U_q(\mathrm{sl}(2,\mathbb{C}))$ toroidal colored Jones polynomial is defined only for links in $T^2 \times I$, but the SU(2) toroidal colored Jones polynomial can be defined skein-theoretically for links in any thickened surface. Similar to above, we have:

Theorem 8.2 *If L' is a virtual link and $L \subset \Sigma \times I$ is a minimal representative of L', define $J_n^T(L') = J_n^T(L)$. Then J_n^T is an invariant of virtual links.*

Here Σ is a closed, orientable surface and J_n^T is the SU(2) toroidal colored Jones polynomial, defined skein-theoretically as in Corollary 5.6. To prove Theorem 8.2, we need only show the skein-theoretic J_n^T is preserved by orientation-preserving homeomorphisms of surfaces. This is done in the lemma below.

Lemma 8.3 *Let L and L' be links in $\Sigma \times I$ with respective diagrams $D, D' \subset \Sigma$, Σ a closed, orientable surface. If f is an orientation-preserving homeomorphism of Σ satisfying $f(D) = D'$, then $J_n^T(L;q) = J_n^T(L';q)$ for all $n \in \mathbb{N}$. Here J_n^T is the* SU(2) *toroidal Jones polynomial, defined skein-theoretically.*

Proof Because f preserves orientation, D and D' have the same writhe. Thus it suffices to prove the result for $\hat{J}^T_{\boldsymbol{n},q}$, which follows from the case of $\hat{J}^T_{\boldsymbol{2},q}$. Equivalently, we show the bracket $\langle * \rangle_\tau$ defined in Section 5 is invariant under orientation-preserving homeomorphisms of Σ.

The claim follows by induction on crossing number, noting f induces a bijection on the crossings of D and D'. If D has no crossings, $\langle D \rangle_\tau$ is determined by whether or not D is contractible, which is preserved by f. For an arbitrary diagram D, we can "resolve" a crossing using the relation (c) of Definition 5.2. Since f commutes with both types of crossing resolution in relation (c), the claim follows inductively. $\square$

As Remark 3.5 discusses, J_n^T is distinct from existing quantum invariants of virtual links. To our knowledge, it is the first invariant of virtual links to exhibit volume conjecture behavior for genus one virtual links, ie links in the thickened torus. Continuing our discussion from Section 7.2, it is interesting to ask what kind of volume conjecture behavior emerges in higher-genus virtual links.

Appendix The toroidal colored Jones polynomial and rotation number

The following generalization of property (b) of Lemma 5.1 is not hard to prove, using Proposition 3.8 and a direct computation.

Proposition A.1 *Let $K \subset T^2 \times I$ be a knot projecting to a simple closed curve in T^2.*

(a) *If K is nullhomotopic, the* SU(2) *toroidal colored Jones polynomial J_n^T satisfies*

$$J_n^T(K;q) = -[n],$$

and the $U_q(\mathrm{sl}(2,\mathbb{C}))$ toroidal colored Jones polynomial J_n^T satisfies

$$J_n^T(K;q) = [n].$$

(b) *If K is not nullhomotopic, the* $\mathrm{SU}(2)$ *and* $U_q(\mathrm{sl}(2,\mathbb{C}))$ *toroidal colored Jones polynomials both satisfy*

$$J_n^T(K;q) = n.$$

Proposition A.1 says, in a sense, that contractible, simple closed curves in T^2 are "quantized" by the toroidal colored Jones polynomial while essential, simple closed curves are not. We would like to motivate geometrically why this striking phenomenon occurs.

To accomplish this, we recall Lin and Wang's definition of the Jones polynomial [21], adapted from work in [31]. As we will see, their construction extends in a natural way to define J_2^T and, by cabling, J_n^T for all $n > 2$. Its use of rotation number provides insight into Proposition A.1, at least for $n = 2$.

We briefly recall Lin and Wang's definition. First, fix the preferred basis $\{e_0, e_1\}$ of V^2 we used in Section 7. In this basis the R–matrix coefficients are

$$\begin{aligned} R^{0,0}_{0,0} &= R^{1,1}_{1,1} = q^{1/4}, & R^{1,0}_{0,1} &= R^{0,1}_{1,0} = q^{-1/4}, \\ R^{0,1}_{0,1} &= q^{1/4} - q^{-3/4}, & (R^{-1})^{0,0}_{0,0} &= (R^{-1})^{1,1}_{1,1} = q^{-1/4}, \\ (R^{-1})^{1,0}_{0,1} &= (R^{-1})^{0,1}_{1,0} = q^{1/4}, & (R^{-1})^{1,0}_{1,0} &= q^{-1/4} - q^{3/4}, \end{aligned}$$

and all other entries of R and R^{-1} are zero.

