Geometry & Topology

Volume 21 (2017)

Issue 4 (pages 1931–2555)

msp

GEOMETRY & TOPOLOGY

msp.org/gt

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

The subscription price for 2017 is US $495/year for the electronic version, and $730/year (+$60, if shipping outside the US) for print and electronic. Subscriptions, requests for back issues and changes of subscriber address should be sent to MSP. Geometry & Topology is indexed by Mathematical Reviews, Zentralblatt MATH, Current Mathematical Publications and the Science Citation Index.

Geometry & Topology (ISSN 1465-3060 printed, 1364-0380 electronic) at Mathematical Sciences Publishers, c/o Department of Mathematics, University of California, 798 Evans Hall #3840, Berkeley, CA 94720-3840, is published 6 times per year and continuously online. Periodical rate postage paid at Berkeley, CA 94704, and additional mailing offices. POSTMASTER: send address changes to Mathematical Sciences Publishers, c/o Department of Mathematics, University of California, 798 Evans Hall #3840, Berkeley, CA 94720-3840.

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

PUBLISHED BY

mathematical sciences publishers

nonprofit scientific publishing

http://msp.org/

Geometry & Topology 21 (2017) 1931–1968

On representation varieties of 3–manifold groups

MICHAEL KAPOVICH
JOHN J MILLSON

We prove universality theorems ("Murphy's laws") for representation varieties of fundamental groups of closed 3–dimensional manifolds. We show that germs of SL(2)–representation schemes of such groups are essentially the same as germs of schemes over $\mathbb{Q}$ of finite type.

14B12, 20F29, 57M05

1 Introduction

In this paper we will prove that there are no restrictions on local geometry of representation schemes of 3–manifold groups to PO(3) and SL(2). Note that both groups $H = \mathrm{PO}(3)$ and $H = \mathrm{SL}(2)$ are affine algebraic group schemes defined over $\mathbb{Q}$; thus, for every finitely generated group Γ, the representation schemes

$$\mathrm{Hom}(\Gamma, H)$$

and character schemes

$$X(\Gamma, H) = \mathrm{Hom}(\Gamma, H) /\!\!/ H$$

are affine algebraic schemes over $\mathbb{Q}$. Our goal is to show that, *to some extent*, these are the only restrictions on local geometry of the representation and character schemes of fundamental groups of closed 3–manifolds. The universality theorem we thus obtain is one of many universality theorems about moduli spaces of geometric objects; see Mnëv [11], Richter-Gebert [15], Kapovich and Millson [6; 7; 8], Vakil [18], Payne [13], Rapinchuk [14].

Below is the precise formulation of our universality theorem. In what follows we use the notation $G = \mathrm{PO}(3)$ and $\widetilde{G} = \mathrm{Spin}(3)$.

Theorem 1.1 *Let $X \subset \mathbb{C}^N$ be an affine algebraic scheme over $\mathbb{Q}$ and let $x \in X$ be a rational point. Then there exist*

1. *an open subscheme $X' \subset X$ containing x,*
2. *a closed 3–dimensional manifold M with fundamental group π,*

Published: 19 May 2017 DOI: 10.2140/gt.2017.21.1931

3. *a representation* $\rho_0 \colon \pi \to \mathrm{PO}(3,\mathbb{R})$*, such that the image of* ρ_0 *is dense in* $\mathrm{PO}(3,\mathbb{R})$*,*
4. *an open* G*–invariant subscheme* $R' \subset \mathrm{Hom}(\pi, G)$*, whose set of real points contains* ρ_0 *and a closed subscheme (over* $\mathbb{Q}$*)* $R'_c \subset R'$ *that is a cross-section for the action*
$$G \times R' \to R',$$
5. *an isomorphism of schemes over* $\mathbb{Q}$
$$f \colon R'_c \to X' \times G^k \quad \text{with } f(\rho_0) = (x, 1),$$
for some k*, and*
6. *an isomorphism* $F \colon R' \to X' \times G^{k+1}$ *and the composition of isomorphisms*
$$R' \cong R'_c \times G \cong X' \times G^{k+1}$$
(since R'_c *is a cross-section).*

Remark 1.2 One can show that the same theorem holds for a homomorphism ρ_1 whose image is a finite group with trivial centralizer in $\mathrm{PO}(3,\mathbb{R})$.

Theorem 1.1 is proven in Section 6. In Section 7 we prove various corollaries of our main theorem.

Corollary 1.3 *With the notation of Theorem 1.1, there exists an open embedding of schemes*
$$X' \times G^k \hookrightarrow X(\pi, G) = \mathrm{Hom}(\pi, G) /\!\!/ G$$
which sends $(x,1)$ *to* $[\rho_0]$*. In particular, the analytic germ* $(X \times \mathbb{C}^{3k}, x \times 0)$ *is isomorphic to the analytic germ* $(X(\pi, G), [\rho_0])$*.*

Since the groups $\mathrm{PSL}(2,\mathbb{C})$ and $\mathrm{PO}(3,\mathbb{C}) = G(\mathbb{C})$ are isomorphic, and
$$\widetilde{G}(\mathbb{C}) = \mathrm{Spin}(3,\mathbb{C}) \cong \mathrm{SL}(2,\mathbb{C}),$$
the "universality theorem" for $\mathrm{PO}(3)$–representations leads to the one for $\mathrm{SL}(2)$–representations. In the next corollary, π and ρ_0 are the 3–manifold group and its representation to $\mathrm{PO}(3,\mathbb{R})$ constructed in Theorem 1.1 given X and x. We note that the action of $\mathrm{SL}(2)$ on $\mathrm{Hom}(\widetilde{\pi}, \mathrm{SL}(2))$ factors through an action of $\mathrm{PSL}(2)$.

Corollary 1.4 *Let* $X \subset \mathbb{C}^N$ *be an affine algebraic scheme over* $\mathbb{Q}$ *and* $x \in X$ *be a rational point. Then there exist*

1. *an open subscheme $X' \subset X$ containing x,*
2. *a closed 3–dimensional manifold $\widetilde{M}$ with fundamental group $\widetilde{\pi}$,*
3. *a representation $\widetilde{\rho}_0\colon \widetilde{\pi} \to \mathrm{SU}(2) < \mathrm{SL}(2,\mathbb{C})$, such that the image of $\widetilde{\rho}_0$ is dense in $\mathrm{SU}(2)$,*
4. *an open $\mathrm{SL}(2)$–invariant subscheme $\widetilde{R}' \subset \mathrm{Hom}(\widetilde{\pi}, \mathrm{SL}(2))$, such that $\widetilde{\rho}_0 \in \widetilde{R}'(\mathbb{C})$,*
5. *a cross-section $\widetilde{R}'_c$ for the action of $\mathrm{PSL}(2)$ on $\widetilde{R}'$,*
6. *a regular étale covering of schemes over $\mathbb{C}$, equivariant with respect to the action of $\mathrm{SL}(2)$,*
$$\widetilde{q}\colon \widetilde{R}' \to R' \quad \text{with } \widetilde{q}(\widetilde{\rho}_0) = \rho_0,$$
such that the deck-transformation group of this cover is isomorphic to $\mathbb{Z}_2^{k+r}$, for some k and r.

By combining part 6 with Theorem 1.1, we also obtain an $\mathrm{SL}(2)$–equivariant regular étale covering
$$\widetilde{R}' \to X' \times \mathrm{PSL}(2)^{k+1},$$
and the étale covering
$$\widetilde{R}'_c \to X' \times \mathrm{PSL}(2)^{k}$$
sending $\widetilde{\rho}_0$ to $(x,1)$. The latter yields the regular étale covering
$$h\colon \widetilde{R}' /\!\!/ \mathrm{SL}(2) \to X' \times \mathrm{PSL}(2)^{k}.$$
In particular, the morphisms $\widetilde{q}$ and F (from Theorem 1.1) and h (as above) yield isomorphisms of analytic germs
$$\big(\mathrm{Hom}(\widetilde{\pi}, \mathrm{SL}(2)), \widetilde{\rho}_0\big) \to (X' \times \mathbb{C}^{3k+3}, x \times 0),$$
$$\big(X(\widetilde{\pi}, \mathrm{SL}(2)), [\widetilde{\rho}_0]\big) \to (X' \times \mathbb{C}^{3k}, x \times 0),$$
for some $k \geq 0$. Thus, if the scheme X' is nonreduced at x, so are $\mathrm{Hom}(\widetilde{\pi}, \mathrm{SL}(2))$ and $X(\widetilde{\pi}, \mathrm{SL}(2))$.

Remark 1.5 Despite our efforts, we were unable to replace an étale covering with an isomorphism in Corollary 1.4. This is strangely reminiscent of the finite abelian coverings appearing in our universality theorem for planar linkages; see Kapovich and Millson [8]. Note that a relation between universality theorems for projective arrangements and spherical linkages was established in Kapovich and Millson [7], where a finite abelian covering appeared for essentially the same reason as in the present paper.

Example 1.6 Pick a natural number ℓ. Then there exists a closed 3–dimensional manifold $\widetilde{M}$, an integer n and a representation $\rho\colon \pi_1(\widetilde{M}) \to \mathrm{SU}(2)$ with dense image, such that the completed local ring of the germ

$$X(\pi_1(\widetilde{M}), \mathrm{SL}(2)), [\rho])$$

is isomorphic to the completion of the ring

$$\mathbb{C}[t, t_1, \dots, t_{3k}]/(t^\ell).$$

This shows that the representation and character schemes of 3–manifold groups can be nonreduced (at points of Zariski density), which is why we refrain from referring to these schemes as "varieties", as is commonly done in the literature.

Remark 1.7 Recently Igor Rapinchuk [14, Theorem 3] proved a universality theorem for character schemes of groups Γ satisfying Kazhdan's property (T): it involves representations of such groups Γ into $\mathrm{SL}(n,\mathbb{C})$. Unlike the results in Kapovich and Millson [6] and this paper, Rapinchuk's theorem applies to the entire character variety $X^{\mathrm{red}}(\Gamma, \mathrm{SL}(n,\mathbb{C}))$ minus the trivial representation (which is an isolated point). In Rapinchuk's theorem, the number n (and the group Γ) depend on the given affine variety X over $\mathbb{Q}$.

Acknowledgements Partial financial support to Kapovich was provided by the NSF grant DMS-12-05312 and to Millson by the NSF grant DMS-15-18657. Kapovich is also grateful to the Korea Institute for Advanced Study for its hospitality and excellent working conditions. We are grateful to the referee and Michael Heusener for useful remarks and corrections.

2 Preliminaries

2.1 Representation and character schemes

We will say that a subscheme $Y \subset X$ is *clopen* if it is both closed and open. We will use the topologist's notation

$$\mathbb{Z}_m := \mathbb{Z}/m\mathbb{Z}$$

for the cyclic group of order m. Given a subset S of a group G we will use the notation $\langle\langle S\rangle\rangle$ for the normal closure of S in G.

Let G be an algebraic group scheme over a field $\boldsymbol{k}$ of characteristic zero (this will be the default assumption through the rest of the paper) with Lie algebra $\mathfrak{g}$. Let Γ be a finitely presented group with presentation

$$\langle s_1, \dots, s_p | r_1 = 1, \dots, r_q = 1\rangle.$$

(In fact, one needs Γ only to be finitely generated, but all finitely generated groups in this paper will be also finitely presented.) Every word w in the generators s_i and their inverses s_i^{-1}, for $i = 1, \dots, p$, defines a morphism

$$w\colon G^p \to G,$$

obtained by substituting elements $g_1^{\pm 1}, \dots, g_p^{\pm 1} \in G$ in the word w for the letters $s_1^{\pm 1}, \dots, s_p^{\pm 1}$. We then obtain the *representation scheme*

$$\mathrm{Hom}(\Gamma, G) = \{(g_1, \dots, g_p) \in G^p \mid r_j(g_1, \dots, g_p) = 1,\ j = 1, \dots, q\},$$

as every homomorphism $\Gamma \to G$ is determined by its values on the generators of Γ. We will thus think of points of this scheme as homomorphisms $\rho\colon \Gamma \to G$. The representation scheme is known to be independent of the presentation of the group Γ. We refer the reader to [10; 17] for more details. We also refer the reader to [16; 17] for detailed discussion of character varieties/schemes and a survey of their applications to 3–dimensional topology.

We will frequently use the following two facts about representation schemes; see eg [17]:

1. $\mathrm{Hom}(\Gamma_1 \star \cdots \star \Gamma_k, G) \cong \prod_{i=1}^k \mathrm{Hom}(\Gamma_i, G)$.
2. For each $\rho \in \mathrm{Hom}(\Gamma, G(\boldsymbol{k}))$ satisfying $\mathrm{H}^1(\Gamma, \mathfrak{g}_{\mathrm{Ad}\,\rho}) = 0$, the point ρ is a smooth point of the scheme $\mathrm{Hom}(\Gamma, G)$. The G–orbit through ρ is open in $\mathrm{Hom}(\Gamma, G)$.

In what follows we will use the simplified notation $\mathrm{H}^q(\Gamma, \mathrm{Ad}\,\rho)$ instead of $\mathrm{H}^q(\Gamma, \mathfrak{g}_{\mathrm{Ad}\,\rho})$.

We assume from now on that the group G is affine; in particular, $\mathrm{Hom}(\Gamma, G)$ is also an affine scheme. The group G acts naturally on this scheme:

$$(g, \rho) \mapsto \rho^g \quad \text{where } \rho^g(\gamma) = g\rho(\gamma)g^{-1}.$$

Assuming, in addition, that G is reductive, we obtain the GIT quotient

$$X(\Gamma, G) = \mathrm{Hom}(\Gamma, G) /\!\!/ G,$$

which is a scheme of finite type known as the *character scheme* (or, more commonly, as the *character variety*). However, as we will see, both representation and character schemes can be nonreduced, so we will avoid the traditional *representation/character variety* terminology.

We will use the notation

$$\mathrm{Hom}^{\mathrm{red}}(\Gamma, G) \quad \text{and} \quad X^{\mathrm{red}}(\Gamma, G)$$

to denote the varieties which are the reductions of the schemes

$$\mathrm{Hom}(\Gamma, G) \quad \text{and} \quad X(\Gamma, G).$$

Recall that for every $\rho \in \mathrm{Hom}(\Gamma, G)$, the vector space of cocycles

$$Z^1(\Gamma, \mathrm{Ad}\,\rho)$$

is isomorphic to the Zariski tangent space $T_\rho \mathrm{Hom}(\Gamma, G)$ and this isomorphism carries the subspace of coboundaries $B^1(\Gamma, \mathrm{Ad}\,\rho)$ to the tangent space of the G–orbit through ρ. Note, however, that $\mathrm{H}^1(\Gamma, \mathrm{Ad}\,\rho)$ is *not* always isomorphic to the Zariski tangent space of $[\rho] \in X(\Gamma, G)$; see [2, Section 6] as well as [17].

Suppose now that the group Φ is finite. Then for every $\rho \in \mathrm{Hom}(\Phi, G)$,

$$\mathrm{H}^1(\Phi, \mathrm{Ad}\,\rho) = 0.$$

(Furthermore, $\mathrm{H}^i(\Phi, \mathrm{Ad}\,\rho) = 0$ for $i \geq 1$.) In particular, the G–orbit of ρ is a clopen (closed and open) subscheme

$$\mathrm{Hom}_\rho(\Phi, G) \subset \mathrm{Hom}(\Phi, G).$$

This subscheme is isomorphic to the quotient $G/\zeta_G(\rho(\Phi))$, where $\zeta_G(H)$ denotes the centralizer of the subgroup H in G. (Note that if $\zeta_G(\rho(\Phi))$ equals the center of G, then the point $[\rho] \in X(\Phi, G)$ is a reduced isolated point in the character scheme and the entire character scheme is smooth.) We obtain:

Lemma 2.1 *For every finite group Φ and connected affine group G, the scheme $\mathrm{Hom}(\Phi, G)$ is smooth and each of its irreducible components is G–homogeneous. These irreducible components are the open subschemes $\mathrm{Hom}_\rho(\Phi, G)$. If the representation ρ is trivial, then $\mathrm{Hom}_\rho(\Phi, G)$ is a single point.*

The following lemma is also immediate:

Lemma 2.2 *Let $\phi\colon \Gamma \to \Gamma'$ be a group homomorphism. Then the pull-back map $\phi^*(\rho) = \rho \circ \phi$ is a morphism of schemes*

$$\mathrm{Hom}(\Gamma', G) \to \mathrm{Hom}(\Gamma, G).$$

Lemma 2.3 *Let Γ be a finitely presented group and let $\Theta \subset \Gamma$ be a finite subset with the quotient group*

$$\Gamma' := \Gamma/\langle\langle \Theta \rangle\rangle.$$

Let $\phi\colon \Gamma \to \Gamma'$ denote the projection homomorphism. Then the pull-back morphism

$$\phi^*\colon \mathrm{Hom}(\Gamma', G) \to \mathrm{Hom}_\Theta(\Gamma, G)$$

is an isomorphism, where

$$\mathrm{Hom}_\Theta(\Gamma, G) \subset \mathrm{Hom}(\Gamma, G)$$

is the closed subscheme defined by

$$\mathrm{Hom}_\Theta(\Gamma, G) = \{\rho \in \mathrm{Hom}(\Gamma, G) \mid \rho(\theta) = 1 \textit{ for all } \theta \in \Theta\}.$$

Proof Given a finite presentation P of Γ let P' be the presentation of Γ' obtained from P by adding words representing elements of Θ as the relators. Then the assertion follows immediately from the definition of the representation scheme of a group using a group presentation. □

Corollary 2.4 *Suppose that every element $\theta \in \Theta$ has finite order. Then the isomorphism $\phi_*\colon \mathrm{Hom}(\Gamma', G) \to \mathrm{Hom}_\Theta(\Gamma, G)$ sends $\mathrm{Hom}(\Gamma', G)$ to the open subscheme $\mathrm{Hom}_\Theta(\Gamma, G) \subset \mathrm{Hom}(\Gamma, G)$.*

Proof Consider an element $\theta \in \Theta$ and the trivial representation $\rho_{0,\theta}\colon \langle\theta\rangle \to G$. By Lemma 2.1, the singleton

$$\{\rho_{0,\theta}\} = \mathrm{Hom}_{\rho_{0,\theta}}(\langle\theta\rangle, G) \subset \mathrm{Hom}(\langle\theta\rangle, G)$$

is a reduced isolated point in the scheme $\mathrm{Hom}(\langle\theta\rangle, G)$. In particular, this point is open in $\mathrm{Hom}(\langle\theta\rangle, G)$. We have the pull-back morphism

$$\phi_\theta^*\colon \mathrm{Hom}(\Gamma, G) \to \mathrm{Hom}(\langle\theta\rangle, G),$$

induced by the inclusion homomorphism $\phi_\theta\colon \langle\theta\rangle \hookrightarrow \Gamma$. Therefore,

$$(\phi_\theta^*)^{-1}\big(\mathrm{Hom}_{\rho_{0,\theta}}(\langle\theta\rangle, G)\big) \subset \mathrm{Hom}(\Gamma, G)$$

is an open subscheme. Furthermore, by the definition of $\mathrm{Hom}_\Theta(\Gamma, G)$,

$$\mathrm{Hom}_\Theta(\Gamma, G) = \bigcap_{\theta\in\Theta} (\phi_\theta^*)^{-1}\big(\mathrm{Hom}_{\rho_{0,\theta}}(\langle\theta\rangle, G)\big).$$

(A homomorphism belongs to $\mathrm{Hom}_\Theta(\Gamma, G)$ if and only if it sends each $\theta \in \Theta$ to $1 \in G$.) Therefore, $\mathrm{Hom}_\Theta(\Gamma, G)$ is also open in $\mathrm{Hom}(\Gamma, G)$. □

2.2 Coxeter groups

We refer the reader to [3] for the basics of Coxeter groups.

Let Δ be a finite simplicial graph with the vertex and edge sets denoted $V = V(\Delta)$ and $E = E(\Delta)$, respectively. We will use the notation $e = [v, w]$ for the edge of Δ connecting v and w, if it exists. We assume also that we are given a function

$$m\colon E \to \mathbb{N} \quad \text{with } m(e) \geq 2 \text{ for all } e \in E$$

labeling the edges of Δ. We will call the pair (Δ, m) a *labeled graph* or a *Coxeter* *graph*. Given a labeled graph, we define the associated *Coxeter group* $\Gamma = \Gamma_\Delta$ by the presentation

$$\Big\langle \{g_v\}_{v\in V} \mid g_v^2 = 1\ \forall v\in V,\ \underbrace{g_v g_w \cdots}_{m(e)\text{ terms}} = \underbrace{g_w g_v \cdots}_{m(e)\text{ terms}}\ \ \forall v, w\in V \text{ for which } e=[v,w]\in E\Big\rangle.$$

Alternatively, one can describe the relators of this group as $g_v^2 = 1$ and

$$(g_v g_w)^m = 1$$

where $m = m(e),\ e = [v,w]$.

Remark 2.5 Note that the notation we use here is different from the one in Lie theory, where two generators commute whenever the corresponding vertices are not connected by an edge. In our notation, every such pair of elements of Γ generates an infinite dihedral subgroup of Γ.

We also define the *canonical central extension*

$$1 \to \mathbb{Z}_2 \to \widetilde{\Gamma} \xrightarrow{\eta} \Gamma \to 1 \tag{1}$$

of the group Γ, with the *extended Coxeter group* $\widetilde{\Gamma} = \widetilde{\Gamma}_\Delta$ given by the presentation

$$\Big\langle z, \{g_v\}_{v\in V} \mid z^2 = 1, [g_v, z] = 1, g_v^2 = z\ \forall v \in V,$$
$$\underbrace{g_v g_w \cdots}_{m(e)\text{ terms}} = z^{m(e)+1} \underbrace{g_w g_v \cdots}_{m(e)\text{ terms}}\ \ \forall v, w \in V \text{ for which } e = [v,w] \in E\Big\rangle.$$

The number $r = |V|$ (the cardinality of V) is called the *rank* of Γ and $\widetilde{\Gamma}$. We will refer to the generator z of the group $\widetilde{\Gamma}$ as *the central element* of $\widetilde{\Gamma}$, even though, the center of $\widetilde{\Gamma}$ might be larger than $\mathbb{Z}_2$: this happens precisely when Δ consists of a single vertex.

A subgraph $\Sigma \subset \Delta$ is called *full* if for every pair of vertices $v, w \in \Sigma$, the edge $[v,w]$ in Δ also belongs to Σ. Every subgraph $\Sigma \subset \Delta$ inherits labels from Δ. For the new labeled graph (which we still denote Σ), we have the natural homomorphism

$$\iota_\Sigma\colon \Gamma_\Sigma \to \Gamma_\Delta$$

sending each generator $g_v \in \Gamma_\Sigma$, where $v \in V(\Sigma)$, to the generator of Γ_Δ with the same name. It is immediate that the homomorphism ι_Σ lifts to a homomorphism

$$\widetilde{\iota}_\Sigma\colon \widetilde{\Gamma}_\Sigma \to \widetilde{\Gamma}_\Delta$$

sending each g_v to itself ($v \in V(\Sigma)$) and the central element $z \in \widetilde{\Gamma}_\Sigma$ to the central element $z \in \widetilde{\Gamma}_\Delta$. We will use this construction in two special cases:

a. $\Sigma := \Delta_\varnothing$ is the subgraph which has the same vertex set as Γ, but empty edge set. Then

$$\Gamma_\Sigma \cong F_r \quad \text{and} \quad \widetilde{\Gamma}_\Sigma \cong F_r \times \mathbb{Z}_2,$$

where F_r is the free group on r generators.

b. $\Sigma \subset \Delta$ is a full subgraph. In this case, the homomorphism ι_Σ is injective; see eg [3, page 113]. It follows that the homomorphism $\widetilde{\iota}_\Sigma$ is injective as well.

For full subgraphs $\Sigma \subset \Delta$, the subgroups $\iota_\Sigma(\Gamma_\Sigma) < \Gamma_\Delta$ and $\widetilde{\iota}_\Sigma(\widetilde{\Gamma}_\Sigma) < \widetilde{\Gamma}_\Delta$ are called *parabolic* subgroups of Γ_Δ and $\widetilde{\Gamma}_\Delta$, respectively. We say that a parabolic subgroup of Γ_Δ or $\widetilde{\Gamma}_\Delta$ is *elementary*, if it is a finite parabolic subgroup of rank ≤ 2. The latter requirement simply means that Σ consists of at most two vertices; the finiteness condition means that if Σ consists of two vertices, then these vertices are connected by an edge. We will refer to such subgraphs Σ as *elementary* as well.

Example 2.6

1. If Δ consists of a single edge e labeled 2, then $\Gamma_\Delta \cong \mathbb{Z}_2 \times \mathbb{Z}_2$ and
$$\widetilde{\Gamma}_\Delta \cong Q_8,$$
the finite quaternion group.
2. If the edge e is labeled 4, then Γ_Δ is the dihedral group $I_2(4)$ of order 8; it admits an epimorphism
$$\Gamma_\Delta \to \mathbb{Z}_2 \times \mathbb{Z}_2$$
with kernel the center of Γ_Δ, which is generated by the involution $g_v g_w g_v g_w$.

3 Representations of Coxeter groups and extended Coxeter groups

In this section we prove some basic facts about representations of Coxeter and extended Coxeter groups to $\mathrm{PSL}(2,\mathbb{C})$ and $\mathrm{SL}(2,\mathbb{C})$, respectively.

3.1 Representations of elementary Coxeter groups

Recall that the quotient map

$$p\colon \widetilde{G}(\mathbb{C}) = \mathrm{SL}(2,\mathbb{C}) \to G(\mathbb{C}) = \mathrm{PSL}(2,\mathbb{C}) = \mathrm{SL}(2,\mathbb{C})/\{\pm 1\},$$

is a 2–fold covering. The extended Coxeter groups appear naturally in the context of lifting homomorphisms of Coxeter groups from $\mathrm{PSL}(2,\mathbb{C})$ to $\mathrm{SL}(2,\mathbb{C})$.

Consider the labeled graph Δ, consisting of two vertices v, w and the edge $[v, w]$ labeled $m \geq 2$. The corresponding Coxeter group Γ_Δ is a finite dihedral group, usually denoted $I_2(m)$. This group is isomorphic to the subgroup of the group of symmetries of a regular planar $2m$–gon, acting simply transitively on the set of edges of this polygon. Hence, this group embeds canonically (up to conjugation) into $O(2)$ and thus into $\mathrm{PO}(3,\mathbb{R}) \cong \mathrm{SO}(3,\mathbb{R}) < \mathrm{PSL}(2,\mathbb{C})$. If m is odd, then such a group of symmetries will lift isomorphically to a subgroup of $\mathrm{SU}(2) < \mathrm{SL}(2,\mathbb{C})$. In contrast, we will be interested (only) in the case when m is even; in fact, we will be using only Coxeter graphs with the labels $m = 2$ and $m = 4$ in this paper.

Below we will prove several lemmas about faithful representations of elementary Coxeter (and extended Coxeter) groups into $G(\mathbb{C})$ (and $\widetilde{G}(\mathbb{C})$).

Lemma 3.1 1. *There exists, unique up to conjugation, a faithful representation* $\rho\colon \mathbb{Z}_2 \times \mathbb{Z}_2 \to G(\mathbb{C})$.

2. *There are no faithful representations* $\mathbb{Z}_2 \times \mathbb{Z}_2 \to \widetilde{G}(\mathbb{C})$.

Proof 1. As the image of ρ is finite, it is conjugate to a subgroup of $\mathrm{SO}(3,\mathbb{R}) < G(\mathbb{C})$. Part 1 now follows from the fact that the group $\mathrm{SO}(3,\mathbb{R})$ acts transitively on the set of pairs of orthogonal 1–dimensional subspaces of $\mathbb{R}^3$ (these subspaces, in our setting, are fixed lines of the images of the direct factors of $\mathbb{Z}_2 \times \mathbb{Z}_2$ under ρ).

2. Part 2 follows from the fact that any involution $A \in \mathrm{SL}(2,\mathbb{C})$ has both eigenvalues equal to -1, ie A equals $-1 \in \mathrm{SL}(2,\mathbb{C})$. □

The next lemma and proposition generalize Lemma 3.1 to representations of the dihedral group $I_2(4)$.

Lemma 3.2 *All injective representations* $\rho\colon \Gamma = I_2(4) \to G(\mathbb{C})$ *are conjugate to each other.*

Proof Since the group Γ is finite, its image in $G(\mathbb{C})$ lies in a conjugate of the maximal compact subgroup $\mathrm{SO}(3,\mathbb{R}) < G(\mathbb{C})$. Thus, we will assume that $\rho(\Gamma)$ is contained in $\mathrm{SO}(3,\mathbb{R})$. Since the product of the generating involutions $\rho(g_v)$, $\rho(g_w)$ of $\rho(\Gamma)$ has order 4, the fixed lines of $\rho(g_v)$, $\rho(g_w)$ meet at the angle $\frac{\pi}{4}$ in $\mathbb{R}^3$. Now, the assertion follows from the fact that $\mathrm{SO}(3,\mathbb{R})$ acts transitively on the set of 1–dimensional subspaces in $\mathbb{R}^3$ meeting at the given angle. □

Proposition 3.3 *Consider the dihedral group* $I_2(2m) = \Gamma_\Delta$ *and its isomorphism*

$$\rho\colon \Gamma_\Delta \to \Gamma < G(\mathbb{C}).$$

Then:

1. *For every choice of matrices $\widetilde{g}_u \in \widetilde{G}(\mathbb{C})$ projecting to the generators $\rho(g_u) \in \Gamma < G(\mathbb{C})$, the map*
$$g_u \to \widetilde{g}_u \quad \text{for } u \in \{v, w\}$$
extends to a monomorphism $\widetilde{\rho}\colon \widetilde{\Gamma}_\Delta \to \widetilde{G}(\mathbb{C})$.
2. *The centralizer of the group $\widetilde{\rho}(\widetilde{\Gamma}_\Delta)$ in $\widetilde{G}(\mathbb{C})$ equals the center of $\widetilde{G}(\mathbb{C})$.*

Proof The proof amounts to elementary linear algebra; we include the details for the sake of completeness. For notational convenience we will identify the isomorphic groups Γ and Γ_Δ. After conjugating the subgroup Γ in $G(\mathbb{C})$, we can (and will) assume that Γ lies in the subgroup $\mathrm{SO}(3,\mathbb{R}) < G(\mathbb{C})$. The orthogonal subgroup is covered by the unitary subgroup $\mathrm{SU}(2) < \mathrm{SL}(2,\mathbb{C})$. We let $Z(\mathrm{SU}(2)) \cong \mathbb{Z}_2$ denote the center of $\mathrm{SU}(2)$; this center consists of the matrices ± 1.

We begin with several trivial observations. Since ρ is injective, the involutions g_v, g_w are distinct rotations in $\mathrm{SO}(3,\mathbb{R})$. In particular, their fixed-point sets in $\mathbb{CP}^1$ are pairwise disjoint. Suppose the elements $\widetilde{g}_u, \widetilde{g}_v \in \mathrm{SU}(2)$ project to g_v, g_w, respectively. Since the kernel of the covering $\widetilde{G} \to G$ is isomorphic to $\mathbb{Z}_2$, the unitary transformations $\widetilde{g}_v, \widetilde{g}_w \in \mathrm{SU}(2)$ have order at most 4:
$$\widetilde{g}_u^2 \in Z(\mathrm{SU}(2)) \quad \text{for } u \in \{v, w\}.$$
Note that the only involution in $\mathrm{SU}(2)$ is the matrix -1. Since $\widetilde{g}_u$ projects nontrivially to $\mathrm{SO}(3)$, this matrix cannot be an involution. It follows that
$$\widetilde{g}_u^2 = -1 \in \mathrm{SU}(2) \quad \text{for } u \in \{v, w\}.$$
The eigenvalues of the matrices $\widetilde{g}_v$, $\widetilde{g}_w$ have to be roots of unity of the order 4, which implies that the spectrum of each matrix $\widetilde{g}_u$, where $u \in \{v, w\}$, equals $\{i, -i\}$.

We next claim that the eigenspaces of unitary transformations $\widetilde{g}_v$, $\widetilde{g}_w$ are pairwise distinct. If not, then these matrices would be simultaneously diagonalizable, which would imply that their projections to $\mathrm{PSL}(2,\mathbb{C})$ are equal. (Two involutions in $\mathrm{SO}(3,\mathbb{R})$ which have same fixed-point sets have to be the same.)

Suppose now that $A \in \mathrm{SL}(2,\mathbb{C})$ is a matrix centralizing the subgroup $\langle \widetilde{g}_v, \widetilde{g}_w \rangle$ generated by $\widetilde{g}_v$, $\widetilde{g}_w$. We claim that A is a scalar matrix, ie an element of the center of $\mathrm{SL}(2,\mathbb{C})$. Indeed, since A commutes with both $\widetilde{g}_v$, $\widetilde{g}_w$, it has to preserve the eigenspaces of each matrix $\widetilde{g}_v$, $\widetilde{g}_w$. (Here we are using the fact that the eigenvalues of $\widetilde{g}_u$ are distinct for $u \in \{v, w\}$.) However, a nonscalar matrix in $\mathrm{SL}(2,\mathbb{C})$ cannot have three distinct eigenlines. Therefore, A is a scalar matrix. This implies the second claim of the lemma.

The generators g_v, g_w satisfy

$$t = (g_v g_w)^m = (g_w g_v)^m,$$

where t is an order-2 element, which belongs to the center of Γ. (In the geometric realization of Γ_Σ as a group of symmetries of a regular $2m$–gon, the element t corresponds to the order-2 rotation, the central symmetry of the polygon.)

If we had the relation

$$(\tilde{g}_v \tilde{g}_w)^m = (\tilde{g}_w \tilde{g}_w)^m,$$

it would result in the monomorphism

$$\alpha\colon \Gamma \to \mathrm{SU}(2) \quad \text{with } \alpha(g_u) = \tilde{g}_u \text{ for } u \in \{v, w\},$$

lifting the embedding $\rho\colon \Gamma \hookrightarrow \mathrm{SO}(3,\mathbb{R})$. The image of the center $Z(\Gamma)$ of Γ would then be in the center of $\alpha(\Gamma)$, hence, as we noted above, in the center of $\mathrm{SU}(2)$. Then, the composition $\rho = p \circ \alpha$ would send $Z(\Gamma)$ to 1, which is a contradiction.

This leaves us with the only possibility

$$(\tilde{g}_v \tilde{g}_w)^m = -(\tilde{g}_w \tilde{g}_v)^m.$$

To conclude, the map given by

$$g_v \mapsto \tilde{g}_v, \quad z \mapsto -1 \in \mathrm{SL}(2,\mathbb{C}),$$

extends to an homomorphism $\tilde{\Gamma}_\Delta \to p^{-1}(\Gamma)$, sending the central element $z \in \tilde{\Gamma}_\Delta$ to the matrix $-1 \in \mathrm{SL}(2,\mathbb{C})$. Injectivity of this homomorphism follows from injectivity of the representation $\Gamma \to \mathrm{PSL}(2,\mathbb{C})$. □

3.2 Representations faithful on elementary subgroups

For a Coxeter group $\Gamma = \Gamma_\Delta$ we define two subschemes

$$\mathrm{Hom}_o(\Gamma, G) \subset \mathrm{Hom}(\Gamma, G) \quad \text{and} \quad \mathrm{Hom}_o(\tilde{\Gamma}, \tilde{G}) \subset \mathrm{Hom}(\tilde{\Gamma}, \tilde{G}).$$

The former consists of homomorphisms which are injective on every elementary subgroup of Γ; the latter consists of homomorphisms which are injective on every elementary subgroup of $\tilde{\Gamma}$ and send $z \in \tilde{\Gamma}$ to $-1 \in \mathrm{SL}(2,\mathbb{C})$. (In fact, the requirement for z follows from faithfulness on elementary subgroups, except when Δ has no edges.) Since elementary subgroups of Γ and $\tilde{\Gamma}$ are finite, both $\mathrm{Hom}_o(\Gamma, G)$ and $\mathrm{Hom}_o(\tilde{\Gamma}, \tilde{G})$ are open subschemes of the respective representation schemes. We will see later on that these subschemes are also closed. For each

$$\rho \in \mathrm{Hom}_o(\tilde{\Gamma}, G)$$

we have $\rho(z) = 1$, while for each $\widetilde{\rho} \in \mathrm{Hom}_o(\widetilde{\Gamma}, \widetilde{G})$ which projects to $\rho \in \mathrm{Hom}_o(\Gamma, G)$ we have $\widetilde{\rho}(z) = -1$.

In the paper we will be using the labeled graph Ω depicted in Figure 1: This graph has five vertices and nine edges. The edges left unlabeled in the figure all have the label 2. The vertices x, y are the only ones not connected to each other by an edge.

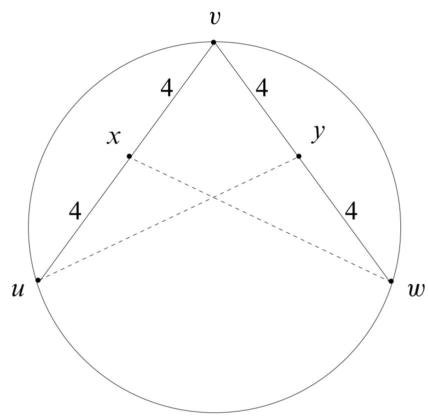


Figure 1: The graph Ω

In what follows, we will also use the subgraph $\Upsilon \subset \Omega$, which is the complete graph on the vertices u, v, w. The parabolic subgroup $\Gamma_\Upsilon < \Gamma_\Omega$ is isomorphic to $\mathbb{Z}_2^3$. Since the group Γ_Υ is finite, the representation scheme $\mathrm{Hom}(\Gamma_\Upsilon, G)$ is smooth.

Lemma 3.4 *Each representation $\rho \in \mathrm{Hom}_o(\Gamma, G(\mathbb{C}))$ of the group $\Gamma = \Gamma_\Upsilon$ satisfies:*

1. *The kernel of ρ is generated by the subgroup*
$$\langle g_u g_v g_w \rangle \cong \mathbb{Z}_2,$$
and the image of ρ is isomorphic to $\mathbb{Z}_2 \times \mathbb{Z}_2$.
2. *The centralizer of the abelian subgroup $\rho(\Gamma) < G(\mathbb{C})$ in the group $G(\mathbb{C})$ equals the subgroup $\rho(\Gamma)$ itself.*
3. $\mathrm{Hom}_o(\Gamma, G(\mathbb{C}))$ *is the $G(\mathbb{C})$–orbit of a singleton $\{\rho_\Upsilon\}$.*

Proof This lemma is also elementary:

1. Consider a homomorphism $\rho \in \mathrm{Hom}_o(\Gamma, G(\mathbb{C}))$. For each element $\gamma \in \Gamma$ we let $\mathrm{Fix}(\gamma)$ denote the fixed-point set of $\rho(\gamma)$ in $\mathbb{CP}^1$. The condition that all three involutions $\rho(g_u)$, $\rho(g_v)$, $\rho(g_w)$ are distinct implies that the three fixed-point sets $\mathrm{Fix}(g_u)$, $\mathrm{Fix}(g_v)$, $\mathrm{Fix}(g_w)$ are pairwise disjoint.

Remark 3.5 If g_1, g_2 are commuting involutions in $\mathrm{PSL}(2,\mathbb{C})$ with fixed-point sets $\{\xi_1,\eta_1\}$, $\{\xi_2,\eta_2\}$, respectively, and $\xi_1=\xi_2$, then $g_2g_1g_2^{-1}=g_1$ implies that

$$g_2(\{\xi_1,\eta_1\})=\{\xi_1,\eta_1\}.$$

Since $g_1(\xi_1)=\xi_1$, it follows that $g_2(\eta_1)=\eta_1$. However, each involution in $\mathrm{PSL}(2,\mathbb{C})$ is determined by its fixed-point set. Therefore, $g_1=g_2$.

Commutativity of $\rho(\Gamma)$ implies that this group preserves the six-point set

$$F=\mathrm{Fix}(g_u)\cup\mathrm{Fix}(g_v)\cup\mathrm{Fix}(g_w)\subset\mathbb{CP}^1.$$

The element $\rho(g_u)$ fixes $\mathrm{Fix}(g_u)$, of course, and defines nontrivial involutions of the other two fixed-point sets

$$\mathrm{Fix}(g_v)\to\mathrm{Fix}(g_v)\quad\text{and}\quad\mathrm{Fix}(g_w)\to\mathrm{Fix}(g_w).$$

The same applies to g_v and g_w. It follows that

$$\rho(g_ug_v)|_F=\rho(g_w)|_F.$$

Hence

$$\rho(g_ug_v)=\rho(g_w),$$

and thus

$$\langle g_ug_vg_w\rangle<\ker(\rho).$$

The equality of these subgroups of Γ follows from the condition that

$$\rho\in\mathrm{Hom}_o(\Gamma,G(\mathbb{C})).$$

This establishes part 1 of the lemma.

2. To prove part 2, note that every $g\in G$ centralizing $\rho(\Gamma)$ has to preserve each of the sets $\mathrm{Fix}(g_u)$, $\mathrm{Fix}(g_v)$, $\mathrm{Fix}(g_w)$. After composing g with elements of $\rho(\Gamma)$, we achieve that g fixes the set $\mathrm{Fix}(g_u)\cup\mathrm{Fix}(g_v)$ pointwise. Therefore, $g\in\rho(\Gamma)$. This proves part 2 of the lemma.

3. To prove part 3 note that, by part 1, the pull-back morphism

$$\mathrm{Hom}_o(\mathbb{Z}^2,G(\mathbb{C}))\to\mathrm{Hom}_o(\Gamma,G(\mathbb{C}))$$

induced by the quotient homomorphism

$$1\to\langle g_ug_vg_w\rangle\to\Gamma\to\mathbb{Z}^2$$

is surjective. Now, the claim follows from Lemma 3.1. □

Lemma 3.6 1. $\mathrm{Hom}_o(\Gamma_\Omega, G(\mathbb{C}))$ *is a single orbit* $G(\mathbb{C}) \cdot \rho_\Omega$.

2. *The representation* ρ_Ω *is infinitesimally rigid:* $\mathrm{H}^1(\Gamma_\Omega, \mathrm{sl}(2,\mathbb{C})) = 0$.
3. *For each* $\rho \in \mathrm{Hom}_o(\Gamma_\Omega, G(\mathbb{C}))$, *the adjoint action* $\mathrm{Ad}\,\rho$ *of* Γ_Ω *on the Lie algebra* $\mathrm{sl}(2,\mathbb{C})$ *has no nonzero fixed vectors.*

Proof 1. Consider $\rho \in \mathrm{Hom}_o(\Gamma_\Omega, G)$. In view of Lemma 3.4, we can assume that the restriction of ρ to the subgroup Γ_Υ equals the representation ρ_Υ. Consider now the dihedral subgroups

$$\langle g_u, g_x\rangle \quad \text{and} \quad \langle g_x, g_v\rangle$$

in the group Γ_Ω. It follows from Lemma 3.2 that there are exactly two extensions (which are faithful on all elementary parabolic subgroups) of the representation

$$\rho|_{\langle g_u, g_v\rangle}$$

to the subgroup $\langle g_u, g_x, g_v\rangle$. We will denote these extensions ρ_+ and ρ_-. For both extensions, $\rho_\pm(g_x)$ lies in $\mathrm{SO}(3,\mathbb{R})$, its fixed line in $\mathbb{R}^3$ is contained in the span of the fixed lines of $\rho_\Upsilon(g_u)$, $\rho_\Upsilon(g_v)$. This fixed line makes the angle $\frac{\pi}{4}$ with the fixed lines of $\rho_\Upsilon(g_u)$, $\rho_\Upsilon(g_v)$ and is orthogonal to the fixed line of $\rho(g_w)$. These representations $\rho_\pm$ are conjugate via the element $\rho(g_v) \in \mathrm{SO}(3)$. Therefore, after such conjugation, we fix the value $\rho(g_x)$. We next repeat this argument for the subgroup of Γ_Ω generated by

$$\{g_v, g_y, g_w\}.$$

Since conjugation via $\rho(g_w)$ does not alter $\rho(g_x)$, we obtain the required uniqueness statement.

2. In what follows we will be using the fact that the adjoint representation of $\mathrm{PSL}(2,\mathbb{C})$ is isomorphic to the complexification V of the standard representation of $\mathrm{SO}(3,\mathbb{R})$ on $\mathbb{R}^3$. We will also use the notation V and $\mathrm{sl}(2,\mathbb{C})$ for the representation $\mathrm{Ad}\,\rho$ of the group Γ_Ω (and its subgroups) in the notation for cocycles and cohomology groups. In particular, for each element a of

$$\{u, v, w, x, y\},$$

the fixed-point set of $\mathrm{Ad}\,\rho(g_a)$ is a line in V, which we will denote by V^a. An elementary but useful geometric observation is that

$$V^x \subset V^u \oplus V^v,$$

while

$$V = V^u \oplus V^v \oplus V^w = V^u \oplus V^x \oplus V^w.$$

Consider a cocycle $\xi \in Z^1(\Gamma_\Omega, V)$. Since Γ_Υ is finite, $\mathrm{H}^1(\Gamma_\Upsilon, V) = 0$. Since the restriction of ξ to the subgroup Γ_Υ is a coboundary, by subtracting off a coboundary from ξ, we can assume that ξ vanishes in Γ_Υ. Similarly, there exist $\alpha, \beta \in V$ such that

$$\xi(h) = \alpha - \operatorname{Ad}\rho(h)\alpha \quad \text{for all } h \in \langle x, u\rangle,$$
$$\xi(h) = \beta - \operatorname{Ad}\rho(h)\beta \quad \text{for all } h \in \langle x, w\rangle.$$

It follows that $\alpha \in V^u$ and $\beta \in V^w$. Moreover, by looking at the value $\xi(x)$, we see that

$$\alpha - \beta \in V^x.$$

Since the lines V^u, V^x, V^w also span V, it follows that $\alpha = \beta = 0$. Therefore, $\xi(x) = 0$. Similarly, $\xi(y) = 0$ and thus $\xi = 0$ on the entire group Γ_Ω.

3. This follows from the fact that $\rho(\Gamma_\Upsilon)$ has no nonzero fixed vectors in $V = \mathrm{sl}(2, \mathbb{C})$. □

Corollary 3.7 *The scheme* $\mathrm{Hom}_o(\Gamma_\Omega, G)$ *is smooth.*

From now on, we will be making the following assumption on the labeled graphs Δ of Coxeter groups Γ:

Assumption 3.8 1. *Every label of the graph* Δ *is even.*

2. Δ *contains as a full subgraph the graph* Ω *above.*

Proposition 3.9 1. *The schemes* $\mathrm{Hom}_o(\Gamma, G)$ *and* $\mathrm{Hom}_o(\widetilde{\Gamma}, \widetilde{G})$ *are clopen subschemes in* $\mathrm{Hom}(\Gamma, G)$ *and* $\mathrm{Hom}(\widetilde{\Gamma}, \widetilde{G})$, *respectively.*

2. *There is a morphism of schemes* $q\colon \mathrm{Hom}_o(\widetilde{\Gamma}, \widetilde{G}) \to \mathrm{Hom}_o(\Gamma, G)$, *such that for every* $\widetilde{\rho} \in \mathrm{Hom}_o(\widetilde{\Gamma}, \widetilde{G})$ *and* $\rho = q(\widetilde{\rho})$ *we have*

$$p \circ \widetilde{\rho} = \rho \circ \eta,$$

where $\eta\colon \widetilde{\Gamma} \to \Gamma$ *is the quotient map from* (1).

3. *The morphism* q *is a regular étale covering with the deck-group* $\mathbb{Z}_2^r$, *where* r *is the rank of* Γ.

Proof 1. We will give a proof for $\mathrm{Hom}_o(\Gamma, G)$, since the other statement is similar. Consider an elementary subgroup $\Gamma_\Sigma \subset \Gamma$; this subgroup is finite. In Lemma 2.1 we proved that each irreducible component of $\mathrm{Hom}(\Gamma_\Sigma, G)$ is a clopen subscheme of $\mathrm{Hom}(\Gamma_\Sigma, G)$; furthermore, each component is a single G–orbit of a representation $\Gamma_\Sigma \to G$. Then

$$\mathrm{Hom}_o(\Gamma_\Sigma, G) = \mathrm{Hom}(\Gamma_\Sigma, G) \setminus \bigcup_{\theta \in \Gamma_\Sigma - \{1\}} \mathrm{Hom}_{\langle\theta\rangle}(\Gamma_\Sigma, G)$$

is an open subscheme in $\mathrm{Hom}(\Gamma_\Sigma, G)$. It is also closed since every subscheme removed was open.

For each elementary subgroup $\Gamma_\Sigma < \Gamma$ and inclusion map $\iota_\Sigma\colon \Gamma_\Sigma \to \Gamma$, we have the pull-back morphism

$$\iota_\Sigma^*\colon \mathrm{Hom}(\Gamma, G) \to \mathrm{Hom}(\Gamma_\Sigma, G).$$

Then we have the finite intersection, taken over all elementary subgraphs $\Sigma \subset \Delta$,

$$\mathrm{Hom}_o(\Gamma, G) = \bigcap_\Sigma (\iota_\Sigma^*)^{-1}(\mathrm{Hom}_o(\Gamma_\Sigma, G)).$$

Therefore, $\mathrm{Hom}_o(\Gamma, G) \subset \mathrm{Hom}(\Gamma, G)$ is clopen as a finite intersection of clopen subschemes.

2. For each $\widetilde{\rho} \in \mathrm{Hom}_o(\widetilde{\Gamma}, \widetilde{G}(\mathbb{C}))$, the reduction modulo centers of $\widetilde{\Gamma}$ and $\widetilde{G}$ yields a homomorphism $\rho \in \mathrm{Hom}_o(\Gamma, G(\mathbb{C}))$. We need to check that the map

$$q\colon \mathrm{Hom}_o(\widetilde{\Gamma}, \widetilde{G}(\mathbb{C})) \to \mathrm{Hom}_o(\Gamma, G(\mathbb{C})), \quad q(\widetilde{\rho}) = \rho,$$

obtained in this fashion comes from a morphism of schemes. First, the composition $\widetilde{\rho} \mapsto p \circ \rho$ is clearly a morphism of schemes

$$\mathrm{Hom}(\widetilde{\Gamma}, \widetilde{G}) \to \mathrm{Hom}(\widetilde{\Gamma}, G).$$

For $\Theta = \{z\}$, we obtain an isomorphism of schemes

$$\mathrm{Hom}_\Theta(\widetilde{\Gamma}, G) \to \mathrm{Hom}(\Gamma, G)$$

(see Lemma 2.3), and $\mathrm{Hom}_\Theta(\widetilde{\Gamma}, G)$ contains the image of $\mathrm{Hom}_o(\widetilde{\Gamma}, \widetilde{G})$. Therefore, q is a composition of two morphisms.

We next verify surjectivity. Let $\rho \in \mathrm{Hom}_o(\Gamma, G)$. Define $\widetilde{\rho}\colon \widetilde{\Gamma} \to \widetilde{G}$ by sending generators g_v to arbitrary elements of $p^{-1}(\rho(g_v))$ and sending the central element $z \in \widetilde{\Gamma}$ to $-1 \in \mathrm{SL}(2, \mathbb{C})$. In view of Proposition 3.3, for each elementary subgroup Γ_Σ in Γ, the restriction of $\widetilde{\rho}$ to the generators of $\widetilde{\Gamma}_\Sigma$ extends to a faithful homomorphism $\widetilde{\rho}|_{\Gamma_\Sigma}$.

Since all the relators of the group $\widetilde{\Gamma}$ come from elementary subgroups, it follows that our map of the generators of $\widetilde{\Gamma}$ to $\mathrm{SL}(2)$ extends to a homomorphism $\widetilde{\rho}\colon \widetilde{\Gamma} \to \mathrm{SL}(2)$. This homomorphism belongs to $\mathrm{Hom}_o(\widetilde{\Gamma}, \widetilde{G})$ since it is faithful on each elementary subgroup.

Thus, we obtained a surjective morphism

$$q\colon \mathrm{Hom}_o(\widetilde{\Gamma}, \widetilde{G}) \to \mathrm{Hom}_o(\Gamma, G), \quad q(\widetilde{\rho}) = \rho.$$

The group $\mathbb{Z}_2$ is the group of automorphisms of the covering $\widetilde{G} \to G$; therefore, the product of r copies of $\mathbb{Z}_2$ acts naturally on the product of r copies of $\widetilde{G}$ as the group of automorphisms of the (regular) cover

$$\widehat{p} = p \times \cdots \times p \colon \prod_{i=1}^{r} \widetilde{G} \to \prod_{i=1}^{r} G.$$

Since the rank r of the group Γ is the number of its generators g_v, we have the morphism

$$\prod_{i=1}^{r} \widetilde{G} \cong \mathrm{Hom}(F_r, \widetilde{G}) \xrightarrow{\widehat{p}} \prod_{i=1}^{r} G \cong \mathrm{Hom}(F_r, G),$$

where F_r is the free group on r generators. We also have the commutative diagram

$$\begin{tikzcd}
\mathrm{Hom}(\widetilde{\Gamma}, \widetilde{G}) \arrow[r, "q"] \arrow[d] & \mathrm{Hom}(\Gamma, G) \arrow[d] \\
\mathrm{Hom}(F_r, \widetilde{G}) \arrow[r, "\widehat{p}"] & \mathrm{Hom}(F_r, G)
\end{tikzcd}$$

where the vertical arrows are the inclusions of representation schemes induced by the epimorphisms

$$F_r \to \widetilde{\Gamma} \quad \text{and} \quad F_r \to \Gamma$$

sending the free generators of F_r to the generators g_r of the extended Coxeter and Coxeter groups. It is elementary and left to the reader to verify that the group $\mathbb{Z}_2^r$ of automorphisms of $\widehat{p}$ preserves the subscheme $\mathrm{Hom}_o(\widetilde{\Gamma}, \widetilde{G})$. Therefore, this finite group acts simply transitively on the fibers of the morphism q.

3. It remains to show that the map q is étale, ie that it is an isomorphism of analytic germs at every point. Let $\widetilde{\rho}$ be in $\mathrm{Hom}_o(\widetilde{\Gamma}, \mathrm{SL}(2, \mathbb{C}))$ and set $\rho := q(\widetilde{\rho})$.

Below is a proof which assumes the reader's familiarity with [1], where the theory of controlling differential graded Lie algebras for various deformation problems was developed.

In view of [1, Theorem 6.8], it suffices to verify that the differential graded Lie algebras controlling these germs are quasi-isomorphic. First, the Lie algebras of G and $\widetilde{G}$ are isomorphic under the covering p, which implies that the covering map p induces isomorphisms

$$\mathrm{H}^i(\widetilde{\Gamma}, \mathrm{Ad} \circ \widetilde{\rho}) \to \mathrm{H}^i(\widetilde{\Gamma}, \mathrm{Ad} \circ p(\widetilde{\rho})) \quad \text{for } i \geq 0.$$

Since the central subgroup $\mathbb{Z}_2$ of $\widetilde{\Gamma}$ is finite,

$$\mathrm{H}^i(\mathbb{Z}_2, \mathrm{sl}(2, \mathbb{C})) = 0 \quad \text{for } i \geq 1.$$

Therefore, applying the Lyndon–Hochschild–Serre spectral sequence to the central extension (1), we obtain isomorphisms

(2) $$\mathrm{H}^i(\Gamma, \mathrm{Ad}\circ\rho) \to \mathrm{H}^i(\widetilde{\Gamma}, \mathrm{Ad}\circ\widetilde{\rho}), \quad \text{where } \rho = q(\widetilde{\rho}), \qquad \text{for } i \geq 1.$$

(Actually, for $i = 0$ both cohomology groups vanish, which implies that they too are isomorphic.) These isomorphisms ensure that the morphism

$$q\colon (\mathrm{Hom}(\widetilde{\Gamma}, \widetilde{G}), \widetilde{\rho}) \to (\mathrm{Hom}(\Gamma, G), \rho)$$

is an isomorphisms of germs. □

Remark 3.10 Below is an alternative argument proving that q is étale, which does not reply upon differential graded Lie algebras. The morphism q is étale if and only if q induces bijections of sets of A–points of representation schemes for all local Artin $\mathbb{C}$–algebras A; see [6, Theorem 2.2]. Let A be a local Artin $\mathbb{C}$–algebra and $\epsilon\colon A \to \mathbb{C}$ be the quotient by the maximal ideal. Then we have natural bijections

$$\mathrm{Hom}(\Gamma, G(A)) \cong \mathrm{Hom}(\Gamma, G)(A) \quad \text{and} \quad \mathrm{Hom}(\widetilde{\Gamma}, \widetilde{G}(A)) \cong \mathrm{Hom}(\widetilde{\Gamma}, \widetilde{G})(A)$$

and the commutative diagram

$$\begin{tikzcd}
\mathrm{Hom}(\widetilde{\Gamma}, \widetilde{K}_A) \arrow[r] \arrow[d, "q_{K_A}"'] & \mathrm{Hom}(\widetilde{\Gamma}, \widetilde{G}(A)) \arrow[r, "\widetilde{u}"] \arrow[d, "q_A"'] & \mathrm{Hom}(\widetilde{\Gamma}, \widetilde{G}(\mathbb{C})) \arrow[d, "q"'] \\
\mathrm{Hom}(\Gamma, K_A) \arrow[r] & \mathrm{Hom}(\Gamma, G(A)) \arrow[r, "u"] & \mathrm{Hom}(\Gamma, G(\mathbb{C}))
\end{tikzcd}$$

where $\widetilde{K}_A$ and K_A are the respective kernels of the group homomorphisms

$$\widetilde{G}(A) \to \widetilde{G}(\mathbb{C}) \quad \text{and} \quad G(A) \to G(\mathbb{C})$$

induced by $\epsilon\colon A \to \mathbb{C}$. We observe that the group $\widetilde{K}_A$ is torsion-free and, since the covering map $p\colon \widetilde{G} \to G$ is étale, the induced map $p_{K_A}\colon \widetilde{K}_A \to K_A$ is an isomorphism. Let us prove that for each $\widetilde{\rho}\colon \widetilde{\Gamma} \to \widetilde{G}(\mathbb{C})$, the restriction of q_A to $\widetilde{u}^{-1}(\widetilde{\rho})$ is injective. Suppose that $\widetilde{\rho}_A, \widetilde{\rho}'_A \in u^{-1}(\widetilde{\rho})$ project via q_A to the same homomorphism $\rho_A\colon \Gamma \to G$. Then

$$\widetilde{\rho}_A(\gamma) = \pm\widetilde{\rho}'_A(\gamma) \quad \text{for each } \gamma \in \widetilde{\Gamma}.$$

Assume that $\widetilde{\rho}_A(\gamma) = gk = -\widetilde{\rho}'_A(\gamma) = -gk'$, where $g \in \mathrm{SL}(2, \mathbb{C})$ and $k, k' \in \widetilde{K}_A$. Then $kk' = -1$ which contradicts the property that $\widetilde{K}_A$ is torsion-free. This proves injectivity. Lastly, we verify surjectivity of the restriction map

$$q_A|_{u^{-1}(\widetilde{\rho})}\colon \widetilde{u}^{-1}(\widetilde{\rho}) \to \{\rho_A\colon \Gamma \to G(A) \mid u(\rho_A) = \rho = q(\widetilde{\rho})\},$$

where $\widetilde{\rho} \in \mathrm{Hom}_o(\widetilde{\Gamma}, \widetilde{G}(\mathbb{C}))$. Let $\rho_A \in \mathrm{Hom}(\Gamma, G(A)) \in u^{-1}(\rho)$, where $q(\widetilde{\rho}) = \rho$. Given $\widetilde{g} \in \mathrm{SL}(2,\mathbb{C})$, $p(\widetilde{g}) = g$ and $gk \in \mathrm{SL}(2, A)$, where $k \in K_A$, we lift gk to the element $\widetilde{g}\widetilde{k} = p_{K_A}^{-1}(k)$. We apply this construction to the images of each generator g_v (of Γ and of $\widetilde{\Gamma}$) and the generator z of $\widetilde{\Gamma}$, under the homomorphisms $\widetilde{\rho}$ and ρ_A. We leave it to the reader to verify that the resulting map of the generators of $\widetilde{\Gamma}$ to $\mathrm{SL}(2, A)$ defines a homomorphism $\widetilde{\rho}_A$ in $\mathrm{Hom}_o(\widetilde{\Gamma}, \mathrm{SL}(2, A))$. By the construction, $\widetilde{u}(\widetilde{\rho}_A) = \widetilde{\rho}$ and $q_A(\widetilde{\rho}_A) = \rho_A$.

3.3 Character schemes of representations faithful on elementary subgroups

In this section we extend the results of the previous section from representation schemes to character schemes.

3.3.1 Stability Given a reductive affine algebraic group H and a finitely generated group Λ, we have the algebraic action of the group H on the homomorphism scheme $\mathrm{Hom}(\Lambda, H)$ given by

$$(h, \rho) \mapsto \mathrm{Inn}(h) \circ \rho,$$

where $\mathrm{Inn}(h)$ is the inner automorphism $g \mapsto hgh^{-1}$ of the group H. Recall that the character scheme is defined as the Mumford quotient

$$X(\Lambda, H) = \mathrm{Hom}(\Lambda, H) /\!/ H.$$

Geometrically speaking, the Mumford quotient is obtained by identifying the semisimple points $\mathrm{Hom}_{\mathrm{ss}}(\Lambda, H)$ of the H–action by the *extended orbit equivalence relation*, while the restriction of the projection

$$\mu\colon \mathrm{Hom}_{\mathrm{ss}}(\Lambda, H) \to X(\Lambda, H)$$

to the stable locus $\mathrm{Hom}_{\mathrm{st}}(\Lambda, H)$ (consisting of stable points) is just the quotient by the H–orbit equivalence. Hence the restriction of the projection to the stable locus has especially simple form. We will use the notation

$$\rho \mapsto [\rho]$$

for the projection μ.

A *sufficient condition* for stability of representations $\rho \in \mathrm{Hom}(\Lambda, H)$ (under the H–action) in terms of the Zariski closure of $\rho(\Lambda)$ in H was established in [4]:

Theorem 3.11 *A representation $\rho \in \mathrm{Hom}(\Lambda, H)$ is semistable provided that the Zariski closure $\overline{\rho(\Lambda)}$ is reductive. A representation is stable provided that the Zariski closure $\overline{\rho(\Lambda)}$ is reductive and the centralizer $Z_H(\rho(\Lambda))$ of the image of ρ is finite.*

In the case of representations into $H = \mathrm{PO}(3)$ and $H = \mathrm{Spin}(3)$, the sufficient condition for stability amounts to requiring that the image of ρ is not contained in a Borel subgroup of H. Our next goal is to verify stability condition and identify centralizers of the images of representations in the context of Coxeter and extended Coxeter groups. Recall that we are using the notation G for PSL(2) and $\widetilde{G}$ for SL(2) (regarded as group schemes).

Lemma 3.12 *Let Γ be a Coxeter group and $\widetilde{\Gamma}$ the corresponding extended Coxeter group, satisfying Assumption 3.8. Then for each $\rho \in \mathrm{Hom}_o(\Gamma, G)$ and $\widetilde{\rho} \in \mathrm{Hom}_o(\widetilde{\Gamma}, \widetilde{G})$ we have:*

1. *The representations ρ, $\widetilde{\rho}$ are stable points in Mumford's sense.*
2. *The centralizers of the images of ρ, $\widetilde{\rho}$ equal the center of the target group.*

Proof Recall that we require the group Γ to contain a subgroup Γ_Ω. It suffices to prove both 1 and 2 for the representations $\rho \in \mathrm{Hom}_o(\Gamma_\Omega, G)$, $\widetilde{\rho} \in \mathrm{Hom}_o(\widetilde{\Gamma}_\Omega, \widetilde{G})$, since we have to verify that the image of the representation is not contained in a Borel subgroup and that its centralizer equals the center of the target group.

(i) First, we consider the case of representations $\widetilde{\rho}\colon \widetilde{\Gamma} = \widetilde{\Gamma}_\Omega \to \mathrm{SL}(2, \mathbb{C})$. We restrict our attention to the subgraph $\Sigma \subset \Omega$, which consists of two vertices x, y and the edge $e = [x, y]$ labeled 4. Each representation $\widetilde{\rho} \in \mathrm{Hom}_o(\widetilde{\Gamma}, \widetilde{G})$ projects to a faithful representation

$$\rho\colon \Gamma_\Sigma \hookrightarrow \mathrm{PSL}(2, \mathbb{C}).$$

By Proposition 3.3, the centralizer of the subgroup $\widetilde{\rho}(\widetilde{\Gamma}_\Sigma)$ equals the center of $\mathrm{SL}(2, \mathbb{C})$. Moreover, the images of the generators of $\widetilde{\Gamma}_\Sigma$ under $\widetilde{\rho}$ have distinct eigenlines. It follows that the subgroup $\widetilde{\rho}(\widetilde{\gamma}_\Sigma)$ cannot have an invariant line in $\mathbb{C}^2$, thereby proving that $\widetilde{\rho}(\widetilde{\Gamma})$ is not contained in a Borel subgroup of $\mathrm{SL}(2, \mathbb{C})$. This proves parts 1 and 2 for representations to $\mathrm{SL}(2, \mathbb{C})$.

(ii) Consider now representations $\rho\colon \Gamma \to G$. By the assumption, ρ sends distinct generators of Γ to distinct elements of G. It follows that the group $\rho(\Gamma)$ cannot fix a point in $\mathbb{CP}^1$. In other words, the group $\rho(\Gamma)$ is not contained in a Borel subgroup of G. This proves part 1.

To prove part 2, we will use subgroups Γ_Υ and Γ_Ω of the group Γ. Since ρ belongs to $\mathrm{Hom}_o(\Gamma, \mathrm{PSL}(2, \mathbb{C}))$, the centralizer of $\rho(\Gamma_\Upsilon)$ in G equals the subgroup $\rho(\Gamma_\Upsilon)$ itself (Lemma 3.4). On the other hand, ρ is faithful on the subgroups generated by $\{g_u, g_x\}$, $\{g_v, g_y\}$, $\{g_w, g_z\}$. Therefore,

$$[\rho(g_u), \rho(g_x)] \neq 1, \quad [\rho(g_v), \rho(g_y)] \neq 1 \quad \text{and} \quad [\rho(g_w), \rho(g_z)] \neq 1,$$

and hence the subgroup $\rho(\Gamma)$ has trivial centralizer in $\mathrm{PSL}(2, \mathbb{C})$. □

3.3.2 Cross-sections Let Y be a quasiaffine scheme and $G \times Y \to Y$ be an algebraic group. Suppose that $C \subset Y$ is a closed subscheme, such that the orbit map

$$G \times C \to Y$$

is an isomorphism. In particular, C projects isomorphically onto $Y /\!\!/ G$, since

$$(G \times C) /\!\!/ G \cong C.$$

Such a subscheme C is called a *cross-section* for the action of G on Y. We leave it to the reader to check that if $A_c \subset A$ is a cross-section for the action of G on A and we have an action $G \curvearrowright B$, then $A_c \times B$ is a cross-section for the product action on $A \times B$.

Lemma 3.13 *Suppose that Y is a (quasiaffine) scheme of finite type, $G \times Y \to Y$ is an (algebraic) action of an affine algebraic group, $C \subset X$ is a cross-section for this action. Suppose that $Y' \subset Y$ is a G–invariant subscheme. Then $C' = Y' \cap C$ is also a cross-section for the action $G \times Y' \to Y'$.*

Proof We need to show that the orbit map $G \times C' \to Y'$ is an isomorphism. It suffices to show that for each commutative ring A, the orbit map

$$\mu'\colon G(A) \times C'(A) \to Y'(A)$$

of A–points is a bijection; see the appendix. We have $C'(A) = C(A) \cap Y(A)$. Since the orbit map

$$\mu\colon G(A) \times C(A) \to Y(A)$$

is a bijection and $Y'(A)$ is $G(A)$–invariant, it follows that μ is a bijection. □

Note that if the scheme Y and its subscheme $C \subset Y$ are both smooth then the condition that C is a cross-section for the action of G is easier to check: It suffices to verify that the set of complex points of C is a set-theoretic cross-section for the action of G on $Y(\mathbb{C})$. Indeed, $\mu\colon G \times C \to Y$, the restriction of the G–action on Y, is a morphism. Our hypothesis amounts to the assumption that μ induces a bijection of $\mathbb{C}$–points. The fact that μ is an isomorphism now follows, for instance, from the Zariski Main Theorem.

We now specialize to the case of representation schemes. Let $\pi' = \pi/N$ be a finitely generated group (where π is a finitely generated group and $N \triangleleft \pi$ is a normal subgroup), G be an affine algebraic group, and $G \times \mathrm{Hom}(\pi, G) \to \mathrm{Hom}(\pi, G)$ be the action of G by conjugation on the representation scheme. We will assume that the scheme $\mathrm{Hom}(\pi, G)$ is smooth. Suppose that $U \subset \mathrm{Hom}(\pi, G)$ is a G–invariant open affine subscheme and $U' = U \cap \mathrm{Hom}(\pi', G)$. We assume that $U_c \subset U$ is a closed smooth subscheme. Then, in view of smoothness, the property that U_c is a cross-section for the G–action on U

amounts to the condition that $U_c(\mathbb{C})$ is a cross-section for the action of $G(\mathbb{C})$ on U. If this is the case, then, according to Lemma 3.13, the subscheme $U'_c = U_c \cap U'$ is a cross-section for the action of G on U'.

3.3.3 Cross-sections of representation schemes We apply the above observations in two situations. First, suppose that Γ is a Coxeter group satisfying Assumption 3.8; we let $G = \mathrm{PO}(3)$. We have the identity embedding $\iota_\Omega\colon \Gamma_\Omega \hookrightarrow \Gamma$ of the finite subgroup Γ_Ω. Recall that, according to Lemma 3.6, $\mathrm{Hom}_o(\Gamma_\Omega, G(\mathbb{C}))$ consists of a single G–orbit $G(\mathbb{C}) \cdot \rho_\Omega$. We then set

$$\mathrm{Hom}_c(\Gamma, G) := (\iota_\Omega^*)^{-1}(\rho_\Omega).$$

The next lemma is an analogue of Corollary 12.11 in [6]:

Lemma 3.14 *The subscheme* $\mathrm{Hom}_c(\Gamma, G)$ *is a cross-section for the action* $G \curvearrowright \mathrm{Hom}_o(\Gamma, G)$. *In particular,*

$$X_o(\Gamma, G) \cong \mathrm{Hom}_c(\Gamma, G).$$

Proof We let $\pi' = \Gamma$ and define the new group π as the Coxeter group whose Coxeter graph is obtained from the one of Γ by removing all the edges which are not in Ω. Define

$$\pi_o := \underbrace{\mathbb{Z}_2 \star \cdots \star \mathbb{Z}_2}_{n \text{ times}}$$

with one free factor for each vertex of the Coxeter graph not contained in Ω. We have

$$\mathrm{Hom}_o(\pi_o) = \prod_{i=1}^{n} \mathrm{Hom}_o(\mathbb{Z}_2, G).$$

Then

$$\pi \cong \pi_o \star \Gamma_\Omega \quad \text{and} \quad \mathrm{Hom}_o(\pi, G) \cong \mathrm{Hom}_o(\pi_o, G) \times \mathrm{Hom}_o(\Gamma_\Omega, G).$$

The scheme $U = \mathrm{Hom}_o(\Gamma_\Omega, G)$ is smooth by Lemma 3.6, while $\mathrm{Hom}_o(\mathbb{Z}_2, G)$ is smooth since $\mathbb{Z}_2$ is finite; therefore, the representation scheme $\mathrm{Hom}_o(\pi, G)$ is smooth as well.

Clearly, $\pi' = \pi/N$ for a normal subgroup $N \lhd \pi$. We again have the inclusion homomorphism $\iota_\Upsilon\colon \Gamma_\Omega \to \pi$; the subscheme

$$C = \mathrm{Hom}_c(\pi, G) := (\iota_\Upsilon^*)^{-1}(\rho_\Omega)$$

is smooth since it is naturally isomorphic to $\mathrm{Hom}(\pi_o, G)$. The fact that C is a cross-section for the action of G on U follows immediately from Lemma 3.6 and observations following Lemma 3.13. Lastly, note that

$$U' = \mathrm{Hom}_c(\Gamma, G) = U \cap \mathrm{Hom}(\Gamma, G).$$

Now, the lemma follows from Lemma 3.13. □

The second situation when we apply our description of cross-sections is the one of representations of extended Coxeter groups $\widetilde{\Gamma}$ (again satisfying Assumption 3.8) to the group $\widetilde{G} \cong \mathrm{SL}(2)$. The group $\widetilde{G}$ does not act faithfully on $\mathrm{Hom}(\widetilde{\Gamma}, \widetilde{G})$; this action factors through the action of the group $G = \mathrm{PO}(3)$.

Earlier, we defined the subscheme $\mathrm{Hom}_o(\widetilde{\Gamma}, \widetilde{G}) \subset \mathrm{Hom}(\widetilde{\Gamma}, \widetilde{G})$. Set

$$\mathrm{Hom}_c(\widetilde{\Gamma}, \widetilde{G}) := q^{-1}(\mathrm{Hom}_c(\Gamma, G)).$$

Lemma 3.15 $\mathrm{Hom}_c(\widetilde{\Gamma}, \widetilde{G})$ *is a cross-section for the action of* G *on* $\mathrm{Hom}_o(\widetilde{\Gamma}, \widetilde{G})$.

Proof We let $\pi' = \widetilde{\Gamma}$. Similarly to the proof of Lemma 3.14, we define the extended Coxeter group π by eliminating all the edges which are not in the subgraph Ω. Then π' is isomorphic to a quotient of π and the same proof as in Lemma 3.14 goes through. □

3.3.4 Character schemes We let $X_o(\Gamma, G)$ and $X_o(\widetilde{\Gamma}, \widetilde{G})$ denote the projections of $\mathrm{Hom}_o(\Gamma, G)$ and $\mathrm{Hom}_o(\widetilde{\Gamma}, \widetilde{G})$ to the corresponding character schemes.

In view of Lemmata 3.14 and 3.15, the projections

$$\mathrm{Hom}_o(\Gamma, G) \to X_o(\Gamma, G) \quad \text{and} \quad \mathrm{Hom}_o(\widetilde{\Gamma}, \widetilde{G}) \to X_o(\widetilde{\Gamma}, \widetilde{G})$$

are trivial principal fiber bundles with structure group $G = \mathrm{PSL}(2, \mathbb{C})$: the center of the group $\widetilde{G}$ acts trivially on $\mathrm{Hom}(\widetilde{\Gamma}, \widetilde{G})$. We record this as:

Corollary 3.16 *There exist natural isomorphisms of germs*

$$\begin{aligned}(\mathrm{Hom}_o(\widetilde{\Gamma}, \widetilde{G}), \widetilde{\rho}) &\cong (\mathrm{Hom}_c(\widetilde{\Gamma}, \widetilde{G}) \times G, \widetilde{\rho} \times 1) \\ &\cong (X_o(\widetilde{\Gamma}, \widetilde{G}) \times G, [\widetilde{\rho}] \times 1)\end{aligned}$$

and

$$\begin{aligned}(\mathrm{Hom}_o(\Gamma, G), \rho) &\cong \mathrm{Hom}_c(\Gamma, G) \times G, \rho \times 1) \\ &\cong (X_o(\Gamma, G) \times G, [\rho] \times 1).\end{aligned}$$

3.3.5 Adding a free factor Let F_k be the free group on k generators. For an arbitrary finitely generated group Λ and an algebraic group H we have an isomorphism of schemes

$$\mathrm{Hom}(\Lambda \star F_k, H) \cong \mathrm{Hom}(\Lambda, H) \times \mathrm{Hom}(F_k, H) \cong \mathrm{Hom}(\Lambda, H) \times H^k. \tag{3}$$

This isomorphism is H–equivariant; here and below the action of H is by conjugation on the left side and the diagonal action (by conjugations) on the product space on the right side.

We will use these isomorphisms in the following two special cases: $\Lambda = \Gamma$, $H = G$ and $\Lambda = \widetilde{\Gamma}$, $H = \widetilde{G}$, where $G = \mathrm{PSL}(2)$, $\widetilde{G} = \mathrm{SL}(2)$, and Γ, $\widetilde{\Gamma}$ are Coxeter and extended Coxeter groups, respectively. Then the isomorphisms (3) for these groups allow us to define clopen subschemes

$$\mathrm{Hom}_o(\Gamma \star F_k, G) \subset \mathrm{Hom}(\Gamma \star F_k, G) \quad \text{and} \quad \mathrm{Hom}_o(\widetilde{\Gamma} \star F_k, \widetilde{G}) \subset \mathrm{Hom}(\widetilde{\Gamma} \star F_k, \widetilde{G})$$

as the images of

$$\mathrm{Hom}_o(\Gamma, G) \times G^k \quad \text{and} \quad \mathrm{Hom}_o(\widetilde{\Gamma}, \widetilde{G}) \times \widetilde{G}^k,$$

respectively.

It follows from Lemmata 3.14 and 3.15 that $\mathrm{Hom}_c(\Gamma, G) \times G^k$ is a cross–section for the action of G on $\mathrm{Hom}_o(\Gamma, G) \times G^k$, while $\mathrm{Hom}_c(\widetilde{\Gamma}, \widetilde{G}) \times \widetilde{G}^k$ is a cross–section for the action of G on $\mathrm{Hom}_o(\widetilde{\Gamma}, \widetilde{G}) \times \widetilde{G}^k$.

We thus obtain:

Lemma 3.17 $\quad (\mathrm{Hom}_o(\Gamma, G) \times G^k)/G \cong X_o(\Gamma, G) \times G^k.$

The étale covering q defined above yields, for each k, the étale covering

$$q_k\colon \mathrm{Hom}_o(\widetilde{\Gamma}, \widetilde{G}) \times \widetilde{G}^k \cong \mathrm{Hom}_o(\widetilde{\Gamma} \star F_k, \widetilde{G}) \to \mathrm{Hom}_o(\Gamma \star F_k, G) \cong \mathrm{Hom}_o(\Gamma, G) \times G^k.$$

Corollary 3.18

1. $X_o(\Gamma \star F_k, G) \cong X_o(\Gamma, G) \times G^k$.
2. $\mathrm{Hom}_o(\widetilde{\Gamma} \star F_k, \widetilde{G}) \cong X_o(\widetilde{\Gamma}, \widetilde{G}) \times \widetilde{G}^k$.
3. *The covering q_k is étale.*

4 Universality theorem of Panov and Petrunin

The proofs of Theorem 1.1 and its corollaries hinge upon two results, the first of which is the following:

Theorem 4.1 (Panov–Petrunin universality theorem [12]) *Let Γ be a finitely presented group. Then there exists a closed 3–dimensional (nonorientable) hyperbolic orbifold O such that $\pi_1(Y) \cong \Gamma$, where Y is the underlying space of O. Furthermore, Y is a 3–dimensional pseudomanifold without boundary.*

Remark 4.2 Examination of the proof in [12] shows that the orbifold O admits a hyperbolic manifold cover $\widetilde{O} \to O$ with deck-transformation group $\mathbb{Z}_2^4$.

The singular set of the pseudomanifold Y consists of singular points y_j for $j=1,\dots,2k$, whose neighborhoods C_j in Y are cones over $\mathbb{RP}^2$. Note that, since $\mathbb{RP}^2$ has Euler characteristic 1, the number of conical singularities has to be even. Observe also that one needs $k \geq 1$ in this theorem, since fundamental groups of 3–dimensional manifolds are very restricted among finitely presented groups. For instance, there are no 3–manifolds M with $\pi_1(M) \cong \mathbb{Z}^4$; therefore, for $\Gamma \cong \mathbb{Z}^4$, one cannot have $k = 0$ in Theorem 4.1.

Problem 4.3 Does Theorem 4.1 hold with $k = 1$?

Given Γ and Y as in Theorem 4.1, we will construct a closed (nonorientable) 3–dimensional manifold $M = M_\Gamma$ as follows. (Formally speaking, this 3–manifold also depends on the choice of an orbifold O in Theorem 4.1, which is very far from being unique, however, in order to simplify the notation, we will suppress this dependence).

Let O be a 3–dimensional orbifold as in Theorem 4.1 and let Y be the underlying space of O. Let Y' be obtained by removing open cones C_j for $j = 1, \dots, 2k$ from Y. Then Y' is a compact 3–dimensional manifold with $2k$ boundary components each of which is a copy of the projective plane $\mathbb{RP}^2$. We let θ_i denote the generator of the fundamental group of the projective plane $P_i \cong \mathbb{RP}^2 \subset \partial M$, which equals the boundary of the cone C_i and let $\alpha_i\colon \langle\theta_i\rangle \to \langle\theta_{k+i}\rangle$ be the (unique) isomorphism. We will regard each θ_i as an element of $\pi_1(Y')$. Set

$$\Theta := \{\theta_1, \dots, \theta_k\}.$$

Then

$$\Gamma = \pi_1(Y) = \pi_1(Y')/\langle\langle\theta_1, \dots, \theta_{2k}\rangle\rangle.$$

Next, let M be the closed 3–dimensional manifold obtained by attaching k copies of the product $\mathbb{RP}^2 \times [0,1]$ to Y' along the boundary projective planes, pairing the projective planes P_i and P_{i+k} for $i = 1, \dots, k$. Then $\pi = \pi_1(M_\Gamma)$ is the iterated HNN extension

$$\Big(\Big(\big(\pi_1(Y')\star_{\langle\alpha_1\rangle}\big)\star_{\langle\alpha_2\rangle}\Big)\cdots\Big)\star_{\langle\alpha_k\rangle}$$

of $\pi_1(Y')$ with stable letters $t_1, \dots, t_k$.

Taking the quotient

$$\phi\colon \pi \to \pi/\langle\langle\Theta\rangle\rangle, \tag{4}$$

we therefore obtain the group $\Gamma \star F_k$, where F_k is the free group on k generators, projections of the stable letters t_i for $i = 1, \dots, k$ in the above HNN extension. We let

$$\psi\colon \Gamma \star F_k \to \Gamma$$

denote the further projection to the first direct factor and set

$$\xi := \psi \circ \phi \colon \pi \to \Gamma. \tag{5}$$

Now, given an algebraic group H, we obtain

$$\mathrm{Hom}_\Theta(\pi, H) = \phi^*(\mathrm{Hom}(\Gamma \star F_k, H)),$$

a clopen subscheme in $\mathrm{Hom}(\pi, H)$ (see Corollary 2.4). The isomorphism

$$\mathrm{Hom}(\Gamma, H) \times H^k \xrightarrow{\cong} \mathrm{Hom}(\Gamma \star F_k, H) \xrightarrow{\phi^*} \mathrm{Hom}_\Theta(\pi, H)$$

restricted to $\mathrm{Hom}(\Gamma, H) \times 1$ equals ξ^*. We thus obtain:

Lemma 4.4 *For each open subscheme $S \subset \mathrm{Hom}(\Gamma, H)$, there exists an open subscheme $R \subset \mathrm{Hom}(\pi, H)$ isomorphic to $S \times H^k$ via the morphism ϕ^*. Furthermore, R contains $\xi^*(S)$.*

Proof Take $R := \phi^*(S \times H^k)$, where we identify $\mathrm{Hom}(\Gamma, H) \times H^k$ with the representation scheme $\mathrm{Hom}(\Gamma \star F_k, H)$. □

Note that, because of the H–equivariance of ϕ^*, if S is H–invariant and there exists a cross-section $S_c \subset S$ for the action $H \curvearrowright S$, then the pull-back $R_c := \phi^*(S_c \times H^k)$ is a cross-section for $H \curvearrowright R$.

We will be using these results for Coxeter groups Γ as well as for the extended Coxeter groups $\widetilde{\Gamma}$, with the group H given respectively by either $H = G = \mathrm{PSL}(2)$ or $H = \widetilde{G} = \mathrm{SL}(2)$. In order to simplify the notation, we will refer to the fundamental group $\pi_1(M_\Gamma)$ as π and the fundamental group $\pi_1(M_{\widetilde{\Gamma}})$ as $\widetilde{\pi}$. We will denote by Θ the subset (defined above) of order-2 elements in π and by $\widetilde{\Theta}$ the similar subset in $\widetilde{\pi}$.

We obtain a commutative diagram

$$\begin{tikzcd}
\mathrm{Hom}_{\widetilde{\Theta}}(\widetilde{\pi}, \widetilde{G}) \arrow[r, "\cong"] \arrow[d, "\widehat{q}"'] & \mathrm{Hom}(\widetilde{\Gamma} \star F_k, \widetilde{G}) \arrow[r, "\cong"] \arrow[d] & \mathrm{Hom}(\widetilde{\Gamma}, \widetilde{G}) \times \widetilde{G}^k \arrow[d, "q_k"'] \\
\mathrm{Hom}_\Theta(\pi, G) \arrow[r, "\cong"] & \mathrm{Hom}(\Gamma \star F_k, G) \arrow[r, "\cong"] & \mathrm{Hom}(\Gamma, G) \times G^k
\end{tikzcd}$$

where q_k is an étale covering (see Section 3.3.5), and hence $\widehat{q}$ also is an étale covering. The groups of covering transformations for both are $\mathbb{Z}_2^{k+r}$, where r is the rank of the Coxeter group Γ.

We let

$$\mathrm{Hom}_o(\pi, G) \subset \mathrm{Hom}_\Theta(\pi, G) \quad \text{and} \quad \mathrm{Hom}_o(\widetilde{\pi}, \widetilde{G}) \subset \mathrm{Hom}_{\widetilde{\Theta}}(\widetilde{\pi}, \widetilde{G})$$

be the subschemes which are the preimages of $\mathrm{Hom}_o(F_k \star \Gamma, G)$ and $\mathrm{Hom}_o(\widetilde{\Gamma} \star F_k, \widetilde{G})$ under the isomorphisms

$$\mathrm{Hom}_\Theta(\pi, G) \to \mathrm{Hom}(\Gamma \star F_k, G) \quad \text{and} \quad \mathrm{Hom}_{\widetilde{\Theta}}(\widetilde{\pi}, \widetilde{G}) \to \mathrm{Hom}(\widetilde{\Gamma} \star F_k, \widetilde{G}),$$

respectively. Note also that the covering q is equivariant with respect to the morphism $p\colon \widetilde{G} \to G$, hence $\widehat{q}$ is equivariant as well. Because of this equivariance, if $R_c \subset R \subset \mathrm{Hom}_\Theta(\pi, G)$ is a cross-section, so is $\widetilde{R}_c := \widehat{q}^{-1}(R_c) \subset \widetilde{R} \subset \mathrm{Hom}_{\widetilde{\Theta}}(\widetilde{\pi}, \widetilde{G})$. The cross-sections we will be using are

$$\mathrm{Hom}_c(\pi, G) := R_c \subset R = \mathrm{Hom}_o(\pi, G) \quad \text{and} \quad \mathrm{Hom}_c(\widetilde{\pi}, \widetilde{G}) := \widetilde{R}_c \subset \widetilde{R} = \mathrm{Hom}_o(\widetilde{\pi}, \widetilde{G}).$$

We obtain:

Lemma 4.5 *Consider a representation $\varphi \in \mathrm{Hom}_o(\Gamma, G(\mathbb{C}))$; let ρ be its image in $\mathrm{Hom}_\Theta(\pi, G)$ and pick*

$$\widetilde{\rho} \in \widehat{q}^{-1}(\rho) \in \mathrm{Hom}_o(\widetilde{\pi}, \widetilde{G}(\mathbb{C})).$$

Then we have isomorphisms of germs

$$(X(\widetilde{\pi}, \widetilde{G}), [\widetilde{\rho}]) \cong (X(\pi, G), [\rho]) \cong (X_o(\Gamma, G) \times G^k, [\varphi] \times 1),$$

where $1 = (1, \dots, 1) \in G^k$.

Proof These isomorphisms follow from the fact that $\widehat{q}$ is an étale covering and the existence of the following cross-sections for the actions of G:

$$R_c = \mathrm{Hom}_c(\pi, G) \subset \mathrm{Hom}_o(\pi, G), \quad \mathrm{Hom}_c(\Gamma, G) \times G^k \subset \mathrm{Hom}_o(\Gamma, G) \times G^k$$

and

$$\widetilde{R}_c = \mathrm{Hom}_c(\widetilde{\pi}, \widetilde{G}) \subset \mathrm{Hom}_o(\widetilde{\pi}, \widetilde{G}), \quad \mathrm{Hom}_c(\widetilde{\Gamma}, \widetilde{G}) \times \widetilde{G}^k \subset \mathrm{Hom}_o(\widetilde{\Gamma}, \widetilde{G}) \times \widetilde{G}^k. \quad \square$$

5 A universality theorem for Coxeter groups

The second key ingredient we need is the following theorem, which is essentially contained in Kapovich and Millson [6]. Before stating the theorem we recall (Lemma 3.14) that the action $G \curvearrowright \mathrm{Hom}_o(\Gamma, G)$ has a cross–section $\mathrm{Hom}_c(\Gamma, G) \subset \mathrm{Hom}_o(\Gamma, G)$, ie $\mathrm{Hom}_o(\Gamma, G)$ is G–equivariantly isomorphic to the product $X_o(\Gamma, G) \times G$, where G acts trivially on $X_o(\Gamma, G)$. As always, $G = \mathrm{PO}(3)$.

Theorem 5.1 *Let X and $x \in X$ be as in Theorem 1.1. Then there exists an open subscheme $X' \subset X$ containing x, a finitely generated Coxeter group Γ (such that every edge of its graph Δ has label 2 or 4) and a representation $\rho_c\colon \Gamma \to \mathrm{PO}(3, \mathbb{R})$ with dense* *image, such that X' is isomorphic to an open subscheme $S' \subset X_o(\Gamma, G)$. The representation ρ_c belongs to* $\mathrm{Hom}_o(\Gamma, \mathrm{PO}(3,\mathbb{R}))$. *Furthermore, under this isomorphism, x corresponds to $[\rho_c]$.*

Remark 5.2 Since $\mathrm{Hom}_o(\Gamma, G) \cong X_o(\Gamma, G) \times G$, with $\mathrm{Hom}_c(\Gamma, G)$ containing ρ_c serving as a cross-section for the action $G \curvearrowright \mathrm{Hom}_o(\Gamma, G)$, the preimage S'_o of S' in $\mathrm{Hom}_o(\Gamma, G)$ is isomorphic to $S'_c \times G \cong S' \times G$. Here and in what follows, $S'_c \subset S'_o$ is the cross-section given by

$$S'_c := \mathrm{Hom}_c(\Gamma, G) \cap S'_o.$$

Furthermore, as we saw in Section 2.2, the representation ρ_c lifts to a representation

$$\widetilde{\rho}_c\colon \widetilde{\Gamma} \to \mathrm{SU}(2)$$

of the canonical central extension $\widetilde{\Gamma}$ of Γ.

Since the universality theorems proven in [6] are somewhat different from the one stated above, we outline the proof of Theorem 5.1. The main differences are that the results of [6] are about representations of Shephard and Artin groups rather than Coxeter groups. Furthermore, the representation to $\mathrm{PO}(3,\mathbb{R})$ constructed in [6] has finite image (which was important for [6]), although the image group does have trivial centralizer in $\mathrm{PO}(3,\mathbb{C})$.

Outline of proof of Theorem 5.1 The arguments below are minor modifications of the ones in [6].

Step 1 (scheme-theoretic version of Mnëv universality theorem) Without loss of generality, we may assume that the rational point x is the origin 0 in the affine space containing X. In [6] we first construct a *based projective arrangement* A, such that an open subscheme $\mathrm{BR}_0(A, \mathbb{P}^2)$ in the space of *based projective realizations* $\mathrm{BR}(A, \mathbb{P}^2)$, is isomorphic to X as a scheme over $\mathbb{Q}$, and, moreover, the *geometrization* isomorphism

$$X \xrightarrow{\text{geo}} \mathrm{BR}_0(A, \mathbb{P}^2)$$

sends $x \in X$ to a based realization $\psi_0\colon A \to \mathbb{P}^2$ whose image is the *standard triangle*. Furthermore, the images of the points and lines in A under ψ_0 are real.

Remark 5.3 Subsequently, a proof of this result was also given by Lafforgue in [9], who was apparently unaware of [6].

Step 2 An arrangement A is a certain bipartite graph containing a subgraph T (the "base") which is isomorphic to the incidence graph of the "standard triangle" (also known as "standard quadrangle"); see [6, Figure 7]. The subgraph T has five vertices

v_{00}, v_x, v_y, c_{10}, v_{01}, v_{11} corresponding to the "points" of the standard triangle and six vertices l_x, l_d, l_y, l_{x1}, l_{y1}, l_∞ which correspond to the "lines" of the standard triangle. In [6, Section 11] we further modify the bipartite graph A: we make the identifications of vertices

$$v_{00} \sim l_\infty, \quad v_x \sim l_y \quad \text{and} \quad v_y \sim l_x,$$

and we also add to A the edges

$$[v_{10}, v_{00}] \quad \text{and} \quad [v_{01}, v_{00}].$$

We will use the upper-case notation $V_{00} = \psi_0(v_{00})$, $V_x = \psi_0(v_x)$, etc to denote vectors in $\mathbb{C}^3$ which project to the images under ψ_0 of the point-vertices of T. The choice of this vectors is not unique, of course; we assume that V_{00}, V_x, V_y form a basis and

$$V_{10} = V_{00} + V_x, \quad V_{01} = V_{00} + V_y \quad \text{and} \quad V_{11} = V_{00} + V_x + V_y. \tag{6}$$

This is possible because of the incidences in $\psi_0(T)$.

However, here, unlike in [6], *we will not add the edge* $[v_{00}, v_{11}]$. (The purpose of this edge in [6] was to ensure that certain representation of a Shephard group is finite.) We let A' denote the resulting graph (no longer bipartite). We assign labels to the edges of A' as follows: all edges are labeled 2 except for the two edges

$$[v_{10}, v_{00}] \quad \text{and} \quad [v_{01}, v_{00}],$$

which have the label 4. We then let Γ denote the Coxeter group corresponding to this labeled graph. We let T' denote the labeled subgraph of A', whose vertices are the images of the vertices of the arrangement T.

The labeled graph Ω as in Figure 1 embeds into T' via the map given by

$$v \mapsto v_{00}, \quad x \mapsto v_{10}, \quad y \mapsto v_{01}, \quad u \mapsto v_x, \quad w \mapsto v_y.$$

We equip the vector space $\mathbb{C}^3$ with a nondegenerate bilinear form, so that:

1. All subspaces which appear in the image $\psi_0(T) = \psi_0(A)$ are anisotropic (the bilinear form has nondegenerate restriction to these subspaces).
2. The vectors $V_{00}, V_x, V_y \in \mathbb{C}^3$ are pairwise orthogonal and have unit norm.

We let PO(3) denote the projectivization of the orthogonal group $O(3)$ preserving this bilinear form.

A realization $\psi \in R(A, \mathbb{P}^2)$ is *anisotropic* if for each vertex $v \in A$, the image $\psi(v)$ is an anisotropic subspace in $\mathbb{C}^3$. We will use the notation $R_a(A, \mathbb{P}^2) \subset R(A, \mathbb{P}^2)$ and $\mathrm{BR}_a(A, \mathbb{P}^2) \subset \mathrm{BR}(A, \mathbb{P}^2)$ for open schemes of anisotropic realizations and anisotropic based realizations. By condition 1 on the inner product above, $\mathrm{BR}_a(A,\mathbb{P}^2)$ contains ψ_0.

To every anisotropic realization $\psi \in R(A,\mathbb{P}^2)$, we associate a representation of the group Γ by sending every generator $g_v \in \Gamma$ to the isometric involution in PO(3) fixing the subspace $\psi(v)$ in $\mathbb{P}^2$. As in [6], this map of generators of Γ to PO(3) defines a representation

$$\rho_\psi\colon \Gamma \to \mathrm{PO}(3,\mathbb{C}).$$

We define

$$\rho_c := \rho_{\psi_0}.$$

By the construction, each representation ρ_ψ is faithful on elementary subgroups: For the edges $[v,w]$ in A (where v is a point and w is a line), the incidence condition $\psi(v) \in \psi(v)$ in P^2 forces the point reflection in $\psi(v)$ be distinct from the line reflection in $\psi(w)$. For the edges

$$[v_{10}, v_{00}] \quad \text{and} \quad [v_{01}, v_{00}],$$

condition (6) forces the point reflections in $\psi(v_{00})$, $\psi(v_{10})$, $\psi(v_{01})$ to be pairwise noncommuting and hence both subgroups

$$\rho_\psi(\langle g_{v_{00}}, g_{v_{10}}\rangle) < \mathrm{PO}(3,\mathbb{C}) \quad \text{and} \quad \rho_\psi(\langle g_{v_{00}}, g_{v_{01}}\rangle) < \mathrm{PO}(3,\mathbb{C})$$

are isomorphic to $I_2(4)$. We also note that

$$\rho_\psi|_{\Gamma_\Omega} = \rho_\Omega := \rho_{\psi_0}\colon \Gamma_\Omega \to \mathrm{PO}(3,\mathbb{C}). \tag{7}$$

We thus obtain the *algebraization* morphism of schemes

$$\mathrm{alg}\colon \mathrm{BR}_a(A,\mathbb{P}^2) \to \mathrm{Hom}(\Gamma,\mathrm{PO}(3)) \quad \text{given by } \psi \mapsto \rho_\psi.$$

As in [6], the morphism alg is an isomorphism to its image. It follows from Lemma 3.14 and (7) that the subscheme

$$S_c := \mathrm{alg}(\mathrm{BR}_a(A,\mathbb{P}^2)) \subset \mathrm{Hom}_c(\Gamma,\mathrm{PO}(3)) \subset \mathrm{Hom}_o(\Gamma,\mathrm{PO}(3))$$

is a cross-section for the action of G on the G–orbit of S_c, which we denote

$$S \subset \mathrm{Hom}_o(\Gamma,\mathrm{PO}(3)).$$

Let $\Sigma \subset A'$ denote the complete subgraph whose vertices are the vertices (points and lines) of the standard triangle in A, except for the vertex v_{11}. As in [6], the image under ρ_c of the corresponding parabolic Coxeter subgroup $\Gamma_\Sigma \subset \Gamma$ is isomorphic to the finite Coxeter group B_3 (the symmetry group of the regular octahedron) divided by the center $\mathbb{Z}_2$. Such a group is a maximal finite subgroup of $\mathrm{PO}(3,\mathbb{R})$. However, the involution $\rho_c(g_{v_{11}})$ does not belong to the group $\rho_c(\Gamma_\Sigma)$ (this would be an order-2 rotation in the center of a face of the octahedron). Thus, the group $\rho_c(\Gamma)$ has to be dense in $\mathrm{PO}(3,\mathbb{R})$, as it contains (actually, is equal to) the dense subgroup $\rho_c(\Gamma_{T'})$. This is the only essential difference between the construction in this paper and in [6], where it was important for the group $\rho_c(\Gamma)$ to be finite.

We let

$$\mu\colon \mathrm{Hom}_o(\Gamma, G) \to X_o(\Gamma, G)$$

denote the restriction of the GIT quotient $\mathrm{Hom}(\Gamma, G) \to X(\Gamma, G)$. Since $\mathrm{Hom}_c(\Gamma, G)$ is a cross-section for the G–action on $\mathrm{Hom}_o(\Gamma, G)$, the morphism μ is a trivial principal G–bundle.

Theorem 5.4 alg: $\mathrm{BR}_a(A, \mathbb{P}^2) \to \mathrm{Hom}_c(\Gamma, \mathrm{PO}(3,\mathbb{C}))$ *is an isomorphism.*

Proof We will only sketch the proof since it follows closely the argument in [6, Theorem 12.14] and the latter is quite long. One verifies that alg induces a natural isomorphism of functors of points. For instance, over the complex numbers, each representation $\rho \in \mathrm{Hom}_c(\Gamma, \mathrm{PO}(3,\mathbb{C}))$ gives rise to an anisotropic realization: $\psi(v) \in \mathbb{P}^2(\mathbb{C})$ is the point fixed by $\rho(g_v)$ (if v is a point-vertex) and $\psi(v) \in \mathbb{P}^2(\mathbb{C})$ is the line fixed by $\rho(g_v)$ (if v is a line-vertex). □

Corollary 5.5
1. S_c *is a cross-section for the action of* G *on* $\mathrm{Hom}_o(\Gamma, G)$.
2. $\mu \circ \mathrm{alg}\colon \mathrm{BR}_a(A, \mathbb{P}^2) \to X_o(\Gamma, G)$ *is an isomorphism.*
3. $S = \mu \circ \mathrm{alg}(\mathrm{BR}_a(A, \mathbb{P}^2)) \subset X(\Gamma, G)$ *is an open subscheme.*

Proof Part 1 follows from the fact that

$$S_c = \mathrm{alg}(\mathrm{BR}_a(A, \mathbb{P}^2(\mathbb{C})) = \mathrm{Hom}_c(\Gamma, \mathrm{PO}(3,\mathbb{C}))$$

and the latter is a cross-section for the G–action on $\mathrm{Hom}_o(\Gamma, G)$ (Lemma 3.14). Part 2 is immediate from Theorem 5.4 and part 1. Part 3 follows from the fact that $X_o(\Gamma, G)$ is an open subscheme in $X(\Gamma, G)$. □

We define $X' := \mathrm{geo}^{-1}(\mathrm{BR}_a(A, \mathbb{P}^2)) \subset X$, an open subscheme in X. The composition of geo, alg and μ

$$X \supset X' \xrightarrow{\mathrm{geo}} \mathrm{BR}_a(A, \mathbb{P}^2) \cap \mathrm{BR}_0(A, \mathbb{P}^2) \xrightarrow{\mathrm{alg}} S'_c \xrightarrow{\mu} S' \subset X_o(\Gamma, G),$$

where $X' \subset X$ and $S' \subset S \subset X_o(\Gamma, G)$ are open subschemes, yields an isomorphism

$$\kappa\colon X' \to S' \subset S.$$

The isomorphism κ sends the point $x \in X'$ to $[\rho_c] \in X_o(\Gamma, G)$. This concludes the proof of Theorem 5.1. □

We let S'_o denote the preimage of S' in $\mathrm{Hom}_o(\Gamma, G)$; then

$$S'_c = S_c \cap S'_o \quad \text{and} \quad S'_o = G \cdot S'_c \cong S'_c \times G.$$

Inverting the isomorphism κ and multiplying with the identity map $G \to G$, we obtain:

Corollary 5.6 *There exists isomorphisms of schemes over* $\mathbb{Q}$

$$\omega\colon S'_o \to X' \times G \quad \textit{and} \quad \kappa\colon S'_c \to X'$$

and a commutative diagram

$$\begin{array}{ccc} S'_o & \xrightarrow{\ \omega\ } & X' \times G \\ \downarrow & & \downarrow \\ S'_c & \xrightarrow{\ \kappa^{-1}\ } & X' \end{array}$$

where the vertical arrows are quotients by the G*–action.*

6 Proof of Theorem 1.1

We continue with notation introduced in the previous sections. Given an affine scheme X over $\mathbb{Q}$ and a rational point $x \in X$, we use Theorem 5.1 to construct a Coxeter group Γ and a representation $\rho_c\colon \Gamma \to \mathrm{PO}(3, \mathbb{R}) < \mathrm{PO}(3, \mathbb{C})$. Then, as in Section 4, we will construct a closed 3–manifold $M = M_\Gamma$ with fundamental group π, and a clopen subscheme $\mathrm{Hom}_o(\pi, G) \subset \mathrm{Hom}(\pi, G)$ which is isomorphic to the product $\mathrm{Hom}_o(\Gamma, G) \times G^k$. In (5) we defined an epimorphism

$$\xi\colon \pi \xrightarrow{\phi} \Gamma \star F_k \xrightarrow{\psi} \Gamma.$$

Set $\rho_0 := \xi^*(\rho_c) \in \mathrm{Hom}(\pi, G)$. The subgroup $\rho_0(\pi) = \rho_c(\Gamma) < G(\mathbb{R})$ is dense according to Theorem 5.1.

We next "convert" the open subscheme $S'_o \subset \mathrm{Hom}_o(\Gamma, G)$ (from the end of the previous section) to an open subscheme $R' \subset \mathrm{Hom}(\pi, G)$. Namely, given S'_o, we let R' be the subscheme as in Lemma 4.4, namely, $\phi^*(S'_o \times G^k)$.

By combining the isomorphism

$$R' \to S'_o \times G^k \subset \mathrm{Hom}(\Gamma, G) \times G^k$$

with the isomorphism

$$\omega \times \mathrm{id}\colon S'_o \times G^k \to X' \times G \times G^k$$

(where ω is from Corollary 5.6), we obtain an isomorphism

$$f\colon R' \xrightarrow{\cong} S'_o \times G^k \xrightarrow{\cong} X' \times G^{k+1} \subset X \times G^{k+1},$$

sending $\rho_0 \in R'$ to

$$x' = x \times 1 \in X' \times G^{k+1}.$$

By the construction, R' is open in $\mathrm{Hom}(\pi, G)$ and $X' \times G^{k+1}$ is open in $X \times G^{k+1}$. The cross-section $S'_c \subset S'_o \subset \mathrm{Hom}_o(\Gamma, G)$ (see Remark 5.2) yields a cross-section $R'_c \subset R' \subset \mathrm{Hom}(\pi, G)$ for the action $G \curvearrowright R'$:

$$R'_c = \phi^*(\psi^*(S'_c) \times G^k).$$

This concludes the proof of Theorem 1.1. □

7 Corollaries of Theorem 1.1

Theorem 1.1 deals with representation schemes of 3–manifold groups to $G = \mathrm{PO}(3)$; we now consider the corresponding character schemes. Since $R'_c \subset \mathrm{Hom}_o(\pi, G)$ is a cross-section for the action of G on R', part 5 of Theorem 1.1 immediately implies Corollary 1.3.

We next consider representations of 3–manifold groups to the group $\widetilde{G} = \mathrm{SL}(2)$; we work over $\mathbb{C}$ and thus identify $\mathrm{PSL}(2, \mathbb{C})$ with $\mathrm{PO}(3, \mathbb{C})$.

Recall that, according to Theorem 5.1 (and Remark 5.2), for every affine scheme X over $\mathbb{Q}$ and a rational point $x \in X$, there exists an open subscheme $X' \subset X$ containing x, a Coxeter group Γ an open subscheme $S'_o \subset \mathrm{Hom}(\Gamma, G)$, and an isomorphism of schemes over $\mathbb{C}$ (which is the identity on the G–factor)

$$S'_o \cong S'_c \times G \cong S' \times G \to X' \times G,$$

sending $\rho_c \in S'_c$ to $x \times 1$. (S'_c is a certain cross-section for the action $G \curvearrowright S'_o$.)

Next, we consider representations of the corresponding extended Coxeter group $\widetilde{\Gamma}$. Proposition 3.9 gives us a G–equivariant regular étale covering

$$q\colon \mathrm{Hom}_o(\widetilde{\Gamma}, \widetilde{G}) \to \mathrm{Hom}_o(\Gamma, G)$$

with covering group $\mathbb{Z}_2^{k+r}$. Restricting to $S'_o \subset \mathrm{Hom}_o(\Gamma, G)$ we obtain a G–equivariant regular étale covering

$$q'\colon \widetilde{S}'_o \to S'_o \qquad \text{where} \quad \widetilde{S}'_o = q^{-1}(S'_o) \subset \mathrm{Hom}_o(\widetilde{\Gamma}, \widetilde{G}) \quad \text{and} \quad q' = q|_{\widetilde{S}'_o}.$$

We let $\widetilde{\rho}_c\colon \widetilde{\Gamma} \to \widetilde{G}(\mathbb{C})$ be a lift of ρ_c. The subscheme $\widetilde{S}'_o$ is open in $\mathrm{Hom}(\widetilde{\Gamma}, \widetilde{G})$.

Next, as in Section 4, given the group $\widetilde{\Gamma}$ we construct a closed 3–manifold $\widetilde{M}$ with fundamental group $\widetilde{\pi}$. We obtain the epimorphism

$$\widetilde{\xi}\colon \widetilde{\pi} \to \widetilde{\Gamma} \times F_k \to \widetilde{\Gamma}$$

and set $\widetilde{\rho}_0 := \widetilde{\xi}^*(\widetilde{\rho}_c) \in \mathrm{Hom}(\widetilde{\pi}, \widetilde{G})$.

We have an étale covering

$$\widehat{q}\colon \mathrm{Hom}_{\widetilde{\Theta}}(\widetilde{\pi}, \widetilde{G}) \to \mathrm{Hom}_{\Theta}(\pi, G),$$

equivariant with respect to the action of G. Restricting to the open subscheme

$$\widetilde{R} = \mathrm{Hom}_o(\widetilde{\pi}, \widetilde{G}) \subset \mathrm{Hom}_\Theta(\widetilde{\pi}, \widetilde{G}) \subset \mathrm{Hom}(\widetilde{\pi}, \widetilde{G}),$$
$$\text{where } \widetilde{R} = \widehat{q}^{-1}(R) \quad \text{and} \quad R = \mathrm{Hom}_o(\pi, G) \subset \mathrm{Hom}_\Theta(\pi, G) \subset \mathrm{Hom}(\pi, G),$$

we obtain the G–equivariant étale covering

$$\widetilde{q}\colon \widetilde{R} \to R$$

with the group $\mathbb{Z}_2^{k+r}$ of covering transformations. According to Section 4 we also have cross-sections

$$R_c \subset R \quad \text{and} \quad \widetilde{R}_c = \widetilde{q}^{-1}(R_c) \subset \widetilde{R}$$

for the G–actions on the schemes. The open subschemes that appear in Corollary 1.4 are smaller; we let

$$\widetilde{R}' = \widetilde{q}^{-1}(R'),$$

where $R' \subset \mathrm{Hom}(\pi, G)$ is the subscheme appearing in Theorem 1.1. The cross-sections for G–actions on these subschemes are $\widetilde{R}'_c$ and R'_c, respectively, where $\widetilde{R}'_c = \widetilde{q}^{-1}(R'_c)$. By the construction $\widetilde{\rho}_0$ belongs to $\widetilde{R}'_c(\mathbb{C})$. This proves Corollary 1.4. □

8 Orbifold-group representations

Let $\widehat{\Gamma}$ be the fundamental group of the hyperbolic orbifold appearing in Theorem 4.1. This group contains cyclic subgroups $\langle \theta_i \rangle \cong \mathbb{Z}_2$ for $i = 1, \dots, 2k$ corresponding to the singular points y_i. The group Γ is the quotient

$$\widehat{\Gamma} / \langle\langle \widehat{\Theta} \rangle\rangle,$$

where $\widehat{\Theta} = \{\theta_1, \dots, \theta_{2k}\} \subset \widehat{\Gamma}$. Then for every algebraic group H,

$$\mathrm{Hom}(\Gamma, H) \cong \mathrm{Hom}_{\widehat{\Theta}}(\widehat{\Gamma}, H),$$

and the latter is an open subscheme in $\mathrm{Hom}(\widehat{\Gamma}, H)$ (see Corollary 2.4). Now, let Γ be a Coxeter group (as in Theorem 5.1) or its canonical central extension. In view of Theorems 4.1 and 5.1, we obtain:

Corollary 8.1 *Theorem 1.1 and Corollaries 1.3 and 1.4 also hold for groups π which are fundamental groups of (nonorientable) 3–dimensional closed hyperbolic orbifolds.*

By passing to torsion-free subgroups of finite index of π, in view of [5, Theorem 5.1], we obtain new examples of fundamental groups of hyperbolic 3–manifolds and their representations to SO(3) and SU(2) with nonquadratic singularities of character varieties; see [5, Theorem 5.1], where it is proven that nonquadratic singularities of character schemes are inherited by finite index subgroups. (The first such examples were constructed in [5].)

Question 8.2 Do Theorem 1.1 and Corollaries 1.3 and 1.4 also hold for groups π which are fundamental groups of 3–dimensional closed hyperbolic manifolds? Do they hold for 3–dimensional manifolds which are 3–dimensional (integer or rational) homology spheres?

Appendix: Functor of points of affine schemes

The material of this section is standard; we include it for the sake of completeness. While the results follow easily from the Yoneda lemma, we will give a direct proof.

Lemma A.1 *Let $f\colon R \to S$ be a homomorphism of commutative rings, such that for every commutative ring A the induced map*

$$f_A^*\colon \mathrm{Hom}(S, A) \to \mathrm{Hom}(R, A)$$

is a bijection. Then f is an isomorphism.

Proof First, we take $A = R$. Since

$$f_R^*\colon \mathrm{Hom}(S, R) \to \mathrm{Hom}(R, R)$$

is a bijection, there exists $g \in \mathrm{Hom}(S, R)$ such that $f_R^*(g) = \mathrm{id}_R$, ie

$$g \circ f = \mathrm{id}_R,$$

the identity map. For general A we have the composition

$$\mathrm{Hom}(R, A) \xrightarrow{g_A^*} \mathrm{Hom}(S, A) \xrightarrow{f_A^*} \mathrm{Hom}(R, A)$$

which satisfies $f_A^* \circ g_A^* = (g \circ f)_A^* = \mathrm{id}$. Therefore, g_A^* is also a bijection.

Now, we take $A = S$. Since g_S^* is a bijection, there exists $h \in \mathrm{Hom}(R, S)$ such that

$$g^*(\mathrm{id}_S) = h,$$

ie $h \circ g = \mathrm{id}_S$. Thus

$$f = \mathrm{id}_S \circ f = h \circ g \circ f = h \circ \mathrm{id}_R = h.$$

Hence, $f = h$ and the equations $g \circ f = \mathrm{id}_R$ and $h \circ g = \mathrm{id}_S$ show that $g = f^{-1}$. □

Suppose now that X and Y are affine schemes of finite type over a field k, ie schemes associated with quotient rings $R = k[x_1, \dots, x_m]/I$ and $Y = k[y_1, \dots, y_n]/J$. For a commutative ring A the sets $X(A)$ of A–points of X and $Y(A)$ of Y are naturally identified with the sets of homomorphisms $\mathrm{Hom}(R, A)$ and $\mathrm{Hom}(S, A)$.

Corollary A.2 *If $\phi\colon Y \to X$ is a morphism which induces isomorphisms of functors of A–points, ie bijections $\phi_A\colon Y(A) \to X(A)$ for all commutative rings A, then ϕ is an isomorphism of schemes.*

Proof Consider the ring homomorphism $f\colon R \to S$ associated with ϕ. The bijections ϕ_A are identified with the bijections

$$f_A^*\colon \mathrm{Hom}(S, A) \to \mathrm{Hom}(R, A).$$

According to Lemma A.1, f is a ring isomorphism. Hence, ϕ is an isomorphism of schemes. □

References

[1] **W M Goldman**, **J J Millson**, *The deformation theory of representations of fundamental groups of compact Kähler manifolds*, Inst. Hautes Études Sci. Publ. Math. 67 (1988) 43–96 MR

[2] **J Huebschmann**, *Singularities and Poisson geometry of certain representation spaces*, from "Quantization of singular symplectic quotients" (N P Landsman, M Pflaum, M Schlichenmaier, editors), Progr. Math. 198, Birkhäuser, Basel (2001) 119–135 MR

[3] **J E Humphreys**, *Reflection groups and Coxeter groups*, corrected 1st edition, Cambridge Studies in Advanced Mathematics 29, Cambridge University Press (1997) MR

[4] **D Johnson**, **J J Millson**, *Deformation spaces associated to compact hyperbolic manifolds*, from "Discrete groups in geometry and analysis" (R Howe, editor), Progr. Math. 67, Birkhäuser, Boston (1987) 48–106 MR

[5] **M Kapovich**, **J J Millson**, *On the deformation theory of representations of fundamental groups of compact hyperbolic* 3*–manifolds*, Topology 35 (1996) 1085–1106 MR

[6] **M Kapovich**, **J J Millson**, *On representation varieties of Artin groups, projective arrangements and the fundamental groups of smooth complex algebraic varieties*, Inst. Hautes Études Sci. Publ. Math. 88 (1998) 5–95 MR

[7] **M Kapovich**, **J J Millson**, *Moduli spaces of linkages and arrangements*, from "Advances in geometry" (J-L Brylinski, R Brylinski, V Nistor, B Tsygan, P Xu, editors), Progr. Math. 172, Birkhäuser, Boston (1999) 237–270 MR

[8] **M Kapovich**, **J J Millson**, *Universality theorems for configuration spaces of planar linkages*, Topology 41 (2002) 1051–1107 MR

[9] **L Lafforgue**, *Chirurgie des grassmanniennes*, CRM Monograph Series 19, Amer. Math. Soc., Providence, RI (2003) MR

[10] **A Lubotzky**, **A R Magid**, *Varieties of representations of finitely generated groups*, Mem. Amer. Math. Soc. 336, Amer. Math. Soc., Providence, RI (1985) MR

[11] **N E Mnëv**, *The universality theorems on the classification problem of configuration varieties and convex polytopes varieties*, from "Topology and geometry—Rohlin Seminar" (O Y Viro, editor), Lecture Notes in Math. 1346, Springer, Berlin (1988) 527–543 MR

[12] **D Panov**, **A Petrunin**, *Telescopic actions*, Geom. Funct. Anal. 22 (2012) 1814–1831 MR

[13] **S Payne**, *Moduli of toric vector bundles*, Compos. Math. 144 (2008) 1199–1213 MR

[14] **I A Rapinchuk**, *On the character varieties of finitely generated groups*, Math. Res. Lett. 22 (2015) 579–604 MR

[15] **J Richter-Gebert**, *Realization spaces of polytopes*, Lecture Notes in Mathematics 1643, Springer, Berlin (1996) MR

[16] **P B Shalen**, *Representations of 3–manifold groups*, from "Handbook of geometric topology" (R J Daverman, R B Sher, editors), North-Holland, Amsterdam (2002) 955–1044 MR

[17] **A S Sikora**, *Character varieties*, Trans. Amer. Math. Soc. 364 (2012) 5173–5208 MR

[18] **R Vakil**, *Murphy's law in algebraic geometry: badly-behaved deformation spaces*, Invent. Math. 164 (2006) 569–590 MR

MK: *Department of Mathematics, University of California, Davis*
1 Shields Ave, Davis, CA 95616, United States

MK: *Korea Institute for Advanced Study*
207-43 Cheongnyangri-dong, Dongdaemun-gu, Seoul, South Korea

JJW: *Department of Mathematics, University of Maryland, College Park*
College Park, MD 20742, United States

kapovich@math.ucdavis.edu, jjm@math.umd.edu

http://www.math.ucdavis.edu/~kapovich,
http://www.math.umd.edu/~millson

Proposed: Walter Neumann Received: 13 September 2013
Seconded: Ciprian Manolescu, Bruce Kleiner Revised: 31 May 2016

Geometry & Topology 21 (2017) 1969–2014

Growth and order of automorphisms of free groups and free Burnside groups

RÉMI COULON
ARNAUD HILION

We prove that an outer automorphism of the free group is exponentially growing if and only if it induces an outer automorphism of infinite order of free Burnside groups with sufficiently large odd exponent.

20E05, 20E36, 20F28, 20F50, 20F65; 68R15

1 Introduction

Let n be an integer. A group G has exponent n if for all $g \in G$, $g^n = 1$. In 1902, W Burnside [7] asked the following question. Is a finitely generated group with finite exponent necessarily finite? In order to study this question, the natural object to look at is the free Burnside group of rank r and exponent n. It is defined to be quotient of the free group $\boldsymbol{F}_r$ of rank r by the (normal) subgroup $\boldsymbol{F}_r^n$ generated by the n^{th} power of all elements. We denote it by $\boldsymbol{B}_r(n)$. Every finitely generated group with finite exponent is a quotient of a free Burnside group.

For a long time, hardly anything was known about free Burnside groups. It was only proved that $\boldsymbol{B}_r(n)$ was finite for some small exponents: $n = 2$, Burnside [7]; $n = 3$, Burnside [7] and Levi and van der Waerden [24]; $n = 4$, Sanov [35]; and $n = 6$, Hall [21]. In 1968, P S Novikov and S I Adian [29; 30; 31] achieved a breakthrough by providing the first examples of infinite Burnside groups. More precisely, they proved the following theorem. Assume that r is at least 2 and n is an odd exponent larger than or equal to 4381; then the free Burnside group of rank r and exponent n is infinite.

This result has been improved in many directions. Adian [1] decreased the bound on the exponent. A Y Ol'shanskiĭ [32] obtained a similar statement using a diagrammatical approach of small cancellation theory. The case of even exponents has been solved by S V Ivanov [22] and I G Lysënok [26]. More recently, T Delzant and M Gromov [17] gave an alternative proof of the infiniteness of Burnside groups. To sharpen our understanding of Burnside groups, we would like to study the symmetries of $\boldsymbol{B}_r(n)$. This leads us to the outer automorphism group of $\boldsymbol{B}_r(n)$.

Published: 19 May 2017 DOI: 10.2140/gt.2017.21.1969

The subgroup F_r^n is characteristic. Hence the projection $F_r \twoheadrightarrow B_r(n)$ induces a natural homomorphism $\mathrm{Out}(F_r) \to \mathrm{Out}(B_r(n))$. This map is neither one-to-one nor onto. However, it provides numerous examples of automorphisms of Burnside groups. For instance, the first author [13] proved that for sufficiently large odd exponents, the image of $\mathrm{Out}(F_r)$ in $\mathrm{Out}(B_r(n))$ contains free subgroups of arbitrary rank and free abelian subgroups of rank $\lfloor \frac{r}{2} \rfloor$. In this article, we are interested in the following question.

Question Which (outer) automorphism of F_r induces an (outer) automorphism of infinite order of $B_r(n)$?

Let G be a finitely generated group endowed with the word-metric. Given $g \in G$, the length $\|g\|$ of its conjugacy class is the length of the smallest word over the generators which represents an element conjugated to g. Given an outer automorphism Φ of G, one says that

- Φ is *exponentially growing* if there exist $g \in G$ and $\lambda > 1$ such that for all integers k, $\|\Phi^k(g)\| \geqslant \lambda^k$,
- Φ is *polynomially growing* if for every $g \in G$, there is a polynomial P such that for all integers k, $\|\Phi^k(g)\| \leqslant P(k)$.

The word-metrics relative to two finite generating sets are bi-Lipschitz equivalent. Therefore, the asymptotic behavior of $\|\Phi^k(g)\|$ does not depend on the choice of generators. Automorphisms of free groups are either exponentially or polynomially growing; see Bestvina and Handel [6] and Bestvina, Feighn and Handel [3]. See also Levitt [25].

We study here the map $\mathrm{Out}(F_r) \to \mathrm{Out}(B_r(n))$. Our main theorem states that an automorphism of F_r has exponential growth if and only if it induces an automorphism of infinite order of $B_r(n)$ for sufficiently large exponents n. From the viewpoint of $\mathrm{Out}(F_r)$, this result provides an unexpected characterization of the growth of automorphisms of free groups. At the level of Burnside groups, it completely describes the automorphisms of F_r that induce automorphisms of infinite order of some Burnside groups.

Remark 1.1 Since $B_r(n)$ is a torsion group, every inner automorphism of $B_r(n)$ has finite order. Therefore, an automorphism $\varphi \in \mathrm{Aut}(B_r(n))$ has finite order if and only if its outer class does also. Hence, for our purpose, we can equivalently work with $\mathrm{Out}(B_r(n))$ or $\mathrm{Aut}(B_r(n))$.

The first examples of automorphisms of $B_r(n)$ with infinite order were given by E A Cherepanov [8]. In particular, he proved that the automorphism φ of $F(a,b)$ given by $\varphi(a) = ab$ and $\varphi(b) = a$ (also called the Fibonacci morphism) induces an automorphism of infinite order of $B_2(n)$ for all odd integers $n \geqslant 665$. In [13], the first author provides a large class of automorphisms with the same property.

Theorem 1.2 (Coulon [13, Theorem 1.3]) *Let φ be a hyperbolic automorphism of $\boldsymbol{F}_r$ (ie the semidirect product $\boldsymbol{F}_r \rtimes_\varphi \mathbb{Z}$ is word-hyperbolic). There exists an integer n_0 such that for all odd exponents $n \geqslant n_0$, the automorphism φ induces an element of infinite order of $\mathrm{Out}(\boldsymbol{B}_r(n))$.*

The Fibonacci morphism φ used by Cherepanov is not hyperbolic. Indeed φ fixes the commutator $[a,b] = aba^{-1}b^{-1}$. Hence the semidirect product $\boldsymbol{F}_2 \rtimes_\varphi \mathbb{Z}$ contains a copy of $\mathbb{Z}^2$ which is an obstruction to being hyperbolic. This observation has a more general topological interpretation. Indeed, any automorphism φ of $\boldsymbol{F}_2$ can be represented by a homeomorphism f of the punctured torus (if φ is the Fibonacci morphism, then f is even pseudo-Anosov). This map f necessarily preserves the boundary component of the torus — which corresponds to the commutator $[a,b]$. Hence the mapping torus induced by f contains an embedded torus. Its fundamental group $\boldsymbol{F}_2 \rtimes_\varphi \mathbb{Z}$ is therefore not hyperbolic.

Nevertheless, like hyperbolic automorphisms, the Fibonacci morphism is exponentially growing. On the other hand, we also know that a polynomially growing automorphism of $\boldsymbol{F}_r$ induces an automorphism of finite order of $\boldsymbol{B}_r(n)$ for every exponent n [13]. It suggests a link between the growth of an automorphism of $\boldsymbol{F}_r$ and its order as automorphism of $\boldsymbol{B}_r(n)$. More precisely, we prove the following statement.

Theorem 1.3 *Let $\Phi \in \mathrm{Out}(\boldsymbol{F}_r)$ be an outer automorphism of $\boldsymbol{F}_r$. The following assertions are equivalent:*

(1) *Φ has exponential growth;*

(2) *there exists $n \in \mathbb{N}$ such that Φ induces an outer automorphism of $\boldsymbol{B}_r(n)$ of infinite order;*

(3) *there exist $\kappa, n_0 \in \mathbb{N}$ such that for all odd integers $n \geqslant n_0$, the automorphism Φ induces an outer automorphism of $\boldsymbol{B}_r(\kappa n)$ of infinite order.*

Remark In this article, we adopt the following convention. The notation $\mathbb{N}$ stands for the set of nonnegative integers, whereas $\mathbb{N}^*$ represents $\mathbb{N} \setminus \{0\}$.

In the statement of Theorem 1.3, (3) $\Longrightarrow$ (2) is easy whereas (2) $\Longrightarrow$ (1) follows from the work of the first author [13, Theorem 1.6]. The new result of this article is the implication (1) $\Longrightarrow$ (3). Before sketching this proof, let us have a look at the arguments used by Cherepanov [8]. The proof of the infiniteness of $\boldsymbol{B}_r(n)$ by Novikov and Adian is based on the following important fact [1].

Proposition 1.4 *Let w be a reduced word of $\boldsymbol{F}_r$. If w does not contain a subword of the form u^{16}, then w induces a nontrivial element of $\boldsymbol{B}_r(n)$ for all odd exponents $n \geqslant 655$.*

In particular, two distinct reduced words without an 8^{th} power define distinct elements of $\boldsymbol{B}_r(n)$. Compute now the orbit of b under the automorphism φ of $\boldsymbol{F}(a,b)$ defined by $\varphi(a) = ab$ and $\varphi(b) = a$. It leads to the following sequence of words:

$$\begin{aligned} &\varphi^1(b) = a, && \varphi^5(b) = abaababa, \\ &\varphi^2(b) = ab, && \varphi^6(b) = abaababaabaab, \\ &\varphi^3(b) = aba, && \varphi^7(b) = abaababaabaababaababa, \\ &\varphi^4(b) = abaab, && \quad\vdots \end{aligned}$$

None of these words contains a 4^{th} power; see Karhumäki [23]. Therefore, they induce pairwise distinct elements of $\boldsymbol{B}_r(n)$. In particular, φ seen as an automorphism of $\boldsymbol{B}_r(n)$, has infinite order.

This argument can be generalized for any exponentially growing automorphism of $\boldsymbol{F}_2$ using an appropriate train track representative. However, it does not work anymore in higher rank. Consider, for instance, the exponentially growing automorphism ψ of $\boldsymbol{F}(a,b,c,d)$ defined by $\psi(a) = a$, $\psi(b) = ba$, $\psi(c) = cbcd$ and $\psi(d) = c$. As previously, we compute the orbit of d under ψ:

$$\begin{aligned} &\psi^1(d) = c, \\ &\psi^2(d) = c\mathbf{b}cd, \\ &\psi^3(d) = cbcd\mathbf{ba}cbcdc, \\ &\psi^4(d) = cbcdbacbcdc\mathbf{ba}^2cbcdbacbcdccbcd, \\ &\psi^5(d) = cbcdbacbcdcba^2cbcdbacbcdc^2bcd\mathbf{ba}^3cbcdbacbcdcba^2 \ldots \\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad \ldots cbcdbacbcdc^2bcdcbcdbacbcdc. \end{aligned}$$

This orbit is exponentially growing. Note that if $\psi^p(d)$ contains a subword ba^m then $\psi^{p+1}(d)$ contains ba^{m+1}. Hence as p tends to infinity, $\psi^p(d)$ contains arbitrarily large powers of a. This cannot be avoided by choosing the orbit of another element. Proposition 1.4 is no more sufficient to tell us whether or not the $\psi^p(d)$ are pairwise distinct in $\boldsymbol{B}_r(n)$. Therefore, we need a more accurate criterion to distinguish two different elements of $\boldsymbol{B}_r(n)$. This is done using elementary moves.

Let $n \in \mathbb{N}$ and $\xi \in \mathbb{R}_+$. An (n,ξ)*–elementary move* consists in replacing a reduced word of the form $pu^m s \in \boldsymbol{F}_r$ by the reduced representative of $pu^{m-n}s$, provided m is an integer larger than $\frac{n}{2} - \xi$. The word u is called the *support* of the elementary move. Note that an elementary move may increase the length of the word.

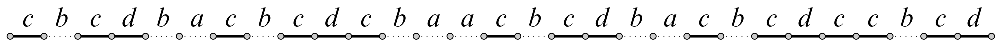


Figure 1: The yellow-red decomposition of $\psi^4(d)$

Theorem 1.5 (Coulon [12]) *There exist integers n_1 and ξ such that for all odd exponents $n \geqslant n_1$, we have the following property. Let w and w' be two reduced words of $\boldsymbol{F}_r$. If w and w' define the same element of $\boldsymbol{B}_r(n)$, then there are two sequences of (n,ξ)–elementary moves which respectively send w and w' to the same word.*

Remark As will be detailed in Section 6.1, this statement is a direct application of the main theorem of Coulon [12]. Its proof relies on the geometric approach of the Burnside problem developed by Delzant and Gromov [17]. Although Theorem 1.5 is not explicitly mentioned in their articles, it should be possible to deduce an analogue statement from the work of Adian [1] and Ol'shanskiĭ [32]. For the convenience of the reader who would be more familiar with Ol'shanskiĭ's techniques, these analogies and differences are discussed in Section 6.1 and in the Appendix.

Thanks to this tool, we can now explain using the example ψ how the implication (1) $\Longrightarrow$ (3) of Theorem 1.3 works. We need to understand the effect of elementary moves on a word $\psi^p(d)$. To that end, we assign colors to the letters. Let us say that a and b are *yellow* letters (dotted lines on Figure 1) whereas c and d are *red* letters (thick lines on the figure). The word $\psi^p(d)$ is the concatenation of maximal yellow and red subwords. To any word w over the alphabet $\{a,b,c,d\}$ we associate its red part $\mathrm{Red}(w)$ obtained by removing from w all the yellow letters. We start with two observations, one on the red words, the other on the yellow ones.

Red words We claim that the support of elementary moves that can be performed on $\psi^p(d)$ only contains yellow letters. Since the orbit of d grows exponentially, one can prove that $\mathrm{Red}(\psi^p(d))$ does not contain large powers. More precisely, there is an integer n_2 such that for all $p \in \mathbb{N}$, the word $\mathrm{Red}(\psi^p(d))$ does not contain any n_2^{th} power; see Proposition 5.11. This fact can be interpreted in terms of dynamical properties of the attracting laminations associated to the automorphism ψ. Let $n > 2n_2 + 2\xi$. Assume now that the support u of an (n,ξ)–elementary move performed on $\psi^p(d)$ contains a red letter. By definition, there exists $m > n_2$ such that u^m is a subword of $\psi^p(d)$. In particular, $\mathrm{Red}(u)^m$ is a subword of $\mathrm{Red}(\psi^p(d))$, which contradicts the definition of n_2. It follows from this remark that the support of any (n,ξ)–elementary move with $n > 2n_2 + 2\xi$ only contains yellow letters.

Yellow words We now claim that elementary moves with yellow support cannot send a maximal yellow subword of $\psi^p(d)$ to the empty word. This fact is important for the

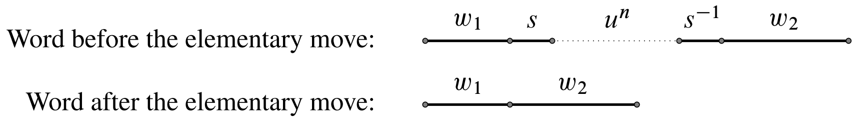


Figure 2: An elementary move collapsing red letters

following reason. We explained that the support of an elementary move performed on $\psi^p(d)$ only contains yellow letters. Such a move could change the red part of $\psi^p(d)$, though. It could indeed completely collapse a maximal yellow subword and thus affect the red letters; see Figure 2.

To prove this second claim, we look at the yellow subwords of $\psi^p(d)$. Notice that the image by ψ of a yellow word is still a yellow word. On the contrary, the image of a red word may contain yellow subwords. Indeed b is a subword of $\psi(c)$. Actually, the yellow subwords of $\psi^p(d)$ can be sorted in two categories: the words that consist in the single letter b which appear as a subword of $\psi(c)$ and the ones which arise as the images by ψ of yellow subwords of $\psi^{p-1}(d)$. In particular, all the maximal yellow subwords of $\psi^p(d)$ belong to the orbit under ψ of b. Consequently, there is an integer n_3 such that for every odd integer $n > n_3$, none of them becomes trivial in $\boldsymbol{B}_r(n)$. In particular, no sequence of (n,ξ)–elementary moves sends a maximal yellow subword of $\psi^p(d)$ to the empty word.