Given a diagram D of an oriented link $L \subset S^3$, let P_c be the set a crossing points of D. In this context, a state s of D is an assignment of 0 or 1 to each component of $D \setminus P_c$. (States are defined differently here than in Section 3 — we ignore local extrema and do not make use of $(V^2)^*$.) If a state s labels a neighborhood of a positive crossing p with $i, j, k, l \in \{0, 1\}$ as in Figure 5, the weight of the crossing is $\omega_p(s) = R^{ij}_{kl}$. If s labels a neighborhood of a negative crossing the same way, the weight of p is $\omega_p(s) = (R^{-1})^{ij}_{kl}$.

Similar to (6), we define the total weight of a state s to be

$$\omega^c(s) = \prod_{p \in P_c} \omega_p(s).$$

A state s is called *admissible* if $\omega^c(s) \neq 0$. Examining the coefficients of R and R^{-1}, we see s is admissible if and only if each crossing of D has one of the patterns of

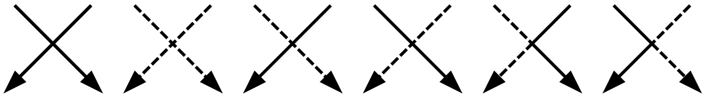


Figure 15: Admissible states near crossings.

labels shown in Figure 15, where dashed and solid lines indicate 0– and 1–labels, respectively. If either of the two rightmost cases in Figure 15 occurs in D, we resolve the given crossing into two vertical lines. This decomposes D into a set of closed curves, each labeled entirely by 0 or entirely by 1 in s. Define $\mathrm{rot}_i(D,s)$ to be the sum of the rotation numbers (the degree of the Gauss map) of all i–labeled curves of D after these resolutions take place. Then:

Proposition A.2 [21] *We have*

$$J_2(L;q) = \frac{1}{[n]}(q^{3/4})^{-w(D)} \sum_{s\in\mathrm{Adm}_c(D)} q^{(\mathrm{rot}_1(D,s)-\mathrm{rot}_0(D,s))/2}\cdot\omega^c(s),$$

where $\mathrm{Adm}_c(D)$ *is the set of admissible states of* D.

Removing the factor of $1/[n]$, this definition extends to a torus with no trouble. It agrees with our definition of J_2^T.

Theorem A.3 *Let* $L\subset T^2\times I$ *be an oriented link with diagram* $D\subset T^2$*. Then*

$$J_2^T(L;q) = (q^{3/4})^{-w(D)} \sum_{s\in\mathrm{Adm}_c(D)} q^{(\mathrm{rot}_1(D,s)-\mathrm{rot}_0(D,s))/2}\cdot\omega^c(s), \tag{28}$$

where all terms are defined as in Proposition A.2.

Proof We sketch the proof. Let P denote the set of crossing points and local extrema of D, as in Section 3, and use σ to denote a state of D in the pseudo-operator invariant context (see (6) and the preceding discussion). Call a state σ *admissible* if $\omega(\sigma)\neq 0$, and let $\mathrm{Adm}(D)$ be the set of admissible states of D in this context.

In the given basis for V^2, the operator $\mu\colon V^2\to V^2$ is defined by

$$\mu_0^0 = q^{-1/2},\quad \mu_1^1 = q^{1/2}, \tag{29}$$

and all other coefficients are zero [17]. Thus, a state σ is admissible only if both sides of every extreme point of D are assigned the same number, either 0 or 1. (Here $i\in\{0,1\}$ might refer to the basis element e_i or the dual element e^i.) It follows that $\mathrm{Adm}(D)$ is in bijection with $\mathrm{Adm}_c(D)$. Furthermore, if $\sigma\in\mathrm{Adm}(D)$, we can perform crossing

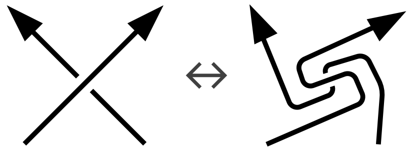


Figure 16: An isotopy which reorients a crossing downward.

resolutions like those preceding Proposition A.2 to decompose D into a set of closed curves, each of which is labeled entirely by 0 or entirely by 1. Therefore it makes sense to write $\mathrm{rot}_i(D, \sigma)$ for an admissible state σ.