We can now argue by contradiction. Let $n > \max\{n_1, 2n_2 + 2\xi, n_3\}$ be an odd integer. Assume that ψ induces an automorphism of finite order of $\boldsymbol{B}_r(n)$. In particular, there exists $p \in \mathbb{N}^*$ such that $\psi^p(d)$ and d have the same image in $\boldsymbol{B}_r(n)$. It follows from Theorem 1.5 that a sequence of (n,ξ)–elementary moves sends $\psi^p(d)$ to d. We claim that performing (n,ξ)–elementary moves on $\psi^p(d)$ does not change its red part. Indeed, $n > 2n_2 + 2\xi$; thus these moves will only change the yellow subwords of $\psi^p(d)$. Moreover, since $n > n_3$, none of the yellow words can completely disappear. In particular, the red word $\mathrm{Red}(\psi^p(d))$ associated to $\psi^p(d)$ should be exactly d. This is a contradiction.

The proof for an arbitrary exponentially growing automorphism of $\boldsymbol{F}_r$ follows the same ideas. One has to replace the words in a, b, c, d by paths in an appropriate relative train track. This leads to a technical difficulty, though. The red and yellow paths that we want to consider do not necessarily represent elements of the free groups. This problem is handled in Sections 4.2 and 5. There we use subtle aspects of the machinery of train-tracks to show that the red words do not contain large powers (Proposition 5.11). In particular, we need to pass to a finite-index subgroup of $\boldsymbol{F}_r$. This operation actually ensures at the same time that no yellow subpath will be removed by elementary moves (see the prior discussion). Beside this fact, the main ingredients are the ones described above.

Acknowledgments Most of this work was done while Coulon was staying at the *Max-Planck-Institut für Mathematik*, Bonn, Germany. He wishes to express his gratitude to all faculty and staff from the MPIM for their support and warm hospitality. Hilion would like to thank Michael Handel and Gilbert Levitt for helpful conversations. Many thanks also go to the referees for their helpful comments and corrections. Hilion is supported by the grant ANR-10-JCJC 01010 of the Agence Nationale de la Recherche.

2 Primitive matrices and substitutions

In this section, we summarize a few properties about primitive integer matrices and substitutions on an alphabet that will be useful later.

2.1 Primitive matrices

A square matrix M of size ℓ whose entries are nonnegative integers is *irreducible* if for each $i, j \in \{1, \dots, \ell\}$, there exists $p \in \mathbb{N}$ such that the (i, j)–entry of M^p is not zero. It is *primitive* when there exists $p \in \mathbb{N}$ such that any entry of M^p is not zero.

The Perron–Frobenius theorem for an irreducible matrix M with nonnegative integer entries states that there exists a unique dominant eigenvalue $\lambda \geqslant 1$ of M associated to an eigenvector with positive coordinates (see for instance Seneta's book [36]). This λ is called the *Perron–Frobenius-eigenvalue* (or simply *PF-eigenvalue*) of M. In addition, if $\lambda = 1$, then M is a transitive permutation matrix.

2.2 Primitive substitutions

Let $\mathcal{A} = \{a_1, \dots, a_\ell\}$ be a finite alphabet. The free monoid generated by $\mathcal{A}$ is denoted by $\mathcal{A}^*$. We write 1 for the empty word, also called the *trivial word*. An infinite word is an element of $\mathcal{A}^{\mathbb{N}}$. Let $m \in \mathbb{N}^*$. A word $w \in \mathcal{A}^*$ is an m^{th} *power* if there exists a nontrivial word $u \in \mathcal{A}^*$ such that $w = u^m$. A nontrivial word $w \in \mathcal{A}^*$ is *primitive* if it is not an m^{th} power with m at least 2 (ie if $w = u^m$, then $u = w$ and $m = 1$). A word $w \in \mathcal{A}^*$ (or an infinite word $w \in \mathcal{A}^{\mathbb{N}}$) *contains an* m^{th} *power*, if there exists a word $u \in \mathcal{A}^* \setminus \{1\}$ such that u^m is a subword of w. The *shift* is the map $S\colon \mathcal{A}^{\mathbb{N}} \to \mathcal{A}^{\mathbb{N}}$ which sends $(w_i)_{i \in \mathbb{N}}$ to $(w_{i+1})_{i \in \mathbb{N}}$. An infinite word w is said to be *shift-periodic* if there exists $q \in \mathbb{N}^*$ such that $S^q(w) = w$. If u stands for the word $w_0 w_1 \cdots w_{q-1}$, then we write $w = u^\infty$. Roughly speaking, it means that w is the infinite power of u.

An endomorphism of the free monoid $\mathcal{A}^*$ is called a *substitution* defined on $\mathcal{A}$. Such a substitution σ is indeed completely determined by the images $\sigma(a) \in \mathcal{A}^*$ of all the letters $a \in \mathcal{A}$. Moreover, it naturally extends to a map $\mathcal{A}^{\mathbb{N}} \to \mathcal{A}^{\mathbb{N}}$. The *matrix* M

of a substitution σ is a square matrix of size ℓ whose (i,j)–entry is the number of occurrences of the letter a_i in the word $\sigma(a_j)$. The substitution σ is said to be *primitive* when M is primitive.

Proposition 2.1 *Let a be a letter of $\mathcal{A}$. Let σ be a primitive substitution on $\mathcal{A}$ such that a is a prefix of $\sigma(a)$.*

(i) *The sequence $(\sigma^p(a))$ converges for the prefix topology to an infinite word $\sigma^\infty(a)$ fixed by σ.*
(ii) *If $\sigma^\infty(a)$ is not shift-periodic, then there exists an integer $m \geqslant 2$ such that for all $p \in \mathbb{N}$, the word $\sigma^p(a)$ does not contain an m^{th} power.*
(iii) *If there exists a nontrivial primitive word u such that $\sigma^\infty(a) = u^\infty$, then there exists an integer $q \geqslant 2$ such that $\sigma(u) = u^q$.*

Remark 2.2 The case covered by Proposition 2.1(iii) is not vacuous. Consider for instance the substitution defined on $\mathcal{A} = \{a,b,c\}$ by $\sigma(a) = ab$, $\sigma(b) = c$ and $\sigma(c) = abc$. The transition matrix M of σ and its square are

$$M = \begin{pmatrix} 1 & 0 & 1 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix}, \quad M^2 = \begin{pmatrix} 1 & 1 & 2 \\ 1 & 1 & 2 \\ 1 & 1 & 2 \end{pmatrix}.$$

In particular, σ is primitive. However, $(\sigma^n(a))$ converges to the infinite shift-periodic word $(abc)^\infty$.

To prove Proposition 2.1, we use the following results due to B Mossé.

Proposition 2.3 (Mossé [27, Théorème 2.4]) *Let σ be a primitive substitution on a finite alphabet $\mathcal{A}$. Let $u \in \mathcal{A}^{\mathbb{N}}$ be an infinite word fixed by σ. Then either*

(i) *u is shift-periodic, or*
(ii) *there exists an integer $m \geqslant 2$ such that u does not contain an m^{th} power.* □

Lemma 2.4 (Mossé [27, Proposition 2.3]) *Let $u \in \mathcal{A}^*$ be a primitive word. Let $m \geqslant 2$ be an integer. If uwu is subword of u^m, then there exists an integer $p \geqslant 0$ such that $w = u^p$.* □

Proof of Proposition 2.1 By assumption, a is a prefix of $\sigma(a)$. Thus there exists $w \in \mathcal{A}^*$ such that $\sigma(a) = aw$. Since σ is primitive, w is not trivial. For every $p \in \mathbb{N}^*$, it follows that $\sigma^p(a)$ is exactly the word

$$\sigma^p(a) = aw\sigma(w)\sigma^2(w)\cdots\sigma^{p-1}(w).$$

In particular, $\sigma^p(a)$ is a prefix of $\sigma^{p+1}(a)$. Therefore, $(\sigma^p(a))$ converges to an infinite word $\sigma^\infty(a)$ fixed by σ:

$$\sigma^\infty(a) = aw\sigma(w)\sigma^2(w)\cdots\sigma^p(w)\cdots,$$

which proves (i). Assume now that this infinite word is not shift-periodic. According to Proposition 2.3, there exists $m \geqslant 2$ such that $\sigma^\infty(a)$ does not contain an m^{th} power. The same holds for the prefixes of $\sigma^\infty(a)$, in particular for all $\sigma^p(a)$, which proves (ii).

Finally, assume that $\sigma^\infty(a) = u^\infty$, where u is a nontrivial primitive word. Since $\sigma^\infty(a)$ is fixed by σ, we obtain that $u^\infty = \sigma(u)^\infty$. In particular, u is a prefix of $\sigma^\infty(u)$. The substitution σ being primitive, $\sigma(u)$ is not shorter than u. We derive that there exists $w_0 \in \mathcal{A}^*$ such that $\sigma(u) = uw_0$. Hence $u^\infty = (uw_0)^\infty$. Lemma 2.4 shows that there exists $p \in \mathbb{N}$ satisfying $w_0 = u^p$. Thus $\sigma(u) = u^{p+1}$. Remember that a is a prefix of u. Hence u can be written $u = au'$. It follows that the length of $\sigma(u) = aw\sigma(u')$ is larger than that of u. Thus $p+1 \geqslant 2$, which proves (iii). □

3 Train-tracks and automorphisms of free groups

In this section, we recollect some material about relative train-track maps. Details can be found in [6], where they have been introduced by Bestvina and Handel. There exist several improvements of relative train-track maps, and we will use here (very few of the) improved relative train-track maps introduced by Bestvina, Feighn and Handel in [4].

3.1 Paths and circuits

The graphs that we consider are metric graphs with oriented edges. By metric graph, we mean a graph equipped with a path metric. If e is an edge of a graph G, then e^{-1} stands for the edge with the reverse orientation. The pair $\{e, e^{-1}\}$ is the *unoriented edge associated to* e (or e^{-1}). By abuse of notation, we will just say the *unoriented edge* e for the pair $\{e, e^{-1}\}$. Let $\Theta\colon \mathcal{E} \to \mathcal{E}$ be the map defined by $\Theta(e) = e^{-1}$. Sometimes, it will be useful to consider a subset $\vec{\mathcal{E}}$ of $\mathcal{E}$ such that $\vec{\mathcal{E}}$ and $\Theta(\vec{\mathcal{E}})$ give rise to a partition of $\mathcal{E}$ (ie we choose a preferred oriented edge for each unoriented edge). We call such a set $\vec{\mathcal{E}}$ a *preferred set of oriented edges for* G.

A *path* in a graph G is a continuous locally injective map $\alpha\colon I \to G$, where $I = [a, b]$ is a segment of $\mathbb{R}$. The *initial point* of α is $\alpha(a)$ and its *terminal point* is $\alpha(b)$; both $\alpha(a)$ and $\alpha(b)$ are the *endpoints* of α. We do not make any distinction between two paths which differ by an orientation-preserving homeomorphism between their domains.

A path is *trivial* if its domain is a point. When the endpoints of α are vertices, α can be viewed as a *path of edges*, ie a concatenation of edges $\alpha = e_1 \cdots e_p$, where the e_i are edges of G such that the terminal vertex of e_i is the initial vertex of e_{i+1} and $e_i \neq e_{i+1}^{-1}$. A *circuit* in G is a continuous locally injective map of an oriented circle into G. We do not make any distinction between two circuits which differ by an orientation-preserving homeomorphism between their domains. A circuit can be viewed as a cyclically ordered sequence of edges without backtracking. If α is a path or a circuit, we denote by α^{-1} the path or circuit, with the reverse orientation.

A continuous map $\alpha\colon I \to G$, where I is segment in $\mathbb{R}$, is homotopic relative to the endpoints to a unique path denoted by $[\alpha]$. A nonhomotopically trivial continuous map $\alpha\colon S^1 \to G$ is homotopic to a unique circuit denoted by $[\alpha]$.

3.2 Topological representatives

Marked graphs and topological representatives Let $r \geqslant 2$. We denote by R_r the *rose* of rank r. It is a graph with one vertex $\star$ and r unoriented edges. The fundamental group $\pi_1(R_r, \star)$ is the free group $\boldsymbol{F}_r$, with basis given by a preferred set of oriented edges. A *marked graph* (G, τ) (often simply denoted by G) is a connected metric graph G having no vertex of valence 1, equipped with a homotopy equivalence $\tau\colon R_r \to G$. This homotopy equivalence τ gives an identification of the fundamental group $\pi_1(G, \tau(\star))$ with $\boldsymbol{F}_r$, well defined up to an inner automorphism. A *topological representative* of an outer automorphism $\Phi \in \mathrm{Out}(\boldsymbol{F}_r)$ is a homotopy equivalence $f\colon G \to G$ of a marked graph (G, τ) such that

- f takes vertices to vertices and edges to paths of edges,
- $\tau^- \circ f \circ \tau\colon R_r \to R_r$ induces Φ on $\boldsymbol{F}_r = \pi_1(R_r, \star)$, where τ^- is a homotopy inverse of τ.

In particular, the restriction of f to an open edge is locally injective.

Induced map on paths and circuits If α is a path or a circuit in G, one defines $f_\#(\alpha)$ as being equal to $[f(\alpha)]$.

Legal turns For any edge e of G, we let $Df(e)$ denote the first edge of $f(e)$. A *turn* is a pair of edges (e_1, e_2) of G which have the same initial vertex. The turn (e_1, e_2) is *degenerate* if $e_1 = e_2$, and nondegenerate otherwise. A turn (e_1, e_2) is *legal* if $((Df)^p(e_1), (Df)^p(e_2))$ is nondegenerate for all $p \in \mathbb{N}$; otherwise, the turn is *illegal*.

3.3 Lifts

Let $f\colon G \to G$ be a topological representative of $\Phi \in \mathrm{Out}(\boldsymbol{F}_r)$. Let $\widetilde{G}$ be the universal cover of G. The theory of covering spaces gives a one-to-one correspondence between the set of the lifts of f to $\widetilde{G}$ and the set of automorphisms in the outer class Φ. More precisely, a lift $\widetilde{f}$ of f is in correspondence with the automorphism $\varphi \in \Phi$ if

$$\widetilde{f} \circ g = \varphi(g) \circ \widetilde{f} \quad \text{for all } g \in \boldsymbol{F}_r, \tag{1}$$

where the elements of $\boldsymbol{F}_r$ are viewed as deck transformations of $\widetilde{G}$.

3.4 Invariant filtrations and transition matrices

Let $f\colon G \to G$ be a topological representative of $\Phi \in \mathrm{Out}(\boldsymbol{F}_r)$.

Filtration, strata and k–legal paths A *filtration* of a topological representative $f\colon G \to G$ is a strictly increasing sequence of f–invariant subgraphs $\varnothing = G_0 \subset G_1 \subset \cdots \subset G_m = G$. The *stratum of height* k denoted by H_k is the closure of $G_k \setminus G_{k-1}$. The edges *of height* k are the edges of H_k. A path *of height* k is a path in G_k which crosses H_k nontrivially; ie its intersection with H_k contains a nontrivial path. A path (of edges) α is k*–legal* if it is a path of G_k and for all subpaths e_1e_2 of α with e_1, e_2 edges of height k, the turn (e_1^{-1}, e_2) is legal.

Transition matrices A *transition matrix* M_k is associated to the stratum H_k. We choose a preferred set of oriented edges $\vec{\mathcal{E}} = \{e_1, \dots, e_\ell\}$ for H_k (where ℓ is the number of unoriented edges of H_k). The transition matrix M_k of H_k is a square matrix of size ℓ whose (i, j)–entry is the number of times the edge e_i or the reverse edge e_i^{-1} occur in the path $f(e_j)$.

The stratum H_k is *irreducible* when its transition matrix M_k is irreducible. Let λ_k be the PF-eigenvalue of M_k; see Section 2.1. If $\lambda_k > 1$, then H_k is called an *exponential stratum*. If M_k is primitive, H_k is said to be *aperiodic*. When the stratum H_k is irreducible and $\lambda_k = 1$, H_k is called a *nonexponential stratum*. When M_k is the zero matrix, the stratum H_k is called a *zero stratum*.

Remark 3.1 Given a topological representative $f\colon G \to G$ and an invariant filtration for f, up to refining the filtration, one can always suppose that any stratum is of one of three possible types: exponential, nonexponential or zero. Moreover, up to replacing Φ by a positive power of Φ, one can assume that Φ admits a topological representative $f\colon G \to G$ with the following properties [4]:

- each exponential stratum is aperiodic,
- each nonexponential stratum H_k consists of a single edge e, and that $f(e) = eu$ where u is loop in G_{k-1} based at the endpoint of e.

3.5 A quick review on relative train-track maps

Relative train-track maps A topological representative $f\colon G \to G$ of an outer automorphism $\Phi \in \mathrm{Out}(\boldsymbol{F}_r)$ with a filtration $\varnothing = G_0 \subset G_1 \subset \cdots \subset G_m = G$ is a *relative train-track map* (RTT) if for every exponential stratum H_k,

(RTT-i) Df maps the set of edges of height k to itself (in particular, each turn consisting of an edge of height k and one of height less than k is legal);

(RTT-ii) if α is a nontrivial path with endpoints in $H_k \cap G_{k-1}$, then $f_\#(\alpha)$ is a nontrivial path with endpoints in $H_k \cap G_{k-1}$;

(RTT-iii) for each k–legal path α, the path $f_\#(\alpha)$ is k–legal.

In particular, an edge e of an exponential stratum H_k is k–legal. Theorem 5.12 in [6] ensures that any outer automorphism Φ of $\boldsymbol{F}_r$ can be represented by an RTT f. By replacing Φ by a positive power of Φ if necessary, one can suppose that Φ satisfies Remark 3.1. In addition, we can ask that all the images of vertices are fixed by f (see Theorem 5.1.5 in [4]). We sum up these facts in the following theorem.

Theorem 3.2 (Bestvina and Handel [6], Bestvina, Feighn and Handel [4]) *Let Φ be an outer automorphism of $\boldsymbol{F}_r$. There exists $p \geqslant 1$ such that Φ^p has a topological representative $f\colon G \to G$ which is an RTT, with the properties that*

- *for all vertices v of G, $f(v)$ is fixed by f,*
- *every exponential stratum of f is aperiodic,*
- *each nonexponential stratum H_k consists of a single edge e, and that $f(e) = eu$, where u is loop in G_{k-1} based at the endpoint of e.*

Splittings Let $f\colon G \to G$ be a topological representative. A *splitting* of a path or a circuit α is a decomposition of α as a concatenation of subpaths $\alpha = \alpha_1\alpha_2\cdots\alpha_q$ (with $q \geqslant 1$ if α is a circuit, and $q \geqslant 2$ if α is a path) such that for all $p \geqslant 0$, $f_\#^p(\alpha) = f_\#^p(\alpha_1) f_\#^p(\alpha_2) \cdots f_\#^p(\alpha_q)$. In that case, one writes $\alpha = \alpha_1 \cdot \alpha_2 \cdots \alpha_q$, and $\alpha_1, \alpha_2, \dots, \alpha_q$ are called the *terms* of the splitting. A basic, but important, property of RTT is given by the following lemma.

Lemma 3.3 (Bestvina and Handel [6, Lemma 5.8]) *Let $f\colon G \to G$ be an RTT. If H_k is an exponential stratum, and if α is a k–legal path, then the decomposition of α as maximal subpaths in H_k or in G_{k-1} is a splitting*

$$\alpha = \alpha_1 \cdot \beta_1 \cdot \alpha_2 \cdots \alpha_{q-1} \cdot \beta_{q-1} \cdot \alpha_q,$$

where the α_i are paths in H_k and the β_i are paths in G_{k-1}, all nontrivial (except possibly α_1 and α_q).

3.6 Growth of automorphisms of free groups

As explained in [3, page 219], the growth of an outer automorphism $\Phi \in \mathrm{Out}(\boldsymbol{F}_r)$ can be detected on an RTT representative [6]. For our purpose, we will use the following observations.

Remark 3.4 [6; 3] Let $\Phi \in \mathrm{Out}(\boldsymbol{F}_r)$.

(1) Φ has either polynomial growth or exponential growth.

(2) Moreover, Φ has exponential growth if and only if one (hence any) RTT $f\colon G \to G$ representing Φ has at least one exponential stratum.

A detailed discussion about the growth of a conjugacy class under iteration of an outer automorphism can be found in [25].

4 Reductions of Theorem 1.3

In this section, we explain how to reduce our main theorem to an easier statement. First, note that given an outer automorphism Φ of the free group, Φ has exponential (polynomial) growth if and only if for every $p \in \mathbb{N}^*$, so does Φ^p. In particular, to prove Theorem 1.3, Φ can be replaced by some positive power of Φ. It will be advantageous to do so, since it allows us to use relative train-track maps with better properties; see Theorem 3.2. We now discuss three reductions.

(1) The first focuses on polynomially growing automorphisms of $\boldsymbol{F}_r$; see Section 4.1. We explain that such an automorphism always induces a finite-order automorphism of Burnside groups. Thus it is sufficient to look at exponentially growing automorphisms.

(2) The second reduction is rather technical. Let $f\colon G \to G$ be an RTT of an exponentially growing automorphism Φ of $\boldsymbol{F}_r$ and H an exponential stratum. The image under f of an edge e in H consists of edges of H and paths contained in lower strata. Later we will need that for every $p \in \mathbb{N}$, maximal subpaths of $f_\sharp^p(e)$ contained in the lower strata are not loops. In Section 4.2, we show that up to passing to a finite-index subgroup, we can always assume that our RTT satisfies this property.

(3) The RTT of an exponentially growing automorphism may contain several exponential strata. In Section 4.3, we prove that it is sufficient to consider automorphisms whose RTT has only one exponential stratum, which is also the top one.

4.1 Polynomially growing automorphisms

Arguing by induction on the rank r of $\boldsymbol{F}_r$, the first author handled the case of polynomially growing automorphisms.

Proposition 4.1 (Coulon [13, Theorem 1.6]) *If $\Phi \in \mathrm{Out}(\boldsymbol{F}_r)$ is polynomially growing, then Φ induces an outer automorphism of finite order of $\boldsymbol{B}_r(n)$ for all positive integers n.*

Remark 4.2 The same proof actually gives a quantitative bound for the order of Φ in $\mathrm{Out}(\boldsymbol{B}_r(n))$. If Φ is an outer polynomially growing automorphism of $\boldsymbol{F}_r$, then $\Phi^{p(r,n)}$ induces a trivial outer automorphism of $\boldsymbol{B}_r(n)$, where

$$p(r,n) = n^{2(2^{r-1}-1)}.$$

Example 4.3 A particular case of polynomially growing automorphisms is given by the automorphisms of $\boldsymbol{F}_2$ induced by a Dehn-twist on a punctured torus. For instance, the automorphism φ defined by $\varphi(a) = a$ and $\varphi(b) = ba$. Here φ^n is trivial in $\mathrm{Aut}(\boldsymbol{B}_r(n))$.

In view of Remark 3.4 (1) and Proposition 4.1, we see that Theorem 1.3 is a consequence of the following proposition.

Proposition 4.4 *If $\Phi \in \mathrm{Out}(\boldsymbol{F}_r)$ has exponential growth, then there exist $\kappa, n_0 \in \mathbb{N}$ such that for all odd integers $n \geqslant n_0$, the automorphism Φ induces an outer automorphism of $\boldsymbol{B}_r(\kappa n)$ of infinite order.*

In the next section, we discuss a second reduction and prove that Proposition 4.4 is a consequence of Proposition 4.8.

4.2 Passing to a finite-index subgroup

Let Φ be an exponentially growing outer automorphism of $\boldsymbol{F}_r$. By replacing Φ by a power of Φ if necessary, we can assume that Φ is represented by an RTT $f\colon G \to G$ with a filtration $\varnothing = G_0 \subset G_1 \subset \cdots \subset G_m = G$, satisfying the properties of Theorem 3.2. We denote by H_k the stratum of height k. Let e be an edge of an exponential stratum H_k. According to Lemma 3.3, $f(e)$ can be split as follows:

$$f(e) = \alpha_1 \cdot \beta_1 \cdot \alpha_2 \cdots \alpha_{q-1} \cdot \beta_{q-1} \cdot \alpha_q,$$

where the α_i are nontrivial paths contained in H_k and the β_i are nontrivial paths contained in G_{k-1}. We denote by $\mathcal{P}_e$ the set $\{\beta_1, \dots, \beta_{q-1}\}$. Let $\mathcal{P}$ be the union of $\mathcal{P}_e$ for all edges e belonging to an exponential stratum. Note that $\mathcal{P}$ is finite.

Let $p \in \mathbb{N}$ and e be an edge of the exponential stratum H_k. Recall that the image under f of an edge of H_k starts and ends by an edge of H_k (property (RTT-i) of relative train-track maps). Thus if β is a maximal subpath of $f_\#^p(e)$ contained in G_{k-1}, then it is the image by some (possibly trivial) power of $f_\#$ of a path in $\mathcal{P}$. Moreover, we assumed that the image by f of any vertex of G is fixed by f. Hence if β is also a loop, there exists a path β' in $\mathcal{P} \cup f_\#(\mathcal{P})$ which is a loop such that β is the image of β' by a (possibly trivial) power of $f_\#$. Since $\boldsymbol{F_r}$ is residually finite, there exists a finite-index normal subgroup $\boldsymbol{H}$ of $\boldsymbol{F_r}$ with the following property. For every path $\beta \in \mathcal{P} \cup f_\#(\mathcal{P})$, if β is a loop, then the conjugacy class of $\boldsymbol{F_r}$ that it represents does not intersect $\boldsymbol{H}$.

Recall that $\widetilde{G}$ stands for the universal cover of G. Let us fix a base point x_0 in G. The fundamental group $\boldsymbol{F_r} = \pi_1(G, x_0)$ can therefore be identified with the deck transformation group acting on the left on $\widetilde{G}$. We fix a lift $\tilde{f}\colon \widetilde{G} \to \widetilde{G}$ of f. It determines an automorphism φ in the outer class of Φ such that for every $g \in \boldsymbol{F_r}$,

$$\varphi(g) \circ \tilde{f} = \tilde{f} \circ g. \tag{2}$$

There are only finitely many subgroups of $\boldsymbol{F_r}$ of a given index. Thus there exists an integer q such that $\varphi^q(\boldsymbol{H}) = \boldsymbol{H}$. Consequently, the intersection $\boldsymbol{L} = \bigcap_{p \in \mathbb{Z}} \varphi^p(\boldsymbol{H})$ is also a normal finite-index subgroup of $\boldsymbol{F_r}$. By definition, $\boldsymbol{L}$ is invariant by both φ and φ^{-1}. It directly follows that φ induces an automorphism of $\boldsymbol{L}$.

We now denote by κ the index of $\boldsymbol{L}$ in $\boldsymbol{F_r}$. Let $\widehat{G}$ be the space $\widehat{G} = \boldsymbol{L} \backslash \widetilde{G}$ and $\rho\colon \widehat{G} \to G$ the natural projection induced by $\widetilde{G} \to G$. The group $\boldsymbol{F_r}$ still acts on the left on $\widehat{G}$ and $\boldsymbol{L}$ is the kernel of this action. The map $\tilde{f}$ induces a map $\hat{f}\colon \widehat{G} \to \widehat{G}$ such that $\rho \circ \hat{f} = f \circ \rho$. Moreover, according to (2), for every $g \in \boldsymbol{F_r}$,

$$\varphi(g) \circ \hat{f} = \hat{f} \circ g. \tag{3}$$

Lemma 4.5 *The map $\hat{f}\colon \widehat{G} \to \widehat{G}$ admits a filtration which makes $\hat{f}$ an RTT representing the outer class of φ restricted to $\boldsymbol{L}$. Moreover, for every exponential stratum $\widehat{H}$ of $\widehat{G}$, there exists an exponential stratum H_k of G such that*

1. *$\widehat{H}$ is contained in $\rho^{-1}(H_k)$, and*
2. *$\hat{f}$ sends $\widehat{H}$ into $\widehat{H} \cup \rho^{-1}(G_{k-1})$.*

Proof We observe that, by construction, $\hat{f}\colon \widehat{G} \to \widehat{G}$ is a topological representative of φ restricted to $\boldsymbol{L}$, and $\varnothing = \rho^{-1}(G_0) \subset \rho^{-1}(G_1) \subset \cdots \subset \rho^{-1}(G_m) = \widehat{G}$ is an invariant filtration for $\hat{f}$. We are going to define a finer filtration $(\widehat{G}_{k,j})$ where the pairs (k, j) are endowed with the lexicographical order such that for every $k \in \{1, \dots, m\}$, we have

$$\rho^{-1}(G_{k-1}) \subset \widehat{G}_{k,1} \subset \widehat{G}_{k,2} \subset \cdots \subset \widehat{G}_{k,s} = \rho^{-1}(G_k).$$

Let $k \in \{0, \dots, m\}$. We focus on the stratum H_k of height k of G. We distinguish three cases.

(1) If H_k is a zero stratum, we just put $\widehat{G}_{k,1} = \rho^{-1}(G_k)$. Since $\rho \circ \widehat{f} = f \circ \rho$, we have $\widehat{f}(\widehat{G}_{k,1}) \subseteq \rho^{-1}(G_{k-1})$. Therefore, the associated stratum is a zero stratum.

(2) If H_k is a nonexponential stratum, it consists of a single edge e with $f(e) = eu$ where u is a loop in G_{k-1}. Then $\rho^{-1}(e)$ is a collection of κ edges: $\widehat{e}_1, \dots, \widehat{e}_\kappa$ (recall that κ is the index of $\boldsymbol{L}$ in $\boldsymbol{F}_r$). Since $\rho \circ \widehat{f} = f \circ \rho$, the map $\widehat{f}$ induces a permutation σ of $\{1, \dots, \kappa\}$ with the following property. For every $j \in \{1, \dots, \kappa\}$, we have $\widehat{f}(\widehat{e}_j) = \widehat{e}_{\sigma(j)}\widehat{u}_j$, where $\widehat{u}_j$ is a path in $\rho^{-1}(G_{k-1})$. We let $\widehat{G}_{k,1} = \rho^{-1}(G_k)$. Note that $\widehat{f}$ leaves $\widehat{G}_{k,1}$ invariant. The corresponding stratum is the closure of $\rho^{-1}(G_k) \setminus \rho^{-1}(G_{k-1})$. Its transition matrix is just the permutation matrix associated to σ. In particular, it is a nonexponential stratum.

(3) Assume now that H_k is an exponential stratum. We define a binary relation on $\rho^{-1}(H_k)$. Given two edges $\widehat{e}_1$ and $\widehat{e}_2$, we say that $\widehat{e}_1 \sim \widehat{e}_2$ if there exists $p \in \mathbb{N}$ such that $\widehat{e}_1$ or $\widehat{e}_1^{-1}$ is an edge of $\widehat{f}^p(\widehat{e}_2)$. This relation is reflexive and transitive. We claim that it is an equivalence relation. Let $\widehat{e}_1$ and $\widehat{e}_2$ be two edges of $\rho^{-1}(H_k)$ such that $\widehat{e}_1 \sim \widehat{e}_2$. We want to prove that $\widehat{e}_2 \sim \widehat{e}_1$. By definition of our relation, there exists $p \in \mathbb{N}$ such that $\widehat{e}_1$ or $\widehat{e}_1^{-1}$ is an edge of $\widehat{f}^p(\widehat{e}_2)$. For simplicity, we assume that $\widehat{e}_1$ belongs to $\widehat{f}^p(\widehat{e}_2)$. The other case works in the same way. We write $e_1 = \rho(\widehat{e}_1)$ and $e_2 = \rho(\widehat{e}_2)$ for their respective images in G. Since the stratum H_k is aperiodic, there exists $q \in \mathbb{N}$ such that e_2 or e_2^{-1} is an edge of $f^q(e_1)$. For simplicity, we assume that e_2 is an edge $f^q(e_1)$. Since $\rho \circ \widehat{f} = f \circ \rho$, there exists a preimage of e_2 in $\widehat{G}$ which is an edge of $\widehat{f}^q(\widehat{e}_1)$. Thus there exists $u \in \boldsymbol{F}_r$ such that $u \cdot \widehat{e}_2$ is an edge of $\widehat{f}^q(\widehat{e}_1)$. We now prove by induction that $u_\ell \cdot \widehat{e}_2$ is an edge of $\widehat{f}^{\ell(p+q)+q}(\widehat{e}_1)$ for every $\ell \in \mathbb{N}$, where

$$u_\ell = \varphi^{\ell(p+q)}(u) \cdots \varphi^{p+q}(u)u.$$

If $\ell = 0$, then the statement follows from the definition of u. Assume that it is true for $\ell \in \mathbb{N}$; ie $u_\ell \cdot \widehat{e}_2$ is an edge of $\widehat{f}^{\ell(p+q)+q}(\widehat{e}_1)$. Using (3) we get that $\varphi^p(u_\ell) \cdot \widehat{f}^p(\widehat{e}_2) = \widehat{f}^p(u_\ell \widehat{e}_2)$ is a subpath of $\widehat{f}^{(\ell+1)(p+q)}(\widehat{e}_1)$. In particular, $\varphi^p(u_\ell) \cdot \widehat{e}_1$ lies in $\widehat{f}^{(\ell+1)(p+q)}(\widehat{e}_1)$. With a similar argument, we get that $\varphi^{p+q}(u_\ell)u \cdot \widehat{e}_2$ lies in $\widehat{f}^{(\ell+1)(p+q)+q}(\widehat{e}_1)$. However,

$$\varphi^{p+q}(u_\ell)u = \varphi^{(\ell+1)(p+q)}(u) \cdots \varphi^{p+q}(u)u = u_{\ell+1}.$$

Thus the statement holds for $\ell + 1$, which completes the proof of the induction.

Since $\boldsymbol{L}$ has finite index in $\boldsymbol{F}_r$, there exist $\ell \in \mathbb{N}$ and $t \in \mathbb{N}^*$ such that u_ℓ and $u_{\ell+t}$ are in the same $\boldsymbol{L}$–coset, ie $u_{\ell+t}u_\ell^{-1} \in \boldsymbol{L}$. However,

$$u_{\ell+t}u_\ell^{-1} = \varphi^{(\ell+t)(p+q)}(u) \cdots \varphi^{(\ell+1)(p+q)}(u) = \varphi^{(\ell+1)(p+q)}(u_{t-1}).$$

Since $\boldsymbol{L}$ is φ–invariant, we derive that u_{t-1} belongs to $\boldsymbol{L}$. Recall that $\boldsymbol{L}$ is the kernel of the action of $\boldsymbol{F}_r$ on $\widehat{G}$; hence $u_{t-1}\cdot\hat{e}_2=\hat{e}_2$. On the other hand, $u_{t-1}\cdot\hat{e}_2$ is an edge of $\hat{f}^{(t-1)(p+q)+q}(\hat{e}_1)$. Consequently, $\hat{e}_2\sim\hat{e}_1$, which completes the proof of our claim.

We denote by $\widehat{H}_{k,1},\dots,\widehat{H}_{k,s}$ the equivalence classes for the relation $\sim$. For every $j\in\{1,\dots,s\}$, we put $\widehat{G}_{k,j}=\widehat{H}_{k,1}\cup\dots\cup\widehat{H}_{k,j}\cup\rho^{-1}(G_{k-1})$. By construction, the filtration

$$\rho^{-1}(G_{k-1})\subset\widehat{G}_{k,1}\subset\dots\subset\widehat{G}_{k,s}=\rho^{-1}(G_k)$$

is $\hat{f}$–invariant. Moreover, the strata $\widehat{H}_{k,1},\dots,\widehat{H}_{k,s}$ associated to this filtration are irreducible. We claim that they are exponential. Let $j\in\{1,\dots,s\}$. Let $M_{k,j}$ be the transition matrix of $\widehat{H}_{k,j}$. It is known that if the PF-eigenvalue of $M_{k,j}$ is 1, then $M_{k,j}$ is a permutation matrix. Thus there exists an edge $\hat{e}\in\widehat{H}_{k,j}$ and a positive integer p such that $\hat{e}$ is the only edge of $\rho^{-1}(H_k)$ in $\hat{f}^p(\hat{e})$. In particular, if e stands for $e=\rho(\hat{e})$, we get that e is the only edge of H_k in $f^p(e)$. This contradicts the fact that H is aperiodic. Hence the PF-eigenvalue of $M_{k,j}$ is larger than 1, and the stratum $\widehat{H}_{k,j}$ is exponential.

Finally, recall that f satisfies properties (RTT-i)–(RTT-iii). It follows from $\rho\circ\hat{f}=f\circ\rho$ that $\hat{f}$ also satisfies these properties. □

Lemma 4.6 *For every edge $\hat{e}$ in an exponential stratum $\widehat{H}$ of $\widehat{G}$, for every $p\in\mathbb{N}$, every maximal subpath of $\hat{f}^p_\#(\hat{e})$ that does not cross $\widehat{H}$ is not a loop.*

Proof Let $\hat{e}$ be an edge of an exponential stratum $\widehat{H}$ of $\widehat{G}$. Let $p\in\mathbb{N}$. Let $\hat{\beta}$ be a maximal subpath of $\hat{f}^p_\#(\hat{e})$ that does not cross $\widehat{H}$. By Lemma 4.5, there exists $k\in\{1,\dots,m\}$ such that H_k is an exponential stratum of G, $\widehat{H}$ is contained in $\rho^{-1}(H_k)$ and $\hat{f}(\widehat{H})$ is a subset of $\widehat{H}\cup\rho^{-1}(G_{k-1})$. We denote by e the image of $\hat{e}$ by ρ. It belongs to H_k. Since ρ is a continuous locally injective map, $\hat{f}^p_\#(\hat{e})$ is a lift of $f^p_\#(e)$. It follows that $\beta=\rho(\hat{\beta})$ is a maximal subpath of $f^p_\#(e)$ contained in G_{k-1}. If β is not a loop, neither is $\hat{\beta}$. Therefore, we can assume that β is a loop in G. By construction of $\mathcal{P}$, there exists a loop β' in $\mathcal{P}\cup f_\#(\mathcal{P})$ such that β is the image of β' by some power of $f_\#$. However, by definition, the conjugacy class of $\boldsymbol{F}_r$ represented by β' does not intersect $\boldsymbol{L}\subset\boldsymbol{H}$. Since $\boldsymbol{L}$ is φ–invariant, neither does the conjugacy class of $\boldsymbol{F}_r$ represented by β. Thus its lift $\hat{\beta}$ in $\widehat{G}$ cannot be a loop. □

Lemma 4.7 *Let n be an integer. Recall that κ is the index of $\boldsymbol{L}$ in $\boldsymbol{F}_r$. If Φ induces an outer automorphism of finite order of $\boldsymbol{B}_r(\kappa n)$, then its restriction to $\boldsymbol{L}$ induces an outer automorphism of finite order of $\boldsymbol{L}/\boldsymbol{L}^n$.*

Proof According to Remark 1.1, the image of φ in $\mathrm{Aut}(\boldsymbol{B}_r(\kappa n))$ has finite order. Hence there exists $p\in\mathbb{N}$ such that for every $g\in\boldsymbol{F}_r$, the element $\varphi^p(g)g^{-1}$ belongs to $\boldsymbol{F}_r^{\kappa n}$. However, $\boldsymbol{L}$ has index κ in $\boldsymbol{F}_r$. It follows that g^κ lies in $\boldsymbol{L}$ for every $g \in \boldsymbol{F}_r$. In particular, $\boldsymbol{F}_r^{\kappa n}$ is a subset of $\boldsymbol{L}^n$. Consequently, $\varphi^p(g)g^{-1}$ belongs to $\boldsymbol{L}^n$ for every $g \in \boldsymbol{L}$. It exactly means that, as an automorphism of $\boldsymbol{L}/\boldsymbol{L}^n$, φ^p is trivial. Hence the restriction of Φ to $\boldsymbol{L}$ induces an automorphism of finite order of $\boldsymbol{L}/\boldsymbol{L}^n$. □

Proposition 4.4 becomes a consequence of the following result.

Proposition 4.8 *Let $\Phi \in \mathrm{Out}(\boldsymbol{F}_r)$ be an outer automorphism represented by an RTT $f\colon G \to G$. Assume that for every edge e in an exponential stratum H, for every $p \in \mathbb{N}$, every maximal subpath of $f_\#^p(e)$ that does not cross H is not a loop. Then there exists $n_0 \in \mathbb{N}$ such that for all odd integers $n \geqslant n_0$, the automorphism Φ induces an outer automorphism of $\boldsymbol{B}_r(n)$ of infinite order.*

In the next section, we discuss a third reduction and prove that Proposition 4.8 is a consequence of Proposition 4.11.

4.3 Automorphisms with only one exponential stratum

The following lemma is proved by the first author in [13] using the structure of free products.

Lemma 4.9 (Coulon [13, Lemma 1.9]) *Let n be an integer. Let φ be an automorphism of $\boldsymbol{F}_r$ which stabilizes a free factor $\boldsymbol{H}$. We assume that φ induces an automorphism of finite order of $\boldsymbol{B}_r(n)$. Then, the restriction of φ to $\boldsymbol{H}$ also induces an automorphism of finite order of $\boldsymbol{H}/\boldsymbol{H}^n$.*

Let $\Phi \in \mathrm{Out}(\boldsymbol{F}_r)$ be an exponentially growing outer automorphism, and let $f\colon G \to G$ be an RTT representing Φ with a filtration $\varnothing = G_0 \subset G_1 \subset \cdots \subset G_m = G$. By Remark 3.4 (2), f has at least one exponential stratum. We assume that f satisfies the additional assumption of Proposition 4.8; ie for every edge e in an exponential stratum H, for every $p \in \mathbb{N}$, every maximal subpath of $f_\#^p(e)$ that does not cross H is not a loop. By replacing Φ by a power of Φ if necessary, we can assume that the exponential strata of the RTT are aperiodic. Note that this operation does not affect the graph G. However, one might need to refine the filtration of the RTT. In particular, the RTT does not necessarily satisfy the additional assumption of Proposition 4.8 anymore. Nevertheless, for every edge e in the *lowest* exponential stratum H_k, for every $p \in \mathbb{N}$, every maximal subpath of $f_\#^p(e)$ that does not cross H_k is not a loop.

Let G' be the connected component of the graph G_k which contains H_k. We assert that G' is f–invariant, ie $f(G') \subseteq G'$. Indeed, $G' \cap f(G')$ is nonempty (it contains H_k) and $f(G')$ is connected. Let $\boldsymbol{H}$ be the free factor of $\boldsymbol{F}_r$ defined by $G' \subseteq G$, and let $\Psi \in \mathrm{Out}(\boldsymbol{H})$ be the outer automorphism induced by the restriction $f' = f|_{G'}\colon G' \to G'$. We note that $f|_{G'}\colon G' \to G'$ is an RTT representing Ψ, which has exactly one exponential stratum, namely H_k, which is aperiodic and the top stratum of f'. In particular, Ψ has exponential growth.

Lemma 4.10 *If Ψ induces an outer automorphism of $\boldsymbol{H}/\boldsymbol{H}^n$ of infinite order, then Φ also induces an outer automorphism of $\boldsymbol{B}_r(n)$ of infinite order.*

Proof There exists an automorphism φ in the class of Φ which stabilizes $\boldsymbol{H}$. Assume that Φ induces an outer automorphism of $\boldsymbol{B}_r(n)$ of finite order. In particular, the image of φ in $\mathrm{Aut}(\boldsymbol{B}_r(n))$ has finite order; see Remark 1.1. It follows from the previous lemma that the restriction to $\boldsymbol{H}$ of φ (and thus Ψ) induces an automorphism (outer automorphism) of finite order of $\boldsymbol{H}/\boldsymbol{H}^n$. □

It follows from our discussion that Proposition 4.8 is a consequence of the following statement.

Proposition 4.11 *Let $\Phi \in \mathrm{Out}(\boldsymbol{F}_r)$ be an outer automorphism represented by an RTT $f\colon G \to G$ with exactly one exponential stratum H, which is aperiodic and the top stratum of f. Assume that for every edge e in H, for every $p \in \mathbb{N}$, every maximal subpath of $f_\#^p(e)$ that does not cross H is not a loop. Then there exists $n_0 \in \mathbb{N}$ such that for all odd integers $n \geqslant n_0$, the automorphism Φ induces an outer automorphism of $\boldsymbol{B}_r(n)$ of infinite order.*

We have seen that Theorem 1.3 can be deduced from Proposition 4.11. The latter will be proved in Sections 5 and 6.

5 Tracking powers

The next two sections are dedicated to the proof of Proposition 4.11. As we explained in the introduction, the goal is to understand to what extent a periodic path can appear in the orbit of a circuit under the iteration of the train-track map. This is the purpose of this section.

The general strategy is the following. We consider an outer automorphism Φ represented by an RTT $f\colon G \to G$ with a single exponential stratum H which is aperiodic. Then, we fix an edge $e_\bullet$ in H. For every $p \in \mathbb{N}$, we look at the path obtained by removing from $f_\#^p(e_\bullet)$ all the edges which are not in H. This sequence can be interpreted as the orbit of $e_\bullet$ under a substitution over the set of oriented edges of H; see Lemma 5.1. It follows from the aperiodicity of H that this substitution is primitive. Therefore, we would like to apply Proposition 2.1. We need to rule out first the case of an infinite shift-periodic word, though (Proposition 2.1 (ii)). The dynamic of the substitution is not sufficient to conclude here. Remark 2.2 provides indeed an example of a primitive substitution σ with an infinite shift-periodic fixed point. However, the particularity of this example is that σ does not represent an automorphism of $\boldsymbol{F}_3$. Our proof (see Proposition 5.2) strongly uses the fact that the substitution we are looking at comes from an automorphism of the free group.

From now on, Φ denotes an outer automorphism of $\boldsymbol{F}_r$ which can be represented by an RTT $f\colon G \to G$ with exactly one exponential stratum H. Moreover, H is aperiodic and the top stratum of f. We denote by $\mathcal{E}$ the set of all the oriented edges of H. In addition, we assume that for every $e \in \mathcal{E}$, for every $p \in \mathbb{N}$, every maximal subpath of $f_\#^p(e)$ that does not cross H is not a loop.

By replacing Φ by a power of Φ if necessary, we can assume that $f(v)$ is fixed by f for every vertex v of G, and that there exists $e_\bullet \in \mathcal{E}$ such that $Df(e_\bullet) = e_\bullet$. Note that this operation does not affect the graph G or the exponential stratum. In particular, H is still the only exponential stratum of f. It is aperiodic and the top stratum. By choice of $e_\bullet$, we have that f fixes the initial vertex x_0 of $e_\bullet$. Thus it naturally defines an automorphism $\varphi \in \mathrm{Aut}(\pi_1(G, x_0))$ in the outer class Φ: if g is an element of $\pi_1(G, x_0)$ represented by a loop α based at x_0, then $\varphi(g)$ is the homotopy class of $f(\alpha)$ (relative to x_0).

5.1 The yellow-red decomposition

We refer to the edges of H as *red edges* and to the edges of $G \setminus H$ as *yellow edges*. Recall that $\widetilde{G}$ denotes the universal cover of G. An edge of $\widetilde{G}$ can be labeled by the edge of G of which it is the lift. In particular, its color is given by the color of its label.

A k–legal path of G (where k is the height of H) will be call a *red-legal* path. A path (in G or in $\widetilde{G}$) is a *yellow path* if it only crosses yellow edges. *Red paths* are defined in the same way. Any path α (in G or in $\widetilde{G}$) can be decomposed as a concatenation of maximal yellow and red subpaths: $\alpha = \alpha_1 \cdots \alpha_q$, where α_i $(1 \leqslant i \leqslant q)$ is a nontrivial subpath of α which is either yellow or red, and α_i and α_{i+1} have not the same color for all $i \in \{1, \dots, q-1\}$. According to Lemma 3.3, this decomposition is a splitting of α.

The red word associated to a path We associate to any path of edges α in G or $\widetilde{G}$ a word $\mathrm{Red}(\alpha)$ over the alphabet $\mathcal{E}$. As a path of edges, α is labeled by a word over the alphabet that consists of all oriented edges of G. The word $\mathrm{Red}(\alpha)$ is obtained from this word by removing all the letters corresponding to yellow edges. We stress on the fact that if α is a reduced path, then $\mathrm{Red}(\alpha)$ is not, in general, a reduced word.

5.2 The induced substitution on red edges

Definition and first properties We associate to the RTT f a substitution σ on $\mathcal{E}$ called the *induced substitution*. It is defined as

$$\sigma(e) = \mathrm{Red}(f(e)) \quad \text{for every } e \in \mathcal{E}.$$

Lemma 5.1 *Let α be a red-legal path in G. For all $p \in \mathbb{N}$, we have*

$$\mathrm{Red}(f_\#^p(\alpha)) = \sigma^p(\mathrm{Red}(\alpha)).$$

Proof We consider a decomposition of α as $\alpha = \alpha_1 e_1 \alpha_2 e_2 \cdots \alpha_q e_q \alpha_{q+1}$ where each $e_i \in \mathcal{E}$ is a red edge, and each α_i is a (possibly trivial) yellow subpath. In particular, $\mathrm{Red}(\alpha) = e_1 e_2 \cdots e_q$. The path α being red-legal, Lemma 3.3 leads to

$$\begin{aligned} f_\#(\alpha) &= f_\#(\alpha_1 e_1 \alpha_2 e_2 \cdots \alpha_q e_q \alpha_{q+1}) \\ &= f_\#(\alpha_1) f(e_1) f_\#(\alpha_2) f(e_2) \cdots f_\#(\alpha_q) f(e_q) f_\#(\alpha_{q+1}). \end{aligned}$$

However, f sends yellow edges to yellow paths. We deduce that

$$\begin{aligned} \mathrm{Red}(f_\#(\alpha)) &= \mathrm{Red}(f(e_1))\,\mathrm{Red}(f(e_2)) \cdots \mathrm{Red}(f(e_q)) \\ &= \sigma(e_1)\sigma(e_2)\cdots\sigma(e_q) = \sigma(e_1 e_2 \cdots e_q) = \sigma(\mathrm{Red}(\alpha)). \end{aligned}$$

The image by $f_\#$ of a red-legal path is still a red-legal path. Therefore, for all $p \in \mathbb{N}$,

$$\mathrm{Red}(f_\#^{p+1}(\alpha)) = \mathrm{Red}(f_\#(f_\#^p(\alpha))) = \sigma(\mathrm{Red}(f_\#^p(\alpha))).$$

The result follows by induction on p. □

Primitivity of the induced substitution The material of this paragraph is widely inspired by the work of P Arnoux et al [2, Section 3]. Recall that $\Theta\colon \mathcal{E} \to \mathcal{E}$ is the map which sends e to e^{-1}. We extend Θ to the free monoid $\mathcal{E}^*$ in the following way. Let w be an element of $\mathcal{E}^*$. By definition, it can be written $w = e_1 e_2 \cdots e_q$ where $e_i \in \mathcal{E}$. We put $\Theta(w) = e_q^{-1} \cdots e_2^{-1} e_1^{-1}$. It defines an involution of $\mathcal{E}^*$ called the *flip map*. Moreover, we observe that $\sigma \circ \Theta(e) = \Theta \circ \sigma(e)$ for all edges $e \in \mathcal{E}$. Thus σ and Θ commute on $\mathcal{E}^*$. The substitution σ is said to be *orientable* with respect to a subset $\vec{\mathcal{E}}$ of $\mathcal{E}$ if

(i) $\vec{\mathcal{E}}$ and $\Theta(\vec{\mathcal{E}})$ make a partition of $\mathcal{E}$,

(ii) $\sigma(\vec{\mathcal{E}}) \subset \vec{\mathcal{E}}^*$.

Note that (i) just says that $\vec{\mathcal{E}}$ is a preferred set of oriented edges for H. In that case, σ induces a substitution of $\vec{\mathcal{E}}^*$, that we still denote by σ.

By assumption, the red stratum H of f is aperiodic. In other words, its transition matrix M is primitive. Applying [2, Proposition 3.7], we know that either

- σ is not orientable, and then σ is a primitive substitution on the alphabet $\mathcal{E}$, or
- there exists a subset $\vec{\mathcal{E}}$ of $\mathcal{E}$ such that σ is orientable with respect to $\vec{\mathcal{E}}$, and then σ induces a primitive substitution on the alphabet $\vec{\mathcal{E}}$.

Thus in both cases, there exists a subset $\mathcal{E}_\bullet$ of $\mathcal{E}$ containing $e_\bullet$ such that $\sigma(\mathcal{E}_\bullet) \subset \mathcal{E}_\bullet^*$, and the substitution $\sigma\colon \mathcal{E}_\bullet^* \to \mathcal{E}_\bullet^*$ is primitive.

5.3 A red word without large powers

The infinite red word $\sigma^\infty(e_\bullet)$ Recall that $e_\bullet$ is a red edge of $\mathcal{E}$ that has been chosen in such a way that $Df(e_\bullet) = e_\bullet$. Because the red stratum is aperiodic, $f(e_\bullet) = e_\bullet \cdot \alpha$ where $\mathrm{Red}(\alpha)$ is nontrivial. In particular, $e_\bullet$ is a prefix of $\sigma(e_\bullet)$. According to Proposition 2.1 the sequence $(\sigma^p(e_\bullet))$ converges to an infinite word $\sigma^\infty(e_\bullet)$ of $\mathcal{E}_\bullet^{\mathbb{N}}$. Note that $f(e_\bullet) = e_\bullet \cdot \alpha$ is a splitting. Hence for every $p \in \mathbb{N}$,

$$f_\#^p(e_\bullet) = e_\bullet \cdot \alpha \cdot f_\#(\alpha) \cdots f_\#^{p-1}(\alpha).$$

Hence $(f_\#^p(e_\bullet))$ also converges to an infinite path

$$f_\#^\infty(e_\bullet) = e_\bullet \cdot \alpha \cdot f_\#(\alpha) \cdots f_\#^p(\alpha) \cdots .$$

Proposition 5.2 *The infinite word $\sigma^\infty(e_\bullet)$ is not shift periodic.*

This proof combines a dynamical argument (σ is a primitive substitution) and a group theoretical one (φ is an automorphism of $\boldsymbol{F_r}$). Let us sketch first the main steps. We assume that the proposition is false. This means that if we restrict our attention to the red edges, the path $f_\#^\infty(e_\bullet)$ is periodic. We construct from G a colored graph Γ on which $f_\#^\infty(e_\bullet)$ coils up. More precisely, its fundamental group $\boldsymbol{H}$ can be decomposed as a free product $\boldsymbol{H} = \boldsymbol{L} * \langle h \rangle$, where $\boldsymbol{L}$ is generated by conjugates of yellow loops, and h is represented by a loop $\hat{\gamma}$ with the following property. If we collapse all the yellow edges of Γ, we obtain a simple (red) loop which is exactly the image of $\hat{\gamma}$ by the same operation. Moreover, this red loop is the period of the red word associated to $f_\#^\infty(e_\bullet)$; see Figure 3. We show that the RTT f induces a homotopy equivalence $\hat{f}\colon \Gamma \to \Gamma$ that catches two conflicting features of Φ:

1. Since the stratum H is exponential, $\hat{f}$ should increase the length of the red word associated to $\hat{\gamma}$; see Proposition 5.7.
2. The yellow components of G are invariant under f. It follows that the automorphism of $\boldsymbol{H}$ induced by $\hat{f}$ sends h to $gh^{\pm 1}$, where g belongs to the normal subgroup generated by $\boldsymbol{L}$.

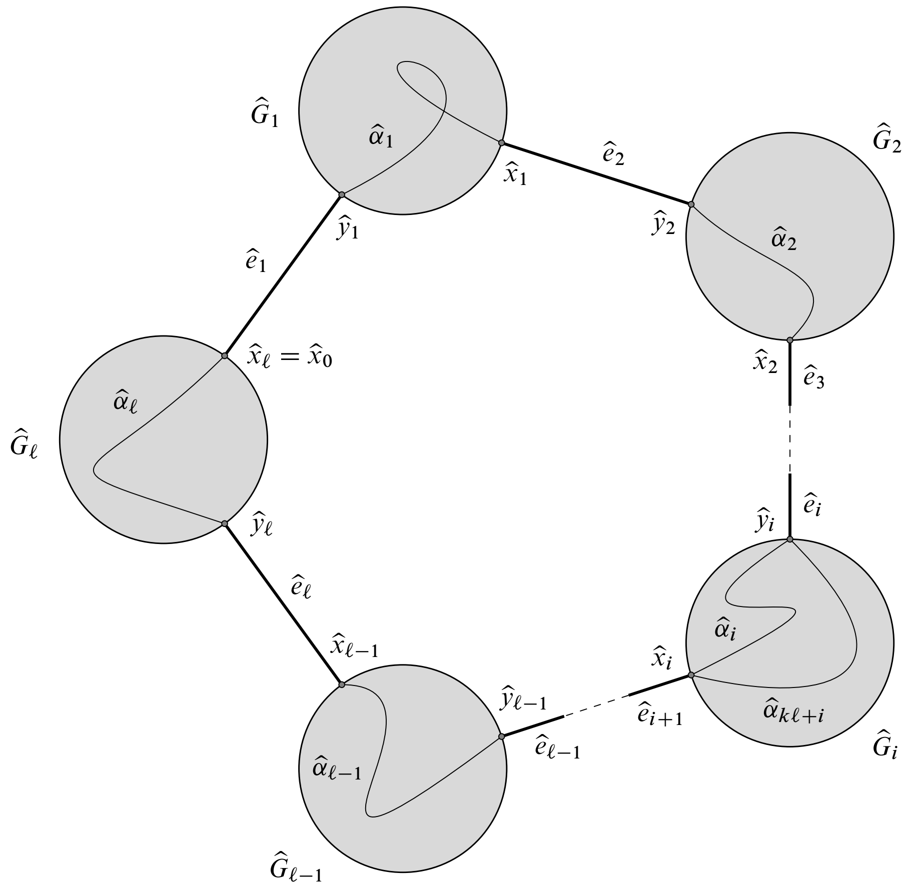


Figure 3: The graph Γ

The key fact is that these two properties can be observed in the abelianization of $\boldsymbol{H}$, which leads to a contradiction.

Proof of Proposition 5.2 Assume that $\sigma^\infty(e_\bullet)$ is shift-periodic. Recall that σ is primitive as a substitution of $\mathcal{E}_\bullet^*$. Proposition 2.1 implies that there exist an integer $q \geqslant 2$ and a primitive word $u = e_1 e_2 \cdots e_\ell$ of $\mathcal{E}_\bullet^*$ such that $\sigma^\infty(e_\bullet) = u^\infty$ and $\sigma(u) = u^q$. Notice that $e_1 = e_\bullet$. This means, in particular, that the infinite path $f_\#^\infty(e_\bullet)$ is obtained as a concatenation

$$f_\#^\infty(e_\bullet) = \gamma_0 \cdot \gamma_1 \cdot \gamma_2 \cdots \gamma_k \cdots$$

of loops

$$\gamma_k = e_1 \cdot \alpha_{k\ell+1} \cdot e_2 \cdot \alpha_{k\ell+2} \cdots e_\ell \cdot \alpha_{(k+1)\ell},$$

where the α_i are (possibly trivial) yellow paths. Moreover, if α_i is nontrivial, then it is a maximal yellow subpath of the image by a power of $f_\#$ of a red edge. By assumption, none of them is a loop. Recall that x_0 is the initial point of $e_1 = e_\bullet$. For every $i \in \{1, \dots, \ell\}$, we have that y_i and x_i respectively stand for the initial and the terminal points of α_i. In particular, $x_0 = x_\ell$. We now focus on the path $\gamma = \gamma_0$:

$$\gamma = e_1 \cdot \alpha_1 \cdot e_2 \cdot \alpha_2 \cdots e_\ell \cdot \alpha_\ell.$$

Lemma 5.3 *The path $f_\#(\gamma)$ is exactly $\gamma_0 \cdots \gamma_{q-1}$. In particular, it is an initial subpath of the infinite path $f_\#^\infty(e_\bullet)$.*

Proof By construction, there exists $p \in \mathbb{N}$ such that $\gamma_0 \cdot \gamma_1$ is a proper initial subpath of $f_\#^p(e_\bullet)$. Moreover, the terminal point of $\gamma_0\gamma_1$, which is also the initial vertex of the red edge $e_1 = e_\bullet$, is a splitting point of the yellow-red splitting of $f_\#^p(e_\bullet)$. Thus $f_\#(\gamma_0\gamma_1) = f_\#(\gamma_0) \cdot f_\#(\gamma_1)$ is an initial subpath of $f_\#^{p+1}(e_\bullet)$, hence of $f_\#^\infty(e_\bullet)$. However, by Lemma 5.1,

$$\mathrm{Red}(f_\#(\gamma_0)) = \sigma(\mathrm{Red}(\gamma_0)) = \sigma(u) = u^q = \mathrm{Red}(\gamma_0 \cdots \gamma_{q-1}).$$

It follows that there exists a subpath α' of $\alpha_{q\ell}$ such that

$$f_\#(\gamma_0) = [\gamma_0 \cdots \gamma_{q-2}][e_1\alpha_{(q-1)\ell+1}e_2\alpha_{(q-1)\ell+2} \cdots e_\ell\alpha'].$$

On the other hand, γ_1 and thus $f_\#(\gamma_1)$ starts with the red edge $e_1 = e_\bullet$. Since $f_\#(\gamma_0) \cdot f_\#(\gamma_1)$ is an initial subpath of $f_\#^{p+1}(e_\bullet)$, the path α' is necessarily the whole $\alpha_{q\ell}$. Consequently, $f_\#(\gamma) = \gamma_0 \cdots \gamma_{q-1}$. □

The graph Γ and the loop $\hat{\gamma}$ Let $i \in \{1, \dots, \ell\}$. We define $\widehat{G}_i$ to be a copy of the largest connected yellow subgraph of G containing y_i. We denote by $\hat{\alpha}_i$ (respectively $\hat{y}_i$ and $\hat{x}_i$) the path α_i (respectively the vertices y_i and x_i) viewed as a path of $\widehat{G}_i$ (respectively as vertices of $\widehat{G}_i$).

We now construct a graph Γ as follows. We start with the disjoint union of the $\widehat{G}_i$ for $i \in \{1, \dots, \ell\}$. Then for every $i \in \{1, \dots, \ell\}$, we add an oriented edge $\hat{e}_i$ whose initial and terminal points are respectively $\hat{x}_{i-1}$ and $\hat{y}_i$. The reverse edge $\hat{e}_i^{-1}$ is attached accordingly. In this process, we think about the indices i as elements of $\mathbb{Z}/\ell\mathbb{Z}$. In particular, $\hat{x}_0$ should be understood as the point $\hat{x}_\ell$ of $\widehat{G}_\ell$. We denote by Γ the graph obtained in this way; see Figure 3. Let ρ be the graph morphism $\rho\colon \Gamma \to G$ such that for every $i \in \{1, \dots, \ell\}$, we have that $\rho(\hat{e}_i) = e_i$ and the restriction of ρ to $\widehat{G}_i$ is the natural embedding $\widehat{G}_i \hookrightarrow G$. We color the edges of Γ by the color of their images under ρ. In other words, the edges $\hat{e}_i$ are red, whereas the edges of the subgraphs $\widehat{G}_i$ are yellow. By construction, the loop $\hat{\gamma}$ defined below is a lift of γ in Γ:

$$\hat{\gamma} = \hat{e}_1\hat{\alpha}_1\hat{e}_2\hat{\alpha}_2 \cdots \hat{e}_\ell\hat{\alpha}_\ell.$$

The subgroup $\boldsymbol{H}$ We denote by $\boldsymbol{H}$ the fundamental group $\pi_1(\Gamma, \hat{x}_0)$. Let us choose a maximal tree T_i in each $\widehat{G}_i$. The union T defined below is a maximal tree of Γ:

$$T = \left(\bigcup_{i=1}^{\ell} T_i\right) \cup \left(\bigcup_{i=1}^{\ell-1} \hat{e}_i\right).$$

For every edge e of Γ not in T, we write β_e for the path contained in T starting at $\hat{x}_0$ and ending at the initial vertex of e. We define h_e as the element of $\boldsymbol{H}$ represented by $\beta_e e \beta_{e^{-1}}^{-1}$. Let $i \in \{1, \dots, \ell\}$. For each unoriented edge of $\widehat{G}_i \setminus T_i$, we chose one of the two corresponding oriented edges. We denote then by $\mathcal{F}_i$ the preferred set of oriented edges obtained in this way. We write $\mathcal{F}$ for the union

$$\mathcal{F} = \bigcup_{i=1}^{\ell} \mathcal{F}_i.$$

Lemma 5.4 *Let h be the element of $\boldsymbol{H}$ represented by $\hat{\gamma}$. The family $\mathcal{B}$ obtained by taking the union of $(h_e)_{e \in \mathcal{F}}$ and $\{h\}$ is a free basis of $\boldsymbol{H}$.*

Proof It follows from the definition of $\mathcal{F}$ that the family $(h_e)_{e \in \mathcal{F} \cup \{\hat{e}_\ell\}}$ is a free basis of $\boldsymbol{H}$. By construction of Γ, we have $h = g \cdot h_{\hat{e}_\ell}$ where g is a product of some h_e with $e \in \mathcal{F} \cup \mathcal{F}^{-1}$. This implies that the family $(h_e)_{e \in \mathcal{F}}$ together with h forms a free basis of $\boldsymbol{H}$. □

Let $k \in \mathbb{N}$ and $i \in \{1, \dots, \ell\}$. The path $\alpha_{k\ell+i}$ and α_i have the same endpoints, namely y_i (the terminal point of e_i) and x_i (the initial point of e_{i+1}). In particular, they are contained in the same maximal yellow connected component of G. We denote by $\hat{\alpha}_{k\ell+i}$ the copy in $\widehat{G}_i$ of $\alpha_{k\ell+i}$; see Figure 3. We put

$$\hat{\gamma}_k = \hat{e}_1 \hat{\alpha}_{k\ell+1} \hat{e}_2 \hat{\alpha}_{k\ell+2} \cdots \hat{e}_\ell \hat{\alpha}_{(k+1)\ell}.$$

By construction, $\hat{\gamma}_k$ is a loop of Γ based at $\hat{x}_0$ lifting γ_k (ie $\rho \circ \hat{\gamma}_k = \gamma_k$).

Lemma 5.5 *Let h be the element of $\boldsymbol{H}$ represented by $\hat{\gamma}$. Let $k \in \mathbb{N}$. There exists g in the normal subgroup generated by $(h_e)_{e \in \mathcal{F}}$ such that the element of $\boldsymbol{H}$ represented by the loop $\hat{\gamma}_k$ is gh.*

Proof It follows from the equality

$$\begin{aligned}\hat{\gamma}_k = {} & \big[\hat{e}_1(\hat{\alpha}_{k\ell+1}\hat{\alpha}_1^{-1})\hat{e}_1^{-1}\big]\big[\hat{e}_1\hat{\alpha}_1\hat{e}_2(\hat{\alpha}_{k\ell+2}\hat{\alpha}_2^{-1})\hat{e}_2^{-1}\hat{\alpha}_1^{-1}\hat{e}_1^{-1}\big]\cdots \\ & \cdots\big[\hat{e}_1\hat{\alpha}_1\hat{e}_2\cdots\hat{e}_{\ell-1}(\hat{\alpha}_{(k+1)\ell-1}\hat{\alpha}_{\ell-1}^{-1})\hat{e}_{\ell-1}^{-1}\cdots\hat{e}_2^{-1}\hat{\alpha}_1^{-1}\hat{e}_1^{-1}\big]\big[\hat{\gamma}(\hat{\alpha}_\ell^{-1}\hat{\alpha}_{(k+1)\ell})\hat{\gamma}^{-1}\big]\hat{\gamma}. \end{aligned}$$ □

Lemma 5.6 *The map $\rho\colon \Gamma \to G$ is locally injective.*

Proof We prove this lemma by contradiction. Let $\widehat{e}$ and $\widehat{e}'$ be two distinct edges of Γ with the same initial vertex $\widehat{v}$. Suppose that $\rho(\widehat{e}) = \rho(\widehat{e}')$. There exists $i \in \{1, \dots, \ell\}$ such that $\widehat{v}$ is a vertex of $\widehat{G}_i$. By construction, ρ preserves the color of the edges, thus $\widehat{e}$ and $\widehat{e}'$ necessarily have the same color. We distinguish two cases. Assume first that $\widehat{e}$ and $\widehat{e}'$ are both yellow edges. Recall that the restriction of ρ to $\widehat{G}_i$ is the inclusion $\widehat{G}_i \hookrightarrow G$. Thus $\widehat{e} = \widehat{e}'$, a contradiction. Assume now that $\widehat{e}$ and $\widehat{e}'$ are red. By construction of Γ, at most two red edges have an initial vertex in $\widehat{G}_i$. Without loss of generality, we can assume that $\widehat{e}^{-1} = \widehat{e}_i$ and $\widehat{e}' = \widehat{e}_{i+1}$ (as previously, if $i = \ell$ then $\widehat{e}_{i+1}$ corresponds to $\widehat{e}_1$). Then $\rho(\widehat{v})$ is the terminal vertex y_i of e_i and the initial vertex x_i of e_{i+1}. Thus the yellow path α_i is either trivial or a loop of G. By assumption, it cannot be a loop; thus α_i is trivial and $e_{i+1} = \rho(\widehat{e}') = \rho(\widehat{e}) = e_i^{-1}$. This contradicts the fact that γ is a path. Consequently, ρ is locally injective. $\square$

If follows from the lemma that ρ induces an embedding ρ_* from $\boldsymbol{H}$ into $\pi_1(G, x_0)$. From now on, we identify $\boldsymbol{H}$ with its image in $\pi_1(G, x_0)$.

The automorphism induced on $\boldsymbol{H}$ Recall that φ is the automorphism of $\pi_1(G, x_0)$ in the outer class Φ induced by f. We now prove that φ induces an automorphism of $\boldsymbol{H}$. To that end, we lift the RTT f into a map $\widehat{f}\colon \Gamma \to \Gamma$.

Proposition 5.7 *There exists a continuous map $\widehat{f}\colon \Gamma \to \Gamma$ satisfying the following:*

1. $f \circ \rho = \rho \circ \widehat{f}$,
2. $\widehat{f}(\widehat{\gamma})$ *is homotopic relative to its endpoints to* $\widehat{\gamma}_0 \cdots \widehat{\gamma}_{q-1}$.

Proof The map $\widehat{f}\colon \Gamma \to \Gamma$ is built step by step. Let us first define some auxiliary objects that will be needed during the construction. Let Γ_ℓ be the graph obtained from Γ by disconnecting $\widehat{e}_1$ from $\widehat{G}_\ell$ at $\widehat{x}_0$; see Figure 4. It comes with a natural map $\Gamma_\ell \to \Gamma$ which is a local isometry. For simplicity, we use the same notation for the paths of Γ_ℓ and their images in Γ. For instance, $\widehat{\gamma}$ can be seen as a subpath of Γ_ℓ. Similarly, we still denote by ρ the locally injective map $\rho\colon \Gamma_\ell \to G$. For every $i \in \{1, \dots, \ell\}$, we denote by Γ_i the subgraph of Γ_ℓ consisting of the red edges $\widehat{e}_1, \dots, \widehat{e}_i$ and the yellow graphs $\widehat{G}_1, \dots, \widehat{G}_i$. By convention, we put $\Gamma_0 = \{\widehat{x}_0\}$.

Let $i \in \{1, \dots, \ell\}$. The path γ can be split as follows:

$$\gamma = (e_1\alpha_1 \cdots e_i\alpha_i) \cdot (e_{i+1}\alpha_{i+1} \cdots e_\ell\alpha_\ell).$$

Therefore, $f_\#(e_1\alpha_1 \cdots e_i\alpha_i)$ is an initial subpath of $f_\#(\gamma) = \gamma_0 \cdots \gamma_{q-1}$; see Lemma 5.3. However, the path $\widehat{\gamma}_0 \cdots \widehat{\gamma}_{q-1}$ is, by construction, the unique lift of $\gamma_0 \cdots \gamma_{q-1}$ in Γ

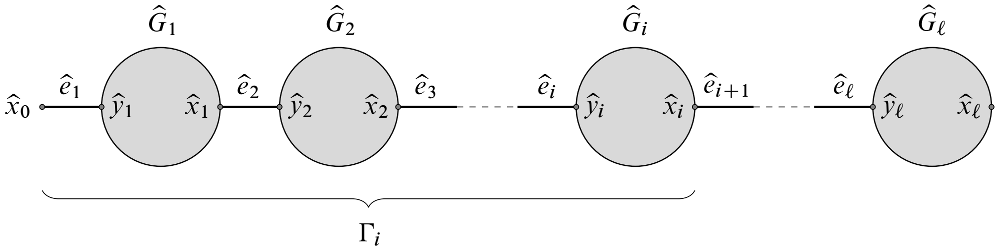


Figure 4: The graph Γ_ℓ

starting at $\widehat{x}_0$. We denote by $\widehat{\beta}_i$ the initial subpath of $\widehat{\gamma}_0 \cdots \widehat{\gamma}_{q-1}$ such that $\rho \circ \widehat{\beta}_i = f_\#(e_1\alpha_1 \cdots e_i\alpha_i)$. In particular, $\widehat{\beta}_\ell = \widehat{\gamma}_0 \cdots \widehat{\gamma}_{q-1}$. By convention, we define $\widehat{\beta}_0$ to be the trivial path equal to $\widehat{x}_0$. We begin with the following claim whose proof is by induction on i.

Claim *For every $i \in \{0, \dots, \ell\}$, there exists a continuous map $\widehat{f}_i\colon \Gamma_i \to \Gamma$ satisfying the following:*

(1) $f \circ \rho = \rho \circ \widehat{f}_i$,

(2) $\widehat{f}_i(\widehat{e}_1\widehat{\alpha}_1 \cdots \widehat{e}_i\widehat{\alpha}_i)$ *is homotopic relative to its endpoints to* $\widehat{\beta}_i$.

The base of induction By assumption, f fixes the vertex x_0. We put $\widehat{f}_0(\widehat{x}_0) = \widehat{x}_0$; hence the claim holds for $i = 0$.

The inductive step Assume now that the claim holds for $i \in \{0, \dots, \ell-1\}$. Our goal is to extend $\widehat{f}_i$ into a map $\widehat{f}_{i+1}\colon \Gamma_{i+1} \to \Gamma$. To that end, we need to define the restriction of $\widehat{f}_{i+1}$ to $\widehat{e}_{i+1}$ and $\widehat{G}_{i+1}$. We start with the following observation: $\widehat{f}_i(\widehat{x}_i)$ is exactly the terminal point of $\widehat{\beta}_i$. Indeed $\widehat{x}_i$ is the terminal point of $\widehat{\alpha}_i$, hence of $\widehat{e}_1\widehat{\alpha}_1 \cdots \widehat{e}_i\widehat{\alpha}_i$. According to the induction hypothesis, $\widehat{f}_i(\widehat{e}_1\widehat{\alpha}_1 \cdots \widehat{e}_i\widehat{\alpha}_i)$ is homotopic relative to its endpoints to $\widehat{\beta}_i$. In particular, they have the same terminal point, namely $\widehat{f}_i(\widehat{x}_i)$.

Let us now focus on $\widehat{e}_{i+1}$. By construction, the path γ splits as follows:

$$\gamma = (e_1\alpha_1 \cdots e_i\alpha_i) \cdot e_{i+1} \cdot \alpha_{i+1} \cdot (e_{i+2}\alpha_{i+2} \cdots e_\ell\alpha_\ell).$$

Therefore, we have

$$f_\#(\gamma) = f_\#(e_1\alpha_1 \cdots e_i\alpha_i) \cdot f(e_{i+1}) \cdot f_\#(\alpha_{i+1}) \cdot f_\#(e_{i+2}\alpha_{i+2} \cdots e_\ell\alpha_\ell).$$

In particular, $f(e_{i+1})$ is a subpath of $f_\#(\gamma)$. As we explained before, $\widehat{\gamma}_0 \cdots \widehat{\gamma}_{q-1}$ is the (unique) lift of $f_\#(\gamma)$ in Γ starting at $\widehat{x}_0$. Moreover, $\widehat{\beta}_i$ is the initial path of $\widehat{\gamma}_0 \cdots \widehat{\gamma}_{q-1}$ lifting $f_\#(e_1\alpha_1 \cdots e_i\alpha_i)$. We denote by $\widehat{\nu}$ the subpath of $\widehat{\gamma}_0 \cdots \widehat{\gamma}_{q-1}$ lifting $f(e_{i+1})$ whose initial point is the terminal point of $\widehat{\beta}_i$. As we noticed above the initial point of $\hat{\nu}$ (ie the terminal point of $\hat{\beta}_i$) is exactly $\hat{f}_i(\hat{x}_i)$. Consequently, we can extend $\hat{f}_i\colon \Gamma_i \to \Gamma$ to a continuous map $\hat{f}_{i+1}\colon \Gamma_i \cup \hat{e}_{i+1} \to \Gamma$ by sending $\hat{e}_{i+1}$ to $\hat{\nu}$.

The next step is to define the map $\hat{f}_{i+1}$ on $\hat{G}_{i+1}$. Since e_{i+1} is a red edge, its image under f starts and ends by a red edge. In particular, there exists $j \in \{1, \dots, \ell\}$ such that the last edge of $\hat{\nu}$ is $\hat{e}_j$. It follows that f maps y_{i+1} (the terminal point of e_{i+1}) to y_j (the terminal point of e_j). On the other hand, f is continuous and sends yellow edges to yellow paths. Therefore, it maps the largest yellow connected component of G containing y_{i+1} to the largest yellow connected component of G containing y_j. It provides a continuous map from $\hat{f}_{i+1}\colon \hat{G}_{i+1} \to \hat{G}_j$ such that $\hat{f}_{i+1}(\hat{y}_{i+1}) = \hat{y}_j$ and $\rho \circ \hat{f}_{i+1} = f \circ \rho$. This completes the construction for $i+1$. We end the proof of the claim with the following lemma.

Lemma 5.8 *The path $\hat{f}_{i+1}(\hat{e}_1\hat{\alpha}_1 \cdots \hat{e}_{i+1}\hat{\alpha}_{i+1})$ is homotopic relative to its endpoints to $\hat{\beta}_{i+1}$.*

Proof By construction,

$$\hat{f}_{i+1}(\hat{e}_1\hat{\alpha}_1 \cdots \hat{e}_{i+1}\hat{\alpha}_{i+1}) = \hat{f}_i(\hat{e}_1\hat{\alpha}_1 \cdots \hat{e}_i\hat{\alpha}_i)\,\hat{f}_{i+1}(\hat{e}_{i+1})\,\hat{f}_{i+1}(\hat{\alpha}_{i+1}).$$

In particular, it is homotopic relative to its endpoints to $\hat{\beta}_i\,\hat{f}_{i+1}(\hat{e}_{i+1})\,\hat{f}_{i+1}(\hat{\alpha}_{i+1})$. By construction, we also have that $\hat{\beta}_i\,\hat{f}(\hat{e}_i) = \hat{\beta}_i\hat{\nu}$ is the initial path at $\hat{x}_0$ of $\hat{\gamma}_0 \cdots \hat{\gamma}_{q-1}$ lifting $f_\#(e_1\alpha_1 \cdots e_i\alpha_i e_{i+1})$. Note also that $\hat{\beta}_i\hat{\nu}$ ends where $\hat{f}_{i+1}(\hat{\alpha}_{i+1})$ starts, namely at the point $\hat{f}_{i+1}(\hat{y}_{i+1}) = \hat{y}_j$. Thus it is sufficient to prove that $\hat{f}_{i+1}(\hat{\alpha}_{i+1})$ is homotopic relative to its endpoints to the lift starting at $\hat{f}_{i+1}(\hat{y}_{i+1})$ of $f_\#(\alpha_{i+1})$. However, these last paths all belong to $\hat{G}_j$. Moreover, the restriction of ρ to $\hat{G}_j$ is the natural embedding $\hat{G}_j \hookrightarrow G$. The conclusion follows then from the fact that $f(\alpha_{i+1})$ and $f_\#(\alpha_{i+1})$ are homotopic relative to their endpoints in the yellow connected component of G to which they belong. $\square$

Lemma 5.9 *The map $\hat{f}_\ell\colon \Gamma_\ell \to \Gamma$ induces a continuous map $\hat{f}\colon \Gamma \to \Gamma$ such that $f \circ \rho = \rho \circ \hat{f}$.*

Proof By definition, Γ is obtained from Γ_ℓ by attaching the initial point $\hat{x}_0$ of $\hat{e}_1$ to the point $\hat{x}_\ell$ of $\hat{G}_\ell$. Therefore, it is sufficient to prove that $\hat{f}_\ell(\hat{x}_\ell) = \hat{f}_\ell(\hat{x}_0)$. It follows from the first step of the construction that $\hat{f}_\ell(\hat{x}_0) = \hat{x}_0$. On the other hand, $\hat{f}_\ell(\hat{\gamma})$ and $\hat{\beta}_\ell = \hat{\gamma}_0 \cdots \hat{\gamma}_{q-1}$ are homotopic relative to their endpoints. Thus the terminal point of $\hat{\gamma}$ (ie $\hat{x}_\ell$) is sent to the terminal point of $\hat{\gamma}_{q-1}$, ie $\hat{x}_0$. Hence $\hat{f}_\ell(\hat{x}_\ell) = \hat{f}_\ell(\hat{x}_0) = \hat{x}_0$. $\square$

We can now complete the proof of Proposition 5.7. Lemma 5.9 provides the map we are looking for. The second point becomes a consequence of Lemma 5.8. $\square$

Lemma 5.10 *The map φ induces an automorphism of $\boldsymbol{H}$.*

Proof It follows from Proposition 5.7 that $\varphi(\boldsymbol{H})$ is a subgroup of $\boldsymbol{H}$. It follows from Lemma 6.0.6 in [4] that the restriction of φ to $\boldsymbol{H}$ is an automorphism. □

The abelianization of H Now we complete the proof of Proposition 5.2. Let d be the rank of the free group $\boldsymbol{H}$. We consider the abelianization morphism $\boldsymbol{H} \to \mathbb{Z}^d$. In particular, φ induces an automorphism φ_{ab} of $\mathbb{Z}^d$. We denote by $\mathcal{C}$ the image in $\mathbb{Z}^d$ of the free basis $\mathcal{B}$ of $\boldsymbol{H}$ given by Lemma 5.4. The first $(d-1)$ elements of $\mathcal{B}$ (the ones corresponding to oriented edges in $\mathcal{F}$) are conjugates of yellow loops of Γ. However, f, and thus $\widehat{f}$, maps yellow edges to yellow edges. Hence the subgroup $\mathbb{Z}^{d-1}$ generated by the first $(d-1)$ elements of $\mathcal{C}$ is invariant under φ_{ab}. By Proposition 5.7, $\widehat{f}(\widehat{\gamma})$ is homotopic relative to $\{\widehat{x}_0\}$ to $\widehat{\gamma}_0\widehat{\gamma}_1\cdots\widehat{\gamma}_{q-1}$. It follows from Lemma 5.5 that the matrix R of φ_{ab} in the basis $\mathcal{C}$ has the following shape:

$$R = \left(\begin{array}{ccc|c} \star & \cdots & \star & \star \\ \vdots & \ddots & \vdots & \vdots \\ \star & \cdots & \star & \star \\ \hline 0 & \cdots & 0 & q \end{array}\right).$$

Since $q \geqslant 2$, the determinant of R cannot be invertible in $\mathbb{Z}$, which contradicts the fact that φ_{ab} is an automorphism. We have thus proved that $\sigma^\infty(e_\bullet)$ is not shift-periodic. □

Proposition 5.11 *There exists an integer $m \geqslant 2$ such that for every $p \in \mathbb{N}$, as a word over $\mathcal{E}_\bullet$, $\mathrm{Red}(f_\#^p(e_\bullet))$ does not contains an m^{th} power.*

Proof According to Lemma 5.1, for every $p \in \mathbb{N}$, the red word associated to $f_\#^p(e_\bullet)$ is exactly $\sigma^p(e_\bullet)$. However, the substitution σ is primitive and the infinite word $\sigma^\infty(e_\bullet)$ is not shift-periodic. Hence the result follows from Proposition 2.1. □

6 The automorphism of $B_r(n)$ induced by φ

6.1 A criterion of nontriviality in $B_r(n)$

Let us have a pause in order to introduce a key ingredient for the sequel of the proof of Proposition 4.11. As explained in the introduction, we need a tool to decide whether two elements in a free Burnside group are distinct. The main theorem of [12] will play that role. In [12], Coulon considers a more general situation than the one we are interested in. Given a nonelementary torsion-free hyperbolic group G, he studies the natural projection $G \to G/G^n$, where G^n stands for the (normal) subgroup of G

generated by the n^{th} power of all elements of G. He provides a criterion to decide whether two elements $g, g' \in G$ have the same image in the quotient G/G^n. For our purpose, we focus on the case where G is a free group. This is the situation that we describe below.

Let (X, x_0) be a pointed simplicial tree. Given two points x and x' of X, we denote by $|x-x'|$ the distance between them, whereas $[x, x']$ stands for the geodesic joining x and x'. Let g be an isometry of X. Its *translation length*, denoted by $\|g\|$, is the quantity $\|g\| = \inf_{x \in X} |gx-x|$. If X is the Cayley graph of $\boldsymbol{F}_r$, then $\|g\|$ is exactly the length of the conjugacy class of $g \in \boldsymbol{F}_r$. The set of points $A_g = \{x \in X \mid |gx-x| = \|g\|\}$ is called the *axis* of g. It is a subtree of X. It is known that either $\|g\| = 0$ and A_g is the set of fixed points of g, or $\|g\| > 0$ and A_g is a bi-infinite geodesic on which g acts by translation of length $\|g\|$. In the first case, g is said to be *elliptic*, in the second one *hyperbolic*. For more details, we refer the reader to [16]. We now assume that $\boldsymbol{F}_r$ acts by isometries on X.

Definition 6.1 Let $n \in \mathbb{N}$ and $\xi \in \mathbb{R}_+$. Let y and z be two points of X. We say that z is the image of y by an *(n, ξ)–elementary move* (or simply *elementary move*) if there is a hyperbolic element $u \in \boldsymbol{F}_r$ such that

(1) $\operatorname{diam}\big([x_0, y] \cap A_u\big) \geqslant \big(\frac{n}{2} - \xi\big)\|u\|$,

(2) $z = u^{-n} y$.

The point z is the image of y by a *sequence of (n, ξ)–elementary moves* if there is a finite sequence $y = y_0, y_1, \dots, y_\ell = z$ such that for all $i \in \{0, \dots, \ell - 1\}$, the point y_{i+1} is the image of y_i by an (n, ξ)–elementary move.

Knowing that the hyperbolicity constant of a tree is zero, this notion of (n, ξ)–elementary move is exactly the one defined in [12]. The next statement is a particular case of the main theorem of [12] when the group G is free and the underlying space X is a tree.

Proposition 6.2 (Coulon [12]) *Assume that $\boldsymbol{F}_r$ acts properly cocompactly by isometries on (X, x_0). There exist $n_1 \in \mathbb{N}$ and $\xi \in \mathbb{R}_+$ such that for every odd exponent $n \geqslant n_1$ the following holds: two isometries $g, g' \in \boldsymbol{F}_r$ have the same image in $\boldsymbol{B}_r(n)$ if and only if there exist two finite sequences of (n, ξ)–elementary moves which respectively send gx_0 and $g'x_0$ to the same point of X.*

Remark 6.3 Roughly speaking, an elementary move allows us to replace a subword of the form v^m by v^{m-n} provided m is sufficiently large. Assume indeed that (X, x_0) is the Cayley graph of $\boldsymbol{F}_r$ pointed at the identity element of $\boldsymbol{F}_r$. There is a natural one-to-one correspondence between reduced words and geodesics of X starting at x_0. More precisely, given an element $g \in \boldsymbol{F}_r$, the reduced word w which represents g labels the geodesic between x_0 and gx_0. Let us suppose now that w can be written (as a reduced word) $w = pv^m s$ with $m \geqslant \frac{n}{2} - \xi$. It follows that

$$\operatorname{diam}\big([x_0, gx_0] \cap A_u\big) \geqslant \|u^m\| \geqslant \big(\tfrac{n}{2} - \xi\big)\|u\|,$$

where u is the element of $\boldsymbol{F}_r$ represented by pvp^{-1}. Thus $u^{-n}g$, which is represented by $pv^{m-n}s$, is the image of g by an elementary move. With this dictionary in mind, Theorem 1.5 becomes a direct application of Proposition 6.2, where (X, x_0) is the Cayley graph of $\boldsymbol{F}_r$ based at 1.

Later in the proof, the tree X will be the universal cover of an RTT. Therefore, this formulation, which extends the idea of substituting subwords, is more appropriate for our purpose.

Proposition 6.2 provides in particular a criterion for detecting trivial elements in $\boldsymbol{B}_r(n)$.

Corollary 6.4 *Assume that $\boldsymbol{F}_r$ acts properly cocompactly by isometries on (X, x_0). There exist $n_1 \in \mathbb{N}$ and $\xi \in \mathbb{R}_+$ such that for every odd exponent $n \geqslant n_1$, the following holds: an element $g \in \boldsymbol{F}_r$ is trivial in $\boldsymbol{B}_r(n)$ if and only if there exists a finite sequence of (n, ξ)–elementary moves which sends gx_0 to x_0.*

However, despite the similarity with the word problem in a group, Corollary 6.4 is not equivalent to Proposition 6.2. This comes from the fact that (n, ξ)–elementary moves are not symmetric. One first has to see a large power along the geodesic $[x_0, gx_0]$ before performing an elementary move. For instance, if a and b are two distinct primitive elements of $\boldsymbol{F}_r$, there is no sequence of elementary moves that sends a^n to b^n. Corollary 6.4 only implies a weaker form of Proposition 6.2 in the sense that we need to allow a larger class of elementary moves: those of the form $pv^m s \to pv^{m-n}s$ with $m \geqslant \frac{n}{4} - \frac{\xi}{2}$.

In our situation, we will apply this criterion with two elements of the form g and $\varphi^p(g)$, where $g \in \boldsymbol{F}_r$ and φ is the automorphism we want to study. The theory of train-track provides much information about the path $[x_0, \varphi^p(g)x_0]$. Therefore, it is also more natural to have a criterion that uses conditions on $[x_0, gx_0]$ and $[x_0, \varphi^p(g)x_0]$ rather than $[gx_0, \varphi^p(g)x_0]$.

Remark Proposition 6.2 is "well known" to the experts of Burnside's groups. To our knowledge, it has never been formulated in such a level of simplicity, though. The reader can, for instance, compare our definition of elementary moves with the one of *simple r–reversal of rank α* used by Adian; see [1, Section 4.18, pages 8–16] for the prerequisites.

After consulting Adian and Ol'shanskiĭ, it seems that the closest published statements to Proposition 6.2 are [1, Chapter VI, Lemma 2.8] and [32, Lemma 5.5]. They should lead to a similar result, but with a weaker requirement to perform elementary moves. Adian's approach would provide an analogue of Theorem 1.5 with a sharper critical exponent ($n_1 = 667$) but where an elementary move is allowed as soon as $m \geqslant 90$ (instead of $m \geqslant \frac{n}{2} - \xi$). This is unfortunately not enough for our purpose. In the Appendix, we explain how our results on $\mathrm{Out}(\boldsymbol{B}_r(n))$ can be proved using Ol'shanskiĭ's work instead of Proposition 6.2. In particular, we prove an analogue of Theorem 1.5 where elementary moves are allowed as soon as $m \geqslant \frac{n}{3}$; see Proposition A.2.

6.2 Performing elementary moves in $\widetilde{G}$

We get back to the proof of Proposition 4.11. The notation is the same as in Section 5.

Metrics on $\widetilde{G}$ For our purpose, the pointed tree (X, x_0) that appears in Proposition 6.2 will be the universal cover $(\widetilde{G}, \widetilde{x}_0)$ of G where $\widetilde{x}_0$ is preimage of x_0. By declaring that any edge of $\widetilde{G}$ is isometric to the unit real segment $[0,1]$, we obtain an $\boldsymbol{F}_r$–invariant length metric on $\widetilde{G}$: the *combinatorial metric*. We denote by $|\alpha|$ the resulting *combinatorial length* of a path α in $\widetilde{G}$.

We also define a pseudolength metric on $\widetilde{G}$ in the following way. We first consider that any yellow edge has length zero. Recall that $\mathcal{E}$ is the set of all the oriented red edges of G. We chose a preferred set of oriented edges $\vec{\mathcal{E}}$. Recall that the transition matrix M of the red stratum of f is aperiodic. We denote by $\lambda > 1$ the Perron–Frobenius dominant eigenvalue of M, and we consider a positive right eigenvector $l = (l_e)_{e \in \vec{\mathcal{E}}}$ associated to λ. We declare the lifts of e isometric to the real segment $[0, l_e]$. The resulting pseudometric is called the *PF-pseudometric*. We denote by $|\alpha|_{\mathrm{PF}}$ the resulting length of the path α in $\widetilde{G}$: this is called the *PF-length* of α. This length only depends on the red word $\mathrm{Red}(\alpha) \in \mathcal{E}^*$. If α is a red-legal path, we thus get that for all $p \in \mathbb{N}$,

$$|f_\#^p(\alpha)|_{\mathrm{PF}} = \lambda^p |\alpha|_{\mathrm{PF}}.$$

Unless stated otherwise we will work with $\widetilde{G}$ endowed with the combinatorial metric.

The element g and its orbit Recall that $e_\bullet$ is the red edge fixed at the beginning of Section 5. For all $p \in \mathbb{N}$, we have that $f_\#^p(e_\bullet)$ is a path starting by $e_\bullet$. Its yellow-red decomposition is a splitting. The red stratum H is aperiodic. Thus if p is a sufficiently large integer, one can find another occurrence of $e_\bullet$ in $f_\#^p(e_\bullet)$: namely $f_\#^p(e_\bullet) = e_\bullet \nu_0 e_\bullet \nu_1$. The path $\nu = e_\bullet \nu_0$ is a red-legal circuit, and the yellow-red decomposition of ν is a splitting. We denote by g the element of $\pi_1(G, x_0)$ represented by ν. By construction, the geodesic $[\widetilde{x}_0, g\widetilde{x}_0]$ is the lift in $\widetilde{G}$ of ν starting at $\widetilde{x}_0$.

Lemma 6.5 *There exists an integer n_2 with the following property. Let $p \in \mathbb{N}$. Let β be a path of $\widetilde{G}$ such that the red words respectively associated to β and $f_{\#}^{p}(\nu)$ are the same. For all $u \in \boldsymbol{F}_r \setminus \{1\}$, if*

$$\operatorname{diam}(\beta \cap A_u) > n_2 \|u\|,$$

then the axis of u only contains yellow edges.

Proof By construction, there exists $p_0 \in \mathbb{N}$ such that ν is a prefix of $f_{\#}^{p_0}(e_\bullet)$. More generally, $f_{\#}^{p}(\nu)$ is a prefix of $f_{\#}^{p+p_0}(e_\bullet)$ for every $p \in \mathbb{N}$. According to Proposition 5.11, there exists $m \in \mathbb{N}$ such that for every $p \in \mathbb{N}$, the red word associated to $f_{\#}^{p}(\nu)$ does not contain an m^{th} power. Put $n_2 = m + 2$. Note that n_2 does not depend on the path β. Let u be a nontrivial element of $\boldsymbol{F}_r$ such that

$$\operatorname{diam}(\beta \cap A_u) > n_2 \|u\|.$$

In particular, there is a vertex $x \in A_u$ such that for every $j \in \{0, \dots, m\}$, the point $u^j x$ belongs to β. Assume now that A_u contains a red edge e. Since A_u is a u–invariant bi-infinite geodesic, the geodesic $[x, ux]$ contains some red edges. In particular, if α stands for the path $[x, u^m x]$, then $\operatorname{Red}(\alpha)$ contains an m^{th} power. However, β is a path of $\widetilde{G}$. Consequently, $[x, u^m x]$ is a subset, hence a subpath, of β. Therefore, the red word associated to β, and thus to $f_{\#}^{p}(\nu)$, contains an m^{th} power. This is a contradiction. □

We finish this section with the proof of Proposition 4.11.

Proof of Proposition 4.11 Recall that g is the element of $\pi_1(G, x_0)$ represented by the red legal circuit $\nu = e_\bullet \nu_0$. Our goal is to prove that for sufficiently large odd integers n, the sequence $(\varphi^p(g))_{p\in\mathbb{N}}$ of elements of $\boldsymbol{F}_r$ is embedded in $\boldsymbol{B}_r(n)$. Since φ is an automorphism, it is sufficient to check that $\varphi^p(g) \not\equiv g$ in $\boldsymbol{B}_r(n)$ for all $p \in \mathbb{N}^*$. We are going to use the criterion of Section 6.1. Recall that the geodesic $[\tilde{x}_0, g\tilde{x}_0]$ is a lift in $\widetilde{G}$ of ν. We denote by n_1, ξ and n_2 the constants given, accordingly, by Proposition 6.2 and Lemma 6.5. For the rest of the proof, we fix an odd integer n larger than

$$n_0 = \max\bigl\{n_1,\ 2n_2 + 2\xi + 1,\ 2|\tilde{x}_0 - g\tilde{x}_0| + 2\xi + 1\bigr\}.$$

Note that this lower bound only depends on the outer automorphism Φ and the RTT f.

Let $p \in \mathbb{N}^*$. By construction, the path $\beta = [\tilde{x}_0, \varphi^p(g)\tilde{x}_0]$ is a lift of $f_{\#}^{p}(\nu)$. Assume now that $\varphi^p(g) \equiv g$ in $\boldsymbol{B}_r(n)$. By Proposition 6.2, there exist two sequences of (n,ξ)–elementary moves which respectively send $g\tilde{x}_0$ and $\varphi^p(g)\tilde{x}_0$ to the same point of $\widetilde{G}$. However, we fixed $n > 2|\tilde{x}_0 - g\tilde{x}_0| + 2\xi$. Therefore, no (n,ξ)–elementary move can be performed on $[\tilde{x}_0, g\tilde{x}_0]$. It follows that there exists a sequence of (n,ξ)–elementary moves which sends $\varphi^p(g)\tilde{x}_0$ to $g\tilde{x}_0$. We denote by β_i the reduced path obtained from β after the i^{th} (n,ξ)–elementary move. In particular, $\beta_0 = \beta$. Note that the initial point of β_i is always $\tilde{x}_0$. Recall that $\boldsymbol{F}_r^n$ is the normal subgroup of $\boldsymbol{F}_r$ generated by the n^{th} power of every element. We are going to show, by induction on i, that

(H1) the endpoints of a maximal yellow subpath of β_i are not in the same $\boldsymbol{F}_r^n$–orbit,

(H2) $\operatorname{Red}(\beta_i) = \operatorname{Red}(\beta)$.

The base of induction Let α be a maximal yellow subpath of β_0. Recall that β_0 is a lift of $f_\#^p(\nu)$. On the other hand, $f_\#^p(\nu)$ is a subpath of $f_\#^q(e_\bullet)$ for some $q \in \mathbb{N}$. It follows from our assumption that α is not mapped by $\widetilde{G} \twoheadrightarrow G$ to a loop. In particular, its endpoints are not in the same $\boldsymbol{F}_r$–orbit, which provides (H1). Assertion (H2) is obvious.

The inductive step Assume that these two conditions hold for i. For simplicity, we denote by $\tilde{z}_i$ the terminal point of β_i; hence $\beta_i = [\tilde{x}_0, \tilde{z}_i]$. We focus on the $(i+1)^{\text{st}}$ elementary move. Let us denote by A_u the axis of the elementary move performed on β_i. In particular, $\operatorname{diam}(\beta_i \cap A_u) \geqslant \left(\frac{n}{2} - \xi\right)\|u\|$ in $\widetilde{G}$. By hypothesis (H2), the red words associated to β_i, β and $f_\#^p(\nu)$ are the same. By Lemma 6.5, the axis A_u only contains yellow edges. In particular, A_u crosses β_i along (a part of) a maximal yellow subpath of β_i that we denote by α; see Figure 5.

Let $\tilde{y}$ and $\tilde{y}'$ be the respective initial and terminal points of α. By (H1), $\tilde{y} \neq u^{-n}\tilde{y}'$. Recall that the action of $\boldsymbol{F}_r$ on $\widetilde{G}$ respects the yellow-red decomposition. Consequently, the path β_{i+1} is exactly

$$\beta_{i+1} = [\tilde{x}_0, \tilde{y}] \cup [\tilde{y}, u^{-n}\tilde{y}'] \cup [u^{-n}\tilde{y}', u^{-n}\tilde{z}_i].$$

In particular, $\operatorname{Red}(\beta_{i+1}) = \operatorname{Red}(\beta_i)$. Combined with (H2), we get $\operatorname{Red}(\beta_{i+1}) = \operatorname{Red}(\beta)$, which corresponds to (H2) at step $i+1$. The maximal yellow subpaths of β_{i+1} are of three kinds:

- the ones of $[\tilde{x}_0, \tilde{y}]$ which are actually maximal yellow subpaths of β_i,
- the ones of $[u^{-n}\tilde{y}', u^{-n}\tilde{z}_i]$ which are translates of maximal yellow subpaths of β_i,
- the geodesic $[\tilde{y}, u^{-n}\tilde{y}']$.

By (H1), the endpoints of any maximal yellow path of the first two kinds are not in the same $\boldsymbol{F}_r^n$–orbit. Thus $\tilde{y}$ and $\tilde{y}'$ are not in the same $\boldsymbol{F}_r^n$–orbit, being the endpoints of α. Hence neither are $\tilde{y}$ and $u^{-n}\tilde{y}'$. This gives (H2) at step $i+1$, which completes the induction.

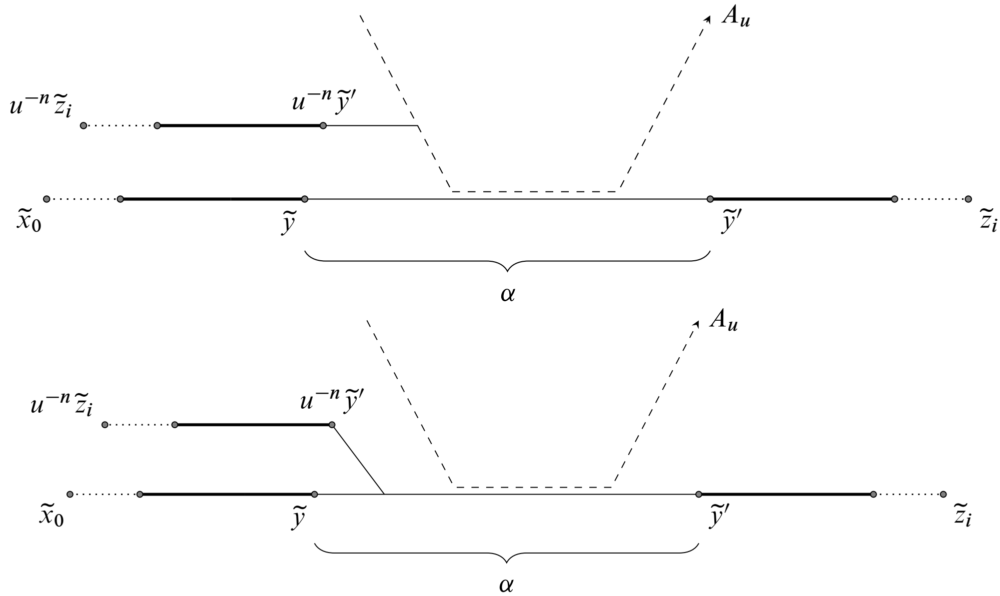


Figure 5: Performing a move on β_i, the two possible configurations. The thin lines refer to yellow paths, the thick ones to red paths. Top: α does not contain the full n^{th} power of u. Bottom: α contains the full n^{th} power of u, but cannot be totally removed.

Recall that (β_i) is the collection of paths obtained by the sequence of elementary moves which sends $\varphi^p(g)$ to g. It follows from the previous discussion that at each step i, $|\beta_i|_{\mathrm{PF}} = |\beta|_{\mathrm{PF}}$. In particular, $|f_\#^p(\nu)|_{\mathrm{PF}} = |\beta|_{\mathrm{PF}} = |\nu|_{\mathrm{PF}}$. However, we build ν in such a way that $|f_\#^p(\nu)|_{\mathrm{PF}} = \lambda^p|\nu|_{\mathrm{PF}}$. This contradicts our original assumption. Therefore, $\varphi^p(g) \not\equiv g$ in $\boldsymbol{B}_r(n)$ for every $p \in \mathbb{N}$. In particular, φ (respectively Φ) induces an automorphism (respectively outer automorphism) of $\boldsymbol{B}_r(n)$ of infinite order. $\square$

7 Comments and questions

7.1 About other possible strategies of proof

In the introduction, we recalled the argument given by Cherepanov. It is easy to elaborate a generalization to a wider class of automorphisms which does not require the criterion stated in Proposition 6.2.

An outer automorphism $\Phi \in \mathrm{Out}(\boldsymbol{F}_r)$ is *irreducible with irreducible powers* (or simply *iwip*) if there is no (conjugacy class of a) proper free factor of $\boldsymbol{F}_r$ which is invariant by some positive power of Φ. An iwip outer automorphism can be represented by an (absolute) train-track map $f\colon G \to G$ with a primitive transition matrix [6]. Roughly speaking, it implies that there are no "yellow strata" which were the ones responsible for having large powers in our words. As a particular case of Proposition 5.11, there exists a loop γ in G and an integer n_2 with the following property. For every $p \in \mathbb{N}$, the word labeling the loop $f_\#^p(\gamma)$ does not contain an n_2^{th} power (as a complete word, not just its red part). Consequently, Proposition 1.4 is sufficient to conclude. Note also that, in this context, Proposition 5.11 can be proved in a much easier way by using either the action of $\boldsymbol{F}_r$ on the stable tree associated to Φ [18, Theorem 2.1] or the fact that the attracting laminations of Φ cannot be carried by a subgroup of rank 1 [3, Proposition 2.4].

However, as we explained in the introduction, there exist automorphisms for which one cannot use the same strategy. Consider, for instance, the automorphism ψ of $\boldsymbol{F}_4 = \boldsymbol{F}(a,b,c,d)$ defined in the introduction by $\psi(a) = a$, $\psi(b) = ba$, $\psi(c) = cbcd$ and $\psi(d) = c$. One can view ψ as a relative train-track map on the rose: there is only one exponential stratum (the "red stratum" which corresponds to the free factor $\langle c, d\rangle$) and the restriction of ψ to $\langle a, b\rangle$ has polynomial growth (and $\langle a, b\rangle$ gives rise to a "yellow stratum"). We saw that a^{p-1} occurs as subword of $\psi^p(d)$. Nevertheless, we still do not need Proposition 6.2 to conclude here that the automorphism ψ satisfies the statement of Theorem 1.3. It is sufficient to pass to the quotients of $\boldsymbol{F}_r$ and $\boldsymbol{B}_r(n)$ by the normal subgroup generated by a and b, and then to argue as previously.

Nevertheless, given an arbitrary automorphism, this trick (passing to a well chosen quotient) seems to be less easy to run. Look at the automorphism ψ of $\boldsymbol{F}_4 = \boldsymbol{F}(a,b,c,d)$ defined by

$$\psi\colon a \mapsto a, \quad b \mapsto ba, \quad c \mapsto cd^{-1}bd, \quad d \mapsto dcd^{-1}bd.$$

This automorphism grows exponentially. However, if one considers the quotient of $\boldsymbol{F}_4$ by the normal subgroup generated by a and b, it induces the Dehn twist $c \mapsto c$, $d \mapsto dc$, which has finite order as an automorphism of $\boldsymbol{B}_2(n)$.

Let φ be an automorphism of $\boldsymbol{F}_r$. The geometry of the suspension $\boldsymbol{F}_r \rtimes_\varphi \mathbb{Z}$ might provide an alternative proof of Theorem 1.3. In [13], the first author solved indeed the case where $\boldsymbol{F}_r \rtimes_\varphi \mathbb{Z}$ is a hyperbolic group. Generalizing the Delzant–Gromov approach of the Burnside Problem, he constructed a sequence of groups $\boldsymbol{H}_j$ with $\varinjlim \boldsymbol{H}_j = \boldsymbol{B}_r(n)$ such that for every j,

- φ induces an automorphism of infinite order of $\boldsymbol{H}_j$,
- $\boldsymbol{H}_j \rtimes_\varphi \mathbb{Z}$ is a hyperbolic group obtained from $\boldsymbol{H}_{j-1} \rtimes_\varphi \mathbb{Z}$ by small cancellation.

It follows from the hyperbolicity that φ induces an automorphism of infinite order of $\boldsymbol{B}_r(n)$.

If φ is an arbitrary exponentially growing automorphism, then $\boldsymbol{F}_r \rtimes_\varphi \mathbb{Z}$ is no more hyperbolic. However, F Gautero and M Lustig proved that $\boldsymbol{F}_r \rtimes_\varphi \mathbb{Z}$ is hyperbolic relatively to a family of subgroups which consists of conjugacy classes that grow polynomially under iteration by φ [19; 20]. Therefore, one could use a generalization of the iterated small cancellation theory to relative hyperbolic groups. We refer the reader to [14] for a detailed presentation of the Delzant–Gromov approach to the Burnside problem and to [15] for a generalization. See also [34] for a theory of small cancellation in relatively hyperbolic groups.

7.2 Quotients of $\mathrm{Out}(\boldsymbol{F}_r)$

The following remark is due to M Sapir. Proposition 4.1 says that for every integer $n \geqslant 1$, polynomially growing automorphisms of $\boldsymbol{F}_r$ induce automorphisms of finite order of $\boldsymbol{B}_r(n)$. More precisely, their orders divide

$$p(r,n) = n^{2(2^{r-1}-1)}.$$

Let us denote by $\mathcal{Q}_{r,n}$ the quotient of $\mathrm{Out}(\boldsymbol{F}_r)$ by the (normal) subgroup generated by

$$\{\Phi^{p(r,n)} \mid \Phi \in \mathrm{Out}(\boldsymbol{F}_r) \text{ polynomially growing}\}.$$

In particular, the $p(r,n)^{\mathrm{th}}$ power of the Nielsen transformations which generate $\mathrm{Out}(\boldsymbol{F}_r)$ are trivial in $\mathcal{Q}_{r,n}$. It follows from Proposition 4.1 that the map $\mathrm{Out}(\boldsymbol{F}_r) \to \mathrm{Out}(\boldsymbol{B}_r(n))$ induces a natural map $\mathcal{Q}_{r,n} \to \mathrm{Out}(\boldsymbol{B}_r(n))$. Therefore, we have the following results:

Theorem 7.1 *Let $r \geqslant 3$. There exists n_0 such that for all odd integers $n \geqslant n_0$, the group $\mathcal{Q}_{r,n}$ contains copies of $\boldsymbol{F}_2$ and $\mathbb{Z}^{\lfloor r/2 \rfloor}$.*

Proof This is a consequence of [13] Theorems 1.8 and 1.10. □

Theorem 7.2 *Let Φ be an outer automorphism of $\boldsymbol{F}_r$. The following assertions are equivalent:*

(1) *Φ has exponential growth;*

(2) *there exists $n \in \mathbb{N}$ such that the image of Φ in $\mathcal{Q}_{r,n}$ has infinite order;*

(3) *there exist $\kappa, n_0 \in \mathbb{N}$ such that for all odd integers $n \geqslant n_0$, the image of Φ in $\mathcal{Q}_{r,\kappa n}$ has infinite order.*

Proof This is a consequence of our main theorem. □

7.3 Exponentially growing automorphisms of the free group can have finite order in a free Burnside group

The constant n_0 in Theorem 1.3 does depend on the outer automorphism $\Phi \in \mathrm{Out}(\boldsymbol{F}_r)$. Indeed, we give in this section explicit examples of automorphisms in the kernel of the natural map $\mathrm{Aut}(\boldsymbol{F}_r) \to \mathrm{Aut}(\boldsymbol{B}_r(n))$ which have exponential growth. In particular, there are iwip automorphisms in this kernel.

7.3.1 A first family of examples An outer automorphism $\Phi \in \mathrm{Out}(\boldsymbol{F}_r)$ induces, by abelianization, an automorphism of $\mathbb{Z}^r$. This defines a homomorphism $\mathrm{Out}(\boldsymbol{F}_r) \to \mathrm{GL}(r, \mathbb{Z})$, $\Phi \mapsto M_\Phi$. Nielsen proved that for $r = 2$, this morphism is an isomorphism [28]. Moreover, Φ has exponential growth if and only if the absolute value of the trace of $M_\Phi^2 \in \mathrm{GL}(2, \mathbb{Z})$ is larger than 2.

Examples Let $\{a, b\}$ be a basis of the free group $\boldsymbol{F}_2$. For $n \in \mathbb{N}^*$, we define $\varphi_n \in \mathrm{Aut}(\boldsymbol{F}_2)$ by $\varphi_n(a) = a(ba^n)^n$, $\varphi_n(b) = ba^n$. We denote by Φ_n the corresponding outer class in $\mathrm{Out}(\boldsymbol{F}_2)$. The outer class Φ_n has exponential growth since the trace of $M_{\Phi_n}^2$ equals $n^4 + 4n^2 + 2$. However, the outer automorphism of $\boldsymbol{B}_2(n)$ induced by Φ_n is the identity.

For $r > 2$, we consider a splitting of $\boldsymbol{F}_r$ as a free product $\boldsymbol{F}_r = \boldsymbol{F}_2 * \boldsymbol{F}_{r-2}$. For $n \in \mathbb{N}^*$, we consider the automorphism $\psi_n = \varphi_n * \mathrm{Id}$ which is equal to φ_n (defined in the previous paragraph) when restricted to the first factor of the splitting and to the identity when restricted to the second factor. Again, the outer class Ψ_n of ψ_n has exponential growth (since Φ_n has), but the outer automorphism of $\boldsymbol{B}_r(n)$ induced by Ψ_n is the identity.

These examples show that the constant n_0 in Theorem 1.3 is not uniform: it does depend on the outer class $\Phi \in \mathrm{Out}(\boldsymbol{F}_r)$. The automorphisms Φ_n are iwip automorphisms. But this is not the case of the automorphisms Ψ_n. We fix this point in the next subsection.

7.3.2 Iwip automorphisms of $\boldsymbol{F}_r$ trivial in $\mathrm{Out}(\boldsymbol{B}_r(n))$ To produce iwip automorphisms in the kernel of the canonical map $\mathrm{Out}(\boldsymbol{F}_r) \to \mathrm{Out}(\boldsymbol{B}_r(n))$, one can follow the idea of W Thurston to generically produce pseudo-Anosov homeomorphisms of a surface by composing well chosen Dehn twist homeomorphisms [37].

In the context of automorphisms of free groups, there is a notion of a Dehn twist (outer) automorphism (see for instance [10]) which generalizes the notion of a Dehn twist homeomorphism of a surface: Example 4.3 provides such a Dehn twist automorphism. In [9], M Clay and A Pettet explain how to generate iwip automorphisms of $\boldsymbol{F}_r$ by composing two Dehn twist automorphisms associated to a filling pair of cyclic splittings of $\boldsymbol{F}_r$. We will not explicitly state these definitions here. For our purpose, we only need to know that

- Dehn twist automorphisms have polynomial growth (in fact linear growth), and
- there exist Dehn twist automorphisms $\Delta_1, \Delta_2 \in \mathrm{Out}(\boldsymbol{F_r})$ satisfying the hypothesis of the following theorem.

Theorem 7.3 (Clay and Pettet [9, Theorem 5.3]) *Let $\Delta_1, \Delta_2 \in \mathrm{Out}(\boldsymbol{F_r})$ be the Dehn twist outer automorphisms for a filling pair of cyclic splittings of $\boldsymbol{F_r}$. There exists $N \in \mathbb{N}$ such that for every $p, q > N$,*

- *the subgroup of $\mathrm{Out}(\boldsymbol{F_r})$ generated by Δ_1^p and Δ_2^q is a free group of rank 2,*
- *if $\Phi \in \langle \Delta_1^p, \Delta_2^q \rangle$ is not conjugate to a power of either Δ_1^p or Δ_2^q, then Φ is an iwip outer automorphism.*

We fix an exponent $n \in \mathbb{N}$. We consider two such Dehn twist outer automorphisms Δ_1 and Δ_2, and the integer $N \in \mathbb{N}$ given by Theorem 7.3. Since Δ_1 and Δ_2 have polynomial growth, they induce an automorphism of finite order of $\boldsymbol{B_r}(n)$. In particular, there exists $p > N$ such that $\Phi = \Delta_1^p \Delta_2^p$ is in the kernel of the map $\mathrm{Out}(\boldsymbol{F_r}) \to \mathrm{Out}(\boldsymbol{B_r}(n))$. However, Theorem 7.3 ensures that Φ is an iwip outer automorphism of $\boldsymbol{F_r}$.

7.4 Growth rates in $\mathrm{Out}(\boldsymbol{F_r})$ and $\mathrm{Out}(\boldsymbol{B_r}(n))$

Let Φ be an exponentially growing automorphism of $\boldsymbol{F_r}$. Our study in Section 6 seems to indicate that for odd exponents n large enough, some structure of Φ is preserved in $\boldsymbol{B_r}(n)$. Therefore, we wonder how much information could be carried through the map $\mathrm{Out}(\boldsymbol{F_r}) \to \mathrm{Out}(\boldsymbol{B_r}(n))$. In particular, what can we say about the growth rate of Φ?

Let $\boldsymbol{G}$ be a group generated by a finite set S. We endow $\boldsymbol{G}$ with the word-metric with respect to S. The length of the conjugacy class of $g \in \boldsymbol{G}$, denoted by $\|g\|$, is the length of the shortest element conjugated to g. An outer automorphism Φ of $\boldsymbol{G}$ naturally acts on the set of conjugacy classes of $\boldsymbol{G}$. Consequently, as in the free group, one can define the *(exponential) growth rate* of Φ by

$$\mathrm{EGR}(\Phi) = \sup_{g \in \boldsymbol{G}} \limsup_{p \to +\infty} \sqrt[p]{\|\Phi^p(g)\|}.$$

Since the word-metrics for two distinct finite generating sets of $\boldsymbol{G}$ are bi-Lipschitz equivalent, this rate does not depend on S. The automorphism Φ is said to have *exponential growth* if $\mathrm{EGR}(\Phi) > 1$.

By our knowledge, it is not known if there exist outer automorphisms of Burnside groups with exponential growth. We would like to ask the following questions:

- Are there automorphisms of $\boldsymbol{B}_r(n)$ with exponential growth?
- Let $\Phi \in \mathrm{Out}(\boldsymbol{F}_r)$ with exponential growth. Is there an integer n_0 such that for all (odd) exponents $n \geqslant n_0$, the automorphism $\widehat{\Phi}_n$ of $\boldsymbol{B}_r(n)$ induced by Φ has exponential growth? Such that $\mathrm{EGR}(\widehat{\Phi}_n) = \mathrm{EGR}(\Phi)$?
- Are there automorphisms of $\boldsymbol{B}_r(n)$ of infinite order which do not have exponential growth?

On the other hand, it could be very interesting to understand to what extent the structure of the attracting laminations associated to an outer automorphism of $\boldsymbol{F}_r$ is preserved in $\boldsymbol{B}_r(n)$. Recall that theses laminations are the fundamental tool used by Bestvina, Feighn and Handel to prove that $\mathrm{Out}(\boldsymbol{F}_r)$ satisfies the Tits alternative [4; 5].

Appendix

Proposition 6.2 can be seen as a weak form of a Dehn algorithm associated to the following presentation of the free Burnside group:

$$\boldsymbol{B}_r(n) = \langle a_1, \dots, a_r \mid x^n = 1 \text{ for all } x \rangle. \tag{4}$$

Let w be a reduced word over the alphabet $\{a_1, \dots a_r\}$. If w contains a subword v corresponding to almost half a relation from (4), we allow v to be replaced, in w, by its complement. Proposition 6.2 states that if w represents the trivial element, then after finitely many steps we will get the trivial word.

For our purpose, we actually do not need such a strong statement. The aim of this appendix is to explain how Theorem 1.3 can be proved using only Ol'shanskiĭ's work on free Burnside groups [32]. It might be possible to proceed in the same way using the Novikov–Adian approach [1]. The exposition would, however, be more technical. We first recall some results of Ol'shanskiĭ, and then list the modifications that need to be made to our original proof of Proposition 4.11.

Let (X, x_0) be a pointed simplicial tree endowed with an action by isometries of $\boldsymbol{F}_r$.

Definition A.1 Let $n \in \mathbb{N}^*$ and $c \in (0,1)$. Let y and z be two points of X. We say that z is the image of y by an (c,n)*–weak elementary move* if there is a hyperbolic element $u \in \boldsymbol{F}_r$ such that

(1) $\mathrm{diam}\big([x_0, y] \cap A_u\big) \geqslant cn\|u\|$,

(2) $z = u^{-n} y$.

The point z is the image of y by a *sequence of* (c,n)*–weak elementary moves* if there is a finite sequence $y = y_0, y_1, \dots, y_\ell = z$ such that for all $i \in \{0, \dots, \ell-1\}$, the point y_{i+1} is the image of y_i by a (c,n)–weak elementary move.

Remark As we recalled previously, in our original framework, we allowed a regular move to be performed if $[x_0, y]$ contained almost half of a relation. Here we relax this condition: one can perform a weak move even if $[x_0, y]$ contains a (much) smaller ratio of a relation. The allowed ratio is given by c. In practice we will always have $c \leqslant \frac{1}{3}$.

Let us focus first on the case where X is the Cayley graph of $\boldsymbol{F}_r$ with respect to the free basis $\{a_1, \dots, a_r\}$. To avoid any ambiguity, we denote it by T. Let t_0 be the vertex of T corresponding to the identity. The following result is a consequence of Ol'shanskiĭ's work [32].

Proposition A.2 *Let $n > 10^{10}$ be an odd integer. An element $g \in \boldsymbol{F}_r$ is trivial in $\boldsymbol{B}_r(n)$ if and only if there exists a finite sequence of $\left(\frac{1}{3}, n\right)$–weak elementary moves which sends gt_0 to t_0.*

Remark The proof below relies on Ol'shanskiĭ's diagrammatical approach of the Burnside problem. To keep the appendix short, we do not recall all the necessary background on diagrams. In particular, we use the vocabulary and notations of [32] without any further explanation. For an extensive introduction to diagrams, we refer the reader to [33].

Proof The "if" part directly follows from the definition of weak elementary moves. Let us focus on the "only if" part. Let (C_j) be the system of independent relations of $\boldsymbol{B}_r(n)$ defined in [32, page 203]:

$$\boldsymbol{B}_r(n) = \langle a_1, \dots, a_r \mid C_1^n, C_2^n, \dots, C_i^n, \dots \rangle.$$

Let $g \in \boldsymbol{F}_r \setminus \{1\}$ whose image in $\boldsymbol{B}_r(n)$ is trivial. Let w be the noncontractible word over the group alphabet $\{a_1^{\pm 1}, \dots, a_r^{\pm 1}\}$ representing g. As g is trivial in $\boldsymbol{B}_r(n)$, there exists $i \geqslant 1$ such that g is trivial in $\langle a_1, \dots, a_r \mid C_1^n, C_2^n, \dots, C_i^n \rangle$. In other words, w labels the contour of a diagram of rank i that we denote by Δ. Without loss of generality, we can assume that Δ is minimal [32, page 205].

The next arguments are a variation of [32, Lemma 5.5]. This lemma covers indeed two cases, where Δ is a *disc* diagram or an *annular* diagram and requires therefore the label of Δ to be cyclically noncontractible. In our case, we only need to consider disc diagrams, hence the assumption that w is noncontractible will be sufficient. Since w is not trivial, Δ contains a least one cell. By [32, Lemma 3.6], Δ admits a θ–cell. Recall that $\theta = 0.985$, while $\gamma = 10^{-6}/7$. Consequently, its degree of contiguity to a section p of the contour of Δ is at least $\frac{1}{3} + 400\gamma$. Let Π be a θ–cell whose corresponding contiguity subdiagram Γ has minimal type $\tau(\Gamma)$. Its contour is decomposed as $p_1 q_1 p_2 q_2$, where $q_1 = \Gamma \wedge p$ and $q_2 = \Gamma \wedge \Pi$. Applying [32, Lemma 2.1] one observes that $|p_i| \leqslant 2\gamma n \min\{|C_k|, |C_l|\}$, where $k = r(\Pi)$ and $l = r(p)$. It follows that $|p_1| + |p_2| \leqslant 400\gamma|\partial\Pi|$; thus $|q_2| \geqslant \frac{1}{3}|\partial\Pi| + 100(|p_1| + |p_2|)$. Applying [32, Lemma 5.4], one gets that q_1 and q_2 have a common subpath q such that $|q| \geqslant |\partial\Pi|/3$. In other words, the contour of Π can be decomposed as $q\bar{q}$ where q^{-1} is a section of the contour of Δ and $|q| \geqslant |\partial\Pi|/3$. As Δ is a diagram of rank i, there exists $j \leqslant i$ and a cyclic permutation D of $C_j^{\pm 1}$ such that the label of $\bar{q}q$ is D^n. On the other hand, the contour of Δ can be written $rq^{-1}s$.

We now rephrase this observation using our geometric point of view. Let v and h be the element of $\boldsymbol{F}_r$ represented by the respective labels of $\bar{q}q$ (ie D) and r. Let $u = hvh^{-1}$. Let γ be the path of T starting at t_0 and labeled as q^{-1}. Recall that the contour $rq^{-1}s$ of Δ is labeled by the noncontractible word w; hence $h\gamma$ is a subpath of the geodesic $[t_0, gt_0]$. On the other hand, the collection of words (C_i) has been chosen in a minimal way. In particular, C_i and thus D are cyclically reduced. As a consequence, γ is contained in $[t_0, v^{-n}t_0]$ which lies in the axis of v. It follows that $h\gamma$ is a path contained in hA_v, ie the axis of u. Hence

$$\operatorname{diam}\bigl([t_0, gt_0] \cap A_u\bigr) \geqslant |\gamma| \geqslant \tfrac{1}{3}|\partial\Pi| = \tfrac{1}{3}|D| = \tfrac{n}{3}\|u\|.$$

In other words, $u^n g$ is obtained from g by performing a $\left(\frac{1}{3}, n\right)$–weak elementary move.

Let Δ' be the diagram obtained from Δ by removing the cell Π. Its contour is exactly $r\bar{q}s$. By our choice of v and h, its label represents $u^n g$. In other words, removing Π is equivalent to performing a $\left(\frac{1}{3}, n\right)$–weak elementary move. By the very definition of diagrams, Δ only contains finitely many cells. An induction on the number of cells in Δ shows that after finitely many weak elementary moves gt_0 is sent to t_0. □

The proof of Proposition 4.11 does not take place in the Cayley graph of $\boldsymbol{F}_r$ but in the universal cover of the underlying graph of an RTT map. Therefore, we need an analogue of Proposition A.2 in an arbitrary tree. From now on, (X, x_0) is a pointed simplicial tree. We assume that $\boldsymbol{F}_r$ acts properly cocompactly by isometries on X. There exists a natural $\boldsymbol{F}_r$–equivariant map $F\colon T \to X$ sending t_0 to x_0. Since $\boldsymbol{F}_r$ acts properly cocompactly on X, there exist $k \geqslant 1$ and $l \geqslant 0$ such that F is a (k, l)–quasiisometry, meaning that for every $t, t' \in T$,

$$k^{-1}|t - t'| - l \leqslant |F(t) - F(t')| \leqslant k|t - t'| + l.$$

We denote by ∂X the boundary at infinity of X.

Lemma A.3 *There exists $B \geqslant 0$ with the following property. Let $t, t' \in T$. Let $u \in \boldsymbol{F}_r \setminus \{1\}$ and $m \in \mathbb{N}$. Assume that we have*

$$\operatorname{diam}\big([t,t'] \cap A_u\big) \geqslant m\|u\|$$

in T. Then the following holds in X:

$$\operatorname{diam}\big([F(t), F(t')] \cap A_u\big) \geqslant m\|u\| - B.$$

Remark By abuse of notation, A_u stands for the axis of u in both T and X. Similarly with $\|u\|$.

Proof Recall first a well known statement of hyperbolic geometry: the stability of quasigeodesics. There exists $d \geqslant 0$ with the following property. The Hausdorff distance between a (k,l)–quasigeodesic of X and any geodesic with the same endpoints (possibly in ∂X) is bounded above by d [11, Chapitre 3, Théorème 1.3].

By assumption, there exists a point s in $[t,t'] \cap A_u$ such that $u^m s$ still belongs to $[t,t'] \cap A_u$. Recall that the axis of u in T is an u–invariant geodesic. Hence its image under F is a u–invariant (k,l)–quasigeodesic of X. It follows from the stability of quasigeodesics that $F(s)$ and $u^m F(s)$ lie in the d–neighborhood of the axis of u in X. In the same way, we see that $F(s)$ and $u^m F(s)$ lie in the d neighborhood of $[F(t), F(t')]$. Consequently, the following holds in X:

$$m\|u\| \leqslant |u^m F(s) - F(s)| \leqslant \operatorname{diam}\big([F(t), F(t')]^{+d} \cap A_u^{+d}\big). \tag{5}$$

Here the notation Y^{+d} stands for the d–neighborhood of $Y \subset X$. However, we observe that

$$\operatorname{diam}\big([F(t), F(t')]^{+d} \cap A_u^{+d}\big) \leqslant \operatorname{diam}\big([F(t), F(t')] \cap A_u\big) + 2d. \tag{6}$$

The result follows from (5) and (6) with $B = 2d$. □

Proposition A.4 *There exists $n_1 \in \mathbb{N}$ such that for every odd integer $n \geqslant n_0$, the following holds: an element $g \in \boldsymbol{F}_r$ is trivial in $\boldsymbol{B}_r(n)$ if and only if there exists a finite sequence of $\big(\frac{1}{4}, n\big)$–weak elementary moves which sends gx_0 to x_0.*

Proof Let B be the parameter given by Lemma A.3. Let $n_1 = \max\{10^{10}, 12B\}$. Let $n \geqslant n_0$. Let g be an element of $\boldsymbol{F}_r$. The tree X being simplicial, the translation length in X of any nontrivial element of $\boldsymbol{F}_r$ is at least 1. It follows from our choice of n_0 that for every $u \in \boldsymbol{F}_r \setminus \{1\}$, we have the following property. If $\operatorname{diam}\big([t_0, gt_0] \cap A_u\big) \geqslant \frac{n}{3}\|u\|$ in T, then $\operatorname{diam}\big([x_0, gx_0] \cap A_u\big) \geqslant \frac{n}{4}\|u\|$ in X. In other words, to any $\big(\frac{1}{3}, n\big)$–weak elementary move in T corresponds a $\big(\frac{1}{4}, n\big)$–weak elementary move in X. Hence the result follows from Proposition A.2. □

Corollary A.5 *There exists $n_1 \in \mathbb{N}$ such that for every odd integer $n \geqslant n_0$, the following holds: two isometries $g, g' \in \boldsymbol{F}_r$ have the same image in $\boldsymbol{B}_r(n)$ if and only if there exist two finite sequences of $\left(\frac{1}{8}, n\right)$–weak elementary moves which respectively send gx_0 and $g'x_0$ to the same point of X.*

Proof We apply Proposition A.4 with $g^{-1}g'$. □

Let us come back to Proposition 4.11. The main idea of the proof was the following. The criterion (Proposition 6.2) gave us a sequence of moves performed in $\widetilde{G}$ to send $\varphi^p(g)\tilde{x}_0$ to $\tilde{x}_0$. However, performing a move required us first to see a large part of a relation along $[\tilde{x}_0, \varphi^p(g)\tilde{x}_0]$. Because of Lemma 6.5, the support of the moves only contained yellow letters. Therefore, the red part was preserved, which led to a contradiction. Note that it does not matter whether the requirement to perform a move is to see one half, one fourth or one tenth of the relation. Therefore, the proof of Proposition 4.11 works in exactly the same way with the following modifications:

(1) Replace Proposition 6.2 by Corollary A.5.

(2) Define the critical exponent n_0 as

$$n_0 = \max\{n_1,\ 8n_2 + 1,\ 8|g\tilde{x}_0 - \tilde{x}_0| + 1\}.$$

(3) Replace every (n, ξ)–elementary move by a $\left(\frac{1}{8}, n\right)$–weak elementary move.