Finally, we may assume all crossings of D have both strands oriented downward — otherwise, we can apply an isotopy as in Figure 16. This isotopy does not change the value of (28), since it does not change the diagram or any rotation numbers. With this assumption, if $p \in D$ is a crossing point, $\omega_p(\sigma) = \omega_p(s)$ for any state $\sigma \in \mathrm{Adm}(D)$ with corresponding state $s \in \mathrm{Adm}_c(D)$.

We now compute

$$\begin{aligned} J_2^T(L;q) &= (q^{3/4})^{-w(D)} \sum_{\sigma \in \mathrm{Adm}(D)} \prod_{p \in P} \omega_p(\sigma) \\ &= (q^{3/4})^{-w(D)} \sum_{\sigma \in \mathrm{Adm}(D)} \prod_{p \in (P \setminus P^c)} \omega_p(\sigma) \prod_{p \in P^c} \omega_p(\sigma) \\ &= (q^{3/4})^{-w(D)} \sum_{\sigma \in \mathrm{Adm}(D)} (\mu_0^0)^{\mathrm{rot}_0(D,\sigma)} (\mu_1^1)^{\mathrm{rot}_1(D,\sigma)} \omega^c(\sigma) \\ &= (q^{3/4})^{-w(D)} \sum_{s \in \mathrm{Adm}_c(D)} q^{(\mathrm{rot}_1(D,s) - \mathrm{rot}_0(D,s))/2} \cdot \omega^c(s). \end{aligned}$$

The key observation of the third equality is that μ counts rotation numbers. Examining Theorem 2.1, we see that a weight of μ_i^i is assigned to each left-oriented, i–colored cap and a weight of $(\mu^{-1})_i^i = (\mu_i^i)^{-1}$ is assigned to each left-oriented, i–colored cup. Thus, if C is a curve of D (after crossing resolution) labeled entirely by i, the exponent of the product of the μ_i^i's gives the rotation number of C. (See Figure 17.) □

Having defined J_2^T as in Theorem A.3, the higher invariants J_n^T, for $n > 2$, can be recovered using the cabling formula, Theorem 4.6.

As promised, we only needed to normalize the formula in Proposition A.2 to define J_2^T as in Theorem A.3. From this perspective, J_n and J_n^T become two instances of the same formula, and the definition of the latter is forced by the definition of the former. In other words, from this point of view, there is no other way we could have defined J_n^T.

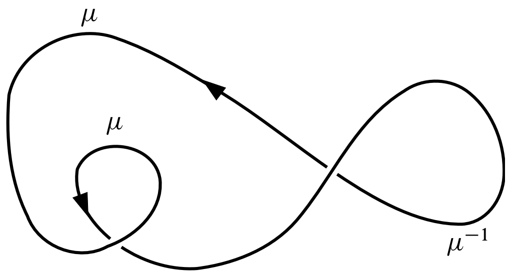


Figure 17: The exponent of the product of the μ's is the rotation number of the curve (in this case 1).

Additionally, (28) provides insight into Proposition A.1. Let $K \subset T^2 \times I$ be a knot which projects to a simple, closed curve $C \subset T^2$. Then C has no crossings, and only two state assignments as defined in (28). If C is contractible, it has rotation number ± 1 and

$$J_2^T(K;q) = q^{1/2} + q^{-1/2} = [2].$$

If C is not contractible, it has rotation number 0 and

$$J_2^T(K;q) = q^0 + q^0 = 2.$$

While we cannot fully explain why the toroidal colored Jones polynomial "quantizes" contractible curves and not essential ones, this discussion suggests a relationship with the curvature of a link.

Remark A.4 The exact R–matrix used here is slightly different than the one used in [21, Section 2.3]. To recover that matrix from ours, first multiply R by $q^{1/4}$ (and multiply R^{-1} by $q^{-1/4}$), then make the variable substitution $q' = -q^{1/2}$. We also use downward-oriented crossings rather than upward-oriented ones — these two convention changes result in a slightly different formula for J_2.