References

[1] **S I Adian**, *The Burnside problem and identities in groups*, Ergeb. Math. Grenzgeb. 95, Springer, New York (1979) MR

[2] **P Arnoux**, **V Berthé**, **A Hilion**, **A Siegel**, *Fractal representation of the attractive lamination of an automorphism of the free group*, Ann. Inst. Fourier (Grenoble) 56 (2006) 2161–2212 MR

[3] **M Bestvina**, **M Feighn**, **M Handel**, *Laminations, trees, and irreducible automorphisms of free groups*, Geom. Funct. Anal. 7 (1997) 215–244 MR

[4] **M Bestvina**, **M Feighn**, **M Handel**, *The Tits alternative for* $\mathrm{Out}(F_n)$*, I: Dynamics of exponentially-growing automorphisms*, Ann. of Math. 151 (2000) 517–623 MR

[5] **M Bestvina**, **M Feighn**, **M Handel**, *The Tits alternative for* $\mathrm{Out}(F_n)$*, II: A Kolchin type theorem*, Ann. of Math. 161 (2005) 1–59 MR

[6] **M Bestvina**, **M Handel**, *Train tracks and automorphisms of free groups*, Ann. of Math. 135 (1992) 1–51 MR

[7] **W Burnside**, *On an unsettled question in the theory of discontinuous groups*, Quart. J. 33 (1902) 230–238

[8] **E A Cherepanov**, *Free semigroup in the group of automorphisms of the free Burnside group*, Comm. Algebra 33 (2005) 539–547 MR

[9] **M Clay**, **A Pettet**, *Twisting out fully irreducible automorphisms*, Geom. Funct. Anal. 20 (2010) 657–689 MR

[10] **M M Cohen**, **M Lustig**, *Very small group actions on* **R***–trees and Dehn twist automorphisms*, Topology 34 (1995) 575–617 MR

[11] **M Coornaert**, **T Delzant**, **A Papadopoulos**, *Géométrie et théorie des groupes*, Lecture Notes in Mathematics 1441, Springer, New York (1990) MR

[12] **R Coulon**, *A criterion for detecting trivial elements of Burnside groups*, preprint (2012) arXiv To appear in Bull. Soc. Math. France

[13] **R Coulon**, *Outer automorphisms of free Burnside groups*, Comment. Math. Helv. 88 (2013) 789–811 MR

[14] **R Coulon**, *On the geometry of Burnside quotients of torsion free hyperbolic groups*, Internat. J. Algebra Comput. 24 (2014) 251–345 MR

[15] **R B Coulon**, *Partial periodic quotients of groups acting on a hyperbolic space*, Ann. Inst. Fourier (Grenoble) 66 (2016) 1773–1857 MR

[16] **M Culler**, **J W Morgan**, *Group actions on* **R***–trees*, Proc. London Math. Soc. 55 (1987) 571–604 MR

[17] **T Delzant**, **M Gromov**, *Courbure mésoscopique et théorie de la toute petite simplification*, J. Topol. 1 (2008) 804–836 MR

[18] **D Gaboriau**, **A Jaeger**, **G Levitt**, **M Lustig**, *An index for counting fixed points of automorphisms of free groups*, Duke Math. J. 93 (1998) 425–452 MR

[19] **F Gautero**, *Combinatorial mapping-torus, branched surfaces and free group automorphisms*, Ann. Sc. Norm. Super. Pisa Cl. Sci. 6 (2007) 405–440 MR

[20] **F Gautero**, **M Lustig**, *The mapping-torus of a free group automorphism is hyperbolic relative to the canonical subgroups of polynomial growth*, preprint (2007) arXiv

[21] **M Hall, Jr**, *Solution of the Burnside problem for exponent six*, Illinois J. Math. 2 (1958) 764–786 MR

[22] **S V Ivanov**, *The free Burnside groups of sufficiently large exponents*, Internat. J. Algebra Comput. 4 (1994) 1–308 MR

[23] **J Karhumäki**, *On cube-free ω–words generated by binary morphisms*, Discrete Appl. Math. 5 (1983) 279–297 MR

[24] **F Levi**, **B L van der Waerden**, *Über eine besondere Klasse von Gruppen*, Abh. Math. Sem. Univ. Hamburg 9 (1933) 154–158 MR

[25] **G Levitt**, *Counting growth types of automorphisms of free groups*, Geom. Funct. Anal. 19 (2009) 1119–1146 MR

[26] **I G Lysënok**, *Infinite Burnside groups of even period*, Izv. Ross. Akad. Nauk Ser. Mat. 60 (1996) 3–224 MR In Russian; translated in Izv. Math. 60 (1996) 453–654

[27] **B Mossé**, *Puissances de mots et reconnaissabilité des points fixes d'une substitution*, Theoret. Comput. Sci. 99 (1992) 327–334 MR

[28] **J Nielsen**, *Die Isomorphismen der allgemeinen, unendlichen Gruppe mit zwei Erzeugenden*, Math. Ann. 78 (1917) 385–397 MR

[29] **P S Novikov**, **S I Adjan**, *Infinite periodic groups, I*, Izv. Akad. Nauk SSSR Ser. Mat. 32 (1968) 212–244 MR In Russian; translated in Math. USSR - Izv. 2 (1968) 209–236

[30] **P S Novikov**, **S I Adjan**, *Infinite periodic groups, II*, Izv. Akad. Nauk SSSR Ser. Mat. 32 (1968) 251–524 MR In Russian; translated in Math. USSR - Izv. 2 (1968) 241–479

[31] **P S Novikov**, **S I Adjan**, *Infinite periodic groups, III*, Izv. Akad. Nauk SSSR Ser. Mat. 32 (1968) 709–731 MR In Russian; translated in Math. USSR - Izv. 2 (1968) 665–685

[32] **A Y Ol'shanskiĭ**, *The Novikov–Adyan theorem*, Mat. Sb. 118(160) (1982) 203–235, 287 MR In Russian; translated in Math. USSR - Sb. 46 (1983) 203–236

[33] **A Y Ol'shanskiĭ**, *Geometry of defining relations in groups*, Mathematics and its Applications (Soviet Series) 70, Kluwer, Dordrecht (1991) MR

[34] **D Osin**, *Small cancellations over relatively hyperbolic groups and embedding theorems*, Ann. of Math. (2) 172 (2010) 1–39

[35] **I N Sanov**, *Solution of Burnside's problem for exponent* 4, Leningrad State Univ. Annals Math. Ser. 10 (1940) 166–170 MR

[36] **E Seneta**, *Non-negative matrices and Markov chains*, 2nd edition, Springer, New York (2006) MR

[37] **W P Thurston**, *On the geometry and dynamics of diffeomorphisms of surfaces*, Bull. Amer. Math. Soc. 19 (1988) 417–431 MR

Institut de Recherche Mathématique de Rennes - UMR 6625
CNRS - Université de Rennes 1, Campus de Beaulieu, bâtiments 22 et 23
263 avenue du Général Leclerc, CS 74205, 35042 Rennes Cedex, France

I2M UMR 7373 - Centre de Mathématiques et Informatique, Aix Marseille Université
CNRS, Centrale Marseille, I2M, Marseille, France, Technopôle Château-Gombert
39, rue F Joliot Curie, 13453 Marseille 13, France

remi.coulon@univ-rennes1.fr, arnaud.hilion@univ-amu.fr

http://rcoulon.perso.math.cnrs.fr, http://junon.u-3mrs.fr/hilion/

Proposed: Martin Bridson Received: 15 October 2013
Seconded: Walter Neumann, Jean-Pierre Otal Revised: 10 May 2016

msp

Geometry & Topology 21 (2017) 2015–2048

Relations among characteristic classes of manifold bundles

ILYA GRIGORIEV

We study relations among characteristic classes of smooth manifold bundles with highly connected fibers. For bundles with fiber the connected sum of g copies of a product of spheres $S^d \times S^d$, where d is odd, we find numerous algebraic relations among so-called "generalized Miller–Morita–Mumford classes". For all $g > 1$, we show that these infinitely many classes are algebraically generated by a finite subset.

Our results contrast with the fact that there are no algebraic relations among these classes in a range of cohomological degrees that grows linearly with g, according to recent homological stability results. In the case of surface bundles ($d = 1$), our approach recovers some previously known results about the structure of the classical "tautological ring", as introduced by Mumford, using only the tools of algebraic topology.

55R40, 55T10, 57R22

1 Introduction

Let M be a $2d$–dimensional closed oriented smooth manifold. We denote by $\operatorname{Diff} M$ the topological group of *orientation-preserving* diffeomorphisms of M. The bar construction can be used to construct the space $\operatorname{BDiff}(M)$ that classifies bundles with fiber M. For any characteristic class of vector bundles $p \in H^{*+2d}(\mathrm{BSO}_{2d}; \mathbb{Q})$, we will define a *generalized Miller–Morita–Mumford* (MMM) *class* (or just *kappa class*) $\kappa_p \in H^*(\operatorname{BDiff}(M); \mathbb{Q})$. These are the simplest examples of characteristic classes of bundles[1] with fiber M and structure group $\operatorname{Diff} M$.

We are mainly interested in the case where the fiber is $\sharp_g S^d \times S^d$, the connected sum of g copies of $S^d \times S^d$. More generally, we let the fiber be a highly connected manifold (see Definition 2.5) of genus g and dimension $2d$, denoted M_g^{2d} or M_g. Recall that $H^*(\mathrm{BSO}_{2d}; \mathbb{Q}) = \mathbb{Q}[p_1, \dots, p_{d-1}, e]$, where p_i is the Pontryagin class of degree $4i$ and e is the Euler class of degree $2d$. Let $\mathcal{S} \subset H^*(\mathrm{BSO}_{2d}; \mathbb{Q})$ consist of

[1]A geometric example of such a bundle is a proper submersion $f : E \to B$ of smooth, oriented manifolds that has M as its fiber.

Published: 19 May 2017 DOI: 10.2140/gt.2017.21.2015

the monomials in the Pontryagin classes and the Euler class. For each such monomial, there is a corresponding MMM class in $H^*(\mathrm{BDiff}\, M_g; \mathbb{Q})$, which gives rise to a map

$$\mathcal{R}_d \colon \mathbb{Q}[\kappa_p \mid p \in \mathcal{S}] \to H^*(\mathrm{BDiff}\, M_g; \mathbb{Q}).$$

This paper presents a large family of polynomials in the MMM classes that lies in the kernel of the map $\mathcal{R}_d$, in the case that d is odd. In the $d > 1$ case, ours are the first results of this kind. In the $d = 1$ case, we recover previously known results, but using purely homotopy theoretic methods. Our first main result is the following.

Theorem 1.1 *The image of $\mathcal{R}_d$ is finitely generated as a $\mathbb{Q}$–algebra when d is odd and $g > 1$.*

In Proposition 5.8, we also show that for all odd d, the Krull dimension of the image of $\mathcal{R}_d$ is at most $2d$.

Our methods generalize the technique Randal-Williams developed for the $d = 1$ case in [22], which in turn is based on the work of Morita [18]. They allow us to present many specific elements in $\ker \mathcal{R}_d$. For instance, Randal-Williams found various relations among the images of the classes

$$\kappa_i := \kappa_{e^{i+1}} \in H^{2di}(\mathrm{BDiff}\, M_g; \mathbb{Q})$$

under the map R_d in the case when $d = 1$. We find that the same relations hold for any odd d (see Section 5.6 for details and examples). This is surprising, as no map between subrings of $H^*(\mathrm{BDiff}\, M_g^{2d})$ for different d that takes κ_i to κ_i can preserve the grading on the cohomology.

1.1 Manifolds with a fixed disk and homological stability

Let $\mathcal{S}' \subset \mathcal{S}$ be the set of monomials in the classes[2] $p_{\lceil (d+1)/4 \rceil}, p_{\lceil (d+1)/4 \rceil + 1}, \dots, p_{d-1}$, and e of total degree greater than $2d$. Let $\mathcal{R}'_d$ denote the map $\mathcal{R}_d$ restricted to $\mathbb{Q}[\kappa_p \mid p \in \mathcal{S}']$. Our second main result is:

Theorem 1.2 *If $d \equiv 3 \pmod 4$, the map $\mathcal{R}'_d$ has nontrivial kernel in degree $2g + 2$. If $d \equiv 1 \pmod 4$, the map $\mathcal{R}'_d$ has nontrivial kernel in degree $6g + 6$.*

By contrast, the map $\mathcal{R}'_d$ is known to be injective in a range of cohomological degrees $* \le (g-4)/2$ when the fiber is $\sharp_g\, S^d \times S^d$ and $d \neq 2$. This fact and the related phenomenon of *homological stability* are a large part of the motivation for our work. We now describe them in more detail.

[2] We use the notation $\lceil \cdot \rceil$ and $\lfloor \cdot \rfloor$ for rounding up and down (respectively) to the nearest integer.

Let $\mathrm{Diff}(M_g, D^{2d}) \subset \mathrm{Diff}(M_g)$ be the subgroup of those diffeomorphisms that fix pointwise a chosen disk in M_g, and let $f\colon \mathrm{BDiff}(M_g, D^{2d}) \to \mathrm{BDiff}(M_g)$ be the map induced on the bar constructions by the inclusion of groups. We define the map $\mathcal{R}_{\delta,d}\colon \mathbb{Q}[\kappa_p \mid p \in \mathcal{S}'] \to H^*(\mathrm{BDiff}(M_g, D^{2d}); \mathbb{Q})$ as the map that makes the following diagram commute:

$$\begin{tikzcd}
\mathbb{Q}[\kappa_p \mid p \in \mathcal{S}] \arrow[rr, "\mathcal{R}_d"] & & H^*(\mathrm{BDiff}\, M_g; \mathbb{Q}) \arrow[d, "f^*"] \\
\mathbb{Q}[\kappa_p \mid p \in \mathcal{S}']_i \arrow[u, hook] \arrow[urr, "\mathcal{R}'_d"] \arrow[rr, "\mathcal{R}_{\delta,d}"'] & & H^*(\mathrm{BDiff}(M_g, D^{2d}); \mathbb{Q})
\end{tikzcd} \tag{1.1.1}$$

(The δ stands for "fixed disk". See Appendix A for a comparison of the images of the various maps in the diagram.)

The next fact, in the $d = 1$ case, is a consequence of the Madsen–Weiss theorem [15] and the Harer stability theorem [11], with the improved stability range by Boldsen [3]. In the case when $d > 2$, the fact is a consequence of two theorems of Galatius and Randal-Williams [10; 9].

Fact 1.3 *If $M_g = \sharp_g S^d \times S^d$ and $d \neq 2$, the map $\mathcal{R}_{\delta,d}$ is an isomorphism in the range of cohomological degrees $* \leq (g-4)/2$. Thus the map $\mathcal{R}'_d$ is injective in the same range of degrees.*

For $d = 1$, the range of degrees can be improved to $* \leq 2g/3$.

In particular, the ring $H^*\big(\mathrm{BDiff}\big(\sharp_g S^d \times S^d, D^{2d}\big); \mathbb{Q}\big)$ satisfies *homological stability*: it is independent of g in a range of cohomological degrees. Theorem 1.2 implies that this range of cohomological degrees cannot be improved beyond $* \leq 2g+1$.

In Appendix A, we prove another version of Theorem 1.1.

Theorem A.4 *The image of $\mathcal{R}_{\delta,d}$ is finitely generated as a $\mathbb{Q}$–algebra when d is odd and $g > 1$.*

1.2 Comparison with known results for surface bundles

In the $d = 1$ case, the fiber of our bundle is an oriented genus-g surface $\Sigma_g = M_g^2 = \sharp_g S^1 \times S^1$ and the generalized Miller–Morita–Mumford classes correspond to the classical ones, with $\kappa_i = \kappa_{e^{i+1}} \in H^{2i}(\mathrm{BDiff}(\Sigma_g, D^2); \mathbb{Q})$. The map $\mathcal{R}_1$ takes the form

$$\mathcal{R}_1\colon \mathbb{Q}[\kappa_1, \kappa_2, \dots] \to H^*(\mathrm{BDiff}\, \Sigma_g; \mathbb{Q}).$$

The ring of characteristic classes of surface bundles in rational cohomology coincides with the cohomology of the moduli space of Riemann surfaces $\mathcal{M}_g$ since

$$H^*(\mathrm{BDiff}\,\Sigma_g;\mathbb{Q}) = H^*(B\Gamma_g;\mathbb{Q}) = H^*(\mathcal{M}_g;\mathbb{Q}),$$

where Γ_g is the orientation-preserving mapping class group. (The first equality follows from the theorem of Earle and Eells [5], which implies that the natural group homomorphism $\mathrm{Diff}\,\Sigma_g \to \Gamma_g$ is a homotopy equivalence, and thus the bar constructions are weakly homotopy equivalent. The second is true only in rational cohomology and follows from Teichmüller theory; see Farb and Margalit [8, Section 12.6] for an overview.)

The image of $\mathcal{R}_1$ can therefore be thought of as a subring of $H^*(\mathcal{M}_g;\mathbb{Q})$. This subring coincides with the classical *tautological ring*, as defined in Mumford [20]. Techniques of algebraic geometry and low-dimensional topology (hyperbolic geometry, in particular) have been used to obtain many results about the structure of this ring. For example, since $\mathcal{M}_g$ is a $(6g-6)$–dimensional orbifold, the image of $\mathcal{R}_1$ must vanish above that degree, and thus be a finite-dimensional vector space over $\mathbb{Q}$.

More precise results are known; we list the most relevant ones. The image of the map $\mathcal{R}_1$ is trivial above degree $2(g-2)$ by a theorem of Looijenga [13], and in degree $2(g-2)$ it is one-dimensional; see Faber [7] and Looijenga [13]. Morita [19] showed that the kernel of $\mathcal{R}_1$ is nontrivial in degree $2\lfloor g/3\rfloor + 2$. However, $\mathcal{R}_1$ is an isomorphism in degrees $\leq 2\lfloor g/3\rfloor$ according to Fact 1.3 together with the fact that the map $f^*\colon H^*(\mathrm{BDiff}\,\Sigma_g;\mathbb{Z}) \to H^*(\mathrm{BDiff}(\Sigma_g, D^2);\mathbb{Z})$ is an isomorphism in the same range of degrees; see Boldsen [3] and Harer [11]. For two conjectural, complete descriptions of $\ker \mathcal{R}_1$, which differ for $g > 23$ but are known to be true for $g \leq 23$, see Faber [7] and Pandharipande and Pixton [21].

Since the relations in Theorems 1.2 and 1.1 have high cohomological degree, they follow from Looijenga's theorem in the $d = 1$ case. We provide a new proof for the relations of lower degree obtained by Randal-Williams in [22], including all of the existing relations for $g \leq 5$. It is unclear whether our strengthening of Randal-Williams' methods can result in genuinely new relations in the $d = 1$ case.

1.3 Outline of the paper

In Section 2, we define the generalized Miller–Morita–Mumford classes. We then state the main technical result of the paper and the primary source of our relations, Theorem 2.7. We outline its proof and apply it to prove Theorem 1.2.

The details of the proof of Theorem 2.7 take up Sections 3 and 4. In the special case of surface bundles, this work leads to a stronger statement and a new proof of a result of Morita [18, Section 3].

In Section 5, we use Theorem 2.7 to prove Theorem 1.1 and our other results. These calculations use methods Randal-Williams developed for surface bundles in [22], originally based on Morita's result.

Appendix A discusses the relationship between the maps $\mathcal{R}_d$, $\mathcal{R}'_d$ and $\mathcal{R}_{\delta,d}$, and proves Theorem A.4. Appendix B discusses alternative definitions of the pushforward map on cohomology, which is a crucial ingredient in defining the MMM classes.

Acknowledgements I am deeply grateful to my thesis advisor, Søren Galatius, for his constant support and sage advice. I am also very grateful for the numerous conversations with and the insight provided by Oscar Randal-Williams. This work would not exist without their support. I thank Alexander Kupers and Jeremy Miller for numerous helpful conversations, as well as suggestions about the content of Appendix B. I thank Sam Lichtenstein for his advice on linear algebra that improved Section 4.2. I thank the reviewers for their many insightful comments, corrections, and help with the organization of the paper.

This work was supported by an NSF Graduate Research Fellowship. I was also supported in part by the Danish National Research Foundation and the Centre for Symmetry and Deformation through a Nordic Research Opportunity grant, by the NSF grant DMS-1105058 and, while I was in residence at the MSRI in Berkeley, CA during the Spring 2014 semester, by the NSF grant no. 0932078000. I am very grateful to the Stanford University Department of Mathematics for all the support during my years as a graduate student. Part of this work was done during my very pleasant visit to the University of Copenhagen, which I thank for its warm hospitality.

2 Definitions and our main technical result

In this section, we give more precise definitions for terms used in the introduction. We then state the main technical result of this paper and give an informal outline of its proof. Finally, we apply it to prove Theorem 1.2.

Let M be an oriented smooth closed connected manifold and $\operatorname{Diff} M$ is the topological group of *orientation-preserving* diffeomorphisms of M with the C^∞ topology.

Definition 2.1 By an *oriented manifold bundle* (or just *manifold bundle*), we mean a bundle $E \to B$ with fiber M and structure group $\operatorname{Diff} M$.

2.1 Pushforward maps

For an oriented manifold bundle $\pi\colon E \to B$ with fiber M, there is a map of abelian groups $\pi_!\colon H^{*+\dim M}(E;\mathbb{Z}) \to H^*(B;\mathbb{Z})$ called the *pushforward map*, also known as the *umkehr map* or the *Gysin homomorphism*. Note that $\pi_!(1) = 0$ when $\dim M \neq 0$

because of the change of cohomological degree, and thus $\pi_!$ is not a ring map. We will give its definition (originally from [4]) in a more general setting in Section 3.3, Definition 3.5.

To give a little substance to this notion, we mention that in the special case when E and B are closed oriented manifolds, the map $\pi_!$ coincides with the composition of Poincaré duality in E, the natural map on homology induced by π, and Poincaré duality in B. When restricted to de Rham cohomology, the map coincides with integration along the fiber (these equivalences are discussed in detail in [2]).

For our present purposes, it is sufficient to recall one nontrivial property of $\pi_!$. The pushforward map is natural in the sense that, if we form a pullback diagram of manifold bundles

(2.1.1)
$$\begin{tikzcd} f^*(E) \arrow[r, "f'"] \arrow[d, "\pi'"] & E \arrow[d, "\pi"] \\ A \arrow[r, "f"] & B \end{tikzcd}$$

then for any $a \in H^*(E)$, we have $f^*(\pi_!(a)) = \pi'_!(f'^*(a))$.

Further properties of the pushforward map are discussed in Section 5.1.

2.2 Definition of the Miller–Morita–Mumford classes

Let $P \to B$ be the principal $\mathrm{Diff}\, M$–bundle corresponding to the manifold bundle $E \to B$. The group $\mathrm{Diff}\, M$ acts on the total space of the tangent bundle TM as well as on M, and the bundle map $TM \to M$ is equivariant with respect to this action. So the map
$$P \times_{\mathrm{Diff}\, M} TM \to P \times_{\mathrm{Diff}\, M} M = E$$
can be given the structure of a bundle over E with the same fiber and structure group as the bundle $TM \to M$.

Definition 2.2 The *vertical tangent bundle* $T_\pi E$ is the vector bundle of rank $\dim M$ over E defined by the above map.

Remark 2.3 In the special case when the bundle map $E \to B$ is a smooth map between smooth manifolds, the vertical tangent bundle coincides with the subbundle of TE that is the kernel of the derivative $Df\colon TE \to TB$.

As we only consider orientation-preserving diffeomorphisms, $T_\pi E$ is an *oriented* vector bundle. Its characteristic classes determine a map $\gamma\colon H^*(\mathrm{BSO}_{\dim M}; \mathbb{Z}) \to H^*(E; \mathbb{Z})$.

Definition 2.4 Let $E \to B$ be a manifold bundle with m–dimensional fiber and $p \in H^{l+m}(\mathrm{BSO}_m;\mathbb{Z})$. The corresponding *generalized Miller–Morita–Mumford class* or *kappa class* is defined by

$$\kappa_p\begin{pmatrix} E \\ \downarrow \\ B \end{pmatrix} := \pi_!(\gamma^*(p)) \in H^l(B;\mathbb{Z}).$$

The kappa classes are natural with respect to pullbacks of bundles because of the naturality property of pushforwards. To be more precise, the following diagram will commute in the context of the pullback diagram (2.1.1):

$$\begin{tikzcd} H^{*+m}(\mathrm{BSO}_m;\mathbb{Z}) \arrow[r, "p\mapsto\kappa_p\left(\begin{smallmatrix} E \\ \downarrow \\ B \end{smallmatrix}\right)"] \arrow[dr, "p\mapsto\kappa_p\left(\begin{smallmatrix} f^*(E) \\ \downarrow \\ A \end{smallmatrix}\right)"'] & H^*(B;\mathbb{Z}) \arrow[d, "f^*"] \\ & H^*(A;\mathbb{Z}) \end{tikzcd}$$

Every manifold bundle is a pullback of the universal bundle over $\mathrm{BDiff}\, M$. So the kappa classes for any bundle are pullbacks of universal classes $\kappa_p \in H^*(\mathrm{BDiff}\, M;\mathbb{Z})$. Similarly, for $p \in H^{*+m}(\mathrm{BSO}_m;\mathbb{Q})$ there are classes

$$\kappa_p\begin{pmatrix} E \\ \downarrow \\ B \end{pmatrix} \in H^*(B;\mathbb{Q}) \quad \text{and} \quad \kappa_p \in H^*(\mathrm{BDiff}\, M;\mathbb{Q}).$$

2.3 Key source of the relations

Let us state the main technical result that underlies the relations discussed in this paper. We will give an informal outline of the proof at the end of this section and postpone all details to Sections 3 and 4.

We will consider bundles with fiber in the following class of manifolds.

Definition 2.5 By a *highly connected manifold of genus* g, we mean a $2d$–dimensional $(d-1)$–connected smooth oriented closed manifold with middle cohomology isomorphic to $\mathbb{Z}^{2g}$. Throughout the paper, M_g represents such a manifold.

This class includes the connected sum of g copies of $S^d \times S^d$. To give another example, let Q be the total space of a bundle such that the fiber and the base spaces are smooth homotopy d–spheres. A connected sum of g copies of Q will be a highly connected manifold of genus g.

Remark 2.6 If M is an oriented closed smooth $2d$–dimensional $(d-1)$–connected manifold, the universal coefficient theorem implies that $H^d(M;\mathbb{Z}) \cong \mathrm{Hom}(H_d(M),\mathbb{Z})$, which is a free group. Poincaré duality and the fact that d is odd imply that the rank of this group must be even. So M is a highly connected manifold of genus g for some integer g.

Theorem 2.7 *Let d be an* odd *natural number and M_g be a $2d$–dimensional highly connected manifold of genus g. Let $\pi\colon E \to B$ be an oriented manifold bundle with fiber M_g^{2d} and let $a, b \in H^*(E;\mathbb{Z})$ be two classes such that $\pi_!(a) = 0$, $\pi_!(b) = 0$, and* $\deg(a)$ *is* even.

Then the classes

$$\pi_!(a \cup a) \in H^{2\deg(a)-2d}(B;\mathbb{Z}) \quad \text{and} \quad \pi_!(a \cup b) \in H^{\deg(a)+\deg(b)-2d}(B;\mathbb{Z})$$

satisfy the two relations

$$(2g+1)! \cdot \pi_!(a \cup a)^{g+1} = 0, \tag{2.3.1}$$

$$(2g+1)! \cdot \pi_!(a \cup b)^{2g+1} = 0. \tag{2.3.2}$$

(Note the larger power in the second relation.)

Remark 2.8 Because of the $(2g+1)!$ factor in the statement, the theorem is most useful to give relations for cohomology with rational coefficients. It is likely that this factor can be improved somewhat. In [18, Section 3], Morita proved the relation (2.3.1) in the special case of $d = 1$ and $\deg a = 2$ with a factor of $(2g+2)!/(2^{g+1}(g+1)!)$ instead of $(2g+1)!$.

2.4 An application: proof of Theorem 1.2

In this section, we illustrate Theorem 2.7 by proving Theorem 1.2 as an application. Further applications of Theorem 2.7 that result in more elaborate relations are discussed in Section 5.

Proposition 2.9 *Suppose $d \neq 1$ is an odd integer. Let $s = \lceil (d+1)/4 \rceil$, and let p_s be the $4s$–dimensional Pontryagin class. Then*

$$\kappa_{p_s^2}^{g+1} = 0 \in H^{(2 \text{ or } 6)(g+1)}(\mathrm{BDiff}\, M_g;\mathbb{Q}), \quad \textit{where } \deg \kappa_{p_s^2} = \begin{cases} 2 & \textit{if } d \equiv 3 \pmod 4, \\ 6 & \textit{if } d \equiv 1 \pmod 4. \end{cases}$$

Proof Let $d \geq 3$ be odd. Let $\pi\colon E \to (B = \mathrm{BDiff}\, M_g)$ be the universal manifold bundle with fiber M_g^{2d}. The $4s$–dimensional Pontryagin class of the vertical tangent bundle gives rise to the class $p_s \in H^{4s}(E;\mathbb{Q})$.

Our choice of s insures that, depending on $d \bmod 4$, either $4s = d+1$ or $4s = d+3$. Since under our assumptions $4s < 2d$, we have $\pi_!(p_s) = 0$. Also, $\deg p_s$ is even.

Thus we can apply Theorem 2.7 to obtain the following relation concerning the class $\pi_!(p_s^2)$, which is either 2– or 6–dimensional:

$$(2g+1)!\pi_!(p_s^2)^{g+1} = 0 \in H^{(2 \text{ or } 6)(g+1)}(B;\mathbb{Q}).$$

The class $\pi_!(p_s^2)$ coincides with the class $\kappa_{p_s^2} \in H^{(2 \text{ or } 6)}(\mathrm{BDiff}\, M_g;\mathbb{Q})$ by definition. So rationally $\kappa_{p_s^2}^{g+1} = 0 \in H^{(2 \text{ or } 6)(g+1)}(\mathrm{BDiff}\, M_g;\mathbb{Q})$ as desired. □

When $d \neq 1$, Proposition 2.9 immediately implies Theorem 1.2 since, in the terminology of the introduction, $p_s^2 \in \mathcal{S}'$.

Fact 1.3 implies that the class $\kappa_{p_s^2} \in H^*\bigl(\mathrm{BDiff}\bigl(\sharp_g\, S^d \times S^d, D^{2d}\bigr);\mathbb{Q}\bigr)$ is not zero when g is large enough. Therefore we also have $\kappa_{p_s^2} \neq 0 \in H^*\bigl(\mathrm{BDiff}\bigl(\sharp_g\, S^d \times S^d\bigr);\mathbb{Q}\bigr)$, even though we just showed $\kappa_{p_s^2}^{g+1} = 0$.

When $d = 1$ and $g > 1$, Theorem 1.2 follows from Corollary 5.18, which in this case is due to Morita (Looijenga's theorem [13] is even stronger). The $S^1 \times S^1$ case can be done by replacing p_s with the class $e\kappa_{e^2}$ in the above proof. The $g = 0$ case follows from the fact that $\mathrm{BDiff}\, S^2 \simeq \mathrm{BSO}_3$ by a theorem of Smale.

2.5 Outline of the proof of Theorem 2.7

We aim to prove that a certain power of the class $\pi_!(a \cup b)$ is torsion. If we wanted to prove that $2\alpha^2 = 0$ for some integral cohomology class α, it would be sufficient to decompose it as product of a integral cohomology class of *odd* degree β and another class: $\alpha = \beta \cup \gamma$. Our proof is loosely analogous.

In Section 3, we will use the Serre spectral sequence for the fibration $\pi\colon E \to B$ to define the pushforward map on cohomology $\pi_!$. The key result of Section 3 is that, under the assumptions of Theorem 2.7, the cohomology class $\pi_!(a \cup b)$ is the product of two terms on the E_2 page of the spectral sequence, at least one of which — we call it ι — has odd degree (Proposition 3.8).

The class ι turns out to be a cohomology class with a $2g$–dimensional, twisted coefficient system. In Section 4, we prove Proposition 4.1, which implies that since $\deg \iota$ is odd, ι^{2g+1} is torsion. We then relate various notions of cup product to conclude that $\pi_!(a \cup b)^{2g+1}$ and $\pi_!(a \cup a)^{g+1}$ are both torsion.

3 Spectral sequence argument

In this section, we begin the detailed proof of Theorem 2.7. A reader more interested in applications might want to skip directly to Section 5.

The proof of Theorem 2.7 is most naturally stated in the setting of *oriented Serre fibrations*. This setting is more general than the setting of manifold bundles. We first define the pushforward map in this generality. Then our goal is to prove Proposition 3.8, which in certain cases allows us to decompose cohomology classes of the form $\pi_!(a \cup b)$.

3.1 Oriented Serre fibrations and twisted coefficient systems

By a *twisted coefficient system* over B, we will mean a bundle of abelian groups over B with some fiber A and the discrete group $\operatorname{Aut} A$ as its structure group. Given a basepoint $* \in B$, twisted coefficient systems correspond bijectively to $\mathbb{Z}[\pi_1(B,*)]$–modules (see eg [16, Section 5.3]). Moreover, maps and tensor products of twisted coefficient systems correspond to maps and tensor products of $\mathbb{Z}[\pi_i(B,*)]$–modules, respectively.

Let $E \to B$ be a Serre fibration, $* \in B$ be a chosen basepoint, and M be the homotopy fiber at the basepoint. The homotopy-lifting property of Serre fibrations gives rise to an action of $\pi_1(B,*)$ on the cohomology groups $H^i(M;\mathbb{Z})$ for all i. This gives rise to a twisted coefficient system that we denote $\mathcal{H}^i(M)$. The cup product on cohomology $H^i(M;\mathbb{Z}) \otimes H^j(M;\mathbb{Z}) \to H^{i+j}(M;\mathbb{Z})$ is a map of $\mathbb{Z}[\pi_i(B,*)]$–modules. So there is a well-defined cup product on twisted coefficient systems

(3.1.1) $$\cup\colon \mathcal{H}^i(M) \otimes \mathcal{H}^j(M) \to \mathcal{H}^{i+j}(M).$$

We are interested in the case in which the homotopy fiber is a closed, connected manifold M^{2d}. An *orientation* for such a Serre fibration $E \to B$ is a choice of a trivialization for the twisted coefficient system corresponding to the top cohomology, ie a choice of an isomorphism or: $\mathcal{H}^{2d}(M) \xrightarrow{\sim} \mathbb{Z}$, where the right-hand side is the untwisted coefficient system over B. An *oriented Serre fibration* is a Serre fibration $E \to B$ that is equipped with a choice of an orientation.

Example 3.1 Any (oriented) manifold bundle in the sense of Section 2 is an example of an oriented Serre fibration, since the structure group of the manifold bundle preserves the given orientation of the fiber M.

3.2 Convergence of Serre spectral sequences

In this section, we recall the features of the convergence theorem for the cohomological Serre spectral sequence that we will need.

As we will discuss in more detail in Section 4.1, for any coefficient systems $\mathcal{A}$ and $\mathcal{B}$ over B, there is a notion of *cohomology with twisted coefficients* and a *cup product* (different from the one defined in (3.1.1))

(3.2.1) $$\cup\colon \; H^p(B;\mathcal{A}) \otimes H^q(B;\mathcal{B}) \to H^{p+q}(B;\mathcal{A} \otimes \mathcal{B}).$$

Moreover, any map of coefficient systems $f\colon \mathcal{A} \to \mathcal{B}$ determines a map on cohomology that we will denote $f_{\mathrm{coeff}}\colon H^*(B;\mathcal{A}) \to H^*(B;\mathcal{B})$.

The Serre spectral sequence for a Serre fibration $\pi\colon E \to B$ with fiber M (which, for the purposes of the convergence theorem, can be any CW complex) relates the following two objects:

(1) The cohomology of the total space $H^*(E;\mathbb{Z})$ together with the cup product and a filtration

$$H^*(E;\mathbb{Z}) = \cdots = F^{-1} = F^0 H^*(E;\mathbb{Z}) \supset F^1 H^*(E;\mathbb{Z}) \supset \cdots \tag{3.2.2}$$

defined as follows. Let $B^{(j)}$ denote the j–skeleton of the CW complex B, $J^{(j)} = \pi^{-1}(B^{(j)}) \subset E$, and $J^{(-1)} = \varnothing$. We set

$$F^i H^*(E) := \ker\bigl(H^*(E) \to H^*(J^{(i-1)})\bigr) = \mathrm{image}\bigl(H^*(E, J^{(i-1)}) \to H^*(E)\bigr).$$

Note that this filtration respects the cup product, ie the cup product restricts to a map $F^p H^*(E;\mathbb{Z}) \otimes F^{p'} H^*(E;\mathbb{Z}) \to F^{p+p'} H^*(E;\mathbb{Z})$.

(2) The E_2 page of the spectral sequence which is the bigraded ring

$$E_2^{p,q} := H^p(B;\mathcal{H}^q(M))$$

with the product specified by the composition of maps

$$\begin{aligned} \bullet\colon E_2^{p,q} \otimes E_2^{p',q'} &= H^p(B;\mathcal{H}^q(M)) \otimes H^{p'}(B;\mathcal{H}^{q'}(M)) \\ &\xrightarrow[(3.2.1)]{\cup} H^{p+p'}(B;\mathcal{H}^p(M) \otimes \mathcal{H}^{q'}(M)) \\ &\xrightarrow[(3.1.1)]{\cup_{\mathrm{coeff}}} H^{p+p'}(B;\mathcal{H}^{q+q'}(M)) = E_2^{p+p',q+q'}. \end{aligned} \tag{3.2.3}$$

The convergence theorem relates these two objects by way of the E_∞ page of the spectral sequence:

Theorem 3.2 (convergence theorem for the Serre spectral sequence [16, Theorem 5.2]) *There is a spectral sequence with the E_2 page as described above such that the following two definitions of its E_∞ page are equivalent (together with the product structure):*

(a) *Successive quotients of the filtration* (3.2.2) *together with the cup product*

$$E_\infty^{p,q} \cong F^p H^{p+q}(E;\mathbb{Z}) / F^{p+1} H^{p+q}(E;\mathbb{Z}).$$

(b) *A subquotient of the E_2 page obtained by repeatedly taking homology using the differentials in the spectral sequence. Repeatedly taking subquotients of a group results in a subquotient, so there are subgroups $B^{p,q} \subset Z^{p,q} \subset E_2^{p,q}$ such that*

$$E_\infty^{p,q} = Z^{p,q} / B^{p,q}.$$

By a small abuse of language, we write

$$Z^{p,q} = \ker(\textit{differentials out of } (p,q) \textit{ terms}),$$
$$B^{p,q} = \operatorname{image}(\textit{differentials into } (p,q) \textit{ terms}).$$

The product structure is induced from the product structure on the E_2 page. This uses the fact that all the differentials respect the product on their respective pages of the spectral sequence.

3.3 Pushforwards and spectral sequences

In this section, we assume that the fiber M^{2d} is a $2d$–dimensional oriented closed connected manifold.

Lemma 3.3 *If M has dimension $2d$, the filtration on cohomology is such that*

$$F^{n-2d}H^n(E;\mathbb{Z}) = H^n(E;\mathbb{Z}) \quad \textit{for all } n.$$

If M is also $(d-1)$–connected, then we also have

$$F^{n-d}H^n(E;\mathbb{Z}) = F^{n-2d+1}H^n(E;\mathbb{Z}).$$

(For the indices in this and the following arguments, refer to Figure 1.)

Proof Since the fiber M is $2d$–dimensional, $E_2^{n-q,q} = 0$ for $q > 2d$, and therefore $0 = E_\infty^{n-q,q} = F^{n-q}H^n(E;\mathbb{Z})/F^{n-q+1}H^n(E;\mathbb{Z})$ as well.

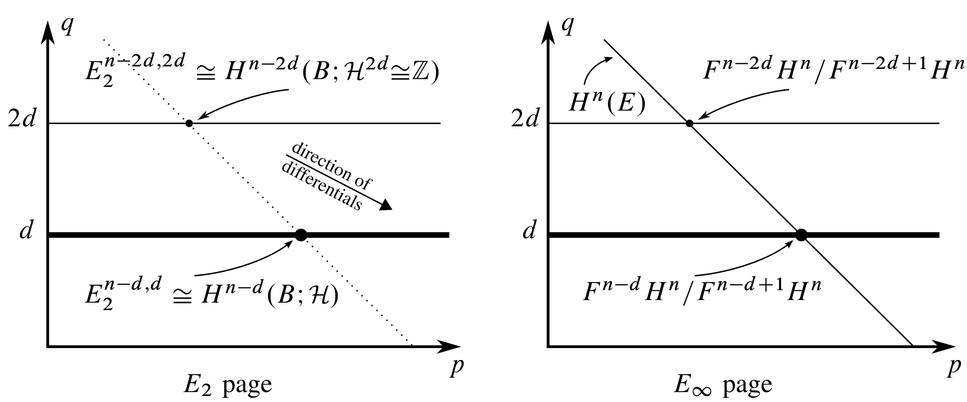


Figure 1: The E_2 and E_∞ pages of the Serre spectral sequence with fiber a $(d-1)$–connected oriented closed $2d$–dimensional manifold M. The entries with total degree n are highlighted. We abbreviate $F^iH^n := F^iH^n(E;\mathbb{Z})$.

If M is $(d-1)$–connected, then $H^q(M;\mathbb{Z})=0$ for $2d>q>d$ by Poincaré duality. Thus $E_2^{n-q,q}=0$ as well in this range. $\square$

From the E_2 page onwards, all the differentials in the spectral sequence go in the down-and-right direction. In particular, there are no differentials *into* the $2d^{\text{th}}$ row of the spectral sequence (ie the $E_i^{n-2d,2d}$ terms for $i\geq 2$). So

$$B^{n-2d,2d}=\text{image(differentials into } (n-2d,2d) \text{ terms})=0.$$

The convergence theorem implies that $E_\infty^{n-2d,2d}\subset E_2^{n-2d,2d}/B^{n-2d,2d}$, so we have:

Lemma 3.4 $$E_\infty^{n-2d,2d}\subset E_2^{n-2d,2d}.$$

By definition, $E_\infty^{n-2d,2d}=F^{n-2d}H^n(E;\mathbb{Z})/F^{n-2d+1}H^n(E;\mathbb{Z})$. We can now define the pushforward map that we use throughout this paper:

Definition 3.5 [4, Section 8] If the Serre fibration $\pi\colon E\to B$ with fiber M^{2d} is *oriented*, we define the *pushforward map on cohomology* $\pi_!\colon H^*(E;\mathbb{Z})\to H^{*-2d}(B;\mathbb{Z})$ to be the composition of maps

$$\begin{tikzcd} \begin{array}{c} H^n(E;\mathbb{Z})= \\ F^{n-2d}H^n(E;\mathbb{Z}) \end{array} \arrow[r, two heads] \arrow[rrr, bend right, "\pi_!"'] & E_\infty^{n-2d,2d} \arrow[r, hook] & E_2^{n-2d,2d} \arrow[r, "\sim", "\mathrm{or}_{\mathrm{coeff}}"'] & H^{n-2d}(B;\mathbb{Z}). \end{tikzcd} \tag{3.3.1}$$

Various properties of the pushforward map (which are not used in this section nor in Section 4) are discussed in Sections 2.1 and 5.1.

3.4 Secondary pushforwards and the decomposition of pushforwards

Let us now assume that our Serre fibration is oriented and that the fiber M is a $(d-1)$–connected $2d$–dimensional oriented closed manifold. Let us consider the kernel of the map $\pi_!$ we just defined.

Lemma 3.6 *Let* $(\ker\pi_!)^n := (\ker\pi_!)\cap H^n(E;\mathbb{Z})\subset H^*(E;\mathbb{Z})$. *If* M *is* $2d$*–dimensional and* $(d-1)$*–connected, then*

$$(\ker\pi_!)^n=F^{n-d}H^n(E;\mathbb{Z}).$$

Proof By examining the map (3.3.1), we see that the quotient map

$$H^n(E;\mathbb{Z})=F^{n-2d}H^n(E;\mathbb{Z})\twoheadrightarrow E_\infty^{n-2d,2d}=F^{n-2d}H^n(E;\mathbb{Z})/F^{n-2d+1}H^n(E;\mathbb{Z})$$

must take $(\ker \pi_!)^n$ to zero and therefore $(\ker \pi_!)^n = F^{n-2d+1} H^n(E;\mathbb{Z})$. Lemma 3.3 states that since M is $(d-1)$–connected, $F^{n-2d+1}H^n(E;\mathbb{Z}) = F^{n-d}H^n(E;\mathbb{Z})$. □

We will now attempt to repeat the construction of the map (3.3.1). The lemma gives us a quotient map $(\ker \pi_!)^n = F^{n-d}H^n(E;\mathbb{Z}) \twoheadrightarrow E_\infty^{n-d,d}$ (see also Figure 1 for indices). It is no longer necessarily true that $E_\infty^{n-d,d}$ is a subset of $E_2^{n-d,d}$, but the convergence theorem states that it is in general a subset of a quotient:

$$E_\infty^{p,q} = \frac{Z^{p,q}}{B^{p,q}} \subset \frac{E_2^{p,q}}{B^{p,q}}.$$

So we have the following sequence of maps:

$$\begin{tikzcd}
(\ker \pi_!)^n \arrow[r, equal] \arrow[drrr, dashed, "\xi"] & F^{n-d} H^n(E;\mathbb{Z}) \arrow[r, two heads] & E_\infty^{n-d,d} \arrow[r, hook] & E_2^{n-d,d}/B^{p,d} \\
 & & & E_2^{n-d,d} = H^{n-d}(B;\mathcal{H}^d) \arrow[u, two heads]
\end{tikzcd} \tag{3.4.1}$$

We use the fact that the wrong-way map in this diagram is surjective to make the following definition:

Definition 3.7 For each $a \in (\ker \pi_!)^n$, we define its *secondary pushforward* $\xi(a)$ to be some element in $E_2^{n-d,d} = H^{n-d}(B;\mathcal{H}^d)$ which maps to the same element of $E_2^{n-d,d}/B^{p,d}$ as a under the maps in (3.4.1). From now on, we assume that we have fixed a choice of such a $\xi(a)$ for every a.

Since there is no reason for $\xi\colon (\ker \pi_!)^n \dashrightarrow H^{n-d}(B;\mathcal{H}^d)$ to be a group homomorphism, we will call it a *correspondence* rather than a *map* and denote it with a dashed arrow.

Proposition 3.8 *Let $a \in (\ker \pi_!)^{p+d}$ and $b \in (\ker \pi_!)^{p'+d}$. The cohomology class $\pi_!(a \cup b) \in H^{p+p'}(B;\mathbb{Z})$ is the image of $\xi(a) \otimes \xi(b)$ under the following map:*

$$\begin{tikzcd}
E_2^{p,d} \otimes E_2^{p',d} \arrow[r, "\bullet"] & E_2^{p+p',2d} \arrow[r, "\sim", "\mathrm{or}_{\mathrm{coeff}}"'] & H^{p+p'}(B;\mathbb{Z}) \\
\xi(a) \otimes \xi(b) \arrow[u, phantom, "\Cup"] \arrow[rr, mapsto] & & \pi_!(a \cup b) \arrow[u, phantom, "\Cup"]
\end{tikzcd} \tag{3.4.2}$$

Proof Since the Serre spectral sequence is multiplicative, every term in the diagram (3.4.1) is a subset of some ring. The following diagram combines the multiplication maps on every term:

$$
\begin{tikzcd}
(a\otimes b)\in (\ker\pi_!)^{p+d}\otimes(\ker\pi_!)^{p'+d} \arrow[r, "\cup"] \arrow[d, equal] & H^{p+p'+2d}(E;\mathbb{Z}) \arrow[d, equal] \arrow[dddd, bend left=60, "\pi_!"] \\
F^pH^{p+d}(E;\mathbb{Z})\otimes F^{p'}H^{p'+d}(E;\mathbb{Z}) \arrow[d, two heads] & F^{p+p'}H^{p+p'+2d}(E;\mathbb{Z}) \arrow[d, two heads] \\
E_\infty^{p,d}\otimes E_\infty^{p',d} \arrow[r, "E_\infty\text{ mult.}"] \arrow[d, hook] & E_\infty^{p+p',2d} \arrow[d, hook, "\text{(a)}"] \\
\frac{E_2^{p,d}}{B^{p,d}}\otimes\frac{E_2^{p',d}}{B^{p',d}} \arrow[r, "E_2\text{ mult.}", "\text{(b)}"'] & \frac{E_2^{p+p',2d}}{B^{p+p',2d}} = E_2^{p+p',2d} \arrow[d, "\wr", "\mathrm{or}_{\mathrm{coeff}}"'] \\
E_2^{p,d}\otimes E_2^{p',d}\ni\xi(a)\otimes\xi(b) \arrow[u, two heads] & \pi_!(a\cup b)\in H^{p+p'}(B;\mathbb{Z})
\end{tikzcd}
$$

We observe the following:

- The convergence theorem implies that the diagram commutes and the map (b) is well defined.
- The composition of maps (a) coincides with the map (3.3.1) from the definition of $\pi_!$.
- In the image of the map (b), the group $B^{p+p',2d}$ is zero as we discussed in the proof of Lemma 3.4.
- The composition of maps from $E_2^{p,d}\otimes E_2^{p',d}$ to $H^{p+p'}(B;\mathbb{Z})$ in the diagram is precisely the map (3.4.2).

By the construction of the secondary pushforward, the image of $\xi(a)\otimes\xi(b)$ in $H^{p+p'}(B;\mathbb{Z})$ is the same as the image of $a\otimes b$, which is precisely $\pi_!(a\cup b)$. □

4 Remainder of the proof of Theorem 2.7

The first goal of this section is to prove the following property of the cup product (3.2.1):

Proposition 4.1 *Let $\mathcal{H}$ be a twisted coefficient system with fiber $\mathbb{Z}^k$ with $k\le 2g$. Let $\iota\in H^*(B;\mathcal{H})$ have odd degree. Then*

$$(2g+1)!\cdot\iota^{2g+1}=0\in H^{(2g+1)\deg(\iota)}(B;\mathcal{H}^{\otimes 2g+1}).$$

This proposition is a generalization of the fact that if $\beta \in H^*(B;\mathbb{Z})$ has odd degree, then $2\beta^2 = 0$. Similarly to that fact, the proof relies on the generalized commutativity of cup product with twisted coefficients.

Once we prove Proposition 4.1, we will relate it with Proposition 3.8 to complete the proof of Theorem 2.7.

4.1 Cup product and twisted coefficients

In this section, we state the formal properties of cup product for cohomology with twisted coefficients that we use. They generalize familiar properties of the usual cup product. See [23] for a reference.

Cohomology with twisted coefficients assigns a graded abelian group $H^*(X;\mathcal{A})$ to the pair $(X,\mathcal{A})$ of a space and a twisted coefficient system. Given two coefficient systems $\mathcal{A}$ and $\mathcal{B}$ over the same space X, the *cup product with twisted coefficients* we mentioned in (3.2.1) is a map $\cup\colon H^*(X,\mathcal{A})\otimes H^*(X,\mathcal{B}) \to H^*(X,\mathcal{A}\otimes\mathcal{B})$. Also, given a map of coefficient systems $f\colon \mathcal{A}\to\mathcal{B}$, there is a corresponding map on cohomology $f_{\mathrm{coeff}}\colon H^*(X;\mathcal{A}) \to H^*(X;\mathcal{B})$.

The following properties of cup products on cohomology with twisted coefficients will be important for us:

- The cup product is *associative* in the sense that the two possible cup products of three terms $H^*(X,\mathcal{A})\otimes H^*(X,\mathcal{B})\otimes H^*(X,\mathcal{C}) \to H^*(X,\mathcal{A}\otimes\mathcal{B}\otimes\mathcal{C})$ are the same.
- The cup product *commutes with change of coefficients* in the following sense: Let $f\colon \mathcal{A}\to\mathcal{B}$ be and $g\colon \mathcal{C}\to\mathcal{D}$ be maps of coefficient systems (all over the same space X). There is a corresponding map $f\otimes g\colon \mathcal{A}\otimes\mathcal{C}\to\mathcal{B}\otimes\mathcal{D}$. The following diagram commutes:

$$\begin{tikzcd}
H^*(X;\mathcal{A})\otimes H^*(X;\mathcal{C}) \arrow[r, "f_{\mathrm{coeff}}\otimes g_{\mathrm{coeff}}"] \arrow[d, "\cup"'] & H^*(X;\mathcal{B})\otimes H^*(X;\mathcal{D}) \arrow[d, "\cup"'] \\
H^*(X;\mathcal{A}\otimes\mathcal{C}) \arrow[r, "(f\otimes g)_{\mathrm{coeff}}"] & H^*(X;\mathcal{B}\otimes\mathcal{D})
\end{tikzcd}$$

- The cup product is *graded-commutative* in the following sense: Denote by $\tau\colon \mathcal{A}\otimes\mathcal{B}\to\mathcal{B}\otimes\mathcal{A}$ the map that swaps the coordinates. For $a\in H^p(X;\mathcal{A})$ and $b\in H^q(X;\mathcal{B})$, we have

$$\alpha\cup\beta = (-1)^{pq}\tau_{\mathrm{coeff}}(\beta\cup\alpha). \tag{4.1.1}$$

These facts can be proven in the same way as the corresponding facts for the regular cup product; we refer to [23, Section 11] for details. As in the regular case, graded commutativity of the cup product doesn't hold in general on the level of chains.

4.2 Powers of odd classes and proof of Proposition 4.1

Before proving Proposition 4.1, we need to state two lemmas.

For any representation V of the symmetric group S_n, we denote by $\operatorname{Alt} V$ the *alternating subrepresentation*

$$\operatorname{Alt} V = \{v \in V \mid \forall \sigma \in S_n,\ \sigma \cdot v = \operatorname{sgn}(\sigma) v\} \subset V.$$

Let $\mathcal{H}$ be a twisted coefficient system. Then $\mathcal{H}^{\otimes t}$, for any t, is an S_t–representation with the action defined by $\sigma \cdot (h_1 \otimes \cdots \otimes h_t) = (h_{\sigma(1)} \otimes \cdots \otimes h_{\sigma(t)})$. This action on coefficients also makes the cohomology $H^*(B; \mathcal{H}^{\otimes t})$ into an S_t–representation.

Lemma 4.2 *If $\iota \in H^{\deg(\iota)}(B; \mathcal{H})$ with $\deg(\iota)$ odd, then $\iota^t \in \operatorname{Alt} H^*(B; \mathcal{H}^{\otimes t})$.*

Proof First, consider the $t = 2$ case. Since ι has odd degree, the formula for commutativity of cup product states that, if $\tau \in S_2$ is the nontrivial transposition,

$$\tau_{\mathrm{coeff}}(\iota \cup \iota) = -\iota \cup \iota = \operatorname{sgn}(\tau) \cdot (\iota \cup \iota) \in H^{2\deg(\iota)}(B; \mathcal{H}^{\otimes 2}).$$

The general case follows from the facts that any permutation $\sigma \in S_t$ can be decomposed into a product of transpositions, and that the number of these transpositions mod 2 is determined by $\operatorname{sgn}(\sigma)$. $\square$

The inclusion $i\colon \operatorname{Alt} \mathcal{H}^{\otimes t} \hookrightarrow \mathcal{H}^{\otimes t}$ is a map of coefficient systems, and therefore induces a map on cohomology. If our coefficient system was a $\mathbb{Q}$–vector space, we would want to prove that all of $\operatorname{Alt} H^*(B; \mathcal{H}_{\mathbb{Q}}^{\otimes t})$ is in the image[3] of the map $i_{\mathrm{coeff}}\colon H^*(B; \operatorname{Alt} \mathcal{H}_{\mathbb{Q}}^{\otimes t}) \to H^*(B; \mathcal{H}_{\mathbb{Q}}^{\otimes t})$. We prove an integral version of the same statement.

Lemma 4.3 *Suppose $\alpha \in \operatorname{Alt} H^{\deg \alpha}(B; \mathcal{H}^{\otimes t})$. Then $t!\alpha$ is contained in the image of the map $i_{\mathrm{coeff}}\colon H^*(B; \operatorname{Alt} \mathcal{H}^{\otimes t}) \to H^*(B; \mathcal{H}^{\otimes t})$. By an abuse of notation, we will denote this fact by $t!\alpha \in H^*(B; \operatorname{Alt} \mathcal{H}^{\otimes t})$.*

Proof Consider the map on coefficient systems $p\colon \mathcal{H}^{\otimes t} \to \operatorname{Alt} \mathcal{H}^{\otimes t}$ defined by

$$(v \in \mathcal{H}^{\otimes t}) \xmapsto{p} \Big(\sum_{\sigma \in S_t} \operatorname{sgn}(\sigma)(\sigma \cdot v) \Big).$$

[3] With a little more work, one can show that i_{coeff} induces an isomorphism $H^*(B; \operatorname{Alt} \mathcal{H}_{\mathbb{Q}}^{\otimes t}) \xrightarrow{\sim} \operatorname{Alt} H^*(B; \mathcal{H}_{\mathbb{Q}}^{\otimes t})$.

(It is easy to check that its image indeed lies in $\operatorname{Alt}\mathcal{H}^{\otimes t} \subset \mathcal{H}^{\otimes t}$.) The map on cohomology p_{coeff} has image in $H^*(B; \operatorname{Alt}\mathcal{H}^{\otimes t})$.

At the same time, if $\alpha \in \operatorname{Alt} H^{\deg\alpha}(B; \mathcal{H}^{\otimes t}) \subset H^*(B; \mathcal{H}^{\otimes t})$, then $\sigma_{\mathrm{coeff}} \cdot \alpha = \operatorname{sgn}(\sigma)\alpha$, and thus

$$p_{\mathrm{coeff}}(\alpha) = \sum_{\sigma \in S_t} \operatorname{sgn}(\sigma)(\sigma_{\mathrm{coeff}} \cdot \alpha) = \sum_{\sigma \in S_t} \operatorname{sgn}(\sigma)^2(\alpha) = t!\alpha.$$

So $t!\alpha \in H^*(B; \operatorname{Alt}\mathcal{H}^{\otimes t})$ as desired. □

Proof of Proposition 4.1 Let $\iota \in H^*(B; \mathcal{H})$ have odd degree and suppose that the twisted coefficient system $\mathcal{H}$ has a free abelian group of rank $\leq 2g$ as fiber. Then we have $\operatorname{Alt}\mathcal{H}^{\otimes 2g+1} = 0$. By the above two lemmas, $t!\iota^t \in H^*(B; \operatorname{Alt}\mathcal{H}^{\otimes t})$. So $(2g+1)!\iota^{2g+1} = 0$ as desired. □

Remark 4.4 In the above proof, the full strength of the assumption that $\mathcal{H}$ is free abelian is unnecessary. If the fiber of $\mathcal{H}$ is any finitely generated abelian group such that $\dim_{\mathbb{Q}}(\mathcal{H} \otimes \mathbb{Q}) \leq 2g$, then $\operatorname{Alt}\mathcal{H}^{\otimes 2g+1}$ will be a torsion group, and so ι^{2g+1} will be torsion. If $\mathcal{H}$ is generated by $2g$ elements and has no 2–torsion, $\operatorname{Alt}\mathcal{H}^{\otimes 2g+1} = 0$.

4.3 Proof of Theorem 2.7

Let d be an *odd* natural number and $\pi\colon E \to B$ be an oriented Serre fibration with fiber M_g^{2d}, a $2d$–dimensional highly connected manifold of genus g.

Remark 4.5 The result we prove is more general than the statement of Theorem 2.7, as we do not need to make any assumptions about smoothness of the bundle or of M_g. However, to apply the theorem to more general bundles, one would need to define some sort of "kappa classes" as pushforwards of some cohomology classes on the total space. The results of Ebert and Randal-Williams from [6] show that this is possible in rational cohomology for topological bundles with fiber M_g. Their results also suggests that some kappa classes can be defined this way for *block bundles* with structure group $\widetilde{\operatorname{Diff} M_g}$. To apply the full strength of our results, one would need also to define *intersection classes* (see Definition 5.9) in such a way that Lemma 5.11 holds.

Let us restate Proposition 3.8 from the last section in a form that does not involve spectral sequences. Let $\mathcal{H}$ denote the twisted coefficient system $\mathcal{H}^d(M_g)$ and ω denote the map

$$\omega\colon \mathcal{H} \otimes \mathcal{H} \xrightarrow{\cup} \mathcal{H}^{2d}(M_g) \xrightarrow{\text{or}} \mathbb{Z}.$$

Proposition 4.6 *Let $a \in H^{\deg(a)}(E)$ and $b \in H^{\deg(b)}(E)$ be two classes such that $\pi_!(a)=0$ and $\pi_!(b)=0$. Then there are $\iota \in H^{\deg(a)-d}(B;\mathcal{H})$ and $\kappa \in H^{\deg(b)-d}(B;\mathcal{H})$ that depend only on a and b (respectively) such that $\pi_!(a\cup b)$ is the image of $\iota\otimes\kappa$ under the following composition of maps, where $i=\deg(a)+\deg(b)-2d$:*

$$\begin{tikzcd}
H^{\deg(a)-d}(B;\mathcal{H})\otimes H^{\deg(b)-d}(B;\mathcal{H}) \arrow[r, "\cup"] & H^i(B;\mathcal{H}\otimes\mathcal{H}) \arrow[r, "\omega_{\mathrm{coeff}}"] & H^i(B;\mathbb{Z}) \\
\iota\otimes\kappa \arrow[u, phantom, sloped, "\in"] \arrow[rr, mapsto] & & \pi_!(a\cup b) \arrow[u, phantom, sloped, "\in"]
\end{tikzcd} \tag{4.3.1}$$

Proof The map (3.4.2) from Proposition 3.8 is the composition of the product on the E_2 page of the spectral sequence (3.2.3) with the orientation isomorphism on coefficients:

$$\begin{aligned}
(\mathrm{or}_{\mathrm{coeff}}\circ\bullet)\colon E_2^{p,d}\otimes E_2^{p',d} = H^p(B;\mathcal{H})\otimes H^{p'}(B;\mathcal{H}) &\xrightarrow{\cup} H^{p+p'}(B;\mathcal{H}\otimes\mathcal{H}) \\
&\xrightarrow{\cup_{\mathrm{coeff}}} H^{p+p'}(B;\mathcal{H}^{2d}(M_g)) \\
&\xrightarrow{\mathrm{or}_{\mathrm{coeff}}} H^{p+p'}(B;\mathbb{Z}).
\end{aligned}$$

The composition of the last two arrows in the above diagram is precisely ω_{coeff}, and thus the maps (3.4.2) and (4.3.1) coincide. □

Note that if $\deg(a)$ is *even* while d is *odd*, then $\deg(\iota)$ will be *odd*.

Now the following proposition implies that the map (4.3.1) commutes with taking further cup products. The point is that one can compute the value of $\pi_!(a\cup b)^l$ from the values of ι^l and κ^l. More precisely, we have:

Proposition 4.7 *The following diagram commutes (only up to sign in the top right corner):*

$$\begin{tikzcd}[column sep=large]
\begin{smallmatrix}(\iota\otimes\kappa)\\ \otimes\\ \vdots\\ \otimes\\ (\iota\otimes\kappa)\end{smallmatrix} \in \begin{pmatrix} H^{\deg(a)-d}(B;\mathcal{H}) \\ \otimes \\ H^{\deg(b)-d}(B;\mathcal{H}) \end{pmatrix}^{\otimes l} \arrow[r, "\pm\left(\substack{\text{permute coord.,}\\ \text{then } \cup\otimes\cup}\right)"] \arrow[d, "\cup"] \arrow[dr, "\cup"] & \begin{pmatrix} H^{(\deg(a)-d)\cdot l}(B;\mathcal{H}^{\otimes l}) \\ \otimes \\ H^{(\deg(b)-d)\cdot l}(B;\mathcal{H}^{\otimes l}) \end{pmatrix} \ni \begin{smallmatrix}\pm(\iota\cup\cdots\cup\iota)\\ \otimes\\ (\kappa\cup\cdots\cup\kappa)\end{smallmatrix} \arrow[d, "\cup"', "\text{then permute coefficients}"] \\
H^i(B;\mathcal{H}\otimes\mathcal{H})^{\otimes l} \arrow[r, "\cup"] \arrow[d, "(\omega_{\mathrm{coeff}})^{\otimes l}"] & H^{il}(B;(\mathcal{H}\otimes\mathcal{H})^{\otimes l}) \arrow[d, "(\omega^{\otimes l})_{\mathrm{coeff}}"] \\
\pi_!(a\cup b)^{\otimes l}\in H^i(B;\mathbb{Z})^{\otimes l} \arrow[r, "\cup"] & H^{il}(B;\mathbb{Z}^{\otimes l}\cong\mathbb{Z}) \ni \pi_!(a\cup b)^l
\end{tikzcd}$$

Proof The commutativity of this diagram follows from repeated applications of the associativity of cup product and the fact that cup product commutes with change of coefficients. In the top right corner, we need to also use the commutativity of cup product, which may insert a sign. □

Proof of Theorem 2.7 Let $a, b \in H^*(E;\mathbb{Z})$ be two classes such that $\pi_!(a) = 0$, $\pi_!(b) = 0$, and $\deg(a)$ is even. By the Proposition 4.7 and the decomposition (4.3.1), we see that there are

$$\iota \in H^{\deg(a)-d}(B;\mathcal{H}) \quad \text{and} \quad \kappa \in H^{\deg(b)-d}(B;\mathcal{H})$$

such that $\pi_!(a \cup b)^{2g+1}$ is the image of $\iota^{2g+1} \cup \kappa^{2g+1}$ under some group homomorphism (the composition of the vertical maps on the right side of the diagram in Proposition 4.7). Since $\deg(a)$ is even and d is odd, ι has odd cohomological degree. Since $\operatorname{rank}\mathcal{H} = \operatorname{rank} H^d(M_g;\mathbb{Z}) = 2g$, Proposition 4.1 states that $(2g+1)! \cdot \iota^{2g+1} = 0$. This proves that $(2g+1)! \cdot \pi_!(a \cup b)^{2g+1} = 0$.

Similarly, $\pi_!(a \cup a)^{g+1}$ is the image of $\iota^{g+1} \cup \iota^{g+1} = \iota^{2g+1} \cup \iota$ under a group homomorphism. Again $(2g+1)! \cdot \iota^{2g+1} = 0$ and thus $(2g+1)! \cdot \pi_!(a \cup a)^{g+1} = 0$. □

5 Generating relations using methods of Randal-Williams

In this section, we apply Theorem 2.7 to obtain the results claimed in the introduction as well as some additional relations in $\ker \mathcal{R}_d$.

5.1 Further properties of pushforwards

To do our calculations, we will use the following properties of the pushforward map.

Proposition 5.1 (properties of the pushforward map) *Let $\pi\colon E \to B$ be an oriented Serre fibration with some closed manifold M as fiber. The pushforward map $\pi_!\colon H^{*+\dim(M)}(E;\mathbb{Z}) \to H^*(B;\mathbb{Z})$, as defined in Definition 3.5, satisfies the following:*

(1) *For any classes $a \in H^*(E;\mathbb{Z})$ and $b \in H^*(B;\mathbb{Z})$, we have*

$$\pi_!(a \cup \pi^*(b)) = \pi_!(a) \cup b.$$

This makes the pushforward into a map of $H^*(B;\mathbb{Z})$–modules, and is sometimes called the push-pull formula.

(2) *As already mentioned in Section 2.1, pushforwards are* natural *with respect to maps $f\colon A \to B$. If $\pi'\colon f^*(E) \to A$ is the pullback of the fibration $\pi\colon E \to B$, then for any $a \in H^*(E;\mathbb{Z})$, we have $f^*(\pi_!(a)) = \pi'_!(f^*(a))$.*

(3) *Suppose both maps $G \xrightarrow{\pi''} E$ and $E \xrightarrow{\pi} B$ are oriented Serre fibrations with (possibly different) closed oriented manifolds as fibers. Then so is the composition $(\pi \circ \pi'')\colon G \to B$. Pushforward maps are* functorial *in the sense that $\pi_! \circ \pi''_! = (\pi \circ \pi'')_!$ as maps from the cohomology of G to the cohomology of B.*

For proofs, we refer to [4, Section 8].

We will also need the following well-known fact:

Lemma 5.2 *Let $\pi\colon E \to B$ be an oriented manifold bundle such that B is connected and the fiber is a closed connected oriented manifold M. Let $e = e(T_\pi E \to E) \in H^{\dim M}(E;\mathbb{Z})$. Then $\pi_!(e) = \chi(M) \in H^0(B;\mathbb{Z})$, where $\chi(M) \in \mathbb{Z}$ is the Euler characteristic of M.*

Proof First consider the case when B is a point and $E = M$. The vertical tangent bundle then coincides with the tangent bundle of M. Its Euler class is $e(TM \to M) = \chi(M) \cdot [M]$, where $[M]$ is the generator of $H^{\dim M}(M;\mathbb{Z})$ determined by the orientation. It follows easily from Definition 3.5 that $\pi_!([M]) = 1$ and therefore, by the push-pull formula, $\pi_!(\chi(M) \cdot [M]) = \chi(M) \in H^0(\{*\})$.

In general, consider the inclusion of a point $\{*\} \hookrightarrow B$. The induced map on H^0 is an isomorphism. The desired statement follows from the fact that the Euler class, the vertical tangent bundle, and the pushforward map are all natural with respect to the pullbacks of bundles. □

Remark 5.3 For manifold bundles, there is a commonly used alternative definition of the pushforward map that uses the Pontryagin–Thom construction (see [2] or [1, Section 4]). It coincides with our definition of the pushforward map rationally and, moreover, the two definitions coincide for integral cohomology as long as B is a CW complex of finite type (see Appendix B). We do not know whether the two definitions coincide nor whether Theorem 2.7 applies integrally to the Pontryagin–Thom pushforward more generally, particularly when $B = \mathrm{BDiff}\, M$.

5.2 Notation and conventions

For the remainder of this section, we assume that all cohomology has rational coefficients. Thus we ignore the integral multiple of Theorem 2.7.

Throughout, M_g^{2d} denotes a $2d$–dimensional highly connected manifold of genus g (Definition 2.5). The most important case is when $M_g = \sharp_g\, S^d \times S^d$.

We assume that $2-2g \neq 0$ throughout, and that $2-2g < 0$ in Section 5.5. By the *tautological ring*, we mean the image of the map $\mathcal{R}_d$. We denote this subring by $\mathcal{R}^* = \mathrm{image}(\mathcal{R}_d) \subset H^*(\mathrm{BDiff}\, M_g^{2d};\mathbb{Q})$.

5.3 Direct applications of Theorem 2.7 and the radical

In this section, we illustrate how one can obtain relations using Theorem 2.7 directly. These calculations can serve as a warm-up for more complicated calculations described in Section 5.5. We prove that the tautological ring modulo nilpotent elements is generated by at most $2d$ elements.

Example 5.4 Consider a manifold bundle $\pi\colon E \to B$ with $2d$–dimensional fiber M_g^{2d} and d odd (for example, the universal bundle). If a Pontryagin class $p_i \in H^{4i}(E)$ satisfies $4i < \dim M_g$ then $\pi_!(p_i) = 0$. So the argument of Proposition 2.9 applies to it and we have the following relation concerning $\kappa_{p_i^2} = \pi_!(p_i^2) \in H^{4i\cdot 2-2d}(B)$:

$$(\kappa_{p_i^2})^{g+1} = 0 \in H^{(8i-2d)(g+1)}(B) \quad \text{for} \quad i < \tfrac{1}{2}d = \tfrac{1}{4}\dim M.$$

Example 5.5 More generally, let $p \in H^{2\cdot *}(E)$ be any characteristic class of even degree. Assuming that the Euler characteristic $\chi = 2-2g$ is not zero, we can use the Euler class of the vertical tangent bundle $e \in H^{2d}(E)$ to construct the class $a = p-(e/\chi)\cdot\pi^*(\pi_!(p)) \in H^*(E)$. Because of the push-pull formula (Proposition 5.1) and Lemma 5.2, this class satisfies $\pi_!(a) = 0$.

Let $q \in H^{2\cdot *}(E)$ be another such class. We apply the procedure just described and Theorem 2.7 to obtain the following formula (we use the notation $\pi_!(p) = \kappa_p$):

$$\begin{aligned} 0 &= \left(\pi_!\left(\left(p - \frac{e}{\chi}\kappa_p\right)\left(q - \frac{e}{\chi}\kappa_q\right)\right)\right)^{2g+1} \\ &= \left(\kappa_{pq} - \frac{\kappa_{ep}}{\chi}\kappa_q - \frac{\kappa_{eq}}{\chi}\kappa_p + \frac{\kappa_{e^2}}{\chi^2}\kappa_p\kappa_q\right)^{2g+1}. \end{aligned} \tag{5.3.1}$$

Let $\sqrt{0} \subset \mathcal{R}^*$ denote the *radical* of the tautological ring (that is, the ideal consisting of all the nilpotent element, also known as the *nilradical*). The following easy fact, together with our finite-generation result (Theorem 1.1), provides motivation to consider it.

Lemma 5.6 *If a graded commutative ring A^* is finitely generated as an A^0–algebra and A^0 is a field, then the following statements are equivalent:*

(1) A^* *is finite-dimensional,* (2) $A^*/\sqrt{0} = A^0$, (3) $\dim_{\mathrm{Krull}} A^* = 0$.

Example 5.5 implies:

Lemma 5.7 *In the ring $\mathcal{R}^*/\sqrt{0}$, the class κ_{pq} is in the ideal generated by κ_p and κ_q.*

Proof The expression (5.3.1) implies that

$$\kappa_{pq} - \frac{\kappa_{ep}}{\chi}\kappa_q - \frac{\kappa_{eq}}{\chi}\kappa_p + \frac{\kappa_{e^2}}{\chi^2}\kappa_p\kappa_q \in \sqrt{0}. \qquad \square$$

Proposition 5.8 *If $g \neq 1$, the ring $\mathcal{R}^*/\sqrt{0}$ is generated by the $2d$ elements in the set $E = \{\kappa_{p_i}, \kappa_{p_i \cdot e} \mid 1 \le i \le d\}$. So the Krull dimension of the ring $\mathcal{R}^*$ is at most $2d$.*

Proof Every generator of $\mathcal{R}^*$ that is not in E can be written as κ_{pq} so that $p, q \neq e$. This uses the fact that $p_d = e^2$. It follows that whenever either κ_p or κ_q is not zero, it has strictly positive cohomological degree. By Lemma 5.7, κ_{pq} is decomposable in $\mathcal{R}^*/\sqrt{0}$ as a polynomial in classes of smaller degree. It follows that $\mathcal{R}^*/\sqrt{0}$ is generated by the elements of E. $\square$

5.4 The classifying spaces of manifolds with marked points

To get additional relations, we will use the methods of [22]. Those methods involve certain natural bundles with structure group $\mathrm{Diff}\, M_g$ and fiber $(M_g)^{\times n} = M_g \times \cdots \times M_g$. In this section, we introduce these bundles and the special characteristic classes they possess. The discussion is completely analogous to the two-dimensional case, as described in [22, Section 2.1].

Notation In this section, we denote the universal bundle $\mathrm{EDiff}\, M_g \times_{\mathrm{Diff}\, M_g} M_g \to \mathrm{BDiff}\, M_g$ with fiber M_g as $\mathcal{E}_g^{2d} \to \mathcal{M}_g^{2d}$. The notation refers to the fact that in the case when $d = 1$, the space $\mathcal{M}_g^2$ has the same rational cohomology as the moduli space of Riemann surfaces. We will also use the notation $/\!\!/$ for homotopy quotients: $(- /\!\!/ \mathrm{Diff}\, M) := (- \times_{\mathrm{Diff}\, M} \mathrm{EDiff}\, M)$. For example, $\mathcal{M}_g = * /\!\!/ \mathrm{Diff}\, M_g$ and $\mathcal{E}_g = M_g /\!\!/ \mathrm{Diff}\, M_g$.

For a finite set I, we let $\mathrm{Map}(I; M_g)$ be the space of maps $I \to M_g$,

$$\mathcal{M}_g(I) := \mathrm{Map}(I; M_g) /\!\!/ \mathrm{Diff}\, M_g \quad \text{and} \quad \mathcal{M}_g(n) := \mathcal{M}_g(\{1, \dots, n\}).$$

The fiber of the natural map $\mathcal{M}_g(n) \to \mathcal{M}_g$ is $(M_g)^{\times n}$. So a map from any space B to $\mathcal{M}_g(n)$ gives rise to a manifold bundle over B with fiber M_g together with a choice of n ordered points in each fiber.

For $J \subset I$, there are natural projections $\pi^I_J \colon \mathcal{M}_g(I) \to \mathcal{M}_g(J)$ and $\pi^I_\varnothing \colon \mathcal{M}_g(I) \to \mathcal{M}_g$. We can identify the bundle $\mathcal{M}_g(1) \to \mathcal{M}_g$ with the universal bundle $\mathcal{E}_g \to \mathcal{M}_g$. More generally, the pullback of the universal bundle $(\pi^I_\varnothing)^*(\mathcal{E}_g)$ and $\mathcal{M}_g(I \sqcup \{\star\})$ are canonically isomorphic as bundles over $\mathcal{M}_g(I)$.

Definition 5.9 By the *tautological subring of the cohomology of* $\mathcal{M}_g(I)$ we mean the subring $\mathcal{R}^*(\mathcal{M}_g(I)) \subset H^*(\mathcal{M}_g(I))$ generated by the following three types of classes that we call the *fundamental tautological classes*:

- The generalized MMM classes $\kappa_c \in H^*(\mathcal{M}_g(I))$ that are pulled back from $H^*(\mathcal{M}_g)$ using the canonical map $\mathcal{M}_g(I) \to \mathcal{M}_g$ (there is one such class for each $c \in H^*(\mathrm{BSO}_{2d})$).
- For each choice of $i \in I$, there is a canonical map $\pi_i^I\colon \mathcal{M}_g(I) \to \mathcal{M}_g(\{i\}) \cong \mathcal{E}_g$. The vertical tangent bundle determines a classifying map $\gamma\colon \mathcal{E}_g \to \mathrm{BSO}_{2d}$. For each $c \in H^*(\mathrm{BSO}_{2d})$ and $i \in I$, we define the class $c_{(i)} \in H^*(\mathcal{M}_g(I))$ as the pullback of c via the composition of the above-mentioned maps.[4]
  Note that given $c, d \in H^*(\mathrm{BSO}_{2d})$, we clearly have $(cd)_{(i)} = c_{(i)} d_{(i)}$.
- For each subset $S \subset I$, we consider the *intersection class*
  $$\nu_{(S)} \in H^{2d\cdot(|S|-1)}(\mathcal{M}_g(I))$$
  defined below. We will write simply $\nu_{(1,2)}$ for $\nu_{(\{1,2\})}$.

Definition 5.10 For $S \subset I$, let $\mathrm{Map}(I/S; M_g) \subset \mathrm{Map}(I; M_g)$ be those maps that send all elements of S to the same point. Note that this inclusion has codimension $(|S|-1)\cdot \dim M$. Let $\mathcal{M}_g(I/S) = \mathrm{Map}(I/S; M_g) /\!\!/ \mathrm{Diff}\, M_g$. There is an inclusion $i_S\colon \mathcal{M}_g(I/S) \hookrightarrow \mathcal{M}_g(I)$. As shown in [22, Lemma 2.1], this inclusion has a Thom class
$$\nu'_{(S)} \in H^{2d(|S|-1)}\big(\mathcal{M}_g(I), \mathcal{M}_g(I) - \mathcal{M}_g(I/S); \mathbb{Z}\big).$$
We define the *intersection class* $\nu_{(S)}$ to be the image of $\nu'_{(S)}$ in $H^*(\mathcal{M}_g(I))$.

Lemma 5.11 *The classes $\nu_{(S)}$ satisfy the following:*

(i) *For $S \subset I' \subset I$, the class $\nu_{(S)} \in H^*(\mathcal{M}_g(I))$ is a pullback of the corresponding class $\nu_{(S)} \in H^*(\mathcal{M}_g(I'))$ via the map $(\pi^I_{I'})^*$.*

(ii) *If S and S' intersect at a single point, then $\nu_{(S)}\nu_{(S')} = \nu_{(S\cup S')}$. For example, in $\mathcal{M}_g(\{1,2,\star\})$, we have $\nu_{(1,\star)}\nu_{(2,\star)} = \nu_{(1,\star)}\nu_{(1,2)}$.*

(iii) *In $\mathcal{M}_g(2)$, we have $\nu_{(1,2)}^2 = \nu_{(1,2)} \cdot e_{(1)}$, where e is the Euler class.*

(iv) *For any characteristic class c, we have $\nu_{(1,2)} \cdot c_{(1)} = \nu_{(1,2)} \cdot c_{(2)}$.*

(v) *The pushforward of the class $\nu_{(1,2)} \in H^{2d}(\mathcal{M}_g(2))$ is 1, ie*
$$(\pi^{\{1,2\}}_{\{1\}})_!(\nu_{(1,2)}) = 1 \in H^0(\mathcal{M}_g(1)).$$

[4] We use parentheses in the notation to prevent confusion with the notation p_i for the i^{th} Pontryagin class.

The proof of this lemma is similar to the arguments in [17, Section 11]; see also [22, Lemma 2.1]. The proof of part (v) is very similar to the proof of Lemma 5.2.

Our next goal is to be able to compute the pushforward of any tautological class in $H^*(\mathcal{M}_g(I))$ via the projection maps π^I_J. We will use the properties of the pushforward described in Section 5.1.

Lemma 5.11 and the naturality of the pushforward imply the following.

Lemma 5.12 *For any finite set I, we have*

$$(\pi_I^{I\sqcup\{\star\}})_!(\nu_{(i\star)}) = 1 \quad \text{and} \quad (\pi_I^{I\sqcup\{\star\}})_!(c_{(\star)}) = \kappa_c$$

for all $i\in I$ and $c\in H^*(\mathrm{BSO}_{2d})$. We use the convention $\kappa_e=\chi=2-2g$.

Furthermore, it is possible to rewrite a tautological class in $H^*(\mathcal{M}_g(I\sqcup\{\star\}))$ in terms of a tautological classes in $H^*(\mathcal{M}_g(I))$ as follows:

Lemma 5.13 *We can simplify any monomial in the fundamental tautological classes $m\in H^*(\mathcal{M}_g(I\sqcup\{\star\}))$ in one of the following ways:*

- *If the monomial contains $\nu_{(i,\star)}$ for some $i\in I$, then it can be rewritten as $m=\nu_{(i,\star)}\cdot n'$, where n' is a monomial in classes that do not involve the marked point $\star$. That is, $n'=(\pi_I^{I\sqcup\{\star\}})^*(n)$, where n is a monomial in tautological classes of $\mathcal{M}_g(I)$.*
- *Otherwise, the monomial can be rewritten as $m=c_{(\star)}\cdot n'$, where c is a product (possibly empty) of characteristic classes of the vertical tangent bundle and n' is as before.*

Proof If m does not contain any $\nu_{(i,\star)}$, reordering its terms will put it in the required form. Otherwise, we use the relations

$$\nu_{(i,\star)}\nu_{(j,\star)} = \nu_{(i,\star)}\nu_{(i,j)} \quad \text{and} \quad \nu_{(i,\star)}c_{(\star)} = \nu_{(i,\star)}c_{(i)}$$

from Lemma 5.11 to get rid of any classes that involve $\star$ except for the single $\nu_{(i,\star)}$. □

The push-pull formula and the above lemmas give us the following procedure to compute the pushforward of a general tautological class:

Procedure 5.14 The result of applying the pushforward map

$$(\pi_I^{I\sqcup\{\star\}})_!\colon H^*(\mathcal{M}_g(I\sqcup\{\star\}))\to H^*(\mathcal{M}_g(I))$$

to a tautological class can be computed as follows, one monomial at a time. First, simplify the monomial $m \in H^*(\mathcal{M}_g(I \sqcup \{\star\}))$ using Lemma 5.13. Then apply the push-pull formula and Lemma 5.12 to get one of the following results:

$$(\pi_I^{I\sqcup\{\star\}})_!(m) = (\pi_I^{I\sqcup\{\star\}})_!(\nu_{(i\star)}) \cdot n = n \qquad \text{if } m = \nu_{(i,\star)} \cdot (\pi_I^{I\sqcup\{\star\}})^*(n),$$
$$(\pi_I^{I\sqcup\{\star\}})_!(m) = (\pi_I^{I\sqcup\{\star\}})_!(c_{(\star)}) \cdot n = \kappa_c \cdot n \quad \text{if } m = c_{(\star)} \cdot (\pi_I^{I\sqcup\{\star\}})^*(n).$$

In the second case above, if we have $c_{(\star)} = 1$, then the pushforward will be zero.

Example 5.15 We can compute a pushforward as follows:

$$(\pi_{\{i,j\}}^{\{i,j,\star\}})_!\big(\nu_{(i,\star)}^3 \nu_{(j,\star)}^2 d_{(\star)}\kappa_e\big) = (\pi_{\{i,j\}}^{\{i,j,\star\}})_!\big(\nu_{(i,\star)} e_{(i)}^2 \nu_{(i,j)}^2 d_{(i)}\kappa_e\big) = e_{(i)}^2 \nu_{(i,j)}^2 d_{(i)}\kappa_e.$$

Since pushforward maps are functorial, we can apply Procedure 5.14 several times to calculate $(\pi_J^I)_!$ for any $J \subset I$. There also exist formulas for calculating $(\pi_\varnothing^I)_!$ of a tautological monomial in $H^*(\mathcal{M}_g(I))$ in one step. See [22, Section 2.7] for details.

5.5 Randal-Williams' method and proof of Theorem 1.1

We can obtain numerous relations in the cohomology of $\mathcal{M}_g$ by applying the following idea of [22].

Procedure 5.16 First, we construct some tautological class $c \in \mathcal{R}^*(\mathcal{M}_g(I \sqcup \{\star\}))$ such that $(\pi_I^{I\sqcup\{\star\}})_!(c) = 0$. Applying Theorem 2.7 to one or two such classes will tell us that some polynomial in the ring $\mathcal{R}^*(\mathcal{M}_g(I))$ is equal to zero. We may multiply this relation by any other polynomial and apply $(\pi_\varnothing^I)_!$ to the result to get a relation among the tautological classes of $\mathcal{M}_g$.

We can obtain more relations than were obtained in [22] because the version of our Theorem 2.7 used in [22] (from [18]) only applies when the cohomological degree of c is 2 and does not allow using two cohomology classes at once.

Example 5.17 We illustrate this procedure by repeating, with our notation, the following example from [22, Section 2.2]. Consider the bundle $\pi\colon \mathcal{M}_g(\{1,\star\}) \to \mathcal{M}_g(1)$ (which has fiber M_g). The following class pushes forward to 0:

$$\chi\nu_{(1\star)} - e_{(\star)} \in H^*(\mathcal{M}_g(\{1,\star\})).$$

Theorem 2.7 applies to give us the following relation in the ring $\mathcal{R}^*(\mathcal{M}_g(1))$, which we then simplify using Procedure 5.14 and related lemmas:

$$
\begin{aligned}
(5.5.1)\qquad 0 &= (\pi_!((\chi\nu_{(1\star)} - e_{(\star)})^2))^{g+1} \\
&= (\pi_!(\chi^2\nu_{(1\star)}e_{(1)} - 2\chi\nu_{(1\star)}e_{(1)} + e_{(\star)}^2))^{g+1} \\
&= ((\chi-2)\chi e_{(1)} + \kappa_{e^2})^{g+1} = \sum_{i=0}^{g+1}\binom{g+1}{i}((\chi-2)\chi e_{(1)})^i(\kappa_{e^2})^{g+1-i}.
\end{aligned}
$$

Let us now assume that $\chi = 2-2g < 0$. For each integer k, we can multiply both sides of the formula by $e_{(1)}^k/((\chi-2)\chi)^{g+1}$ and apply $(\pi_{\varnothing}^{\{1\}})_!$ to both sides to get the following relation in the cohomology of $\mathcal{M}_g$:

$$
(5.5.2)\qquad 0 = \sum_{i=0}^{g+1}\binom{g+1}{i}\kappa_{e^{i+k}}\left(\frac{\kappa_{e^2}}{(\chi-2)\chi}\right)^{g+1-i} \in H^{2d(g+k)}(\mathcal{M}_g)
$$

(where we should keep in mind that $\kappa_{e^0} = 0$ and $\kappa_{e^1} = \chi$).

Corollary 5.18 *From the above example, we can see that for $k \geq 0$, the degree $2d(g+k)$ class $\kappa_{g+k} = \kappa_{e^{k+g+1}}$ can be written as a polynomial in lower kappa classes.*

Example 5.19 Assume that $\chi \neq 0$ and fix any $p \in H^{2i}(\mathrm{BSO}_{2d})$. We obtain a relation in the cohomology of $\mathcal{M}_g(1)$ by applying the second part of Theorem 2.7 to the classes $a = \nu_{(1\star)} - e_{(\star)}/\chi \in H^{2d}(\mathcal{M}_g(\{1,\star\}))$ and $b = p_{(\star)} - (e_{(\star)}/\chi)\kappa_p \in H^{2i}(\mathcal{M}_g(\{1,\star\})$ (both classes push down to zero in $\mathcal{M}_g(1)$). The theorem gives us the following formula:

$$
\begin{aligned}
(5.5.3)\qquad 0 &= \big((\pi_{\{1\}}^{\{1,\star\}})_!\big((p_{(\star)} - (e_{(\star)}/\chi)\kappa_p)(\nu_{(1\star)} - (e_{(\star)}/\chi))\big)\big)^{2g+1} \\
&= \left(p_{(1)} - \frac{\kappa_{ep}}{\chi} - \frac{e_{(1)}\kappa_p}{\chi} + \frac{\kappa_{e^2}\kappa_p}{\chi^2}\right)^{2g+1} \in H^*(\mathcal{M}_g(1)).
\end{aligned}
$$

We will use the above example to prove Theorem 1.1. First, we need the following lemma.

Let $\mathcal{A} \subset \mathcal{R}^*(\mathcal{M}_g)$ be the *augmentation ideal* generated by all the elements of the tautological subring that have a nonzero cohomological degree, and let $\mathcal{D} = \mathcal{A}\cdot\mathcal{A}$ be the ideal of the decomposable elements.

Lemma 5.20 *Assume $g > 1$. There is an integer $N > 0$ that depends only on g and d such that for all $p, q \in H^*(\mathrm{BSO}_{2d})$ with $\deg p > 0$,*

$$\kappa_{(p^N q)} \in \mathcal{D} \subset \mathcal{R}^*(\mathcal{M}_g).$$

Proof If $1 \leq \deg p < 2d$, we replace p with p^{2d+1}. This allows us to assume that $\deg p > 2d$.

Let $\mathcal{A}', \mathcal{B}', \mathcal{D}' \subset \mathcal{R}^*(\mathcal{M}_g(1))$ be the following ideals:

$$\mathcal{A}' = \big(\kappa_t \mid t \in H^{>2d}(\mathrm{BSO}_{2d})\big), \quad \mathcal{B}' = \big(t_{(1)} \mid t \in H^{>2d}(\mathrm{BSO}_{2d})\big), \quad \mathcal{D}' = \mathcal{A}' \cdot (\mathcal{A}' + \mathcal{B}').$$

We observe that:

(1) $p_{(1)}^{2g+1} \in \mathcal{D}'$. To see this, note that $e_{(1)}\kappa_p$ and $\kappa_{e^2}\kappa_p$ are in $\mathcal{D}'$ since $\deg(p) > 2d$. Under our assumption that $g > 1$, the formula (5.5.3) implies that we have $p_{(1)}^{2g+1} \in \mathcal{D}'$ as well.

(2) The pushforward operation $(\pi_{\varnothing}^{\{1\}})_!$ takes $\mathcal{D}' \subset \mathcal{M}_g(1)$ into $\mathcal{D} \subset \mathcal{M}_g$.

It follows that $p_{(1)}^{2g+1} q_{(1)} = (p^{2g+1}q)_{(1)} \in \mathcal{D}'$ for all $q \in H^*(\mathrm{BSO}_{2d})$ and, therefore, $\kappa_{(p^{2g+1}q)} \in \mathcal{D}$. □

Now we can finally prove that the tautological ring is finitely generated.

Proof of Theorem 1.1 The infinitely many elements $\kappa_{(e^{a_0}\prod_{i=1}^d p_i^{a_i})}$ (where the a_i are nonnegative integers and the p_i are the Pontryagin classes) generate the tautological ring rationally. By the previous lemma, there is a constant N such that $\kappa_{(e^{a_0}\prod_{i=1}^d p_i^{a_i})}$ is decomposable whenever at least one of the a_i is greater than N. In other words, any such generator is expressible as a polynomial in kappa classes of lower cohomological degree.

So the finitely many generators of cohomological degree less than $\deg(\kappa_{(e^N \prod_{i=1}^d p_i^N)})$ generate the whole tautological subring of $H^*(\mathrm{BDiff}\, M_g; \mathbb{Q})$. □

5.6 Randal-Williams' calculations and high-dimensional manifolds

Using computer calculations, Randal-Williams obtained numerous examples[5] of relations in the $d = 1$ case for $g = 3, 4, 5, 6, 9$ in [22, Section 2] . He also produced a more explicit family of relations in every genus in [22, Section 2.7].

Formally, all the equations and examples from [22] can be interpreted as generators for some ideal $\mathcal{I}_g^{\mathrm{RW}} \subset \mathbb{Q}[\kappa_1, \kappa_2, \dots]$. In this language, the result of [22] is that the ideal $\mathcal{I}_g^{\mathrm{RW}}$ is in the kernel of the map $\mathbb{Q}[\kappa_1, \kappa_2, \dots] \to H^*(\mathrm{BDiff}\, M_g^2)$ in the $d = 1$ case. We will show the following.

Proposition 5.21 *For all odd d, the same ideal $\mathcal{I}_g^{\mathrm{RW}}$ is in the kernel of the corresponding map $\mathbb{Q}[\kappa_1, \kappa_2, \dots] \to H^*(\mathrm{BDiff}\, M_g^{2d})$.*

[5] These include all the relations that exist for $d = 1$ and $g \leq 5$ in degrees $* \leq 2(g-2)$. In higher degrees, the tautological ring vanishes completely according to [13].

As we mentioned in the introduction, this is surprising since the cohomological degree of $\kappa_i = \kappa_{e^{i+1}} \in H^{2di}(\mathrm{BDiff}\, M_g^{2d})$ depends on d.

Example 5.22 ([22, Example 2.5] and Proposition 5.21) For all odd values of d and $g = 4$, we have the following relations in $H^*(\mathrm{BDiff}\, M_4^{2d})$:

$$3\kappa_1^2 = -32\kappa_2 \in H^{4d}(\mathrm{BDiff}\, M_4^{2d}) \quad \text{and} \quad \kappa_2^2 = \kappa_1\kappa_2 = \kappa_3 = 0 \in H^{6d}(\mathrm{BDiff}\, M_4^{2d}).$$

For more examples of relations, see [22, Examples 2.3–2.7].

Proof of Proposition 5.21 First, we repeat the key steps of [22] at our level of generality.

(1) Let $M_g^{2d} \to E \xrightarrow{\pi} B$ be a manifold bundle. Let $c \in H^{2d}(E)$ and $q = \pi_!(c) \in H^0(B) \cong \mathbb{Z}$. The relation (2.3.1) from Theorem 2.7 applied to the cohomology class $(\chi \cdot c - q \cdot e)/\gcd(\chi, q)$ implies that the cohomology class

$$\Omega(E, c) := \frac{1}{(\gcd(\chi, q))^2}(\chi^2 \pi_!(c^2) - 2q\chi\pi_!(e \cdot c) + q^2\kappa_1) \in H^{2d}(B) \tag{5.6.1}$$

has the property that $\Omega(E, c)^{g+1}$ is torsion.

This is precisely the version of [22, Theorem A] that is stated on [22, top of page 1775] for $d = 1$ (we use slightly different notation). Note that the only part of the expression (5.6.1) that depends on d is the cohomological degree.

(2) Consider the bundle $M_g \to \mathcal{E}_g(n) \to \mathcal{M}_g(n)$, defined as the pullback of the universal bundle $\mathcal{E}_g \to \mathcal{M}_g$ to $\mathcal{M}_g(n)$. Following [22], our next step is to apply (5.6.1) to a particular class in the cohomology of its total space.

Recall that $\mathcal{E}_g(n) \cong \mathcal{M}_g(\{1, \dots, n, \star\})$. Given a vector $A = (A_1, \dots, A_n) \in \mathbb{Z}^n$, consider the class

$$c_A := \sum_{i=1}^{n} A_i \nu_{(i\star)} \in H^{2d}(\mathcal{E}_g(n)) = H^{2d}(\mathcal{M}_g(\{1, \dots, n, \star\})).$$

We define the class $\Omega_A := \Omega(\mathcal{E}_g(n), c_A)$ using (5.6.1). It will satisfy $\Omega_A^{g+1} = 0 \in H^{2d(g+1)}(\mathcal{M}_g(n); \mathbb{Q})$. The expression for this class does not depend on d and coincides with [22, (2.1)].

(3) We can now obtain nontrivial examples of relations as follows, repeating the procedure from [22, Section 2.4]. Take the equation $\Omega_A^{g+1} = 0$ for some values of A and n, and perhaps multiply it by another tautological class that doesn't involve Pontryagin classes. Then apply the pushforward $(\pi_{\varnothing}^{\{1,\dots,n\}})_!$ to the result to obtain an element of the kernel of the map $\mathbb{Q}[\kappa_i \mid i \in \mathbb{N}] \to H^*(\mathcal{M}_g^d, \mathbb{Q})$. Every relation obtained in [22] lies in the ideal $\mathcal{I}_g^{\mathrm{RW}} \subset \mathbb{Q}[\kappa_i \mid i \in \mathbb{N}]$ generated by such elements.

To complete the proof, it remains to show that the ideal $\mathcal{I}_g^{\mathrm{RW}}$ does not depend on the value of d. Any tautological class in $H^*(\mathcal{M}_g(n);\mathbb{Q})$ that appears in the above construction (and any tautological class that makes sense for $d=1$) is in the image of the polynomial algebra $\mathbb{Q}[\nu_{(ij)}, e_{(i)}, \kappa_l \mid 1 \le i < j \le n, 1 \le l < \infty]$. The pushforward maps factor through these polynomial algebras. That is to say, there is a map μ that makes the following diagram commute:

$$\begin{tikzcd}
\mathbb{Q}[\nu_{(ij)}, e_{(i)}, \kappa_l \mid 1 \le i < j \le n, 1 \le l < \infty] \arrow[r] \arrow[d, "\mu"] & H^*(\mathcal{M}_g(n);\mathbb{Q}) \arrow[d, "(\pi^{\{1,\dots,n\}}_{\{1,\dots,n-1\}})_!"] \\
\mathbb{Q}[\nu_{(ij)}, e_{(i)}, \kappa_l \mid 1 \le i < j \le n-1, 1 \le l < \infty] \arrow[r] & H^*(\mathcal{M}_g(n-1);\mathbb{Q})
\end{tikzcd}$$

This map μ is determined by Procedure 5.14, and does not depend on the value of d (in fact, only the value of $\kappa_e = \chi = 2-2g$ is at all affected by what the fiber of our bundle is). The expressions for further pushforwards such as $(\pi_{\varnothing}^{\{1,\dots,n\}})_!(b) \in H^*(\mathcal{M}_g)$ also cannot depend on d, since they can be computed by applying Procedure 5.14 repeatedly. It follows that the expressions for the generators of the ideal $\mathcal{I}_g^{\mathrm{RW}}$ do not depend on d, and thus all of Randal-Williams' examples hold verbatim in the $2d$–dimensional case whenever $d \ge 1$ is odd. □

Appendix A: MMM classes related to low Pontryagin classes

In this appendix, we discuss of the images of the maps $\mathcal{R}_d$, $\mathcal{R}'_d$ and $\mathcal{R}_{\delta,d}$ defined in Section 1.1. We prove that the image of $\mathcal{R}_{\delta,d}$ is finitely generated. From now on, we omit the subscript d from the notation.

Proposition A.1 *The maps $\mathcal{R}$, $\mathcal{R}'$ and f^* pictured in diagram* (1.1.1) *are related as follows:*

(1) *There are classes $q_1, \dots, q_{\lceil (d+1/4)\rceil -1} \in \operatorname{image}(\mathcal{R}) \subset H^*(\operatorname{BDiff} M_g;\mathbb{Q})$ that generate $\operatorname{image}(\mathcal{R})$ as an $\operatorname{image}(\mathcal{R}')$–module.*

(2) *$f^*(q_i) = 0 \in H^*(\operatorname{BDiff}(M_g, D^{2d});\mathbb{Q})$ for all i.*

Proof Let $\pi\colon U \to \operatorname{BDiff} M_g$ be the universal bundle and $p_i \in H^*(U;\mathbb{Q})$ be the Pontryagin classes of the vertical tangent bundle. Since M_g is $(d-1)$–connected, the map $\pi^*\colon H^*(\operatorname{BDiff} M_g;\mathbb{Q}) \to H^*(U;\mathbb{Q})$ is an isomorphism in degrees $* < d$ (this can be seen eg using the Serre spectral sequence). It follows that there are classes $q_i \in H^*(\operatorname{BDiff} M_g;\mathbb{Q})$ such that $p_i = \pi^*(q_i)$ for all $i < \lceil (d+1)/4\rceil$.

Now let $m \in \mathcal{S}$. If $\deg m \le 2d$, then $\kappa_m = 0$ or $\kappa_m \in \mathbb{Q}$, so $\kappa_m \in \operatorname{image}\mathcal{R}' \subset \operatorname{image}\mathcal{R}$. If $\deg m > 2d$, then m can be decomposed as a product of some $n \in \mathcal{S}'$ and some Pontryagin classes p_i with $i < \lceil (d+1)/4\rceil$. Since the pushforward is a map of

$H^*(\mathrm{BDiff}\, M_g;\mathbb{Q})$–modules, $\kappa_m = \pi_!\big(n \cdot \prod \pi^*(q_i)\big) = \kappa_n \cdot \prod q_i$ for some indices i. In other words, the q_i generate $\mathrm{image}(\mathcal{R})$ as an $\mathrm{image}(\mathcal{R}')$–module, as desired.

Let us now prove that $f^*(q_i) = 0$ for all i. It is sufficient to consider the universal bundle with a fixed disk and prove that the corresponding universal classes $q_i \in H^*(\mathrm{BDiff}(M_g, D^{2d});\mathbb{Q})$ are zero. We can fix a basepoint $b \in D^{2d} \subset M_g^{2d}$ that determines a section of the universal bundle (which we denote U_δ). The following diagram describes the corresponding map on cohomology:

$$\begin{tikzcd} U_\delta = \mathrm{EDiff}(M_g, D^{2d}) \times_{\mathrm{Diff}(M_g,D^{2d})} M_g \arrow[d, shift left, "\pi"] \\ \mathrm{BDiff}(M_g, D^{2d}) \arrow[u, shift left, "s"] \end{tikzcd} \qquad \begin{tikzcd} H^*(U_\delta;\mathbb{Q}) \arrow[d, shift right, "s^*"'] \\ H^*(\mathrm{BDiff}(M_g, D^{2d});\mathbb{Q}) \arrow[u, shift right, "\pi^*"'] \end{tikzcd}$$

As s is a section we must have $s^*(p_i) = s^*(\pi^*(q_i)) = q_i$ as long as $i < \lceil (d+1)/4 \rceil$. So $q_i = s^*(p_i)$ is a characteristic class of the bundle $s^*(T_\pi U_\delta)$ over $\mathrm{BDiff}(M_g, D^{2d})$. Since a neighborhood of the point b is fixed by the action of $\mathrm{Diff}(M_g, D^{2d})$, this bundle is trivial, and so q_i must be zero. □

Observation A.2 *For $d > 3$, in the notation of the proof above, $p_1 = \pi^*(q_1) \in H^*(U)$. Therefore, for all g,*

$$\chi \kappa_{e^2 p_1} = \chi \pi_!(e^2 \cdot \pi^*(q_1)) = \pi_!(e) \cdot q_1 \cdot \pi_!(e^2) = \kappa_{ep_1} \kappa_{e^2} \in H^*(\mathrm{BDiff}\, M_g;\mathbb{Q}).$$

So the map $\mathcal{R}$ has nontrivial relations in its kernel that do not depend on g. This cannot happen in $\ker \mathcal{R}_\delta$ or $\ker \mathcal{R}'$ by Fact 1.3.

Proposition A.1 implies the following.

Corollary A.3 *$f^*(\kappa_m) = 0$ if $\kappa_m \in \mathrm{image}(\mathcal{R}) - \mathrm{image}(\mathcal{R}')$. So $\mathrm{image}(f^* \circ \mathcal{R}) = \mathrm{image}(\mathcal{R}_\delta)$.*

Theorem A.4 *The image of $\mathcal{R}_{\delta,d}$ is a finitely generated as a $\mathbb{Q}$–algebra when d is odd and $g > 1$.*

Proof By the above corollary, the image of the map $\mathcal{R}_\delta$ is a quotient of the image of the map $\mathcal{R}$, which is finitely generated by Theorem 1.1. □

Remark A.5 If we require that all the Pontryagin classes p_i mentioned in Section 5 satisfy $i \geq \lceil (d+1)/4 \rceil$, all of the arguments in that section will apply to the map $\mathcal{R}' \colon \mathbb{Q}[\kappa_p \mid p \in \mathcal{S}'] \to H^*(\mathrm{BDiff}(M_g);\mathbb{Q})$ without any further modification. This way, one can prove that the image of the map $\mathcal{R}'$ is also finitely generated. That gives another proof that the image of $\mathcal{R}_\delta$ is finitely generated.

Appendix B: The Pontryagin–Thom pushforward

While the definition of the pushforward map used throughout this paper applies to all oriented Serre fibrations, in the case of manifold bundles (M is a smooth closed oriented manifold and $\pi\colon E \to B$ is a bundle with structure group $\operatorname{Diff} M$), there is another commonly used definition of the pushforward map $\pi_{!\mathrm{PT}}\colon H^{*+m}(E;\mathbb{Z}) \to H^*(B;\mathbb{Z})$ that uses the Pontryagin–Thom construction; see [2] or [1, Section 4]. This *Pontryagin–Thom pushforward* has the advantage of being defined even for generalized cohomology theories if the bundle has an appropriate orientation. It is also necessary for constructing the kappa classes as pullbacks of natural classes in the cohomology of the infinite-loop space $\Omega^\infty \mathrm{MTSO}(2d)$ in the manner of [14]. While we do not use that construction explicitly, it is needed in the proof of Fact 1.3.

It is conceivable that the notion of kappa classes depends on which definition of the pushforwards one uses. We do not know whether $\pi_!$ and $\pi_{!\mathrm{PT}}$ coincide for integral cohomology when $B = \mathrm{BDiff}\, M$. However, the following fact applies in most relevant cases. It is accepted in the literature, but we provide a proof for completeness.

Proposition B.1 *If $E \to B$ is a manifold bundle with structure group* $\operatorname{Diff} M$ *and B is a CW complex of finite type, the pushforwards $\pi_{!\mathrm{PT}}$ and $\pi_!$ coincide.*

In rational cohomology, $\pi_{!\mathrm{PT}}$ and $\pi_!$ coincide for any CW complex B.

Proof One can check that the Pontryagin–Thom construction commutes with bundle pullbacks in an appropriate way so that $\pi_{!\mathrm{PT}}$ satisfies the naturality property (2) from Proposition 5.1. If we either work in rational cohomology or assume that B is a CW complex of finite type, we have (see eg [12, Section 3.F] for an overview)

$$H^*(B) = \varprojlim_{\substack{B' \subset B \\ \text{finite subcomplex}}} H^*(B').$$

So we can assume without loss of generality that B is a finite CW complex. Finally, we use the Lemma B.2 below to reduce the case of a finite CW complex to the case of B a closed oriented manifold.

In the case when B is a closed oriented manifold, the fact that $\pi_{!\mathrm{PT}}$ and $\pi_!$ coincide is proven in [2]. Briefly, Boardman proves a multiplicativity property for the *cap* product, similar to property (1) from Proposition 5.1, for both $\pi_!$ and $\pi_{!\mathrm{PT}}$. He then deduces that both pushforwards must coincide with the pushforward determined by Poincaré duality. □

Lemma B.2 *Any finite CW complex B is a retract of a smooth oriented closed manifold D. In particular, there is a map $f\colon D \to B$ such that $f^*\colon H^*(B;\mathbb{Z}) \to H^*(D;\mathbb{Z})$ is injective.*

Proof [6] It is possible to embed B into a Euclidean space. A sufficiently small tubular neighborhood T of such an embedding will be an oriented compact manifold *with boundary* that deformation retracts onto T (see eg the appendix of [12]). In particular, we have maps $B \overset{i}{\hookrightarrow} T \xrightarrow{f'} B$ such that the composition is the identity.

Let $D = T \sqcup_{\partial T} (-T)$ be the double of T. It is a closed oriented manifold. There is an obvious inclusion $T \hookrightarrow D$ and, crucially, the map $f'\colon T \to B$ extends to a map $f\colon D \to B$. So we have our retraction

$$\begin{tikzcd} B \arrow[r, hook, "i"] & T \arrow[r, hook] \arrow[rr, bend right, "f'"'] & D \arrow[r, "f"] & B. \end{tikzcd}$$

The composition is the identity since it coincides with $f' \circ i$. □

References

[1] **J C Becker**, **D H Gottlieb**, *The transfer map and fiber bundles*, Topology 14 (1975) 1–12 MR

[2] **J M Boardman**, *Stable homotopy theory, V: Duality and Thom spectra*, notes, University of Warwick (1966) Available at `http://math.ucr.edu/~res/inprogress/Boardman-V.pdf`

[3] **S K Boldsen**, *Improved homological stability for the mapping class group with integral or twisted coefficients*, Math. Z. 270 (2012) 297–329 MR

[4] **A Borel**, **F Hirzebruch**, *Characteristic classes and homogeneous spaces, I*, Amer. J. Math. 80 (1958) 458–538 MR

[5] **C J Earle**, **J Eells**, *A fibre bundle description of Teichmüller theory*, J. Differential Geometry 3 (1969) 19–43 MR

[6] **J Ebert**, **O Randal-Williams**, *Generalised Miller–Morita–Mumford classes for block bundles and topological bundles*, Algebr. Geom. Topol. 14 (2014) 1181–1204 MR

[7] **C Faber**, *A conjectural description of the tautological ring of the moduli space of curves*, from "Moduli of curves and abelian varieties" (C Faber, E Looijenga, editors), Aspects Math. E33, Friedr. Vieweg, Braunschweig (1999) 109–129 MR

[8] **B Farb**, **D Margalit**, *A primer on mapping class groups*, Princeton Mathematical Series 49, Princeton University Press (2012) MR

[6] We thank Alexander Kupers for a key idea for this proof. This argument is also in [24].

[9] **S Galatius**, **O Randal-Williams**, *Homological stability for moduli spaces of high dimensional manifolds, I*, preprint (2014) arXiv

[10] **S Galatius**, **O Randal-Williams**, *Stable moduli spaces of high-dimensional manifolds*, Acta Math. 212 (2014) 257–377 MR

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

[12] **A Hatcher**, *Algebraic topology*, Cambridge University Press (2002) MR

[13] **E Looijenga**, *On the tautological ring of $\mathcal{M}_g$*, Invent. Math. 121 (1995) 411–419 MR

[14] **I Madsen**, **U Tillmann**, *The stable mapping class group and $Q(\mathbb{C}\mathrm{P}^\infty_+)$*, Invent. Math. 145 (2001) 509–544 MR

[15] **I Madsen**, **M Weiss**, *The stable moduli space of Riemann surfaces: Mumford's conjecture*, Ann. of Math. 165 (2007) 843–941 MR

[16] **J McCleary**, *A user's guide to spectral sequences*, 2nd edition, Cambridge Studies in Advanced Mathematics 58, Cambridge University Press (2001) MR

[17] **J W Milnor**, **J D Stasheff**, *Characteristic classes*, Annals of Mathematics Studies 76, Princeton University Press (1974) MR

[18] **S Morita**, *Families of Jacobian manifolds and characteristic classes of surface bundles, I*, Ann. Inst. Fourier (Grenoble) 39 (1989) 777–810 MR

[19] **S Morita**, *Generators for the tautological algebra of the moduli space of curves*, Topology 42 (2003) 787–819 MR

[20] **D Mumford**, *Towards an enumerative geometry of the moduli space of curves*, from "Arithmetic and geometry, II" (M Artin, J Tate, editors), Progr. Math. 36, Birkhäuser, Boston (1983) 271–328 MR

[21] **R Pandharipande**, **A Pixton**, *Relations in the tautological ring of the moduli space of curves*, preprint (2013) arXiv

[22] **O Randal-Williams**, *Relations among tautological classes revisited*, Adv. Math. 231 (2012) 1773–1785 MR

[23] **N E Steenrod**, *Homology with local coefficients*, Ann. of Math. 44 (1943) 610–627 MR

[24] **R Thom**, *Quelques propriétés globales des variétés différentiables*, Comment. Math. Helv. 28 (1954) 17–86 MR

Department of Mathematics, University of Chicago
5734 S University Ave, Chicago, IL 60637, United States

ilyag@uchicago.edu

http://math.uchicago.edu/~ilyagr/

Proposed: Shigeyuki Morita Received: 30 October 2013
Seconded: Ralph Cohen, Stefan Schwede Revised: 25 May 2016

Geometry & Topology 21 (2017) 2049–2092

The Eynard–Orantin recursion and equivariant mirror symmetry for the projective line

BOHAN FANG
CHIU-CHU MELISSA LIU
ZHENGYU ZONG

We study the equivariantly perturbed mirror Landau–Ginzburg model of $\mathbb{P}^1$. We show that the Eynard–Orantin recursion on this model encodes all-genus, all-descendants equivariant Gromov–Witten invariants of $\mathbb{P}^1$. The nonequivariant limit of this result is the Norbury–Scott conjecture, while by taking large radius limit we recover the Bouchard–Mariño conjecture on simple Hurwitz numbers.

14N35

1 Introduction

The equivariant Gromov–Witten theory of $\mathbb{P}^1$ has been studied extensively. Okounkov and Pandharipande [27; 28] completely solved the equivariant Gromov–Witten theory of the projective line and established a correspondence between the stationary sector of Gromov–Witten theory of $\mathbb{P}^1$ and Hurwitz theory. Givental [20] derived a quantization formula for the all-genus descendant potential of the equivariant Gromov–Witten theory of $\mathbb{P}^1$ (and more generally, $\mathbb{P}^n$). In the nonequivariant limit, these results imply the Virasoro conjecture of $\mathbb{P}^1$.

The Norbury–Scott conjecture [26] relates (nonequivariant) Gromov–Witten invariants of $\mathbb{P}^1$ to Eynard–Orantin invariants [10] of the affine plane curve

$$\left\{x = Y + \frac{1}{Y} \;\middle|\; (x, Y) \in \mathbb{C} \times \mathbb{C}^*\right\}.$$

P Dunin-Barkowski, N Orantin, S Shadrin and L Spitz [5] relate the Eynard–Orantin topological recursion to the Givental formula for the ancestor formal Gromov–Witten potential, and prove the Norbury–Scott conjecture using their main result and Givental's quantization formula for the all-genus descendant potential of the (nonequivariant) Gromov–Witten theory of $\mathbb{P}^1$. It is natural to ask if the Norbury–Scott conjecture can be extended to the equivariant setting, in a way that the original conjecture can be recovered in the nonequivariant limit.

Published: 19 May 2017 DOI: 10.2140/gt.2017.21.2049

1.1 Main results

Our first main result (Theorem A in Section 3.7) relates equivariant Gromov–Witten invariants of $\mathbb{P}^1$ to the Eynard–Orantin invariants [10] of the affine curve

$$\left\{x = t^0 + Y + \frac{Qe^{t^1}}{Y} + \mathsf{w}_1 \log Y + \mathsf{w}_2 \log \frac{Qe^{t^1}}{Y} \;\middle|\; (x, Y) \in \mathbb{C} \times \mathbb{C}^*\right\},$$

where t^0 and t^1 are complex parameters, w_1 and w_2 are equivariant parameters of the torus $T = (\mathbb{C}^*)^2$ acting on $\mathbb{P}^1$, and Q is the Novikov variable encoding the degree of the stable maps to $\mathbb{P}^1$ (see Section 2.2). The superpotential of the T–equivariant Landau–Ginzburg mirror of the projective line is given by

$$W_t^{\mathsf{w}}\colon \mathbb{C}^* \to \mathbb{C}, \quad W_t^{\mathsf{w}}(Y) = t_0 + Y + \frac{Qe^{t^1}}{Y} + \mathsf{w}_1 \log Y + \mathsf{w}_2 \log \frac{Qe^{t^1}}{Y},$$

so Theorem A can be viewed as a version of all-genus equivariant mirror symmetry for $\mathbb{P}^1$. We prove Theorem A using the main result in [5] and a suitable version of Givental's formula [20] for all-genus *equivariant* descendant Gromov–Witten potential of $\mathbb{P}^n$ (see also Lee and Pandharipande [24]).

Our second main result (Theorem B in Section 3.7) gives a precise correspondence between genus-g, n–point descendant equivariant Gromov–Witten invariants of $\mathbb{P}^1$ and Laplace transforms of the Eynard–Orantin invariant $\omega_{g,n}$ along Lefschetz thimbles. This result generalizes the known relation between the A–model, genus-0, 1–point descendant Gromov–Witten invariants and the B–model oscillatory integrals.

1.2 Nonequivariant limit and the Norbury–Scott conjecture

Taking the nonequivariant limit $\mathsf{w}_1 = \mathsf{w}_2 = 0$, we obtain

$$W_t(Y) = t^0 + Y + \frac{Qe^{t^1}}{Y},$$

which is the superpotential of the (nonequivariant) Landau–Ginzburg mirror for the projective line. We obtain all-genus (nonequivariant) mirror symmetry for the projective line.

In the stationary phase $t^0 = t^1 = 0$ and $Q = 1$, the curve becomes

$$\{x = Y + Y^{-1} : (x, Y) \in \mathbb{C} \times \mathbb{C}^*\},$$

and Theorem A specializes to the Norbury–Scott conjecture [26]. (See Section 4.2 for details.)

1.3 Large radius limit and the Bouchard–Mariño conjecture

Let $\mathsf{w}_2 = 0$, $t_0 = 0$ and $q = Qe^{t^1}$; we obtain

$$x = Y + \frac{q}{Y} + \mathsf{w}_1 \log Y,$$

which reduces to

$$x = Y + \mathsf{w}_1 \log Y$$

in the large radius limit $q \to 0$. The $\mathbb{C}^*$–equivariant mirror of the affine line $\mathbb{C}$ is given by

$$W\colon \mathbb{C}^* \to \mathbb{C}, \quad W(Y) = Y + \mathsf{w}_1 \log Y.$$

In the large radius limit, we obtain a version of all-genus $\mathbb{C}^*$–equivariant mirror symmetry of the affine line $\mathbb{C}$.

In particular, letting $\mathsf{w}_1 = -1$ and $X = e^{-x}$, we obtain the Lambert curve

$$X = Ye^{-Y}.$$

In this limit, Theorem A specializes to the Bouchard–Mariño conjecture [2] relating simple Hurwitz numbers (related to linear Hodge integrals by the ELSV formula of Ekedahl, Lando, Shapiro and Vainshtein [6] and Graber and Vakil [21]) to Eynard–Orantin invariants of the Lambert curve. (See Section 5 for details.)

Borot, Eynard, Mulase and Safnuk [1] introduced a new matrix model representation for the generating function of simple Hurwitz numbers, and proved the Bouchard–Mariño conjecture. Eynard, Mulase and Safnuk [9] provided another proof of the Bouchard–Mariño conjecture using the cut-and-joint equation of simple Hurwitz numbers. Recently, new proofs of the ELSV formula and the Bouchard–Mariño conjecture have been given by Dunin-Barkowski, Kazarian, Orantin, Shadrin and Spitz [4].

Acknowledgment We thank P Dunin-Barkowski, B Eynard, M Mulase, P Norbury and N Orantin for helpful conversations. The research of the authors is partially supported by NSF DMS-1206667 and NSF DMS-1159416.

2 A–model

Let $T = (\mathbb{C}^*)^2$ act on $\mathbb{P}^1$ by

$$(t_1, t_2) \cdot [z_1, z_2] = [t_1^{-1} z_1, t_2^{-1} z_2].$$

Let $\mathbb{C}[\mathsf{w}] := \mathbb{C}[\mathsf{w}_1, \mathsf{w}_2] = H_T^*(\text{point}; \mathbb{C})$ be the T–equivariant cohomology of a point.

2.1 Equivariant cohomology of $\mathbb{P}^1$

The T–equivariant cohomology of $\mathbb{P}^1$ is given by

$$H_T^*(\mathbb{P}^1;\mathbb{C}) = \mathbb{C}[H,\mathsf{w}]/\langle (H-\mathsf{w}_1)(H-\mathsf{w}_2)\rangle,$$

where $\deg H = \deg \mathsf{w}_i = 2$. Let $p_1=[1,0]$ and $p_2=[0,1]$ be the T fixed points. Then $H|_{p_i} = \mathsf{w}_i$. The T–equivariant Poincaré dual of p_1 and p_2 are $H-\mathsf{w}_2$ and $H-\mathsf{w}_1$, respectively. Let

$$\phi_1 := \frac{H-\mathsf{w}_2}{\mathsf{w}_1-\mathsf{w}_2},\ \phi_2 := \frac{H-\mathsf{w}_1}{\mathsf{w}_2-\mathsf{w}_1} \in H_T^*(\mathbb{P}^1;\mathbb{C})\otimes_{\mathbb{C}[\mathsf{w}]}\mathbb{C}\Big[\mathsf{w},\frac{1}{\mathsf{w}_1-\mathsf{w}_2}\Big]$$

Then $\deg\phi_\alpha = 0$, and

$$\phi_\alpha\cup\phi_\beta = \delta_{\alpha\beta}\phi_\alpha,$$

So $\{\phi_1,\phi_2\}$ is a canonical basis of the semisimple algebra

$$H_T^*(\mathbb{P}^1;\mathbb{C})\otimes_{\mathbb{C}[\mathsf{w}]}\mathbb{C}\Big[\mathsf{w},\frac{1}{\mathsf{w}_1-\mathsf{w}_2}\Big].$$

We have

$$\phi_1+\phi_2 = 1,$$

$$(\phi_\alpha,\phi_\beta) := \int_{\mathbb{P}^1}\phi_\alpha\cup\phi_\beta = \delta_{\alpha\beta}\int_{\mathbb{P}^1}\phi_\alpha = \frac{\delta_{\alpha\beta}}{\Delta^\alpha},\quad \alpha,\beta\in\{1,2\},$$

where

$$\Delta^1 = \mathsf{w}_1-\mathsf{w}_2,\quad \Delta^2 = \mathsf{w}_2-\mathsf{w}_1.$$

Cup product with the hyperplane class is given by

$$H\cup\phi_\alpha = \mathsf{w}_\alpha\phi_\alpha,\quad \alpha=1,2.$$

2.2 Equivariant Gromov–Witten invariants of $\mathbb{P}^1$

Suppose that $d>0$ or $2g-2+n>0$, so that $\overline{\mathcal{M}}_{g,n}(\mathbb{P}^1,d)$ is nonempty. Given $\gamma_1,\dots,\gamma_n\in H_T^*(\mathbb{P}^1,\mathbb{C})$ and $a_1,\dots,a_n\in\mathbb{Z}_{\geq 0}$, we define genus-$g$, degree-$d$, T–equivariant descendant Gromov–Witten invariants of $\mathbb{P}^1$:

$$\langle \tau_{a_1}(\gamma_1)\dots\tau_{a_n}(\gamma_n)\rangle_{g,n,d}^{\mathbb{P}^1,T} := \int_{[\overline{\mathcal{M}}_{g,n}(\mathbb{P}^1,d)]^{\mathrm{vir}}}\prod_{j=1}^n \psi_j^{a_j}\operatorname{ev}_j^*(\gamma_j)\in\mathbb{C}[\mathsf{w}],$$

where $\operatorname{ev}_j\colon \overline{\mathcal{M}}_{g,n}(\mathbb{P}^1,d)\to\mathbb{P}^1$ is the evaluation at the j^{th} marked point, which is a T–equivariant map. We define genus-g, degree-d primary Gromov–Witten invariants:

$$\langle\gamma_1,\dots,\gamma_n\rangle_{g,n,d}^{\mathbb{P}^1,T} := \langle\tau_0(\gamma_1)\cdots\tau_0(\gamma_n)\rangle_{g,n,d}^{\mathbb{P}^1,T}.$$

Let $\boldsymbol{t} = t^0 1 + t^1 H$. If $2g-2+n>0$, define

$$\langle\!\langle \tau_{a_1}(\gamma_1), \dots, \tau_{a_n}(\gamma_n) \rangle\!\rangle_{g,n}^{\mathbb{P}^1,T} := \sum_{d\geq 0} Q^d \sum_{l=0}^{\infty} \frac{1}{l!} \langle \tau_{a_1}(\gamma_1) \cdots \tau_{a_n}(\gamma_n) \underbrace{\tau_0(\boldsymbol{t}) \cdots \tau_0(\boldsymbol{t})}_{l \text{ times}} \rangle_{g,n+l,d}^{\mathbb{P}^1,T}.$$

Suppose that $2g-2+n+m>0$. Given $\gamma_1, \dots, \gamma_{n+m} \in H_T^*(\mathbb{P}^1)$, we define

$$\begin{aligned}\Big\langle \frac{\gamma_1}{z_1-\psi_1}, \dots, \frac{\gamma_n}{z_n-\psi_n}, \gamma_{n+1}, \dots, \gamma_{n+m} \Big\rangle_{g,n+m,d}^{\mathbb{P}^1,T} \\ := \sum_{a_1,\dots,a_n\geq 0} \langle \tau_{a_1}(\gamma_1) \cdots \tau_{a_n}(\gamma_n) \tau_0(\gamma_{n+1}) \cdots \tau_0(\gamma_{n+m}) \rangle_{g,n+m,d}^{\mathbb{P}^1,T} \prod_{j=1}^{n} z_j^{-a_j-1}.\end{aligned}$$

In particular, if $n+m\geq 3$ then

$$\begin{aligned}(1) \quad & \Big\langle \frac{\gamma_1}{z_1-\psi_1}, \dots, \frac{\gamma_n}{z_n-\psi_n}, \gamma_{n+1}, \dots, \gamma_{n+m} \Big\rangle_{0,n+m,0}^{\mathbb{P}^1,T} \\ & \qquad = \frac{1}{z_1 \cdots z_n} \Big(\frac{1}{z_1} + \cdots + \frac{1}{z_n} \Big)^{n+m-3} \int_{\mathbb{P}^1} \gamma_1 \cup \cdots \cup \gamma_{n+m},\end{aligned}$$

where we use the fact $\overline{\mathcal{M}}_{0,n+m}(\mathbb{P}^1, 0) = \overline{\mathcal{M}}_{0,m+n} \times \mathbb{P}^1$, and the identity

$$\int_{\overline{\mathcal{M}}_{0,k}} \psi_1^{a_1} \cdots \psi_k^{a_k} = \begin{cases} \dfrac{(k-3)!}{\prod_{j=1}^k a_j!} & \text{if } a_1 + \cdots + a_k = k-3, \\ 0 & \text{otherwise.} \end{cases}$$

We use the second line of (1) to extend the definition of the correlator in the first line of (1) to the unstable cases $(n,m) = (1,0)$, $(1,1)$, $(2,0)$:

$$\Big\langle \frac{\gamma_1}{z_1-\psi_1} \Big\rangle_{0,1,0}^{\mathbb{P}^1,T} := z_1 \int_{\mathbb{P}^1} \gamma_1,$$

$$\Big\langle \frac{\gamma_1}{z_1-\psi_1}, \gamma_2 \Big\rangle_{0,2,0}^{\mathbb{P}^1,T} := \int_{\mathbb{P}^1} \gamma_1 \cup \gamma_2,$$

$$\Big\langle \frac{\gamma_1}{z_1-\psi_1}, \frac{\gamma_2}{z_2-\psi_2} \Big\rangle_{0,2,0}^{\mathbb{P}^1,T} := \frac{1}{z_1+z_2} \int_{\mathbb{P}^1} \gamma_1 \cup \gamma_2.$$

Suppose that $2g-2+n+m>0$ or $n>0$. Given $\gamma_1, \dots, \gamma_{n+m} \in H_T^*(\mathbb{P}^1)$, we define

$$\left\langle\!\!\left\langle \frac{\gamma_1}{z_1-\psi_1},\dots,\frac{\gamma_n}{z_n-\psi_n},\gamma_{n+1},\dots,\gamma_{n+m}\right\rangle\!\!\right\rangle_{g,n+m}^{\mathbb{P}^1,T}$$
$$:=\sum_{d\ge 0}\sum_{l\ge 0}\frac{Q^d}{l!}\left\langle \frac{\gamma_1}{z_1-\psi_1},\dots,\frac{\gamma_n}{z_n-\psi_n},\gamma_{n+1},\dots,\gamma_{n+m},\underbrace{\boldsymbol{t},\dots,\boldsymbol{t}}_{l\text{ times}}\right\rangle_{g,n+m+l,d}^{\mathbb{P}^1,T}.$$

Let $q=Qe^{t^1}$. Then, for $m\ge 3$,

$$\langle\!\langle \gamma_1,\dots,\gamma_m\rangle\!\rangle_{0,m}^{\mathbb{P}^1,T}=\sum_{d\ge 0}q^d\langle\gamma_1,\dots,\gamma_m\rangle_{0,m,d}^{\mathbb{P}^1,T}=\delta_{m,3}\int_{\mathbb{P}^1}\gamma_1\cup\dots\cup\gamma_m+q\prod_{i=1}^m\int_{\mathbb{P}^1}\gamma_i.$$

2.3 Equivariant quantum cohomology of $\mathbb{P}^1$

As a $\mathbb{C}[\mathsf{w}]$–module, $QH_T^*(\mathbb{P}^1;\mathbb{C})=H_T^*(\mathbb{P}^1;\mathbb{C})$. The ring structure is given by the quantum product $*$ defined by

$$(\gamma_1\star\gamma_2,\gamma_3)=\langle\!\langle\gamma_1,\gamma_2,\gamma_3\rangle\!\rangle_{0,3}^{\mathbb{P}^1,T},$$

or equivalently,

$$\gamma_1\star\gamma_2=\gamma_1\cup\gamma_2+q\left(\int_{\mathbb{P}^1}\gamma_1\right)\left(\int_{\mathbb{P}^1}\gamma_2\right),$$

where $\cup$ is the product in $H_T^*(\mathbb{P}^1)$ and $q=Qe^{t^1}$. In particular,

$$H\star H=(\mathsf{w}_1+\mathsf{w}_2)H-\mathsf{w}_1\mathsf{w}_2+q.$$

The T–equivariant quantum cohomology of $\mathbb{P}^1$ is

$$QH_T^*(\mathbb{P}^1;\mathbb{C})=\mathbb{C}[H,\mathsf{w},q]/\langle (H-\mathsf{w}_1)\star(H-\mathsf{w}_2)-q\rangle,$$

where $\deg H=\deg\mathsf{w}_i=2$ and $\deg q=4$.

The (nonequivariant) quantum cohomology of $\mathbb{P}^1$ is

$$\mathbb{C}[H,q]/\langle H\star H-q\rangle.$$

Let

$$\phi_1(q)=\frac{1}{2}+\frac{H-\frac{1}{2}(\mathsf{w}_1+\mathsf{w}_2)}{(\mathsf{w}_1-\mathsf{w}_2)\sqrt{1+4q/(\mathsf{w}_1-\mathsf{w}_2)^2}},$$
$$\phi_2(q)=\frac{1}{2}+\frac{H-\frac{1}{2}(\mathsf{w}_1+\mathsf{w}_2)}{(\mathsf{w}_2-\mathsf{w}_1)\sqrt{1+4q/(\mathsf{w}_1-\mathsf{w}_2)^2}}.$$

Then

$$\phi_\alpha(q)\star\phi_\beta(q)=\delta_{\alpha\beta}\phi_\alpha(q),$$

so $\{\phi_1(q), \phi_2(q)\}$ is a canonical basis of the semisimple algebra

$$QH_T^*(\mathbb{P}^1;\mathbb{C}) \otimes \mathbb{C}\Big[\mathsf{w}, \frac{1}{\Delta^1(q)}\Big],$$

where $\Delta^1(q)$ is defined by (2). We also have

$$\begin{aligned}(\phi_\alpha(q), \phi_\beta(q)) &= (1 \star \phi_\alpha(q), \phi_\beta(q)) = (1, \phi_\alpha(q) \star \phi_\beta(q)) \\ &= \delta_{\alpha\beta}(1, \phi_\alpha(q)) = \delta_{\alpha\beta}\int_{\mathbb{P}^1} \phi_\alpha(q) = \frac{\delta_{\alpha\beta}}{\Delta^\alpha(q)},\end{aligned}$$

where

$$\begin{aligned}\Delta^1(q) &= (\mathsf{w}_1 - \mathsf{w}_2)\sqrt{1 + \frac{4q}{(\mathsf{w}_1 - \mathsf{w}_2)^2}}, \\ \Delta^2(q) &= (\mathsf{w}_2 - \mathsf{w}_1)\sqrt{1 + \frac{4q}{(\mathsf{w}_1 - \mathsf{w}_2)^2}} = -\Delta^1(q).\end{aligned} \tag{2}$$

Quantum multiplication by the hyperplane class is given by

$$H \star \phi_\alpha = \frac{\mathsf{w}_1 + \mathsf{w}_2 + \Delta^\alpha(q)}{2}\phi_\alpha, \quad \alpha = 1, 2.$$

Finally, we take the nonequivariant limit $\mathsf{w}_2 = 0$, $\mathsf{w}_1 \to 0^+$. We obtain:

$$\begin{aligned}\phi_1(q) &= \frac{1}{2} + \frac{H}{2\sqrt{q}}, & \phi_2(q) &= \frac{1}{2} - \frac{H}{2\sqrt{q}}, \\ \Delta^1(q) &= 2\sqrt{q}, & \Delta^2(q) &= -2\sqrt{q}, \\ H \star \phi_1(q) &= \sqrt{q}\phi_1(q), & H \star \phi_2(q) &= -\sqrt{q}\phi_2(q).\end{aligned}$$

These nonequivariant limits coincide with the results in [29, Section 2].