References

[1] **C Adams**, **O Eisenberg**, **J Greenberg**, **K Kapoor**, **Z Liang**, **K O'Connor**, **N Pacheco-Tallaj**, **Y Wang**, *tg–hyperbolicity of virtual links*, J. Knot Theory Ramifications 28 (2019) art. id. 1950080 MR Zbl

[2] **J S Carter**, **S Kamada**, **M Saito**, *Stable equivalence of knots on surfaces and virtual knot cobordisms*, J. Knot Theory Ramifications 11 (2002) 311–322 MR Zbl

[3] **A Champanerkar**, **I Kofman**, **J S Purcell**, *Geometrically and diagrammatically maximal knots*, J. Lond. Math. Soc. 94 (2016) 883–908 MR Zbl

[4] **A Champanerkar**, **I Kofman**, **J S Purcell**, *Geometry of biperiodic alternating links*, J. Lond. Math. Soc. 99 (2019) 807–830 MR Zbl

[5] **Q Chen**, **T Yang**, *Volume conjectures for the Reshetikhin–Turaev and the Turaev–Viro invariants*, Quantum Topol. 9 (2018) 419–460 MR Zbl

[6] **F Costantino**, *Coloured Jones invariants of links and the volume conjecture*, J. Lond. Math. Soc. 76 (2007) 1–15 MR Zbl

[7] **R Detcherry**, **E Kalfagianni**, **T Yang**, *Turaev–Viro invariants, colored Jones polynomials, and volume*, Quantum Topol. 9 (2018) 775–813 MR Zbl

[8] **P J Freyd**, **D N Yetter**, *Braided compact closed categories with applications to low-dimensional topology*, Adv. Math. 77 (1989) 156–182 MR Zbl

[9] **S Garoufalidis**, **T T Q Lê**, *Asymptotics of the colored Jones function of a knot*, Geom. Topol. 15 (2011) 2135–2180 MR Zbl

[10] **J Green**, *A table of virtual knots*, electronic resource (2004) Available at `http://www.math.toronto.edu/drorbn/Students/GreenJ`

[11] **S A Grishanov**, **V R Meshkov**, **A V Omelchenko**, *Kauffman-type polynomial invariants for doubly periodic structures*, J. Knot Theory Ramifications 16 (2007) 779–788 MR Zbl

[12] **R M Kashaev**, *The hyperbolic volume of knots from the quantum dilogarithm*, Lett. Math. Phys. 39 (1997) 269–275 MR Zbl

[13] **C Kassel**, *Quantum groups*, Graduate Texts in Math. 155, Springer (1995) MR Zbl

[14] **L H Kauffman**, *State models and the Jones polynomial*, Topology 26 (1987) 395–407 MR Zbl

[15] **L H Kauffman**, *Virtual knot theory*, European J. Combin. 20 (1999) 663–690 MR Zbl

[16] **L H Kauffman**, *Rotational virtual knots and quantum link invariants*, J. Knot Theory Ramifications 24 (2015) art. id. 1541008 MR Zbl

[17] **R Kirby**, **P Melvin**, *The 3–manifold invariants of Witten and Reshetikhin–Turaev for* $\mathrm{sl}(2,\mathbb{C})$, Invent. Math. 105 (1991) 473–545 MR Zbl

[18] **A N Kirillov**, **N Y Reshetikhin**, *Representations of the algebra* $U_q(\mathrm{sl}(2))$, *q–orthogonal polynomials and invariants of links*, from "Infinite-dimensional Lie algebras and groups" (V G Kac, editor), Adv. Ser. Math. Phys. 7, World Sci., Teaneck, NJ (1989) 285–339 MR Zbl

[19] **V Krushkal**, *Graphs, links, and duality on surfaces*, Combin. Probab. Comput. 20 (2011) 267–287 MR Zbl

[20] **G Kuperberg**, *What is a virtual link?*, Algebr. Geom. Topol. 3 (2003) 587–591 MR Zbl

[21] **X-S Lin**, **Z Wang**, *Random walk on knot diagrams, colored Jones polynomial and Ihara–Selberg zeta function*, from "Knots, braids, and mapping class groups — papers dedicated to Joan S. Birman" (J Gilman, W W Menasco, X-S Lin, editors), AMS/IP Stud. Adv. Math. 24, Amer. Math. Soc., Providence, RI (2001) 107–121 MR Zbl