2.4 The A–model canonical coordinates and the Ψ–matrix

Let $\{t^0, t^1\}$ be the flat coordinates with respect to the basis $\{\mathbf{1}, H\}$, and let $\{u^1, u^2\}$ be the canonical coordinates with respect to the basis $\{\phi_1(q), \phi_2(q)\}$. Then

$$\begin{aligned}\frac{\partial}{\partial u^1} &= \frac{1}{2}\Big(1 - \frac{\mathsf{w}_1 + \mathsf{w}_2}{\Delta^1(q)}\Big)\frac{\partial}{\partial t^0} + \frac{1}{\Delta^1(q)}\frac{\partial}{\partial t^1}, \\ \frac{\partial}{\partial u^2} &= \frac{1}{2}\Big(1 - \frac{\mathsf{w}_1 + \mathsf{w}_2}{\Delta^2(q)}\Big)\frac{\partial}{\partial t^0} + \frac{1}{\Delta^2(q)}\frac{\partial}{\partial t^1}, \\ du^1 &= dt^0 + \frac{1}{2}(\Delta^1(q) + \mathsf{w}_1 + \mathsf{w}_2)dt^1, \\ du^2 &= dt^0 + \frac{1}{2}(\Delta^2(q) + \mathsf{w}_1 + \mathsf{w}_2)dt^1.\end{aligned}$$

The above equations determine the canonical coordinates u^1 and u^2 up to a constant in $\mathbb{C}[\mathsf{w}_1,\mathsf{w}_2,1/(\mathsf{w}_1-\mathsf{w}_2)]$. Givental's A–model canonical coordinates (u^1,u^2) are characterized by their large radius limits

$$\lim_{q\to 0}(u^1-t^0-\mathsf{w}_1t^1)=0,\qquad \lim_{q\to 0}(u^2-t^0-\mathsf{w}_2t^1)=0. \tag{3}$$

For $\alpha\in\{1,2\}$ and $i\in\{0,1\}$, define ${\Psi_i}^\alpha$ by

$$\frac{du^\alpha}{\sqrt{\Delta^\alpha(q)}}=\sum_{i=0}^{1}dt^i{\Psi_i}^\alpha,$$

and define the Ψ–matrix to be

$$\Psi:=\begin{bmatrix}{\Psi_0}^1 & {\Psi_0}^2\\ {\Psi_1}^1 & {\Psi_1}^2\end{bmatrix}.$$

Then

$$\begin{bmatrix}\dfrac{du^1}{\sqrt{\Delta^1(q)}} & \dfrac{du^2}{\sqrt{\Delta^2(q)}}\end{bmatrix}=[\,dt^0\ \ dt^1\,]\Psi,$$

$${\Psi_0}^\alpha=\frac{1}{\sqrt{\Delta^\alpha(q)}},\qquad {\Psi_1}^\alpha=\frac{\Delta^\alpha(q)+\mathsf{w}_1+\mathsf{w}_2}{2\sqrt{\Delta^\alpha(q)}}.$$

Let

$$\Psi^{-1}=\begin{bmatrix}{(\Psi^{-1})_1}^0 & {(\Psi^{-1})_1}^1\\ {(\Psi^{-1})_2}^0 & {(\Psi^{-1})_2}^1\end{bmatrix}$$

be the inverse matrix of Ψ, so that

$$\sum_{i=0}^{1}{(\Psi^{-1})_\alpha}^i{\Psi_i}^\beta={\delta_\alpha}^\beta.$$

Then

$${(\Psi^{-1})_\alpha}^0=\frac{\Delta^\alpha(q)-\mathsf{w}_1-\mathsf{w}_2}{2\sqrt{\Delta^\alpha(q)}},\qquad {(\Psi^{-1})_\alpha}^1=\frac{1}{\sqrt{\Delta^\alpha(q)}}.$$

Let $Q=1$, ie $q=e^{t^1}$. We take the nonequivariant limit $\mathsf{w}_2=0$, $\mathsf{w}_1\to 0^+$:

$$u^1=t^0+2\sqrt{q},\qquad u^2=t^0-2\sqrt{q},$$

$$\Psi=\frac{1}{\sqrt{2}}\begin{pmatrix}e^{-\frac14t^1} & -\sqrt{-1}e^{-\frac14t^1}\\ e^{\frac14t^1} & \sqrt{-1}e^{\frac14t^1}\end{pmatrix},$$

$$\Psi^{-1}=\frac{1}{\sqrt{2}}\begin{pmatrix}e^{\frac14t^1} & e^{-\frac14t^1}\\ \sqrt{-1}e^{\frac14t^1} & -\sqrt{-1}e^{-\frac14t^1}\end{pmatrix}.$$

These nonequivariant limits agree with the results in [29, Section 2].

2.5 The $\mathcal{S}$–operator

The $\mathcal{S}$–operator is defined as follows: for any cohomology classes $a, b \in H_T^*(\mathbb{P}^1;\mathbb{C})$,

$$(a, \mathcal{S}(b)) = \left\langle\!\left\langle a, \frac{b}{z-\psi} \right\rangle\!\right\rangle_{0,2}^{\mathbb{P}^1,T}.$$

The T–equivariant J–function is characterized by

$$(J, a) = (1, \mathcal{S}(a))$$

for any $a \in H_T^*(\mathbb{P}^1)$.

Let

$$\chi^1 = \mathsf{w}_1 - \mathsf{w}_2, \quad \chi^2 = \mathsf{w}_2 - \mathsf{w}_1.$$

We consider several different (flat) bases for $H_T^*(\mathbb{P}^1;\mathbb{C})$:

- The canonical basis: $\phi_1 = (H-\mathsf{w}_2)/(\mathsf{w}_1-\mathsf{w}_2)$ and $\phi_2 = (H-\mathsf{w}_1)/(\mathsf{w}_2-\mathsf{w}_1)$.
- The basis dual to the canonical basis with respect to the T–equivariant Poincaré pairing: $\phi^1 = \chi^1\phi_1$ and $\phi^2 = \chi^2\phi_2$.
- The normalized canonical basis $\widehat{\phi}_1 = \sqrt{\chi^1}\phi_1$ and $\widehat{\phi}_2 = \sqrt{\chi^2}\phi_2$, which is self-dual: $\widehat{\phi}^1 = \widehat{\phi}_1$ and $\widehat{\phi}^2 = \widehat{\phi}_2$.
- The natural basis: $T_0 = 1$ and $T_1 = H$.
- The basis dual to the natural basis: $T^0 = H - \mathsf{w}_1 - \mathsf{w}_2$ and $T^1 = 1$.

For $\alpha, \beta \in \{1, 2\}$, define

$$S^\alpha{}_\beta(z) := (\phi^\alpha, \mathcal{S}(\phi_\beta)).$$

Then $S(z) = (S^\alpha{}_\beta(z))$ is the matrix[1] of the $\mathcal{S}$–operator with respect to the ordered basis (ϕ_1, ϕ_2):

$$\mathcal{S}(\phi_\beta) = \sum_{\alpha=1}^{2} \phi_\alpha S^\alpha{}_\beta(z). \tag{4}$$

For $i \in \{0, 1\}$ and $\alpha \in \{1, 2\}$, define

$$S_i{}^{\widehat{\alpha}}(z) := (T_i, \mathcal{S}(\widehat{\phi}^\alpha)).$$

Then $(S_i{}^{\widehat{\alpha}})$ is the matrix of the $\mathcal{S}$–operator with respect to the ordered bases $(\widehat{\phi}^1, \widehat{\phi}^2)$ and (T^0, T^1):

$$\mathcal{S}(\widehat{\phi}^\alpha) = \sum_{i=0}^{1} T^i S_i{}^{\widehat{\alpha}}(z). \tag{5}$$

[1] We use the convention that the *left* superscript/subscript is the *row* number and the *right* superscript/subscript is the *column* number.

We have

$$z\frac{\partial J}{\partial t^i}=\sum_{\alpha=1}^{2} S_i^{\,\hat{\alpha}}(z)\hat{\phi}_\alpha.$$

By [17; 25], the equivariant J–function is

$$J=e^{(t^0+t^1H)/z}\biggl(1+\sum_{d=1}^{\infty}\frac{q^d}{\prod_{m=1}^{d}(H-\mathsf{w}_1+mz)\prod_{m=1}^{d}(H-\mathsf{w}_2+mz)}\biggr).$$

For $\alpha=1,2$, define

$$J^\alpha:=J|_{p_\alpha}=e^{(t^0+t^1\mathsf{w}^\alpha)/z}\sum_{d=0}^{\infty}\frac{q^d}{d!z^d}\frac{1}{\prod_{m=1}^{d}(\chi^\alpha+mz)}.$$

Then

$$z\frac{\partial J}{\partial t^0}=J=\sum_{\alpha=1}^{2}J^\alpha\phi_\alpha,\quad z\frac{\partial J}{\partial t^1}=z\sum_{\alpha=1}^{2}\frac{\partial J^\alpha}{\partial t^1}\phi_\alpha,$$

so

$$S_i^{\,\hat{\alpha}}(z)=\frac{z}{\sqrt{\chi^\alpha}}\cdot\frac{\partial J^\alpha}{\partial t^i}.$$

Following Givental, we define

$$\widetilde{S}_i^{\,\hat{\alpha}}(z):=S_i^{\,\hat{\alpha}}(z)\exp\biggl(-\sum_{n=1}^{\infty}\frac{B_{2n}}{2n(2n-1)}\Bigl(\frac{z}{\chi^\alpha}\Bigr)^{2n-1}\biggr).$$

Then

$$\begin{aligned}\widetilde{S}_0^{\,\hat{\alpha}}(z)&=\frac{1}{\sqrt{\chi^\alpha}}\exp\biggl(\frac{t^0+t^1\mathsf{w}_\alpha}{z}-\sum_{n=1}^{\infty}\frac{B_{2n}}{2n(2n-1)}\Bigl(\frac{z}{\chi^\alpha}\Bigr)^{2n-1}\biggr)\\&\qquad\cdot\biggl(\sum_{d=0}^{\infty}\frac{q^d}{d!z^d}\frac{1}{\prod_{m=1}^{d}(\chi^\alpha+mz)}\biggr),\\ \widetilde{S}_1^{\,\hat{\alpha}}(z)&=\frac{1}{\sqrt{\chi^\alpha}}\exp\biggl(\frac{t^0+t^1\mathsf{w}_\alpha}{z}-\sum_{n=1}^{\infty}\frac{B_{2n}}{2n(2n-1)}\Bigl(\frac{z}{\chi^\alpha}\Bigr)^{2n-1}\biggr)\\&\qquad\cdot\biggl(\mathsf{w}_\alpha\sum_{d=0}^{\infty}\frac{q^d}{d!z^d}\frac{1}{\prod_{m=1}^{d}(\chi^\alpha+mz)}+\sum_{d=1}^{\infty}\frac{q^d}{(d-1)!z^d}\frac{1}{\prod_{m=1}^{d}(\chi^\alpha+mz)}\biggr).\end{aligned}$$

2.6 The A–model R–matrix

By Givental [20], the matrix $(\widetilde{S}_i{}^{\widehat{\beta}})(z)$ is of the form

$$\widetilde{S}_i{}^{\widehat{\beta}}(z) = \sum_{\alpha=1}^{2} \Psi_i{}^{\alpha} R_{\alpha}{}^{\beta}(z) e^{u^{\beta}/z} = (\Psi R(z))_i{}^{\beta} e^{u^{\beta}/z},$$

where $R(z) = (R_{\alpha}{}^{\beta}(z)) = I + \sum_{k=1}^{\infty} R_k z^k$ and is unitary, and

$$\lim_{q\to 0} R_{\alpha}{}^{\beta}(z) = \delta_{\alpha\beta} \exp\biggl(-\sum_{n=1}^{\infty} \frac{B_{2n}}{2n(2n-1)} \Bigl(\frac{z}{\chi^{\beta}}\Bigr)^{2n-1}\biggr).$$

2.7 Gromov–Witten potentials

Introducing formal variables

$$\boldsymbol{u} = \sum_{a\geq 0} u_a z^a, \quad \text{where } u_a = \sum_{\alpha=1}^{2} u_a^{\alpha} \phi_{\alpha}(q),$$

we define

$$\begin{aligned} F_{g,n}^{\mathbb{P}^1,T}(\boldsymbol{u},\boldsymbol{t}) &:= \sum_{\substack{a_1,\dots,a_n\\ a_i\in\mathbb{Z}_{\geq 0}}} \frac{1}{n!} \langle\!\langle \tau_{a_1}(u_{a_1})\cdots\tau_{a_n}(u_{a_n})\rangle\!\rangle_{g,n}^{\mathbb{P}^1,T} \\ &= \sum_{\substack{a_1,\dots,a_n\\ a_i\in\mathbb{Z}_{\geq 0}}} \sum_{m=0}^{\infty} \sum_{d=0}^{\infty} \frac{Q^d}{n!m!} \int_{[\overline{\mathcal{M}}_{g,n+m}(\mathbb{P}^1,d)]^{\mathrm{vir}}} \prod_{j=1}^{n} \mathrm{ev}_j^*(u_{a_j})\psi_j^{a_j} \prod_{i=1}^{m} \mathrm{ev}_{i+n}^*(\boldsymbol{t}). \end{aligned}$$

We define the total descendent potential of $\mathbb{P}^1$ to be

$$D^{\mathbb{P}^1,T}(\boldsymbol{u}) = \exp\biggl(\sum_{n,g} \hbar^{g-1} F_{g,n}^{\mathbb{P}^1,T}(\boldsymbol{u},0)\biggr).$$

Consider the map $\pi\colon \overline{\mathcal{M}}_{g,n+m}(\mathbb{P}^1,d) \to \overline{\mathcal{M}}_{g,n}$ which forgets the map to the target and the last m marked points. Let $\overline{\psi}_i := \pi^*(\psi_i)$ be the pull-backs of the classes ψ_i for $i = 1,\dots,n$ from $\overline{\mathcal{M}}_{g,n}$. Then we can define

$$\overline{F}_{g,n}^{\mathbb{P}^1,T}(\boldsymbol{u},\boldsymbol{t}) := \sum_{\substack{a_1,\dots,a_n\\ a_i\in\mathbb{Z}_{\geq 0}}} \sum_{m=0}^{\infty} \sum_{d=0}^{\infty} \frac{Q^d}{n!m!} \int_{[\overline{\mathcal{M}}_{g,n+m}(\mathbb{P}^1,d)]^{\mathrm{vir}}} \prod_{j=1}^{n} \mathrm{ev}_j^*(u_{a_j})\overline{\psi}_j^{a_j} \prod_{i=1}^{m} \mathrm{ev}_{i+n}^*(\boldsymbol{t}).$$

Let the ancestor potential of $\mathbb{P}^1$ be

$$A^{\mathbb{P}^1,T}(\boldsymbol{u},\boldsymbol{t}) = \exp\biggl(\sum_{n,g} \hbar^{g-1} \overline{F}_{g,n}^{\mathbb{P}^1,T}(\boldsymbol{u},\boldsymbol{t})\biggr).$$

2.8 Givental's formula for equivariant Gromov–Witten potential and the A–model graph sum

The quantization of the $\mathcal{S}$–operator relates the ancestor potential and the descendent potential of $\mathbb{P}^1$ via Givental's formula. Concretely, we have (see [19])

$$D^{\mathbb{P}^1,T}(\boldsymbol{u}) = \exp(F_1^{\mathbb{P}^1,T})\widehat{\mathcal{S}}^{-1} A^{\mathbb{P}^1,T}(\boldsymbol{u},\boldsymbol{t}),$$

where $F_1^{\mathbb{P}^1,T}$ denotes $\sum_n F_{1,n}^{\mathbb{P}^1,T}(\boldsymbol{u},0)$ at $u_0 = u$ and $u_1 = u_2 = \cdots = 0$, and $\widehat{\mathcal{S}}$ is the quantization [19] of $\mathcal{S}$. For our purpose, we need to describe a formula for a slightly different potential: $F_{g,n}^{\mathbb{P}^1,T}(\boldsymbol{u},\boldsymbol{t})$ — the descendent potential with arbitrary primary insertions.

Now we first describe a graph sum formula for the ancestor potential $A^{\mathbb{P}^1,T}(\boldsymbol{u},\boldsymbol{t})$. Given a connected graph Γ, we introduce the following notation:

- $V(\Gamma)$ is the set of vertices in Γ.
- $E(\Gamma)$ is the set of edges in Γ.
- $H(\Gamma)$ is the set of half-edges in Γ.
- $L^o(\Gamma)$ is the set of ordinary leaves in Γ.
- $L^1(\Gamma)$ is the set of dilaton leaves in Γ.

With the above notation, we introduce the following labels:

- **Genus** $g\colon V(\Gamma) \to \mathbb{Z}_{\geq 0}$.
- **Marking** $\beta\colon V(\Gamma) \to \{1,2\}$. This induces $\beta\colon L(\Gamma) = L^o(\Gamma) \cup L^1(\Gamma) \to \{1,2\}$, as follows: if $l \in L(\Gamma)$ is a leaf attached to a vertex $v \in V(\Gamma)$, define $\beta(l) = \beta(v)$.
- **Height** $k\colon H(\Gamma) \to \mathbb{Z}_{\geq 0}$.

Given an edge e, let $h_1(e)$ and $h_2(e)$ be the two half-edges associated to e. The order of the two half-edges does not affect the graph sum formula in this paper. Given a vertex $v \in V(\Gamma)$, let $H(v)$ denote the set of half-edges emanating from v. The valency of the vertex v is equal to the cardinality of the set $H(v)$, written $\mathrm{val}(v) = |H(v)|$. A labeled graph $\vec{\Gamma} = (\Gamma, g, \beta, k)$ is *stable* if

$$2g(v) - 2 + \mathrm{val}(v) > 0$$

for all $v \in V(\Gamma)$.

Let $\mathbf{\Gamma}(\mathbb{P}^1)$ denote the set of all stable labeled graphs $\vec{\Gamma} = (\Gamma, g, \beta, k)$. The genus of a stable labeled graph $\vec{\Gamma}$ is defined to be

$$g(\vec{\Gamma}) := \sum_{v \in V(\Gamma)} g(v) + |E(\Gamma)| - |V(\Gamma)| + 1 = \sum_{v \in V(\Gamma)} (g(v) - 1) + \Big(\sum_{e \in E(\Gamma)} 1\Big) + 1.$$

Define

$$\mathbf{\Gamma}_{g,n}(\mathbb{P}^1) = \{\vec{\Gamma} = (\Gamma, g, \beta, k) \in \mathbf{\Gamma}(\mathbb{P}^1) : g(\vec{\Gamma}) = g,\ |L^o(\Gamma)| = n\}.$$

Given $\alpha \in \{1,2\}$, define

$$\boldsymbol{u}^\alpha(z) = \sum_{a \geq 0} u_a^\alpha z^a.$$

We assign weights to leaves, edges, and vertices of a labeled graph $\vec{\Gamma} \in \mathbf{\Gamma}(\mathbb{P}^1)$ as follows:

(1) **Ordinary leaves** To each ordinary leaf $l \in L^o(\Gamma)$ with $\beta(l) = \beta \in \{1,2\}$ and $k(l) = k \in \mathbb{Z}_{\geq 0}$, we assign

$$(\mathcal{L}^{\boldsymbol{u}})_k^\beta(l) = [z^k]\Bigg(\sum_{\alpha=1,2} \frac{\boldsymbol{u}^\alpha(z)}{\sqrt{\Delta^\alpha(q)}} R_\alpha{}^\beta(-z)\Bigg).$$

(2) **Dilaton leaves** To each dilaton leaf $l \in L^1(\Gamma)$ with $\beta(l) = \beta \in \{1,2\}$ and $2 \leq k(l) = k \in \mathbb{Z}_{\geq 0}$, we assign

$$(\mathcal{L}^1)_k^\beta(l) = [z^{k-1}]\Bigg(-\sum_{\alpha=1,2} \frac{1}{\sqrt{\Delta^\alpha(q)}} R_\alpha{}^\beta(-z)\Bigg).$$

(3) **Edges** To an edge connecting a vertex marked by $\alpha \in \{1,2\}$ to a vertex marked by $\beta \in \{1,2\}$ and with heights k and l at the corresponding half-edges, we assign

$$\mathcal{E}_{k,l}^{\alpha,\beta}(e) = [z^k w^l]\Bigg(\frac{1}{z+w}\Bigg(\delta_{\alpha,\beta} - \sum_{\gamma=1,2} R_\gamma{}^\alpha(-z) R_\gamma{}^\beta(-w)\Bigg)\Bigg).$$

(4) **Vertices** To a vertex v with genus $g(v) = g \in \mathbb{Z}_{\geq 0}$ and marking $\beta(v) = \beta$, with n ordinary leaves and half-edges attached to it with heights $k_1, \dots, k_n \in \mathbb{Z}_{\geq 0}$ and m more dilaton leaves with heights $k_{n+1}, \dots, k_{n+m} \in \mathbb{Z}_{\geq 0}$, we assign

$$(\sqrt{\Delta^\beta(q)})^{2g-2+n+m} \int_{\overline{\mathcal{M}}_{g,n+m}} \psi_1^{k_1} \cdots \psi_{n+m}^{k_{n+m}}.$$

We define the weight of a labeled graph $\vec{\Gamma} \in \mathbf{\Gamma}(\mathbb{P}^1)$ to be

$$w(\vec{\Gamma}) = \prod_{v \in V(\Gamma)} (\sqrt{\Delta^{\beta(v)}(q)})^{2g(v)-2+\mathrm{val}(v)} \Big\langle \prod_{h \in H(v)} \tau_{k(h)} \Big\rangle_{g(v)} \prod_{e \in E(\Gamma)} \mathcal{E}_{k(h_1(e)),k(h_2(e))}^{\beta(v_1(e)),\beta(v_2(e))}(e)$$
$$\cdot \prod_{l \in L^o(\Gamma)} (\mathcal{L}^{\boldsymbol{u}})_{k(l)}^{\beta(l)}(l) \prod_{l \in L^1(\Gamma)} (\mathcal{L}^1)_{k(l)}^{\beta(l)}(l).$$

Then

$$\log(A^{\mathbb{P}^1,T}(\boldsymbol{u},\boldsymbol{t})) = \sum_{\vec{\Gamma}\in\boldsymbol{\Gamma}(\mathbb{P}^1)} \frac{\hbar^{g(\vec{\Gamma})-1} w(\vec{\Gamma})}{|\mathrm{Aut}(\vec{\Gamma})|} = \sum_{g\geq 0}\hbar^{g-1}\sum_{n\geq 0}\sum_{\vec{\Gamma}\in\boldsymbol{\Gamma}_{g,n}(\mathbb{P}^1)} \frac{w(\vec{\Gamma})}{|\mathrm{Aut}(\vec{\Gamma})|}.$$

Now we describe a graph sum formula for $F^{\mathbb{P}^1,T}_{g,n}(\boldsymbol{u},\boldsymbol{t})$ — the descendant potential with arbitrary primary insertions. For $\alpha = 1, 2$, let

$$\widehat{\phi}_\alpha(q) := \sqrt{\Delta^\alpha(q)}\phi_\alpha(q).$$

Then $\widehat{\phi}_1(q)$, $\widehat{\phi}_2(q)$ is the normalized canonical basis of $QH^*_T(\mathbb{P}^1;\mathbb{C})$, the T–equivariant quantum cohomology of $\mathbb{P}^1$. Define

$$S^{\widehat{\underline{\alpha}}}_{\ \widehat{\underline{\beta}}}(z) := \big(\widehat{\phi}_\alpha(q), \mathcal{S}(\widehat{\phi}_\beta(q))\big).$$

Then this is the matrix of the $\mathcal{S}$–operator with respect to the ordered basis $(\widehat{\phi}_1(q), \widehat{\phi}_2(q))$:

$$\mathcal{S}(\widehat{\phi}_\beta(q)) = \sum_{\alpha=1}^{2} \widehat{\phi}_\alpha(q) S^{\widehat{\underline{\alpha}}}_{\ \widehat{\underline{\beta}}}(z). \tag{6}$$

We define a new weight of the ordinary leaves:

(1′) **Ordinary leaves** To each ordinary leaf $l \in L^o(\Gamma)$ with $\beta(l) = \beta \in \{1,2\}$ and $k(l) = k \in \mathbb{Z}_{\geq 0}$, we assign

$$(\mathring{\mathcal{L}}^{\boldsymbol{u}})^\beta_k(l) = [z^k]\bigg(\sum_{\alpha,\gamma=1,2} \frac{\boldsymbol{u}^\alpha(z)}{\sqrt{\Delta^\alpha(q)}} S^{\widehat{\underline{\gamma}}}_{\ \widehat{\underline{\alpha}}}(z) R(-z)_\gamma^{\ \beta}\bigg).$$

We define a new weight of a labeled graph $\vec{\Gamma} \in \boldsymbol{\Gamma}(\mathbb{P}^1)$ to be

$$\mathring{w}(\vec{\Gamma}) = \prod_{v\in V(\Gamma)} \big(\sqrt{\Delta^{\beta(v)}(q)}\big)^{2g(v)-2+\mathrm{val}(v)} \Big\langle \prod_{h\in H(v)} \tau_{k(h)} \Big\rangle_{g(v)} \prod_{e\in E(\Gamma)} \mathcal{E}^{\beta(v_1(e)),\beta(v_2(e))}_{k(h_1(e)),k(h_2(e))}(e) \cdot \prod_{l\in L^o(\Gamma)} (\mathring{\mathcal{L}}^{\boldsymbol{u}})^{\beta(l)}_{k(l)}(l) \prod_{l\in L^1(\Gamma)} (\mathcal{L}^1)^{\beta(l)}_{k(l)}(l).$$

Then

$$\sum_{g\geq 0}\hbar^{g-1}\sum_{n\geq 0} F^{\mathbb{P}^1,T}_{g,n}(\boldsymbol{u},\boldsymbol{t}) = \sum_{\vec{\Gamma}\in\boldsymbol{\Gamma}(\mathbb{P}^1)} \frac{\hbar^{g(\vec{\Gamma})-1}\mathring{w}(\vec{\Gamma})}{|\mathrm{Aut}(\vec{\Gamma})|} = \sum_{g\geq 0}\hbar^{g-1}\sum_{n\geq 0}\sum_{\vec{\Gamma}\in\boldsymbol{\Gamma}_{g,n}(\mathbb{P}^1)} \frac{\mathring{w}(\vec{\Gamma})}{|\mathrm{Aut}(\vec{\Gamma})|}.$$

We can slightly generalize this graph sum formula to the case where we have n *ordered* variables $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$ and n *ordered* ordinary leaves. Let

$$\boldsymbol{u}_j = \sum_{a\geq 0} (u_j)_a z^a$$

and let

$$F_{g,n}^{\mathbb{P}^1,T}(\boldsymbol{u}_1,\dots,\boldsymbol{u}_n,\boldsymbol{t}) := \sum_{\substack{a_1,\dots,a_n\\ a_i\in\mathbb{Z}_{\geq 0}}}\sum_{m=0}^{\infty}\sum_{d=0}^{\infty}\frac{1}{m!}\int_{[\overline{\mathcal{M}}_{g,n+m}(\mathbb{P}^1,d)]^{\mathrm{vir}}}\prod_{j=1}^{n}\mathrm{ev}_j^*((u_j)_{a_j})\psi_j^{a_j} \cdot\prod_{i=1}^{m}\mathrm{ev}_{i+n}^*(\boldsymbol{t}).$$

Define the set of graphs $\widetilde{\mathbf{\Gamma}}_{g,n}(\mathbb{P}^1)$ as the definition of $\mathbf{\Gamma}_{g,n}(\mathbb{P}^1)$ except that the n ordinary leaves are *ordered*. Let $\{l_1,\dots,l_n\}$ be the ordinary leaves in $\Gamma\in\widetilde{\mathbf{\Gamma}}_{g,n}(\mathbb{P}^1)$ and for $j=1,\dots,n$ let

$$(\mathring{\mathcal{L}}^{\boldsymbol{u}_j})_k^{\beta}(l_j) = [z^k]\Bigg(\sum_{\alpha,\gamma=1,2}\frac{\boldsymbol{u}_j^{\alpha}(z)}{\sqrt{\Delta^{\alpha}(q)}}S^{\hat{\gamma}}_{\;\underline{\hat{\alpha}}}(z)R(-z)_{\gamma}^{\;\beta}\Bigg).$$

Define the weight

$$\mathring{w}(\vec{\Gamma}) = \prod_{v\in V(\Gamma)}(\sqrt{\Delta^{\beta(v)}(q)})^{2g(v)-2+\mathrm{val}(v)}\Big\langle\prod_{h\in H(v)}\tau_{k(h)}\Big\rangle_{g(v)}\prod_{e\in E(\Gamma)}\mathcal{E}^{\beta(v_1(e)),\beta(v_2(e))}_{k(h_1(e)),k(h_2(e))}(e) \cdot\prod_{j=1}^{n}(\mathring{\mathcal{L}}^{\boldsymbol{u}_j})^{\beta(l_j)}_{k(l_j)}(l_j)\prod_{l\in L^1(\Gamma)}(\mathcal{L}^1)^{\beta(l)}_{k(l)}(l).$$

Then

$$\sum_{g\geq 0}\hbar^{g-1}\sum_{n\geq 0}F_{g,n}^{\mathbb{P}^1,T}(\boldsymbol{u}_1,\cdots,\boldsymbol{u}_n,\boldsymbol{t}) = \sum_{\vec{\Gamma}\in\widetilde{\mathbf{\Gamma}}(\mathbb{P}^1)}\frac{\hbar^{g(\vec{\Gamma})-1}\mathring{w}(\vec{\Gamma})}{|\mathrm{Aut}(\vec{\Gamma})|} = \sum_{g\geq 0}\hbar^{g-1}\sum_{n\geq 0}\sum_{\vec{\Gamma}\in\widetilde{\mathbf{\Gamma}}_{g,n}(\mathbb{P}^1)}\frac{\mathring{w}(\vec{\Gamma})}{|\mathrm{Aut}(\vec{\Gamma})|}.$$

3 B–model

3.1 The equivariant superpotential and the Frobenius structure of the Jacobian ring

Let Y be coordinates on $\mathbb{C}^*$. The T–equivariant superpotential $W_t^{\mathsf{w}}\colon \mathbb{C}^*\to\mathbb{C}$ is given by

$$W_t^{\mathsf{w}}(Y) = Y + t_0 + \frac{q}{Y} + \mathsf{w}_1\log Y + \mathsf{w}_2\log\frac{q}{Y},$$

where $q = Qe^{t_1}$ and $Y = e^y$. In this section, we assume $\mathsf{w}_1-\mathsf{w}_2$ is a positive real number. The Jacobian ring of W_t^{w} is

$$\mathrm{Jac}(W_t^{\mathsf{w}}) \cong C[Y,Y^{-1},q,\mathsf{w}]\Big/\Big\langle\frac{\partial W_t^{\mathsf{w}}}{\partial y}\Big\rangle = \mathbb{C}[Y,Y^{-1},q,\mathsf{w}]\Big/\Big\langle Y-\frac{q}{Y}+\mathsf{w}_1-\mathsf{w}_2\Big\rangle.$$

Let
$$B := q\frac{\partial W_t^{\mathsf{w}}}{\partial q} = \frac{q}{Y} + \mathsf{w}_2.$$
The Jacobian ring is isomorphic to $QH_T^*(\mathbb{P}^1;\mathbb{C})$ if one identifies B with H:
$$\mathrm{Jac}(W_t^{\mathsf{w}}) \cong \mathbb{C}[B,q,\mathsf{w}]/\langle (B-\mathsf{w}_1)(B-\mathsf{w}_2)-q\rangle.$$
The critical points of W_t^{w} are P_1 and P_2, where
$$P_\alpha = \frac{\mathsf{w}_2-\mathsf{w}_1+\Delta^\alpha(q)}{2}, \quad \alpha = 1,2.$$
Endow a metric on $\mathrm{Jac}(W_q^{\mathsf{w}})$ by the residue pairing
$$(f,g) = \sum_{\alpha=1}^{2} \mathrm{Res}_{Y=P_\alpha} \frac{f(Y)g(Y)}{\partial W_t^{\mathsf{w}}/\partial y}\,\frac{dY}{Y}.$$
By direct calculation, we have
$$(B,B) = \mathsf{w}_1+\mathsf{w}_2, \quad (B,\mathbf{1}) = (\mathbf{1},B) = 1, \quad (\mathbf{1},\mathbf{1}) = 0.$$
We let $b_0 = \mathbf{1}$, $b_1 = B$ and define b^i by $(b^i,b_j) = \delta^i_j$. These calculations show the following well-known fact:

Proposition 3.1 *There is an isomorphism of Frobenius manifolds*
$$QH_T^*(\mathbb{P}^1;\mathbb{C}) \otimes_{\mathbb{C}[\mathsf{w}]} \mathbb{C}\Big[\mathsf{w}, \frac{1}{\mathsf{w}_1-\mathsf{w}_2}\Big] \cong \mathrm{Jac}(W_t^{\mathsf{w}}) \otimes_{\mathbb{C}[\mathsf{w}]} \mathbb{C}\Big[\mathsf{w}, \frac{1}{\mathsf{w}_1-\mathsf{w}_2}\Big].$$

We denote $\mathrm{Jac}(W_t^{\mathsf{w}}) \otimes_{\mathbb{C}[\mathsf{w}]} \mathbb{C}[\mathsf{w}, 1/(\mathsf{w}_1-\mathsf{w}_2)]$ by H_B. The Dubrovin–Givental connection is denoted by $\nabla^B_v = z\partial_v + v\bullet$ on $\mathcal{H}_B := H_B((z))$.

3.2 The B–model canonical coordinates

The isomorphism of Frobenius structures automatically ensures their canonical coordinates are the same up to a permutation and constants. We fix the B–model canonical coordinates in this subsection by the critical values of the superpotential W_t^{w}, and find the constant difference to the A–model coordinates that we set up in earlier sections.

Let $C_t^{\mathsf{w}} = \{(x,y) \in \mathbb{C}^2 : x = W_t^{\mathsf{w}}(e^y)\}$ be the graph of the equivariant superpotential. It is a covering of $\mathbb{C}^*$, given by $y \mapsto e^y$. Let $\overline{\Sigma} \cong \mathbb{P}^1$ be the compactification of $\mathbb{C}^*$ with $Y \in \mathbb{C}^* \subset \mathbb{P}^1$ as its coordinate. At each branch point $Y = P_\alpha$, we have the expansions
$$x = \check{u}^\alpha - \zeta_\alpha^2, \quad y = \check{v}^\alpha - \sum_{k=1}^{\infty} h_k^\alpha(q)\zeta_\alpha^k,$$

where $h_1^\alpha(q) = \sqrt{2/\Delta^\alpha(q)}$. Note that we define ζ_α by $\zeta_\alpha^2 = \check{u}^\alpha - x$, which differs from the definition of ζ in [7; 11] by a factor of $\sqrt{-1}$.

The critical values are

$$\check{u}^\alpha = t^0 + \mathsf{w}_\alpha t^1 + \Delta^\alpha(q) - \chi^\alpha \log \frac{\chi^\alpha + \Delta^\alpha(q)}{2}.$$

Since

$$\frac{\partial \check{u}^\alpha}{\partial t^0} = 1, \quad \frac{\partial \check{u}^\alpha}{\partial t^1} = \frac{q}{P_\alpha} + \mathsf{w}_2 = \frac{\mathsf{w}_1 + \mathsf{w}_2 + \Delta^\alpha(q)}{2},$$

we have

$$d\check{u}^\alpha = du^\alpha, \quad \alpha = 1, 2. \tag{7}$$

Recall that $\lim_{q\to 0} \Delta^1(q) = \mathsf{w}_1 - \mathsf{w}_2$, so in the large radius limit $q \to 0$ we have

$$\lim_{q\to 0} (\check{u}^\alpha - t^0 - \mathsf{w}_\alpha t^1) = \chi^\alpha - \chi^\alpha \log \chi^\alpha. \tag{8}$$

From (7), (8) and (3), we conclude that

$$\check{u}^\alpha = u^\alpha + a_\alpha, \quad \alpha = 1, 2,$$

where

$$a_\alpha = \chi^\alpha - \chi^\alpha \log \chi^\alpha.$$

3.3 The Liouville form and Bergman kernel

On C_t^{w}, let

$$\lambda = x\, dy$$

be the Liouville form on $\mathbb{C}^2 = T^*\mathbb{C}$. Then $d\lambda = dx \wedge dy$. Let

$$\Phi := \lambda|_{C_t^{\mathsf{w}}} = W_t^{\mathsf{w}}(e^y)\, dy = (e^y + t_0 + qe^{-y} + (\mathsf{w}_1 - \mathsf{w}_2)y + \mathsf{w}_2 \log q)\, dy.$$

Then Φ is a holomorphic 1–form on $\mathbb{C}$. Recall that $q = Qe^{t^1}$ and $Y = e^y$. Define

$$\Phi_0 := \frac{\partial \Phi}{\partial t^0} = \frac{dY}{Y},$$
$$\Phi_1 := \frac{\partial \Phi}{\partial t^1} = \Big(\frac{q}{Y} + \mathsf{w}_2\Big)\frac{dY}{Y}.$$

Then Φ_0 and Φ_1 descend to holomorphic 1–forms on $\mathbb{C}^*$ which extends to meromorphic 1–forms on $\mathbb{P}^1$. We have:

- Φ_0 has simple poles at $Y = 0$ and $Y = \infty$, and

$$\mathrm{Res}_{Y\to 0} \Phi_0 = 1, \quad \mathrm{Res}_{Y\to\infty} \Phi_1 = -1.$$

- $\Phi_1 - \mathsf{w}_2 \Phi_0 = -q d(Y^{-1})$ is an exact 1–form.

Let $B(p_1, p_2)$ be the fundamental normalized differential of the second kind on $\overline{\Sigma}$ (see eg [16]). It is also called the Bergman kernel in [10; 11]. In this simple case with $\overline{\Sigma} \cong \mathbb{P}^1$, we have

$$B(Y_1, Y_2) = \frac{dY_1 \otimes dY_2}{(Y_1 - Y_2)^2}.$$

3.4 Differentials of the second kind

Following [7; 11], given $\alpha = 1, 2$ and $d \in \mathbb{Z}_{\geq 0}$, define

$$d\xi_{\alpha,d}(p) := (2d-1)!!\, 2^{-d} \operatorname{Res}_{p' \to P_\alpha} B(p, p')(\sqrt{-1}\zeta_\alpha)^{-2d-1}.$$

Then $d\xi_{\alpha,d}$ satisfies the following properties:

- $d\xi_{\alpha,d}$ is a meromorphic 1–form on $\mathbb{P}^1$ with a single pole of order $2d+2$ at P_α.
- In the local coordinate ζ_α near P_α,
$$d\xi_{\alpha,d} = \left(\frac{-(2d+1)!!}{2^d \sqrt{-1}^{2d+1} \zeta_\alpha^{2d+2}} + f(\zeta_\alpha) \right) d\zeta_\alpha,$$
where $f(\zeta_\alpha)$ is analytic around P_α. The residue of $d\xi_{\alpha,d}$ at P_α is zero, so $d\xi_{\alpha,d}$ is a differential of the second kind.

The meromorphic 1–form $d\xi_{\alpha,d}$ is characterized by the above properties; $d\xi_{\alpha,d}$ can be viewed as a section in $H^0\big(\mathbb{P}^1, \omega_{\mathbb{P}^1}((2d+2)P_\alpha)\big)$. In particular, $d\xi_{\alpha,0}$ is

$$d\xi_{\alpha,0} = \sqrt{\frac{-2}{\Delta^\alpha(q)}}\, d\left(\frac{P_\alpha}{Y - P_\alpha} \right).$$

Then we have

$$\begin{aligned}
d\left(\frac{\Phi_0}{dW}\right) &= d\left(\frac{Y}{(Y-P_1)(Y-P_2)}\right) = \frac{1}{P_1-P_2} d\left(\frac{P_1}{Y-P_1} - \frac{P_2}{Y-P_2}\right) \\
&= \frac{1}{\sqrt{-1}} \frac{1}{\sqrt{2\Delta^1(q)}} d\xi_{1,0} + \frac{1}{\sqrt{-1}} \frac{1}{\sqrt{2\Delta^2(q)}} d\xi_{2,0} \\
&= \frac{1}{\sqrt{-2}} \sum_{\alpha=1}^{2} {\Psi_0}^\alpha d\xi_{\alpha,0}, \\
d\left(\frac{\Phi_1}{dW}\right) &= d\left(\frac{q + \mathsf{w}_2 Y}{(Y-P_1)(Y-P_2)}\right) \\
&= \frac{1}{P_1-P_2} d\left(\frac{q+P_1\mathsf{w}_2}{Y-P_1} - \frac{q+P_2\mathsf{w}_2}{Y-P_2}\right) \\
&= \frac{1}{\sqrt{-1}} \frac{1}{\Delta^1(q)} \left(\sqrt{\frac{\Delta^1(q)}{2}} \left(\frac{q}{P_1} + \mathsf{w}_2\right) d\xi_{1,0} - \sqrt{\frac{\Delta^2(q)}{2}} \left(\frac{q}{P_2} + \mathsf{w}_2\right) d\xi_{2,0} \right)
\end{aligned}$$

$$= \frac{1}{2\sqrt{-2}}\left(\left(\sqrt{\Delta^1(q)}+\frac{\mathsf{w}_1+\mathsf{w}_2}{\sqrt{\Delta^1(q)}}\right)d\xi_{1,0}+\left(\sqrt{\Delta^2(q)}+\frac{\mathsf{w}_1+\mathsf{w}_2}{\sqrt{\Delta^2(q)}}\right)d\xi_{2,0}\right)$$

$$= \frac{1}{\sqrt{-2}}\sum_{\alpha=1}^{2}\Psi_1{}^{\alpha}d\xi_{\alpha,0},$$

so

$$\begin{pmatrix} d\left(\frac{\Phi_0}{dW}\right) \\ d\left(\frac{\Phi_1}{dW}\right)\end{pmatrix} = \frac{1}{\sqrt{-2}}\Psi\begin{pmatrix} d\xi_{1,0} \\ d\xi_{2,0}\end{pmatrix}, \quad \sqrt{-2}\Psi^{-1}\begin{pmatrix} d\left(\frac{\Phi_0}{dW}\right) \\ d\left(\frac{\Phi_1}{dW}\right)\end{pmatrix} = \begin{pmatrix} d\xi_{1,0} \\ d\xi_{2,0}\end{pmatrix}. \tag{9}$$

3.5 Oscillating integrals and the B–model R–matrix

For $\alpha, \beta \in \{1,2\}$, $i \in \{0,1\}$ and $z > 0$, define

$$\check{S}_i{}^{\boldsymbol{\alpha}}(z) := \int_{y\in\gamma_\alpha} e^{W_q^{\mathsf{w}}(Y)/z}\Phi_i = -z\int_{y\in\gamma_\alpha} e^{W_q^{\mathsf{w}}(Y)/z}\, d\left(\frac{\Phi_i}{dW}\right),$$

where γ_α is the Lefschetz thimble going through P_α such that $W_q^{\mathsf{w}}(Y) \to -\infty$ near its ends. It is straightforward to check that $\sum_{i=0}^{1} b^i \check{S}_i{}^{\boldsymbol{\alpha}}$ is a solution to the quantum differential equation $\nabla^B f = 0$ for $\alpha = 1, 2$. We quote the following theorem:

Theorem 3.2 [3; 18; 20] *Near a semisimple point on a Frobenius manifold of dimension n, there is a fundamental solution S to the quantum differential equation satisfying the following properties:*

(1) *S has the form*

$$S = \Psi R(z) e^{U/z},$$

where $R(z)$ is a matrix of formal power series in z and $U = \operatorname{diag}(u^1, \dots, u^n)$ is a matrix formed by canonical coordinates.

(2) *If S is unitary under the pairing of the Frobenius structure, then $R(z)$ is unique up to a right multiplication of $e^{\sum_{i=1}^{\infty} A_{2i-1} z^{2i-1}}$, where the A_k are constant diagonal matrices.*

Remark 3.3 For equivariant Gromov–Witten theory of $\mathbb{P}^1$, the fundamental solution S in Theorem 3.2 is viewed as a matrix with entries in $\mathbb{C}[\mathsf{w}, 1/(\mathsf{w}_1-\mathsf{w}_2)]((z))[\![q, t^0, t^1]\!]$. We choose the canonical coordinates $\{u^\alpha(t)\}$ such that there is no constant term by (3). Then, if we fix the powers of q, t^0 and t^1, only finitely many terms in the expansion of $e^{U/z}$ contribute. So the multiplication $\Psi R(z) e^{U/z}$ is well-defined and the result matrix indeed has entries in $\mathbb{C}[\mathsf{w}, 1/(\mathsf{w}_1-\mathsf{w}_2)]((z))[\![q, t^0, t^1]\!]$.

Remark 3.4 For a general abstract semisimple Frobenius manifold defined over a ring A, the expression $S = \Psi R(z)e^{U/z}$ in Theorem 3.2 can be understood in the following way. We consider the free module $M = \langle e^{u^1/z}\rangle \oplus \cdots \oplus \langle e^{u^n/z}\rangle$ over the ring $A((z))[\![t^1, \cdots, t^n]\!]$, where $t^1, \dots, t^n$ are the flat coordinates of the Frobenius manifold. We formally define the differential $de^{u^i/z} = e^{u^i/z}\, du^i/z$ and we extend the differential to M by the product rule. Then we have a map $d\colon M \to M\,dt^1 \oplus \cdots \oplus M\,dt^n$. We consider the fundamental solution $S = \Psi R(z)e^{U/z}$ as a matrix with entries in M. The meaning that S satisfies the quantum differential equation is understood by the above formal differential.

In our case, the multiplication in the A–model fundamental solution $S = \Psi R(z)e^{U/z}$ is formal in z, as in Remark 3.3. On the B–model side, we use the stationary phase expansion to obtain a product of the form $\Psi R(z)e^{U/z}$. The multiplications $\Psi R(z)e^{U/z}$ on both the A–model and B–model can be viewed as matrices with entries in M, and their differentials are obviously the same with the formal differential above.

We repeat the argument in Givental [19] and state it as the following fact:

Proposition 3.5 *The fundamental solution matrix $\{\check{S}_i^{\boldsymbol{\alpha}}/\sqrt{-2\pi z}\}$ has the asymptotic expansion, where $\check{R}(z)$ is a formal power series in z,*

$$\frac{\check{S}_i^{\boldsymbol{\alpha}}(z)}{\sqrt{-2\pi z}} \sim \sum_{\gamma=1}^{2} \Psi_i^{\ \gamma} \check{R}_\gamma^{\ \alpha}(z) e^{\check{u}^\alpha/z}.$$

Proof By the stationary phase expansion,

$$\check{S}_i^{\ \alpha}(z) \sim \sqrt{2\pi z} e^{\check{u}^\alpha/z}(1 + a_{i,1}^\alpha z + a_{i,2}^\alpha z^2 + \cdots),$$

it follows that $\{\check{S}_i^{\ \alpha}\}$ can be asymptotically expanded in the desired form (notice that Ψ is a matrix in z–degree 0). In particular, by (9),

$$\check{R}_\beta^{\ \alpha}(z) \sim \frac{\sqrt{z} e^{-\check{u}^\alpha/z}}{2\sqrt{\pi}} \int_{\gamma_\alpha} e^{W_t^{\mathrm{w}}/z}\, d\xi_{\beta,0}.$$

The above B–model R–matrix $\check{R}_\beta^{\ \alpha}(z)$ is related to $f_\beta^\alpha(u)$ in Eynard [8] by

$$f_\beta^\alpha(u) = \check{R}_\beta^{\ \alpha}\Big(-\frac{1}{u}\Big). \tag{10}$$

Following Eynard [8], define the Laplace transform of the Bergman kernel

$$\check{B}^{\alpha,\beta}(u,v,q) := \frac{uv}{u+v}\delta_{\alpha,\beta} + \frac{\sqrt{uv}}{2\pi} e^{u\check{u}^\alpha + v\check{u}^\beta} \int_{p_1\in\gamma_\alpha} \int_{p_2\in\gamma_\beta} B(p_1,p_2) e^{-ux(p_1)-vx(p_2)},$$

where $\alpha, \beta \in \{1,2\}$. By [8, Equation (B.9)] and (10),

$$\check{B}^{\alpha,\beta}(u,v,q) = \frac{uv}{u+v}\Big(\delta_{\alpha,\beta} - \sum_{\gamma=1}^{2} \check{R}_\gamma^{\ \alpha}\Big(-\frac{1}{u}\Big)\check{R}_\gamma^{\ \beta}\Big(-\frac{1}{v}\Big)\Big). \tag{11}$$

Setting $u=-v$, we conclude that

$$\Big(\check{R}^*\Big(\frac{1}{u}\Big)\check{R}\Big(-\frac{1}{u}\Big)\Big)^{\alpha\beta} = \Big\{\sum_{\gamma=1}^{2} \check{R}_\gamma^{\ \alpha}\Big(\frac{1}{u}\Big)\check{R}_\gamma^{\ \beta}\Big(-\frac{1}{u}\Big)\Big\} = \delta^{\alpha\beta}.$$

This shows $\check{R}$ is unitary. □

Following Iritani [22] (with slight modification), we introduce the following definition:

Definition 3.6 (equivariant K–theoretic framing) We define

$$\widetilde{\mathrm{ch}}_z\colon K_T(\mathbb{P}^1) \to H_T^*(\mathbb{P}^1;\mathbb{Q})\Big[\Big[\frac{\mathsf{w}_1-\mathsf{w}_2}{z}\Big]\Big]$$

by the following two properties, which uniquely characterize it:

(a) $\widetilde{\mathrm{ch}}_z$ is a homomorphism of additive groups:

$$\widetilde{\mathrm{ch}}_z(\mathcal{E}_1 \oplus \mathcal{E}_2) = \widetilde{\mathrm{ch}}_z(\mathcal{E}_1) + \widetilde{\mathrm{ch}}_z(\mathcal{E}_2).$$

(b) If $\mathcal{L}$ is a T–equivariant line bundle on $\mathbb{P}^1$ then

$$\widetilde{\mathrm{ch}}_z(\mathcal{L}) = \exp\Big(-\frac{2\pi\sqrt{-1}(c_1)_T(\mathcal{L})}{z}\Big).$$

For any $\mathcal{E} \in K_T(\mathbb{P}^1)$, we define the *K–theoretic framing* of $\mathcal{E}$ by

$$\kappa(\mathcal{E}) := (-z)^{1-(c_1)_T(T\mathbb{P}^1)/z}\,\Gamma\Big(1 - \frac{(c_1)_T(T\mathbb{P}^1)}{z}\Big)\widetilde{\mathrm{ch}}_z(\mathcal{E}),$$

where $(c_1)_T(T\mathbb{P}^1) = 2H - \mathsf{w}_1 - \mathsf{w}_2$.

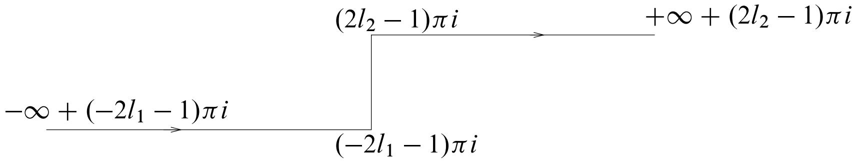
By localization, property (b) in the above definition is characterized by

$$\iota_{p_\alpha}^*\kappa(\mathcal{O}_{\mathbb{P}^1}(l_1p_1 + l_2p_2)) = (-z)^{1-\chi^\alpha/z}\,\Gamma\Big(1-\frac{\chi^\alpha}{z}\Big)e^{-2l_\alpha\pi\sqrt{-1}\chi^\alpha/z}, \quad \alpha = 1,2,$$

where $\iota_{p_\alpha}\colon p_\alpha \to \mathbb{P}^1$ is the inclusion map.

The following definition is motivated by [12; 14]:

Definition 3.7 (equivariant SYZ T–dual) Let $\mathcal{L} = \mathcal{O}_{\mathbb{P}^1}(l_1 p_1 + l_2 p_2)$ be an equivariant ample line bundle on $\mathbb{P}^1$, where l_1 and l_2 are integers such that $l_1 + l_2 > 0$. We define the equivariant SYZ T–dual $\mathrm{SYZ}(\mathcal{L})$ of $\mathcal{L}$ to be the oriented graph



in $\mathbb{C}$. We extend the definition additively to the equivariant K–theory group $K_T(\mathbb{P}^1)$.

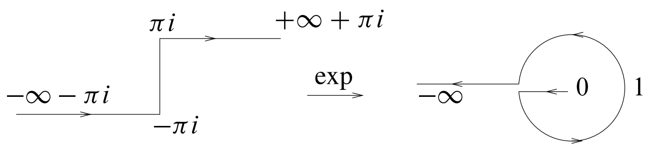


Figure 1: The equivariant SYZ T–dual of $\mathcal{O}_{\mathbb{P}^1}(p_2)$ in $\mathbb{C}$ and the (nonequivariant) SYZ T–dual of $\mathcal{O}_{\mathbb{P}^1}(1)$ in $\mathbb{C}^*$

The following theorem gives a precise correspondence between the B–model oscillatory integrals and the A–model 1–point descendant invariants.

Theorem 3.8 *Suppose that $z, q, \mathsf{w}_1 - \mathsf{w}_2 \in (0, \infty)$. Then, for any $\mathcal{L} \in K_T(\mathbb{P}^1)$,*

$$\int_{y \in \mathrm{SYZ}(\mathcal{L})} e^{W_t^{\mathsf{w}}/z}\, dy = \left\langle\!\left\langle 1, \frac{\kappa(\mathcal{L})}{z - \psi} \right\rangle\!\right\rangle_{0,2}^{\mathbb{P}^1, T}, \tag{12}$$

$$\int_{y \in \mathrm{SYZ}(\mathcal{L})} e^{W_t^{\mathsf{w}}/z}\, y\, dx = -\left\langle\!\left\langle \frac{\kappa(\mathcal{L})}{z - \psi} \right\rangle\!\right\rangle_{0,1}^{\mathbb{P}^1, T}. \tag{13}$$

Here $dx = d(W_t^{\mathsf{w}}(y))$.

Proof The left-hand side of (12) is

$$\int_{y \in \mathrm{SYZ}(\mathcal{L})} e^{W_t^{\mathsf{w}}/z}\, dy = -\frac{1}{z} \int_{y \in \mathrm{SYZ}(\mathcal{L})} e^{W_t^{\mathsf{w}}/z}\, y\, d(W_t^{\mathsf{w}}).$$

By the string equation, the right-hand side of (12) is

$$\left\langle\!\left\langle 1, \frac{\kappa(\mathcal{L})}{z - \psi} \right\rangle\!\right\rangle_{0,2}^{\mathbb{P}^1, T} = \left\langle\!\left\langle \frac{\kappa(\mathcal{L})}{z(z - \psi)} \right\rangle\!\right\rangle_{0,1}^{\mathbb{P}^1, T}.$$

So (12) is equivalent to (13).

It remains to prove (12) for $\mathcal{L} = \mathcal{O}_{\mathbb{P}^1}(l_1 p_1 + l_2 p_1)$, where $l_1 + l_2 \geq 0$. We will express both sides of (12) in terms of (modified) Bessel functions. A brief review of Bessel functions is given in Appendix A. The equivariant quantum differential equation of $\mathbb{P}^1$ is related to the modified Bessel differential equation by a simple transform (see Appendix B).

Let γ_{l_1,l_2} be defined as in Appendix A. Then

$$\begin{aligned}
\int_{\mathrm{SYZ}(\mathcal{L})} e^{W_t^{\mathsf{w}}/z}\,dy &= \int_{\mathrm{SYZ}(\mathcal{L})} \exp\Big(\frac{e^y + t^0 + qe^{-y} + \mathsf{w}_1 y + \mathsf{w}_2(t^1 - y)}{z}\Big)\,dy \\
&= e^{(t^0+\mathsf{w}_2 t^1)/z}\int_{\gamma_{l_1,l_2}} \exp\Big(\frac{e^{y-i\pi} + qe^{i\pi-y} + (\mathsf{w}_1-\mathsf{w}_2)(y-\pi i)}{z}\Big)\,dy \\
&= (-1)^{(\mathsf{w}_1-w_2)/z}\exp\Big(\frac{t^0}{z} + \frac{\mathsf{w}_1+\mathsf{w}_2}{2z}t^1\Big) \\
&\qquad \times \int_{\gamma_{l_1,l_2}} \exp\Big(-\frac{2\sqrt{q}}{z}\cosh\Big(y - \frac{t^1}{2}\Big) + \frac{\mathsf{w}_1-\mathsf{w}_2}{z}\Big(y - \frac{t^1}{2}\Big)\Big)\,dy \\
&= (-1)^{(\mathsf{w}_1-w_2)/z}\exp\Big(\frac{t^0}{z} + \frac{\mathsf{w}_1+\mathsf{w}_2}{2z}t^1\Big) \\
&\qquad \times \int_{\gamma_{l_1,l_2}} \exp\Big(-\frac{2\sqrt{q}}{z}\cosh(y) + \frac{\mathsf{w}_1-\mathsf{w}_2}{z}y\Big)\,dy.
\end{aligned}$$

By Lemma A.1,

$$\begin{aligned}
&\int_{\gamma_{l_1,l_2}} \exp\Big(-\frac{2\sqrt{q}}{z}\cosh(y) + \frac{\mathsf{w}_1-\mathsf{w}_2}{z}y\Big)\,dy \\
&\quad = \frac{\pi}{\sin\big(((\mathsf{w}_2-\mathsf{w}_1)/z)\pi\big)}\Big(e^{-2\pi i l_1(\mathsf{w}_1-\mathsf{w}_2)/z} I_{(\mathsf{w}_1-\mathsf{w}_2)/z}\Big(\frac{2\sqrt{q}}{z}\Big) \\
&\qquad\qquad - e^{-2\pi i l_2(\mathsf{w}_2-\mathsf{w}_1)/z} I_{(\mathsf{w}_2-\mathsf{w}_1)/z}\Big(\frac{2\sqrt{q}}{z}\Big)\Big) \\
&\quad = -\sum_{\alpha=1}^{2} e^{-2\pi i l_\alpha \chi_\alpha/z}\frac{\pi}{\sin((\chi^\alpha/z)\pi)} I_{\chi^\alpha/z}\Big(\frac{2\sqrt{q}}{z}\Big).
\end{aligned}$$

Therefore, the left-hand side of (12) is

$$\begin{aligned}
&\int_{\mathrm{SYZ}(\mathcal{L})} e^{W_t^{\mathsf{w}}/z}\,dy \\
&\quad = -\exp\Big(\frac{t^0}{z} + \frac{\mathsf{w}_1+\mathsf{w}_2}{2z}t^1\Big)\sum_{\alpha=1}^{2} e^{-(2l_\alpha-1)\pi i\chi_\alpha/z}\frac{\pi}{\sin((\chi^\alpha/z)\pi)} I_{\chi^\alpha/z}\Big(\frac{2\sqrt{q}}{z}\Big).
\end{aligned}$$

Recall from Section 2.5 that

$$J^\alpha = \left\langle\!\!\left\langle 1, \frac{\phi^\alpha}{z-\psi} \right\rangle\!\!\right\rangle_{0,2}^{\mathbb{P}^1,T} = \chi^\alpha \left\langle\!\!\left\langle 1, \frac{\phi_\alpha}{z-\psi} \right\rangle\!\!\right\rangle_{0,2}^{\mathbb{P}^1,T}.$$

We have

$$\begin{aligned} J^\alpha &= e^{(t^0+t^1\mathsf{w}_\alpha)/z} \sum_{d=0}^{\infty} \frac{q^d}{d!z^d} \frac{1}{\prod_{m=1}^d (\chi^\alpha + mz)} \\ &= e^{(t^0+t^1\mathsf{w}_\alpha)/z} \sum_{m=0}^{\infty} \left(\frac{2\sqrt{q}}{z}\right)^{2m} \frac{\Gamma(\chi^\alpha/z+1)}{m!\,\Gamma(\chi^\alpha/z+m+1)} \\ &= \exp\left(\frac{t^0}{z} + \frac{\mathsf{w}^1+\mathsf{w}^2}{2z} t^1\right) z^{\chi^\alpha/z} \Gamma\left(\frac{\chi^\alpha}{z}+1\right) I_{\chi^\alpha/z}\left(\frac{2\sqrt{q}}{z}\right), \\ \kappa(\mathcal{L}) &= \sum_{\alpha=1}^{2} (-z)^{\chi^\alpha/(-z)+1} \Gamma\left(1-\frac{\chi^\alpha}{z}\right) e^{-2l_\alpha \pi \sqrt{-1}\chi^\alpha/z} \phi_\alpha. \end{aligned}$$

So the right-hand side of (12) is

$$\begin{aligned} \left\langle\!\!\left\langle 1, \frac{\kappa(\mathcal{L})}{z-\psi} \right\rangle\!\!\right\rangle_{0,2}^{\mathbb{P}^1,T} &= \sum_{\alpha=1}^{2} (-z)^{\chi^\alpha/(-z)+1} \Gamma\left(1-\frac{\chi^\alpha}{z}\right) e^{-2\pi i l_\alpha \chi^\alpha/z} \frac{J^\alpha}{\chi^\alpha} \\ &= -\exp\left(\frac{t^0}{z} + \frac{\mathsf{w}^1+\mathsf{w}^2}{2z} t^1\right) \\ &\qquad \times \sum_{\alpha=1}^{2} (-1)^{\chi^\alpha/(-z)} e^{-2\pi i l_\alpha \chi^\alpha/z} \frac{\pi}{\sin((\chi^\alpha/z)\pi)} I_{\chi^\alpha/z}\left(\frac{2\sqrt{q}}{z}\right) \\ &= -\exp\left(\frac{t^0}{z} + \frac{\mathsf{w}^1+\mathsf{w}^2}{2z} t^1\right) \\ &\qquad \times \sum_{\alpha=1}^{2} e^{-(2l_\alpha-1)\pi i \chi^\alpha/z} \frac{\pi}{\sin((\chi^\alpha/z)\pi)} I_{\chi^\alpha/z}\left(\frac{2\sqrt{q}}{z}\right). \qquad \square \end{aligned}$$

Remark 3.9 Definition 3.6 (equivariant K–theoretic framing) and Definition 3.7 (equivariant SYZ T–dual) can be extended to any projective toric manifold. In [13], the first author uses the mirror theorem [17; 25] and results in [22] to extend Theorem 3.8 to any semi-Fano projective toric manifold. The left-hand side of (12) is known as the central charge of the Lagrangian brane $\mathrm{SYZ}(\mathcal{L})$.

Proposition 3.10 *The A– and B–model R–matrices are equal:*

$$R_\beta^{\ \alpha}(z) = \check{R}_\beta^{\ \alpha}(z).$$

Proof By the asymptotic decomposition theorem of the S–matrix (Theorem 3.2),we only have to compare at the limit $q=0$, $t_0=0$ since both $\widetilde{S}$ and $\check{S}$ are unitary. Notice that Ψ has a nondegenerate limit at $q=0$, so it suffices to show that

$$\widetilde{S}_i^{\hat{\alpha}} e^{-u^\alpha/z}\big|_{q=0,\,t_0=0} \sim \frac{1}{\sqrt{-2\pi z}} \check{S}_i^{\boldsymbol{\alpha}} e^{-\check{u}^\alpha/z}\Big|_{q=0,\,t_0=0}.$$

The Lefschetz thimble γ_2 is $\{Y \mid Y \in (-\infty,0)\}$. While the Lefschetz thimble γ_1 could not be explicitly depicted, we could alternatively consider the thimble $\gamma_1' = \{Y \mid Y \in (0,\infty)\}$ for $z<0$ of the oscillating integral $\int e^{W_t^{\mathrm{w}}/z}\,dy$. The integral yields the same asymptotic answer once we analytically continue $z<0$ to $z>0$, since the stationary phase expansion only depends on the local behavior (higher-order derivatives) of W_t^{w} at the critical points.

So, letting $Y=-Tz$ for $\alpha=2$, or $Y=-q/(Tz)$ for $\alpha=1$,

$$e^{-\check{u}^\alpha/z}\check{S}_0^{\boldsymbol{\alpha}} = e^{-\Delta^\alpha(q)/z}\left(\frac{\chi^\alpha+\Delta^\alpha(q)}{2}\right)^{\frac{\chi^\alpha}{z}}(-z)^{-\chi^\alpha/z}\int_0^\infty e^{-T}e^{-q/(Tz^2)}T^{\chi_\alpha/z-1}\,dT.$$

Taking the limit $q\to 0$,

$$\begin{aligned}\frac{1}{\sqrt{-2\pi z}}e^{-\check{u}^\alpha/z}\check{S}_0^{\boldsymbol{\alpha}}\Big|_{q=0} &= \frac{1}{\sqrt{-2\pi z}}e^{-\chi^\alpha/z}\left(\frac{-\chi^\alpha}{z}\right)^{\frac{\chi^\alpha}{z}}\Gamma\left(\frac{-\chi^\alpha}{z}\right)\\ &\sim \sqrt{\frac{1}{\chi^\alpha}}\exp\left(-\sum_{n=1}^\infty \frac{B_{2n}}{2n(2n-1)}\left(\frac{z}{\chi^\alpha}\right)^{2n-1}\right)\\ &\sim \widetilde{S}_0^{\hat{\alpha}}e^{-u^\alpha/z}\big|_{q=0}.\end{aligned}$$

Here we use the Stirling formula

$$\log\Gamma(z)\sim\frac{1}{2}\log(2\pi)+\left(z-\frac{1}{2}\right)\log z - z + \sum_{n=1}^\infty\frac{B_{2n}}{2n(2n-1)}z^{1-2n}.$$

Notice that

$$\check{S}_1^{\boldsymbol{\alpha}} = z\frac{\partial}{\partial t_1}\check{S}_0^{\alpha} = z\int_{\gamma_\alpha} e^{W_t^{\mathrm{w}}/z}\left(\frac{q}{Y}+\mathrm{w}_2\right)\frac{dY}{Y},$$

and similar calculation shows (letting $Y=-Tz$ if $\alpha=2$ and $Y=-q/(Tz)$ if $\alpha=1$)

$$\begin{aligned}\frac{1}{\sqrt{-2\pi z}}e^{-\check{u}^\alpha/z}\check{S}_1^{\boldsymbol{\alpha}}\Big|_{q=0} &\sim \mathrm{w}^\alpha\sqrt{\frac{1}{\chi^\alpha}}\exp\left(-\sum_{n=1}^\infty\frac{B_{2n}}{2n(2n-1)}\left(\frac{z}{\chi^\alpha}\right)^{2n-1}\right)\\ &\sim \widetilde{S}_1^{\hat{\alpha}}e^{-u^\alpha/z}\big|_{q=0}.\end{aligned}$$

This concludes the proof. ◻

Notice that the matrix $\check{R}$ is given by the asymptotic expansion. This theorem does not imply $\widetilde{S}_i{}^{\hat{\alpha}} e^{-u^\alpha/z} = \check{S}_i{}^{\boldsymbol{\alpha}} e^{-\check{u}^\alpha/z}/\sqrt{-2\pi z}$, which are unequal.