[22] **H R Morton**, *Invariants of links and 3–manifolds from skein theory and from quantum groups*, from "Topics in knot theory" (M E Bozhüyük, editor), NATO Adv. Sci. Inst. Ser. C: Math. Phys. Sci. 399, Kluwer Acad., Dordrecht (1993) 107–155 MR Zbl

[23] **H R Morton**, **P Strickland**, *Jones polynomial invariants for knots and satellites*, Math. Proc. Cambridge Philos. Soc. 109 (1991) 83–103 MR Zbl

[24] **H Murakami**, **J Murakami**, *The colored Jones polynomials and the simplicial volume of a knot*, Acta Math. 186 (2001) 85–104 MR Zbl

[25] **H Murakami**, **Y Yokota**, *Volume conjecture for knots*, SpringerBriefs in Mathematical Physics 30, Springer (2018) MR Zbl

[26] **T Ohtsuki**, *On the asymptotic expansion of the quantum* $\mathrm{SU}(2)$ *invariant at* $q = \exp(4\pi\sqrt{-1}/N)$ *for closed hyperbolic 3–manifolds obtained by integral surgery along the figure-eight knot*, Algebr. Geom. Topol. 18 (2018) 4187–4274 MR Zbl

[27] **J H Przytycki**, *Skein modules of 3–manifolds*, Bull. Polish Acad. Sci. Math. 39 (1991) 91–100 MR Zbl

[28] **N Y Reshetikhin**, **V G Turaev**, *Ribbon graphs and their invariants derived from quantum groups*, Comm. Math. Phys. 127 (1990) 1–26 MR Zbl

[29] **T Soma**, *The Gromov invariant of links*, Invent. Math. 64 (1981) 445–454 MR Zbl

[30] **V G Turaev**, *The Conway and Kauffman modules of a solid torus*, Zap. Nauchn. Sem. Leningrad. Otdel. Mat. Inst. Steklov. 167 (1988) 79–89, 190 MR Zbl

[31] **V G Turaev**, *The Yang–Baxter equation and invariants of links*, Invent. Math. 92 (1988) 527–553 MR Zbl

[32] **V G Turaev**, *Quantum invariants of knots and 3–manifolds*, De Gruyter Studies in Mathematics 18, de Gruyter, Berlin (1994) MR Zbl

[33] **R van der Veen**, *Proof of the volume conjecture for Whitehead chains*, Acta Math. Vietnam. 33 (2008) 421–431 MR Zbl

[34] **R van der Veen**, *The volume conjecture for augmented knotted trivalent graphs*, Algebr. Geom. Topol. 9 (2009) 691–722 MR Zbl

Department of Mathematics, The Graduate Center, CUNY
New York, NY, United States

jboninger@gradcenter.cuny.edu

Received: 12 April 2021 Revised: 1 October 2021

Guidelines for Authors

Submitting a paper to Algebraic & Geometric Topology

Papers must be submitted using the upload page at the AGT website. You will need to choose a suitable editor from the list of editors' interests and to supply MSC codes.

The normal language used by the journal is English. Articles written in other languages are acceptable, provided your chosen editor is comfortable with the language and you supply an additional English version of the abstract.

Preparing your article for Algebraic & Geometric Topology

At the time of submission you need only supply a PDF file. Once accepted for publication, the paper must be supplied in LaTeX, preferably using the journal's class file. More information on preparing articles in LaTeX for publication in AGT is available on the AGT website.

`arXiv` papers

If your paper has previously been deposited on the `arXiv`, we will need its `arXiv` number at acceptance time. This allows us to deposit the DOI of the published version on the paper's `arXiv` page.

References

Bibliographical references should be listed alphabetically at the end of the paper. All references in the bibliography should be cited at least once in the text. Use of BibTeX is preferred but not required. Any bibliographical citation style may be used, but will be converted to the house style (see a current issue for examples).

Figures

Figures, whether prepared electronically or hand-drawn, must be of publication quality. Fuzzy or sloppily drawn figures will not be accepted. For labeling figure elements consider the pinlabel LaTeX package, but other methods are fine if the result is editable. If you're not sure whether your figures are acceptable, check with production by sending an email to graphics@msp.org.

Proofs

Page proofs will be made available to authors (or to the designated corresponding author) in PDF format. Failure to acknowledge the receipt of proofs or to return corrections within the requested deadline may cause publication to be postponed.

ALGEBRAIC & GEOMETRIC TOPOLOGY

Volume 23 Issue 4 (pages 1463–1934) 2023