3.6 The Eynard–Orantin topological recursion and the B–model graph sum

Let $\omega_{g,n}$ be defined recursively by the Eynard–Orantin topological recursion [10]:

$$\omega_{0,1} = 0, \quad \omega_{0,2} = B(Y_1, Y_2) = \frac{dY_1 \otimes dY_2}{(Y_1 - Y_2)^2}.$$

When $2g-2+n>0$,

$$\begin{aligned}\omega_{g,n}(Y_1,\dots,Y_n) = \sum_{\alpha=1}^{2} \mathrm{Res}_{Y\to P_\alpha} &\frac{-\int_{\xi=Y}^{\hat{Y}} B(Y_n,\xi)}{2(\log(Y)-\log(\hat{Y}))dW} \\ &\times \Bigg(\omega_{g-1,n+1}(Y,\hat{Y},Y_1,\dots,Y_{n-1}) + \\ &\sum_{g_1+g_2=g} \sum_{\substack{I\cup J=\{1,\dots,n-1\}\\ I\cap J=\varnothing}} \omega_{g_1,|I|+1}(Y,Y_I)\omega_{g_2,|J|+1}(\hat{Y},Y_J)\Bigg),\end{aligned}$$

where $Y \neq P_\alpha$ is in a small neighborhood of P_α and $\hat{Y} \neq Y$ is the other point in the neighborhood such that $W_q^{\mathsf{w}}(\hat{Y}) = W_q^{\mathsf{w}}(Y)$.

The B–model invariants $\omega_{g,n}$ can be expressed as graph sums [23; 7; 8; 5]. We will use the formula stated in [5, Theorem 3.7], which is equivalent to the formula in [7, Theorem 5.1]. Given a labeled graph $\vec{\Gamma} \in \mathbf{\Gamma}_{g,n}(\mathbb{P}^1)$ with $L^o(\Gamma) = \{l_1,\dots,l_n\}$, we define its weight to be

$$\begin{aligned}w(\vec{\Gamma}) = (-1)^{g(\vec{\Gamma})-1+n} \prod_{v\in V(\Gamma)} \left(\frac{h_1^\alpha}{\sqrt{2}}\right)^{2-2g-\mathrm{val}(v)} \Big\langle \prod_{h\in H(v)} \tau_{k(h)} \Big\rangle_{g(v)} \prod_{e\in E(\Gamma)} \check{B}^{\alpha(v_1(e)),\alpha(v_2(e))}_{k(e),l(e)} \\ \cdot \prod_{j=1}^{n} \frac{1}{\sqrt{-2}} d\xi^{\alpha(l_j)}_{k(l_j)}(Y_j) \prod_{l\in\mathcal{L}^1(\Gamma)} \left(-\frac{1}{\sqrt{-2}}\right) \check{h}^{\alpha(l)}_{k(l)}.\end{aligned}$$

Here,

$$\check{h}_k^\alpha = -\frac{2(2k-1)!!\, h^\alpha_{2k-1}}{\sqrt{-1}^{2k-1}}, \quad \check{B}^{\alpha,\beta}_{k,l} = [u^{-k}v^{-l}]\check{B}^{\alpha,\beta}(u,v,q).$$

Note that the definitions of $\check{B}^{\alpha,\beta}_{k,l}$, $\check{h}^\alpha_k$ and $d\xi^\alpha_k$ in this paper are slightly different from those in [5]; for example, the definition of $\check{B}^{\alpha,\beta}_{k,l}$ in this paper differs from [5, Equation (3.11)] by a factor of 2^{-k-l-1}. In our notation, [5, Theorem 3.7] is equivalent to:

Theorem 3.11 *For $2g-2+n>0$,*

$$\omega_{g,n}=\sum_{\Gamma\in\mathbf{\Gamma}_{g,n}(\mathbb{P}^1)}\frac{w(\vec{\Gamma})}{|\mathrm{Aut}(\vec{\Gamma})|}.$$

3.7 All-genus mirror symmetry

Given a meromorphic function $f(Y)$ on $\mathbb{P}^1$ which is holomorphic on $\mathbb{P}^1\setminus\{P_1,P_2\}$, define

$$\theta(f)=\frac{df}{dW}=\frac{Y^2}{(Y-P_1)(Y-P_2)}\frac{df}{dY}.$$

Then $\theta(f)$ is also a meromorphic function which is holomorphic on $\mathbb{P}^1\setminus\{P_1,P_2\}$. For $\alpha\in\{1,2\}$, let

$$\xi_{\alpha,0}=\frac{1}{\sqrt{-1}}\sqrt{\frac{2}{\Delta^\alpha(q)}}\frac{P_\alpha}{Y-P_\alpha}.$$

Then $\xi_{\alpha,0}$ is a meromorphic function on $\mathbb{P}^1$ with a simple pole at $Y=P_\alpha$ and holomorphic elsewhere. Moreover, the differential of $\xi_{\alpha,0}$ is $d\xi_{\alpha,0}$. For $k>0$, define

$$W_k^\alpha:=d((-1)^k\theta^k(\xi_{\alpha,0})).$$

Define

$$\check{S}^{\boldsymbol{\alpha}}_{\underline{\widehat{\beta}}}(z)=-z\int_{y\in\gamma_\alpha}e^{x/z}\frac{d\xi_{\beta,0}}{\sqrt{-2}},\quad \check{S}^{\kappa(\mathcal{L})}_{\underline{\widehat{\beta}}}(z)=-z\int_{y\in\mathrm{SYZ}(\mathcal{L})}e^{x/z}\frac{d\xi_{\beta,0}}{\sqrt{-2}}. \tag{14}$$

Then

$$\check{S}^{\boldsymbol{\alpha}}_{\underline{\widehat{\beta}}}(z)=-z^{k+1}\int_{y\in\gamma_\alpha}e^{W(y)/z}\frac{W_k^\beta}{\sqrt{-2}},\quad \check{S}^{\kappa(\mathcal{L})}_{\underline{\widehat{\beta}}}(z)=-z^{k+1}\int_{y\in\mathrm{SYZ}(\mathcal{L})}e^{W(y)/z}\frac{W_k^\beta}{\sqrt{-2}}.$$

Therefore,

$$\int_{y\in\mathrm{SYZ}(\mathcal{L})}e^{W(y)/z}\frac{W_k^\beta}{\sqrt{-2}}=-z^{-k-1}\check{S}^{\kappa(\mathcal{L})}_{\underline{\widehat{\beta}}}(z)=-z^{-k-1}\left\langle\!\!\left\langle\hat{\phi}_\alpha(q),\frac{\kappa(\mathcal{L})}{z-\psi}\right\rangle\!\!\right\rangle_{0,2}^{\mathbb{P}^1,T}, \tag{15}$$

where the last equality follows from Theorem 3.8.

For $\alpha=1,2$ and $j=1,\dots,n$, let

$$\tilde{\boldsymbol{u}}_j^\alpha(z)=\sum_{\beta=1}^2 S^{\underline{\hat{\alpha}}}_{\underline{\widehat{\beta}}}(z)\frac{\boldsymbol{u}_j^\beta(z)}{\sqrt{\Delta^\beta(q)}}. \tag{16}$$

Theorem A (all-genus equivariant mirror symmetry for $\mathbb{P}^1$) *When $n>0$ and $2g-2+n>0$, we have*

$$\omega_{g,n}\big|_{W_k^\alpha(Y_j)/\sqrt{-2}=(\tilde{u}_j)_k^\alpha}=(-1)^{g-1+n}F_{g,n}^{\mathbb{P}^1,T}(\boldsymbol{u}_1,\dots,\boldsymbol{u}_n,\boldsymbol{t}). \tag{17}$$

Proof We will prove this theorem by comparing the A–model graph sum in the end of Section 2.8 and the B–model graph sum in Section 3.6.

- **Vertex** By Section 3.2, we have $h_1^\alpha(q) = \sqrt{2/\Delta^\alpha(q)}$. So, in the B–model vertex, $h_1^\alpha/\sqrt{2} = \sqrt{1/\Delta^\alpha(q)}$. Therefore the B–model vertex matches the A–model vertex.

- **Edge** By (11), we know that

$$\begin{aligned}\check{B}^{\alpha,\beta}_{k,l} &= [u^{-k}v^{-l}]\Big(\frac{uv}{u+v}\Big(\delta_{\alpha,\beta} - \sum_{\gamma=1,2} \check{R}_\gamma{}^\alpha\Big(-\frac{1}{u}\Big)\check{R}_\gamma{}^\beta\Big(-\frac{1}{v}\Big)\Big)\Big)\\ &= [z^k w^l]\Big(\frac{1}{z+w}\Big(\delta_{\alpha,\beta} - \sum_{\gamma=1,2} \check{R}_\gamma{}^\alpha(-z)\check{R}_\gamma{}^\beta(-w)\Big)\Big).\end{aligned}$$

By definition,

$$\mathcal{E}^{\alpha,\beta}_{k,l} = [z^k w^l]\Big(\frac{1}{z+w}\Big(\delta_{\alpha,\beta} - \sum_{\gamma=1,2} R_\gamma{}^\alpha(-z)R_\gamma{}^\beta(-w)\Big)\Big).$$

By Proposition 3.10, $\check{R}_\beta{}^\alpha(z) = R_\beta{}^\alpha(z)$, so

$$\check{B}^{\alpha,\beta}_{k,l} = \mathcal{E}^{\alpha,\beta}_{k,l}.$$

- **Ordinary leaf** We have the following expression for $d\xi_k^\alpha$ (see [15]):

$$d\xi_k^\alpha = W_k^\alpha - \sum_{i=0}^{k-1}\sum_\beta \check{B}^{\alpha,\beta}_{k-1-i,0} W_i^\beta.$$

By the calculation for edge above, for $k, l \in \mathbb{Z}_{\geq 0}$,

$$\check{B}^{\alpha,\beta}_{k,l} = [z^k w^l]\Big(\frac{1}{z+w}\Big(\delta_{\alpha,\beta} - \sum_{\gamma=1,2} R_\gamma{}^\alpha(-z)R_\gamma{}^\beta(-w)\Big)\Big).$$

We also have

$$[z^0](R_\beta{}^\alpha(-z)) = \delta_{\alpha,\beta}.$$

Therefore,

$$d\xi_k^\alpha = \sum_{i=0}^{k}\sum_{\beta=1}^{2}([z^{k-i}]R_\beta{}^\alpha(-z))W_i^\beta,$$

so under the identification

$$\frac{1}{\sqrt{-2}} W_k^\alpha(Y_j) = (\tilde{u}_j)_k^\alpha,$$

the B–model ordinary leaf matches the A–model ordinary leaf.

- **Dilaton leaf** We have the following relation between $\check{h}^\alpha_k$ and $f^\alpha_\beta(u,q)$ (see [15]):
$$\check{h}^\alpha_k = [u^{1-k}] \sum_\beta \sqrt{-1} h^\beta_1 f^\alpha_\beta(u,q).$$

By the relation
$$R_\beta{}^\alpha(z) = f^\alpha_\beta\Big(\frac{-1}{z}\Big)$$
and the fact $h^\beta_1(q) = \sqrt{2/\Delta^\beta(q)}$, it is easy to see that the B–model dilaton leaf matches the A–model dilaton leaf. □

Taking Laplace transforms at appropriate cycles to Theorem A produces a theorem concerning descendant potential.

Theorem B (all-genus full descendant equivariant mirror symmetry for $\mathbb{P}^1$) *Suppose that $n > 0$ and $2g-2+n > 0$. For any $\mathcal{L}_1, \dots, \mathcal{L}_n \in K_T(\mathbb{P}^1)$, there is a formal power series identity*

$$\int_{y_1 \in \mathrm{SYZ}(\mathcal{L}_1)} \cdots \int_{y_n \in \mathrm{SYZ}(\mathcal{L}_n)} e^{W(y_1)/z_1 + \cdots + W(y_n)/z_n} \omega_{g,n} = (-1)^{g-1} \left\langle\!\!\left\langle \frac{\kappa(\mathcal{L}_1)}{z_1 - \psi_1}, \dots, \frac{\kappa(\mathcal{L}_n)}{z_n - \psi_n} \right\rangle\!\!\right\rangle_{g,n}. \tag{18}$$

Remark 3.12 By Theorem 3.8,
$$\int_{y_1 \in \mathrm{SYZ}(\mathcal{L})} e^{W(y_1)/z_1} y\,dx = -\left\langle\!\!\left\langle \frac{\kappa(\mathcal{L}_1)}{z_1 - \psi_1} \right\rangle\!\!\right\rangle^{\mathbb{P}^1, T}_{0,1}, \tag{19}$$
which is the analogue of (18) in the unstable case $(g,n) = (0,1)$.

Proof of Theorem B By (16),
$$\tilde{\boldsymbol{u}}^\alpha_j(z) = \sum_{\beta=1}^{2} \sqrt{\Delta^\alpha(q)} \left\langle\!\!\left\langle \phi_\alpha(q), \frac{\phi_\beta(q)}{z - \psi} \right\rangle\!\!\right\rangle^{\mathbb{P}^1, T}_{0,2} \boldsymbol{u}^\beta_j(z).$$

Define the flat coordinates $\bar{\boldsymbol{u}}^\alpha_j$ by
$$\sum_{\alpha=1}^{2} \boldsymbol{u}^\alpha_j(z)\phi_\alpha(q) = \sum_{\alpha=1}^{2} \bar{\boldsymbol{u}}^\alpha_j(z)\phi_\alpha(0),$$
and a power series in $\frac{1}{z}$,
$$S^{\hat{\alpha}}_{\ \beta}(z) = \left\langle\!\!\left\langle \hat{\phi}_\alpha(q), \frac{\phi_\beta(0)}{z - \psi} \right\rangle\!\!\right\rangle_{0,2}.$$

Then

$$\widetilde{\boldsymbol{u}}_j^\alpha(z) = \sum_{\beta=1}^{2} \left(\left\langle\!\!\left\langle \widehat{\phi}_\alpha(q), \frac{\phi_\beta(0)}{z-\psi} \right\rangle\!\!\right\rangle \overline{\boldsymbol{u}}_j^\beta(z) \right)_+ = \sum_{\beta=1}^{2} (S^{\widehat{\alpha}}_{\ \beta}(z)\overline{\boldsymbol{u}}_j^\beta(z))_+ .$$

Notice that $(S^{\widehat{\alpha}}_{\ \beta})$ is unitary, ie $\sum_\gamma S^{\widehat{\gamma}}_{\ \alpha}(z) S^{\widehat{\gamma}}_{\ \beta}(-z) = \delta/\chi^\beta_{\alpha\beta}$. We have

$$\sum_{\alpha=1}^{2} (S^{\widehat{\alpha}}_{\ \gamma}(-z)\widetilde{\boldsymbol{u}}_j^\alpha(z))_+ = \sum_{\alpha=1}^{2} \left(\sum_{\beta=1}^{2} S^{\widehat{\alpha}}_{\ \beta}(z) S^{\widehat{\alpha}}_{\ \gamma}(-z) \overline{\boldsymbol{u}}_j^\beta(z) \right) = \frac{\overline{\boldsymbol{u}}_j^\gamma(z)}{\chi^\gamma}.$$

Taking the Laplace transform of $\omega_{g,n}$,

$$\begin{aligned}
&\int_{y_1\in \mathrm{SYZ}(\mathcal{L}_1)} \cdots \int_{y_n\in \mathrm{SYZ}(\mathcal{L}_n)} e^{W(y_1)/z_1+\cdots+W(y_n)/z_n} \omega_{g,n} \\
&= \int_{y_1\in \mathrm{SYZ}(\mathcal{L}_1)} \cdots \int_{y_n\in \mathrm{SYZ}(\mathcal{L}_n)} e^{\sum_{i=1}^n W(y_i)/z_i} (-1)^{g-1+n} \\
&\qquad\qquad \cdot \left(\sum_{\beta_i,a_i} \left\langle\!\!\left\langle \prod_{i=1}^n \tau_{a_i}(\phi_{\beta_i}(0)) \right\rangle\!\!\right\rangle_{g,n} \prod_{i=1}^n (\overline{u}_i)^{\beta_i}_{a_i} \right) \Bigg|_{(\widetilde{u}_j)^\beta_k = W^\beta_k(y_j)/\sqrt{-2}} \\
&= \int_{y_1\in \mathrm{SYZ}(\mathcal{L}_1)} \cdots \int_{y_n\in \mathrm{SYZ}(\mathcal{L}_n)} e^{\sum_{i=1}^n W(y_i)/z_i} (-1)^{g-1+n} \\
&\quad \cdot \left(\sum_{\beta_i,a_i} \left\langle\!\!\left\langle \prod_{i=1}^n \tau_{a_i}(\phi_{\beta_i}(0)) \right\rangle\!\!\right\rangle_{g,n} \prod_{i=1}^n \left(\chi^{\beta_i} \sum_{\alpha=1}^{2} \sum_{k\in\mathbb{Z}_{\geq 0}} [z_i^{a_i-k}] S^{\widehat{\alpha}}_{\ \beta_i}(-z_i) \frac{W^\alpha_k(y_i)}{\sqrt{-2}} \right) \right).
\end{aligned}$$

Using (15),

$$\begin{aligned}
&\int_{y_1\in \mathrm{SYZ}(\mathcal{L}_1)} \cdots \int_{y_n\in \mathrm{SYZ}(\mathcal{L}_n)} e^{W(y_1)/z_1+\cdots+W(y_n)/z_n} \omega_{g,n} \\
&= (-1)^{g-1+n} \left(\sum_{\beta_i,a_i} \left\langle\!\!\left\langle \prod_{i=1}^n \tau_{a_i}(\phi_{\beta_i}(0)) \right\rangle\!\!\right\rangle_{g,n} \right. \\
&\qquad\qquad \left. \cdot \prod_{i=1}^n \left(\chi^{\beta_i} \sum_{\alpha=1}^{2} \sum_{k\in\mathbb{Z}_{\geq 0}} ([z_i^{a_i-k}] S^{\widehat{\alpha}}_{\ \beta_i}(-z_i)) S_{\underline{\widehat{\alpha}}}^{\ \kappa(\mathcal{L}_i)}(z_i)(-z_i^{-k-1}) \right) \right) \\
&= (-1)^{g-1} \sum_{\beta_i,a_i} \left\langle\!\!\left\langle \prod_{i=1}^n \tau_{a_i}(\phi_{\beta_i}(0)) \right\rangle\!\!\right\rangle_{g,n} \prod_{i=1}^n \chi^{\beta_i} (\phi_{\beta_i}(0), \kappa(\mathcal{L}_i)) z_i^{-a_i-1} \\
&= (-1)^{g-1} \left\langle\!\!\left\langle \frac{\kappa(\mathcal{L}_1)}{z_1-\psi_1}, \ldots, \frac{\kappa(\mathcal{L}_n)}{z_n-\psi_n} \right\rangle\!\!\right\rangle_{g,n}. \qquad \square
\end{aligned}$$

4 The nonequivariant limit and the Norbury–Scott conjecture

In this section, we consider the nonequivariant limit $\mathsf{w}_1 = \mathsf{w}_2 = 0$.

4.1 The nonequivariant R–matrix

By [20, Section 1.3], $R(z) = I + \sum_{n=1}^{\infty} R_n z^n$ is uniquely determined by:

(1) The recursive relation $(d + \Psi^{-1} d\Psi) R_n = [dU, R_{n+1}]$.

(2) The homogeneity of $R(z)$: $R_n q^{n/2}$ is a constant matrix.

The unique solution $R(z)$ satisfying the above conditions was computed explicitly in [29]:

Lemma 4.1 [29, Lemma 3.1] *We have*

$$R_n = q^{-n/2} \frac{(2n-1)!!\,(2n-3)!!}{n!\,2^{4n}} \begin{pmatrix} -1 & 2n\sqrt{-1}(-1)^{n+1} \\ 2n\sqrt{-1} & (-1)^{n+1} \end{pmatrix}.$$

By Proposition 3.10 , $R(z) = \check{R}(z)$. In this subsection, we recover the above lemma by computing the stationary phase expansion of $\check{S}$.

We assume $z, q \in (0, \infty)$, where $q = Qe^{t^1}$. Then

$$\begin{aligned} \check{S}_0{}^2 &= \int_{y=-\infty}^{y=+\infty} e^{(t^0 + e^{y-i\pi} + qe^{-(y-i\pi)})/z}\, dy \\ &= e^{t^0/z} \int_{y=-\infty}^{y=+\infty} e^{-2\sqrt{q}\cosh(y-t^1/2)/z}\, dy \\ &= e^{t^0/z} \int_{y=-\infty}^{y=+\infty} e^{-2\sqrt{q}\cosh(y)/z}\, dy \\ &= 2e^{(t^0-2\sqrt{q})/z} \int_{y=0}^{y=+\infty} e^{-2\sqrt{q}(\cosh(y)-1)/z}\, dy. \end{aligned}$$

Let $T = 2\sqrt{q}(\cosh(y)-1)/z$; then

$$y = \cosh^{-1}\Big(1 + \frac{zT}{2\sqrt{q}}\Big), \quad dy = \frac{1}{2} q^{-\frac{1}{4}} T^{-\frac{1}{2}} \sqrt{\frac{z}{1+zT/(4\sqrt{q})}},$$

$$
\begin{aligned}
\check{S}_0{}^2 &= e^{(t^0-2\sqrt{q})/z}\sum_{n=0}^{\infty}\Big(\frac{z}{\sqrt{q}}\Big)^{n+\frac12}\binom{-1/2}{n}2^{-2n}\int_{T=0}^{T=+\infty}e^{-T}T^{n-\frac12}\,dT\\
&= e^{(t^0-2\sqrt{q})/z}\sum_{n=0}^{\infty}\Big(\frac{z}{\sqrt{q}}\Big)^{n+\frac12}\frac{(-1)^n(2n-1)!!}{n!2^{3n}}\Gamma\Big(n+\frac12\Big)\\
&= \sqrt{\pi}e^{(t^0-2\sqrt{q})/z}\sum_{n=0}^{\infty}\Big(\frac{z}{\sqrt{q}}\Big)^{n+\frac12}\frac{(-1)^n((2n-1)!!)^2}{n!2^{4n}},\\
\check{S}_1{}^2 &= z\frac{\partial}{\partial t_1}\check{S}_0{}^2\\
&= \sqrt{\pi}ze^{(t^0-2\sqrt{q})/z}\sum_{n=0}^{\infty}\Big(\frac{z}{\sqrt{q}}\Big)^{n-\frac12}\Big(1+\Big(\frac14+\frac{n}{2}\Big)\frac{z}{\sqrt{q}}\Big)\frac{(-1)^{n+1}((2n-1)!!)^2}{n!2^{4n}}.
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
\check{S}_0{}^1 &= \sqrt{-\pi}e^{(t^0+2\sqrt{q})/z}\sum_{n=0}^{\infty}\Big(\frac{z}{\sqrt{q}}\Big)^{n+\frac12}\frac{((2n-1)!!)^2}{n!2^{4n}};\\
\check{S}_1{}^1 &= \sqrt{-\pi}ze^{(t^0+2\sqrt{q})/z}\sum_{n=0}^{\infty}\Big(\frac{z}{\sqrt{q}}\Big)^{n-\frac12}\Big(1-\Big(\frac14+\frac{n}{2}\Big)\frac{z}{\sqrt{q}}\Big)\frac{((2n-1)!!)^2}{n!2^{4n}}.
\end{aligned}
$$

Therefore,

$$
\widetilde{S}(z)=\frac{1}{\sqrt{-2\pi z}}\check{S}(z),\quad [z^n](\widetilde{S}(z)e^{-U/z})=\begin{pmatrix}A & B\\ C & D\end{pmatrix},
$$

where

$$
\begin{aligned}
A &= \frac{((2n-1)!!)^2}{\sqrt{2}n!2^{4n}q^{\frac12 n+\frac14}},\quad B=\frac{\sqrt{-1}(-1)^{n+1}((2n-1)!!)^2}{\sqrt{2}n!2^{4n}q^{\frac12 n+\frac14}},\\
C &= \frac{((2n-1)!!)^2}{\sqrt{2}n!2^{4n}q^{\frac12 n-\frac14}}-\Big(\frac{n}{2}-\frac14\Big)\frac{((2n-3)!!)^2}{\sqrt{2}(n-1)!2^{4n-4}q^{\frac12 n-\frac14}},\\
D &= \frac{\sqrt{-1}(-1)^n((2n-1)!!)^2}{\sqrt{2}n!2^{4n}q^{\frac12 n-\frac14}}+\Big(\frac{n}{2}-\frac14\Big)\frac{\sqrt{-1}(-1)^{n+1}((2n-3)!!)^2}{\sqrt{2}(n-1)!2^{4n-4}q^{\frac12 n-\frac14}},
\end{aligned}
$$

and

$$
\begin{aligned}
R_n &= \begin{pmatrix}-\frac{(2n-1)!!\,(2n-3)!!}{n!2^{4n}} & \frac{\sqrt{-1}(-1)^{n+1}(2n-1)!!\,(2n-3)!!}{(n-1)!2^{4n-1}}\\ \frac{\sqrt{-1}(2n-1)!!\,(2n-3)!!}{(n-1)!2^{4n-1}} & \frac{(-1)^{n+1}(2n-1)!!\,(2n-3)!!}{n!2^{4n}}\end{pmatrix}q^{-\frac12 n}\\
&= q^{-\frac12 n}\frac{(2n-1)!!\,(2n-3)!!}{n!2^{4n}}\begin{pmatrix}-1 & 2n\sqrt{-1}(-1)^{n+1}\\ 2n\sqrt{-1} & (-1)^{n+1}\end{pmatrix}.
\end{aligned}
$$

4.2 The Norbury–Scott conjecture

In this subsection, we assume $\mathsf{w}_1 = \mathsf{w}_2 = t^0 = 0$. Then

$$\langle\!\langle \tau_{a_1}(H)\cdots\tau_{a_n}(H)\rangle\!\rangle_{g,n}^{\mathbb{P}^1} = q^{\frac{1}{2}(\sum_{i=1}^n a_i)+1-g}\langle \tau_{a_1}(H)\cdots\tau_{a_n}(H)\rangle_{g,n}^{\mathbb{P}^1}.$$

Note that when $\frac{1}{2}\big(\sum_{i=1}^n a_i\big)+1-g$ is not an nonnegative integer, both sides are zero.

When $2g-2+n>0$, the symmetric n–form $\omega_{g,n}$ is holomorphic near $Y=0$, and one may expand it in the local holomorphic coordinate $\tilde{x} = x^{-1} = (Y+q/Y)^{-1}$.

Theorem 4.2 *Suppose that $2g-2+n>0$. Then, near $Y=0$, the symmetric n–form $\omega_{g,n}$ has the expansion*

$$\omega_{g,n} = (-1)^{g-1+n}\sum_{\substack{a_1,\dots,a_n\\ a_i\in\mathbb{Z}_{\geq 0}}}\langle\!\langle \tau_{a_1}(H)\cdots\tau_{a_n}(H)\rangle\!\rangle_{g,n}^{\mathbb{P}^1}\prod_{j=1}^n\frac{(a_j+1)!}{x^{a_j+2}}dx_j.$$

The Norbury–Scott conjecture corresponds to the specialization $q=1$, ie $t^1=0$ and $Q=1$.

Proof Define $\widetilde{W}_k^\alpha$ by

$$\frac{1}{\sqrt{-2}}\widetilde{W}_k^\alpha = \tilde{\boldsymbol{u}}_k^\alpha\big|_{t_a^0=0,\, t_a^1=(a+1)!x^{-a-2}dx}.$$

By Theorem A, it suffices to show that $\widetilde{W}_k^\alpha$ agrees with the expansion of W_k^α near $Y=0$ in $\tilde{x}=x^{-1}$.

We now compute $\widetilde{W}_k^\alpha$ explicitly:

$$\begin{aligned}
J &= e^{(t^0+t^1H)/z}\bigg(1+\sum_{d=1}^\infty\frac{q^d}{\prod_{m=1}^d(H+mz)^2}\bigg)\\
&= e^{t^0/z}\Big(1+t^1\frac{H}{z}\Big)\bigg(1+\sum_{d=1}^\infty\frac{q^d}{z^{2d}(d!)^2}-2\bigg(\sum_{d=1}^\infty\frac{q^d}{z^{2d}(d!)^2}\sum_{m=1}^d\frac{1}{m}\bigg)\frac{H}{z}\bigg)\\
&= e^{t^0/z}\bigg(1+\sum_{d=1}^\infty\frac{q^d}{z^{2d}(d!)^2}\bigg)\\
&\qquad\qquad + e^{t^0/z}\bigg(t^1\bigg(1+\sum_{d=1}^\infty\frac{q^d}{z^{2d+1}(d!)^2}\bigg)-2\sum_{d=1}^\infty\frac{q^d}{z^{2d+1}(d!)^2}\sum_{m=1}^d\frac{1}{m}\bigg)H,
\end{aligned}$$

$$\begin{aligned} z\frac{\partial J}{\partial t^1} &= e^{t^0/z}\Big(\sum_{d=1}^{\infty}\frac{dq^d}{z^{2d-1}(d!)^2}\Big) \\ &\quad + e^{t^0/z}\bigg(t^1\Big(\sum_{d=1}^{\infty}\frac{dq^d}{z^{2d}(d!)^2}\Big)+1+\sum_{d=1}^{\infty}\frac{q^d}{z^{2d}(d!)^2}\Big(1-2d\sum_{m=1}^{d}\frac{1}{m}\Big)\bigg)H, \end{aligned}$$

$$\begin{aligned} S^0{}_0(z) &= (H,\mathcal{S}(1)) = \Big(1, z\frac{\partial J}{\partial t^1}\Big) \\ &= e^{t^0/z}\bigg(t^1\Big(\sum_{d=1}^{\infty}\frac{dq^d}{z^{2d}(d!)^2}\Big)+1+\sum_{d=1}^{\infty}\frac{q^d}{z^{2d}(d!)^2}\Big(1-2d\sum_{m=1}^{d}\frac{1}{m}\Big)\bigg), \\ S^1{}_0(z) &= (1,\mathcal{S}(1)) = (1,J) \\ &= e^{t^0/z}\bigg(t^1\Big(1+\sum_{d=1}^{\infty}\frac{q^d}{z^{2d+1}(d!)^2}\Big)-2\sum_{d=1}^{\infty}\frac{q^d}{z^{2d+1}(d!)^2}\sum_{m=1}^{d}\frac{1}{m}\bigg), \\ S^0{}_1(z) &= (H,\mathcal{S}(H)) = (H, z\frac{\partial J}{\partial t^1}) = e^{t^0/z}\Big(\sum_{d=0}^{\infty}\frac{q^{d+1}}{z^{2d+1}d!(d+1)!}\Big), \\ S^1{}_1(z) &= (1,\mathcal{S}(H)) = (H,J) = e^{t^0/z}\Big(1+\sum_{d=1}^{\infty}\frac{q^d}{z^{2d}(d!)^2}\Big) \end{aligned}$$

$$S^{\widehat{\alpha}}{}_j(z) = \sum_{i=0}^{1}\Psi_i{}^{\alpha}S^i{}_j(z),$$

$$S^{\hat{1}}{}_1(z) = \frac{1}{\sqrt{2}}e^{t^0/z}\sum_{n=0}^{\infty}\frac{(\sqrt{q})^{n+\frac12}}{z^n}\frac{1}{\lfloor n/2\rfloor!\,\lceil n/2\rceil!},$$

$$S^{\hat{2}}{}_1(z) = \frac{1}{\sqrt{2}}e^{t^0/z}\sum_{n=0}^{\infty}\frac{(-\sqrt{q})^{n+\frac12}}{z^n}\frac{1}{\lfloor n/2\rfloor!\,\lceil n/2\rceil!},$$

$$\widetilde{\boldsymbol{u}}^{\alpha}(z) = \sum_{i=0}^{1}S^{\widehat{\alpha}}{}_i(z)t^i(z),$$

$$\widetilde{u}^1_k\big|_{t^0_a=0} = \frac{1}{\sqrt{2}}\sum_{n=0}^{\infty}\frac{(\sqrt{q})^{n+\frac12}}{\lfloor n/2\rfloor!\,\lceil n/2\rceil!}t^1_{k+n},$$

$$\widetilde{u}^2_k\big|_{t^0_a=0} = \frac{1}{\sqrt{2}}\sum_{n=0}^{\infty}\frac{(-\sqrt{q})^{n+\frac12}}{\lfloor n/2\rfloor!\,\lceil n/2\rceil!}t^1_{k+n}.$$

For $\alpha = 1, 2$,

$$\widetilde{W}^{\alpha}_k = \sqrt{-2}\,\widetilde{u}^{\alpha}_k\big|_{t^0_a=0,\,t^1_a=(a+1)!x^{-a-2}dx} = d\Big(\Big(-\frac{d}{dx}\Big)^k\widetilde{\xi}_{\alpha,0}\Big), \tag{20}$$

where

$$\widetilde{\xi}_{1,0} := -\frac{1}{\sqrt{-1}}\sum_{n=0}^{\infty}(\sqrt{q})^{n+\frac{1}{2}}\binom{n}{\lfloor n/2\rfloor}x^{-n-1}, \tag{21}$$

$$\widetilde{\xi}_{2,0} := -\frac{1}{\sqrt{-1}}\sum_{n=0}^{\infty}(-\sqrt{q})^{n+\frac{1}{2}}\binom{n}{\lfloor n/2\rfloor}x^{-n-1}. \tag{22}$$

Recall that

$$W_k^\alpha = d\Big(\Big(-\frac{d}{dx}\Big)^k\xi_{\alpha,0}\Big). \tag{23}$$

By (20) and (23), to complete the proof it remains to show that $\widetilde{\xi}_{\alpha,0}$ agrees with the expansion of $\xi_{\alpha,0}$ near $Y=0$ in $\widetilde{x}=x^{-1}=\big(Y+\frac{q}{Y}\big)^{-1}$.

Assume that $q\in(0,\infty)$. We have

$$P_1=\sqrt{q},\qquad \Delta^1=2\sqrt{q},\qquad \xi_{1,0}=\frac{1}{\sqrt{-1}}\frac{q^{\frac{1}{4}}}{Y-\sqrt{q}},$$

$$P_2=-\sqrt{q},\quad \Delta^2=-2\sqrt{q},\quad \xi_{2,0}=\frac{q^{\frac{1}{4}}}{Y+\sqrt{q}}.$$

The n^{th} coefficient in the expansion of $\widetilde{x}=\big(Y+\frac{q}{Y}\big)^{-1}$ at $Y=0$ is given by the residue

$$\begin{aligned}
\operatorname{Res}_{Y=0}\widetilde{x}^{-n-1}\xi_{1,0}d\widetilde{x} &= -\frac{1}{\sqrt{-1}}q^{\frac{1}{4}}\operatorname{Res}_{Y=0}\Big(Y+\frac{q}{Y}\Big)^{n-1}\Big(1-\frac{q}{Y^2}\Big)\frac{dY}{Y-\sqrt{q}}\\
&= -\frac{1}{\sqrt{-1}}q^{\frac{1}{4}}\operatorname{Res}_{Y=0}\frac{(Y^2+q)^{n-1}(Y+\sqrt{q})}{Y^{n+1}}dY\\
&= -\frac{1}{\sqrt{-1}}(\sqrt{q})^{n-\frac{1}{2}}\binom{n-1}{\lfloor n/2\rfloor},
\end{aligned}$$

where

$$\begin{aligned}
\xi_{1,0} &= -\frac{1}{\sqrt{-1}}\sum_{n=1}^{\infty}(\sqrt{q})^{n-\frac{1}{2}}\binom{n-1}{\lfloor n/2\rfloor}\widetilde{x}^n = -\frac{1}{\sqrt{-1}}\sum_{n=0}^{\infty}(\sqrt{q})^{n+\frac{1}{2}}\binom{n}{\lfloor (n+1)/2\rfloor}\widetilde{x}^{n+1}\\
&= -\frac{1}{\sqrt{-1}}\sum_{n=0}^{\infty}(\sqrt{q})^{n+\frac{1}{2}}\binom{n}{\lfloor n/2\rfloor}x^{-n-1},
\end{aligned}$$

which agrees with $\widetilde{\xi}_{1,0}$, defined in (21), and

$$\operatorname{Res}_{Y=0}\widetilde{x}^{-n-1}\xi_{2,0}d\widetilde{x} = -q^{\frac{1}{4}}\operatorname{Res}_{Y=0}\Big(Y+\frac{q}{Y}\Big)^{n-1}\Big(1-\frac{q}{Y^2}\Big)\frac{dY}{Y+\sqrt{q}}$$

$$= -q^{\frac{1}{4}}\mathrm{Res}_{Y=0}\frac{(Y^2+q)^{n-1}(Y-\sqrt{q})}{Y^{n+1}}dY$$
$$= -\frac{1}{\sqrt{-1}}(-\sqrt{q})^{n-\frac{1}{2}}\binom{n-1}{\lfloor n/2\rfloor},$$

where

$$\xi_{2,0} = -\frac{1}{\sqrt{-1}}\sum_{n=0}^{\infty}(-\sqrt{q})^{n+\frac{1}{2}}\binom{n}{\lfloor n/2\rfloor}x^{-n-1},$$

which agrees with $\widetilde{\xi}_{2,0}$, defined in (22). □

5 The large radius limit and the Bouchard–Mariño conjecture

In this section, we will specialize Theorem A to the large radius limit case. In this case, Theorem A relates the invariant $\omega_{g,n}$ of the limit curve to the equivariant descendent theory of $\mathbb{C}$. After expanding $\xi_{\alpha,0}$ in suitable coordinates, we can relate the corresponding expansion of $\omega_{g,n}$ to the generation function of Hurwitz numbers and therefore reprove the Bouchard–Mariño conjecture [2] on Hurwitz numbers.

Let $\mathsf{w}_2 = 0$ and $t_0 = 0$, and take the large radius limit $q \to 0$. Then our mirror curve becomes

$$x = Y + \mathsf{w}_1 \log Y.$$

When $\mathsf{w}_1 = -1$, this is just the Lambert curve. Recall that the two critical points P_1 and P_2 of $W_t^{\mathsf{w}}(Y)$ are

$$P_\alpha = \frac{\mathsf{w}_2 - \mathsf{w}_1 + \Delta^\alpha(q)}{2}.$$

Since $\Delta^1(0) = \mathsf{w}_1 - \mathsf{w}_2$, we have $P_1 \to 0$ under the limit $q \to 0$. In other words, P_1 goes out of the curve under the limit $q \to 0$ and $\xi_{1,0} = \sqrt{2/\Delta^\alpha(q)}\,P_1/(Y-P_1) \to 0$. As a result, $W_k^1 = d(\theta^k(\xi_{1,0}))$ also tends to zero under the large radius limit.

Under the identification $W_k^\alpha(Y_j)/\sqrt{-2} = (\widetilde{u}_j)_k^\alpha$ in Theorem A, we have $(\widetilde{u}_j)_k^1 \to 0$ when $q \to 0$. On the A–model side, since $q = 0$, the $S-$matrix $(\mathring{S}^\alpha{}_\beta(z))$ is diagonal. Therefore, we also have $(u_j)_k^1 \to 0$ when $q \to 0$ under the identification in Theorem A. This means that in the localization graph of the equivariant GW invariants of $\mathbb{P}^1$, we can only have a constant map to $p_2 \in \mathbb{P}^1$. Since $H|_{p_2} = \mathsf{w}_2 = 0$ and $t^0 = 0$, we cannot have any primary insertions. Therefore, in the large radius limit, we get

$$F_{g,n}^{\mathbb{P}^1,\mathbb{C}^*}(\boldsymbol{u}_1,\cdots,\boldsymbol{u}_n;\boldsymbol{t}) = \sum_{a_1,\dots,a_n\in\mathbb{Z}_{\geq 0}} \int_{[\overline{\mathcal{M}}_{g,n}(\mathbb{P}^1,0)]^{\mathrm{vir}}} \prod_{j=1}^{n} \mathrm{ev}_j^*((u_j)^2_{a_j}\phi_2(0))\psi_j^{a_j}$$
$$= \sum_{a_1,\dots,a_n\in\mathbb{Z}_{\geq 0}} \frac{1}{-\mathsf{w}_1}\int_{\overline{\mathcal{M}}_{g,n}} \prod_{j=1}^{n}(u_j)^2_{a_j}\psi_j^{a_j}\Lambda_g^\vee(-\mathsf{w}_1),$$

where

$$\Lambda_g^\vee(u) = u^g - \lambda_1 u^{g-1} + \cdots + (-1)^g\lambda_g$$

and $\lambda_j = c_j(\mathbb{E})$ is the j^{th} Chern class of the Hodge bundle. At the same time, we also have $\mathring{S}^2{}_2 = (\widehat{\phi}_2(0),\widehat{\phi}_2(0)) = 1$, so $(u_j)^2_k/\sqrt{-\mathsf{w}_1} = (\widetilde{u}_j)^2_k$. Therefore Theorem A specializes to

$$\omega_{g,n}\big|_{W_k^2(Y_j)/\sqrt{-2}=(u_j)^2_k/\sqrt{-\mathsf{w}_1}} = (-1)^{g-1+n}\sum_{\substack{a_1,\dots,a_n\\ a_i\in\mathbb{Z}_{\geq 0}}} \frac{1}{-\mathsf{w}_1}\int_{\overline{\mathcal{M}}_{g,n}}\prod_{j=1}^{n}(u_j)^2_{a_j}\psi_j^{a_j}\Lambda_g^\vee(-\mathsf{w}_1).$$

Now we study the expansion of $\xi_{2,0}$ near the point $Y=0$ in the coordinate $Z = e^{x/\mathsf{w}_1}$. We have

$$\xi_{2,0} = \frac{1}{\sqrt{-1}}\sqrt{\frac{2}{-\mathsf{w}_1}}\frac{-\mathsf{w}_1}{Y+\mathsf{w}_1}.$$

Since $Z = Ye^{Y/\mathsf{w}_1}$, by taking the differential we have

$$\frac{dZ}{Z} = \frac{Y+\mathsf{w}_1}{Y\mathsf{w}_1}dY.$$

Therefore,

$$\xi_{2,0} = -\frac{1}{\sqrt{-1}}\sqrt{\frac{2}{-\mathsf{w}_1}}\frac{dY}{dZ/Z}\frac{1}{Y}.$$

Let

$$\xi_{2,0} = \sum_{\mu=0}^{\infty} C_\mu Z^\mu$$

near the point $Y=0$. Then we have

$$C_\mu = \mathrm{Res}_{Y\to 0}\,\xi_{2,0}Z^{-\mu}\frac{dZ}{Z} = -\frac{1}{\sqrt{-1}}\sqrt{\frac{2}{-\mathsf{w}_1}}\,\mathrm{Res}_{Y\to 0}\,e^{-\mu Y/\mathsf{w}_1}\frac{dY}{Y^{\mu+1}}$$
$$= -\frac{1}{\sqrt{-1}}\sqrt{\frac{2}{-\mathsf{w}_1}}\frac{(-\mu/\mathsf{w}_1)^\mu}{\mu!}.$$

Therefore,

$$W_k^2 = -\frac{1}{\sqrt{-1}}\sqrt{\frac{2}{-\mathsf{w}_1}}\mathsf{w}_1 \sum_{\mu=0}^{\infty} \frac{(-\mu/\mathsf{w}_1)^{\mu}}{\mu!}\Big(-\frac{\mu}{\mathsf{w}_1}\Big)^{k+1} Z^{\mu-1} dZ.$$

On the A–model side, let

$$(u_j)_{a_j}^2 = \sum_{\mu_j=0}^{\infty} \frac{(-\mu_j/\mathsf{w}_1)^{\mu_j}}{\mu_j!}\Big(\frac{\mu_j}{\mathsf{w}_1}\Big)^{a_j} Z_j^{\mu_j}.$$

Then

$$\begin{aligned}
&F_{g,n}^{\mathbb{C},\mathbb{C}^*}(\boldsymbol{u}_1,\cdots,\boldsymbol{u}_n)\\
&= \sum_{\substack{a_1,\dots,a_n\\ a_i\in\mathbb{Z}_{\geq 0}}} \frac{1}{-\mathsf{w}_1}\int_{\overline{\mathcal{M}}_{g,n}} \prod_{j=1}^{n} \psi_j^{a_j}\Lambda_g^{\vee}(-\mathsf{w}_1)\prod_{j=1}^{n}\Big(\sum_{\mu_j=0}^{\infty}\frac{(-\mu_j/\mathsf{w}_1)^{\mu_j}}{\mu_j!}\Big(-\frac{\mu_j}{\mathsf{w}_1}\Big)^{a_j} Z_j^{\mu_j}\Big)\\
&= \sum_{\substack{a_1,\dots,a_n\\ a_i\in\mathbb{Z}_{\geq 0}}} \frac{1}{-\mathsf{w}_1}\int_{\overline{\mathcal{M}}_{g,n}} \prod_{j=1}^{n}\Big(-\frac{\mu_j\psi_j}{\mathsf{w}_1}\Big)^{a_j}\Lambda_g^{\vee}(-\mathsf{w}_1)\prod_{j=1}^{n}\Big(\sum_{\mu_j=0}^{\infty}\frac{(-\mu_j/\mathsf{w}_1)^{\mu_j}}{\mu_j!}Z_j^{\mu_j}\Big).
\end{aligned}$$

By the ELSV formula [6; 21],

$$\begin{aligned}
H_{g,\mu} &= \frac{(2g-2+|\mu|+n)!}{|\mathrm{Aut}(\mu)|}\prod_{j=1}^{n}\frac{\mu_j^{\mu_j}}{\mu_j!}\int_{\overline{\mathcal{M}}_{g,n}}\frac{\Lambda_g^{\vee}(1)}{\prod_{j=1}^{n}(1-\mu_j)}\\
&= \frac{(2g-2+|\mu|+n)!}{|\mathrm{Aut}(\mu)|}\prod_{j=1}^{n}\frac{\mu_j^{\mu_j}}{\mu_j!}\int_{\overline{\mathcal{M}}_{g,n}}\frac{\Lambda_g^{\vee}(-\mathsf{w}_1)(-\mathsf{w}_1)^{2g-3+2n}}{\prod_{j=1}^{n}(-\mathsf{w}_1-\mu_j)},
\end{aligned}$$

so

$$F^{\mathbb{C},\mathbb{C}^*} = \sum_{l(\mu)=n}\frac{|\mathrm{Aut}(\mu)|}{(2g-2+|\mu|+n)!\,(-\mathsf{w}_1)^{2g-2+|\mu|+n}}H_{g,\mu}\sum_{\sigma\in S_n}\prod_{j=1}^{n} Z_{\sigma(j)}^{\mu_j}.$$

When $\mathsf{w}_1 = -1$, this is just the generating function of the Hurwitz numbers.

Let $W_{g,n}(Z_1,\dots,Z_n)$ be the expansion of $\omega_{g,n}(Y_1,\dots,Y_n)$ in the coordinate Z near $Y=0$. Then we have:

Corollary 5.1 (Bouchard–Mariño conjecture) *For $n>0$ and $2g-2+n>0$, the invariant $W_{g,n}(Z_1,\cdots,Z_n)$ for the curve $x = Y + \mathsf{w}_1\log Y$ satisfies*

$$\begin{aligned}
&\int_0^{Z_1}\cdots\int_0^{Z_n} W_{g,n}(Z_1,\cdots,Z_n)\\
&\quad = (-1)^{g-1+n}\sum_{\substack{a_1,\ldots,a_n\\ a_i\in\mathbb{Z}_{\geq 0}}}\frac{1}{-\mathsf{w}_1}\int_{\overline{\mathcal{M}}_{g,n}}\prod_{j=1}^{n}\psi_j^{a_j}\Lambda_g^{\vee}(-\mathsf{w}_1)\\
&\qquad\qquad\cdot\prod_{j=1}^{n}\Bigl(\sum_{\mu_j=0}^{\infty}\frac{(-\mu_j/\mathsf{w}_1)^{\mu_j+a_j}}{\mu_j!}Z_j^{\mu_j}\Bigr)\\
&\quad = (-1)^{g-1+n}\sum_{l(\mu)=n}\frac{|\mathrm{Aut}(\mu)|H_{g,\mu}}{(2g-2+|\mu|+n)!(-\mathsf{w}_1)^{2g-2+|\mu|+n}}\sum_{\sigma\in S_n}\prod_{j=1}^{n}Z_{\sigma(j)}^{\mu_j}.
\end{aligned}$$

In particular, when $\mathsf{w}_1=-1$, the right-hand side is the generating function of the Hurwitz numbers and the Bouchard–Mariño conjecture is recovered.

Appendix A: Bessel functions

In this section, we give a brief review of Bessel functions.

The Bessel differential equation is

$$x^2\frac{d^2y}{dx^2}+x\frac{dy}{dx}+(x^2-\alpha^2)y=0. \tag{24}$$

The *Bessel function of the first kind* is defined by

$$J_\alpha(x)=\sum_{m=0}^{\infty}\frac{(-1)^m}{m!\,\Gamma(m+\alpha+1)}\Bigl(\frac{x}{2}\Bigr)^{2m+\alpha}.$$

The *Bessel function of the second kind* is defined by

$$Y_\alpha(x)=\frac{J_\alpha(x)\cos(\alpha\pi)-J_{-\alpha}(x)}{\sin(\alpha\pi)}.$$

When n is an integer, $Y_n(x):=\lim_{\alpha\to n}Y_\alpha(x)$.

$J_\alpha(x)$ and $Y_\alpha(x)$ form a basis of the 2–dimensional space of solutions to the Bessel differential equation (24).

Replacing x by ix in (24), one obtains the modified Bessel differential equation

$$x^2\frac{d^2y}{dx^2}+x\frac{dy}{dx}-(x^2+\alpha^2)y=0. \tag{25}$$

The *modified Bessel function of the first kind* is defined by

$$I_\alpha(x) = i^{-\alpha} J_\alpha(ix) = \sum_{m=0}^{\infty} \frac{1}{m!\,\Gamma(m+\alpha+1)} \Big(\frac{x}{2}\Big)^{2m+\alpha}.$$

The *modified Bessel function of the second kind* is defined by

$$K_\alpha(x) = \frac{\pi}{2} \frac{I_{-\alpha}(x) - I_\alpha(x)}{\sin(\alpha\pi)}.$$

The following integral formulas are valid when $\Re(x) > 0$:

$$\begin{aligned} I_\alpha(x) &= \frac{1}{\pi}\int_0^\pi e^{x\cos\theta}\cos(\alpha\theta)\,d\theta - \frac{\sin(\alpha\pi)}{\pi}\int_0^\infty e^{-x\cosh t - \alpha t}\,dt, \\ K_\alpha(x) &= \int_0^\infty e^{-x\cosh t}\cosh(\alpha t)\,dt = \frac{1}{2}\int_{t\in\gamma_{0,0}} e^{-x\cosh t-\alpha t}\,dt, \end{aligned}$$

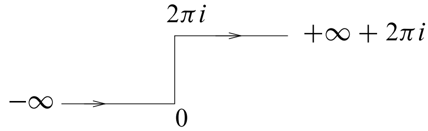
where $\gamma_{0,0}$ is the real line with the standard orientation:

$$-\infty \longrightarrow +\infty$$

We have

$$\begin{aligned} &e^{\alpha\pi i} K_\alpha(x) + i\pi I_\alpha(x) \\ &\quad = \frac{\pi}{2}\frac{e^{\alpha\pi i} I_{-\alpha}(x) - e^{-\alpha\pi i} I_\alpha(x)}{\sin(\alpha\pi)} \\ &\quad = \frac{e^{\alpha\pi i}}{2}\int_{-\infty}^0 e^{-x\cosh t-\alpha t}\,dt + \frac{e^{\alpha\pi i}}{2}\int_0^{2\pi} e^{-x\cos(i\theta)-\alpha(i\theta)}\,d(i\theta) \\ &\qquad\qquad + \frac{e^{-\alpha\pi i}}{2}\int_0^\infty e^{-x\cosh t-\alpha t}\,dt \\ &\quad = \frac{e^{\alpha\pi i}}{2}\int_{\gamma_{0,1}} e^{-x\cosh t-\alpha t}\,dt, \end{aligned}$$

where $\gamma_{0,1}$ is the following contour:

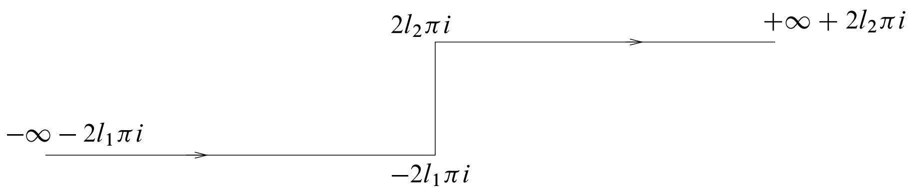


Therefore,

$$\int_{\gamma_{0,0}} e^{-x\cosh t-\alpha t}\,dt=\frac{\pi}{\sin(\alpha\pi)}(I_{-\alpha}(x)-I_{\alpha}(x)),\tag{26}$$

$$\int_{\gamma_{0,1}} e^{-x\cosh t-\alpha t}\,dt=\frac{\pi}{\sin(\alpha\pi)}(I_{-\alpha}(x)-e^{-2\alpha\pi i}I_{\alpha}(x)).\tag{27}$$

For any integers l_1 and l_2 with $l_1+l_2\ge 0$, let γ_{l_1,l_2} be the following contour:



Lemma A.1 *For any $l_1, l_2\in\mathbb{Z}$ such that $l_1+l_2\ge 0$, we have*

$$\int_{\gamma_{l_1,l_2}} e^{-x\cosh t-\alpha t}\,dt=\frac{\pi}{\sin(\alpha\pi)}(e^{2l_1\alpha\pi i}I_{-\alpha}(x)-e^{-2l_2\alpha\pi i}I_{\alpha}(x)).\tag{28}$$

Proof We observe that

$$\int_{\gamma_{l_1-k,l_2+k}} e^{-x\cosh t-\alpha t}\,dt=e^{-2k\alpha\pi i}\int_{\gamma_{l_1,l_2}} e^{-x\cosh t-\alpha t}\,dt.\tag{29}$$

In particular,

$$\begin{aligned}\int_{\gamma_{l_1,-l_1}} e^{-x\cosh t-\alpha t}\,dt&=e^{2l_1\alpha\pi i}\int_{\gamma_{0,0}} e^{-x\cosh t-\alpha t}\,dt\\&=\frac{\pi}{\sin(\alpha\pi)}\big(e^{-2l_1\alpha\pi i}I_{-\alpha}(x)-e^{2l_1\alpha\pi i}I_{\alpha}(x)\big).\end{aligned}$$

This proves (28) in the case $l_1+l_2=0$. If $l_1+l_2>0$ then

$$\gamma_{l_1,l_2}=\sum_{k=-l_1}^{l_2-1}\gamma_{1-k,k}-\sum_{k=1-l_1}^{l_2-1}\gamma_{-k,k}.\tag{30}$$

Equations (29) and (30) imply

$$\begin{aligned}&\int_{\gamma_{l_1,l_2}} e^{-x\cosh t-\alpha t}\,dt\\&=\Big(\sum_{k=-l_1}^{l_2-1}e^{-2k\alpha\pi i}\Big)\int_{\gamma_{0,1}} e^{-x\cosh t-\alpha t}\,dt-\Big(\sum_{k=1-l_1}^{l_2-1}e^{-2k\alpha\pi i}\Big)\int_{\gamma_{0,0}} e^{-x\cosh t-\alpha t}\,dt.\end{aligned}$$

Equation (28) follows from the above equation and (26)–(27). □

Appendix B: The equivariant quantum differential equation for $\mathbb{P}^1$

The equivariant quantum differential equation of $\mathbb{P}^1$ is the vector equation

$$zq\frac{d}{dq}\vec{I} = \begin{pmatrix} 0 & q - \mathsf{w}_1\mathsf{w}_2 \\ 1 & \mathsf{w}_1 + \mathsf{w}_2 \end{pmatrix}\vec{I},$$

which is equivalent to the scalar equation

$$\Big(zq\frac{d}{dq} - \mathsf{w}_1\Big)\Big(zq\frac{d}{dq} - \mathsf{w}_2\Big)I = qI. \tag{31}$$

Let

$$I = \exp\Big(\frac{\mathsf{w}_1 + \mathsf{w}_2}{2z}\log q\Big)y, \quad x = \frac{2\sqrt{q}}{z}.$$

Then (31) is equivalent to

$$x^2\frac{d^2y}{dx^2} + x\frac{dy}{dx} - \Big(x^2 + \Big(\frac{\mathsf{w}_1 - \mathsf{w}_2}{2z}\Big)^2\Big)y = 0,$$

which is the modified Bessel differential equation (25) with $\alpha = (\mathsf{w}_1 - \mathsf{w}_2)/(2z)$. When $\mathsf{w}_1 - \mathsf{w}_2 \neq 0$, any solution to (31) is of the form

$$I = \exp\Big(\frac{\mathsf{w}_1 + \mathsf{w}_2}{2z}\log q\Big)\Big(c_1 I_{\chi^1/z}\Big(\frac{2\sqrt{q}}{z}\Big) + c_2 I_{\chi^2/z}\Big(\frac{2\sqrt{q}}{z}\Big)\Big),$$

where $\chi^1 = \mathsf{w}_1 - \mathsf{w}_2 = -\chi^2$, and c_1 and c_2 are functions of w_1, w_2 and z.

References

[1] **G Borot**, **B Eynard**, **M Mulase**, **B Safnuk**, *A matrix model for simple Hurwitz numbers, and topological recursion*, J. Geom. Phys. 61 (2011) 522–540 MR

[2] **V Bouchard**, **M Mariño**, *Hurwitz numbers, matrix models and enumerative geometry*, from "From Hodge theory to integrability and TQFT $tt*$–geometry" (R Y Donagi, K Wendland, editors), Proc. Sympos. Pure Math. 78, Amer. Math. Soc., Providence, RI (2008) 263–283 MR

[3] **B Dubrovin**, *Geometry of 2D topological field theories*, from "Integrable systems and quantum groups" (M Francaviglia, S Greco, editors), Lecture Notes in Math. 1620, Springer, Berlin (1996) 120–348 MR

[4] **P Dunin-Barkowski**, **M Kazarian**, **N Orantin**, **S Shadrin**, **L Spitz**, *Polynomiality of Hurwitz numbers, Bouchard–Mariño conjecture, and a new proof of the ELSV formula*, Adv. Math. 279 (2015) 67–103 MR

[5] **P Dunin-Barkowski**, **N Orantin**, **S Shadrin**, **L Spitz**, *Identification of the Givental formula with the spectral curve topological recursion procedure*, Comm. Math. Phys. 328 (2014) 669–700 MR

[6] **T Ekedahl**, **S Lando**, **M Shapiro**, **A Vainshtein**, *Hurwitz numbers and intersections on moduli spaces of curves*, Invent. Math. 146 (2001) 297–327 MR

[7] **B Eynard**, *Intersection numbers of spectral curves*, preprint (2011) arXiv

[8] **B Eynard**, *Invariants of spectral curves and intersection theory of moduli spaces of complex curves*, Commun. Number Theory Phys. 8 (2014) 541–588 MR

[9] **B Eynard**, **M Mulase**, **B Safnuk**, *The Laplace transform of the cut-and-join equation and the Bouchard–Mariño conjecture on Hurwitz numbers*, Publ. Res. Inst. Math. Sci. 47 (2011) 629–670 MR

[10] **B Eynard**, **N Orantin**, *Invariants of algebraic curves and topological expansion*, Commun. Number Theory Phys. 1 (2007) 347–452 MR

[11] **B Eynard**, **N Orantin**, *Computation of open Gromov–Witten invariants for toric Calabi–Yau* 3*–folds by topological recursion, a proof of the BKMP conjecture*, Comm. Math. Phys. 337 (2015) 483–567 MR

[12] **B Fang**, *Homological mirror symmetry is T–duality for $\mathbb{P}^n$*, Commun. Number Theory Phys. 2 (2008) 719–742 MR

[13] **B Fang**, *Central charges of T–dual branes for toric varieties*, preprint (2016) arXiv

[14] **B Fang**, **C-C M Liu**, **D Treumann**, **E Zaslow**, *T–duality and homological mirror symmetry for toric varieties*, Adv. Math. 229 (2012) 1875–1911 MR

[15] **B Fang**, **C-C M Liu**, **Z Zong**, *All genus open-closed mirror symmetry for affine toric Calabi–Yau* 3*–orbifolds*, preprint (2013) arXiv

[16] **J D Fay**, *Theta functions on Riemann surfaces*, Lecture Notes in Mathematics 352, Springer, Berlin (1973) MR

[17] **A B Givental**, *Equivariant Gromov–Witten invariants*, Internat. Math. Res. Notices 1996 (1996) 613–663 MR

[18] **A Givental**, *Elliptic Gromov–Witten invariants and the generalized mirror conjecture*, from "Integrable systems and algebraic geometry" (M-H Saito, Y Shimizu, K Ueno, editors), World Sci., River Edge, NJ (1998) 107–155 MR

[19] **A B Givental**, *Gromov–Witten invariants and quantization of quadratic Hamiltonians*, Mosc. Math. J. 1 (2001) 551–568 MR

[20] **A B Givental**, *Semisimple Frobenius structures at higher genus*, Internat. Math. Res. Notices 2001 (2001) 1265–1286 MR

[21] **T Graber**, **R Vakil**, *Hodge integrals and Hurwitz numbers via virtual localization*, Compositio Math. 135 (2003) 25–36 MR

[22] **H Iritani**, *An integral structure in quantum cohomology and mirror symmetry for toric orbifolds*, Adv. Math. 222 (2009) 1016–1079 MR

[23] **I Kostov**, **N Orantin**, *CFT and topological recursion*, J. High Energy Phys. (2010) art. id. 056 MR

[24] **Y-P Lee**, **R Pandharipande**, *Frobenius manifolds, Gromov–Witten theory, and Virasoro constraints*, preprint (2004) Available at `https://people.math.ethz.ch/~rahul/Part1.ps`

[25] **B H Lian**, **K Liu**, **S-T Yau**, *Mirror principle, I*, Asian J. Math. 1 (1997) 729–763 MR

[26] **P Norbury**, **N Scott**, *Gromov–Witten invariants of $\mathbb{P}^1$ and Eynard–Orantin invariants*, Geom. Topol. 18 (2014) 1865–1910 MR

[27] **A Okounkov**, **R Pandharipande**, *Gromov–Witten theory, Hurwitz theory, and completed cycles*, Ann. of Math. 163 (2006) 517–560 MR

[28] **A Okounkov**, **R Pandharipande**, *The equivariant Gromov–Witten theory of* $\mathbf{P}^1$, Ann. of Math. 163 (2006) 561–605 MR

[29] **J S Song**, **Y S Song**, *On a conjecture of Givental*, J. Math. Phys. 45 (2004) 4539–4550 MR

Beijing International Center for Mathematical Research, Peking University
5 Yiheyuan Road, Jingchunyuan 78, Beijing, 100871, China

Department of Mathematics, Columbia University
Room 623, Mail Code 4435, 2990 Broadway, New York, NY 10027, United States

Yau Mathematical Sciences Center, Tsinghua University
Jing Chun Yuan West Building, Tsinghua University, Haidian District, Beijing, 100084, China

bohanfang@gmail.com, ccliu@math.columbia.edu,
zyzong@mail.tsinghua.edu.cn

Proposed: Lothar Göttsche Received: 6 December 2014
Seconded: Richard Thomas, Dan Abramovich Revised: 24 May 2016

Geometry & Topology 21 (2017) 2093–2160

The simplicial suspension sequence in $\mathbb{A}^1$–homotopy

ARAVIND ASOK
KIRSTEN WICKELGREN
BEN WILLIAMS

We study a version of the James model for the loop space of a suspension in unstable $\mathbb{A}^1$–homotopy theory. We use this model to establish an analog of G W Whitehead's classical refinement of the Freudenthal suspension theorem in $\mathbb{A}^1$–homotopy theory: our result refines F Morel's $\mathbb{A}^1$–simplicial suspension theorem. We then describe some E_1–differentials in the EHP sequence in $\mathbb{A}^1$–homotopy theory. These results are analogous to classical results of G W Whitehead. Using these tools, we deduce some new results about unstable $\mathbb{A}^1$–homotopy sheaves of motivic spheres, including the counterpart of a classical rational nonvanishing result.

14F42, 19E15; 55Q15, 55Q20, 55Q25

1 Introduction

If K is an $(n-1)$–connected pointed CW complex, then the suspension map

$$\mathrm{E}\colon \pi_q(K) \to \pi_{q+1}(\Sigma K)$$

fits into a long exact sequence of the form

$$\pi_{3n-2}(K) \xrightarrow{\mathrm{E}} \pi_{3n-1}(\Sigma K) \xrightarrow{\mathrm{H}} \pi_{3n-1}(\Sigma K^{\wedge 2}) \xrightarrow{\mathrm{P}} \pi_{3n-3}(K) \xrightarrow{\mathrm{E}} \cdots$$
$$\cdots \longrightarrow \pi_q(K) \xrightarrow{\mathrm{E}} \pi_{q+1}(\Sigma K) \xrightarrow{\mathrm{H}} \pi_{q+1}(\Sigma K^{\wedge 2}) \xrightarrow{\mathrm{P}} \pi_{q-1}(K) \longrightarrow \cdots .$$

Together with an elementary connectivity estimate for $\Sigma K^{\wedge 2}$, this exact sequence may be viewed as a refinement of the Freudenthal suspension theorem. The exact sequence above was first constructed by G W Whitehead [58, Theorem 1, page 211] if $K = S^n$ and by W D Barcus [9, Proposition 2.9] for K as above (see also Whitehead [59, Theorem XII.2.2, page 543] for a textbook treatment of the general statement).

The morphisms H and P appearing in the above exact sequence were also studied by Whitehead [57, Section 10] in great detail in the case where $K = S^n$. The morphism H is the Hopf invariant, and Whitehead linked the morphism P with Whitehead products. In more detail, begin by observing that the $(n+1)$–fold suspension $\mathrm{E}^{n+1}\colon \pi_{q-n}(S^n) \to \pi_{q+1}(S^{2n+1})$ is an isomorphism for $q < 3n-1$. Define $\mathrm{P}'\colon \pi_{q-n}(S^n) \to \pi_{q-1}(S^n)$

Published: 19 May 2017 DOI: 10.2140/gt.2017.21.2093

by $\mathrm{P}'(\alpha) = [\alpha, \iota_n]$, where ι_n is the identity map on the n–sphere and the bracket denotes the Whitehead product. For $\mathrm{P}\colon \pi_{q+1}(S^{2n+1}) \to \pi_{q-1}(S^n)$, Whitehead observed that

$$\mathrm{P} = \mathrm{P}' \circ (\mathrm{E}^{n+1})^{-1} \quad \text{if} \quad q < 3n-1.$$

While Whitehead established this result for spheres, it has been known for some time that the morphism P is, for general $(n-1)$–connected spaces, still closely related to Whitehead products; see eg I M James [30, Section 2] or Ganea [20, Theorem 3.1 and page 231] for a very general statement. In any case, these kinds of tools were used to great effect in early computations of unstable homotopy groups of spheres, eg, by James [29; 31] and Toda [52].

The goal of this paper, whose title pays homage to the work of James [31], is to establish analogs of the above results in the Morel–Voevodsky unstable $\mathbb{A}^1$–homotopy category [45] and to deduce some consequences of these results. The jumping-off point is to give a James-style model for the loop space of a suspension in $\mathbb{A}^1$–homotopy theory (see Theorem 2.4.2). Using this model, we deduce the following result, which can be thought of as a refinement of the $\mathbb{A}^1$–simplicial suspension theorem of F Morel [44, Theorem 6.61].

Theorem (see Theorem 3.2.1, Remark 3.2.3 and Theorem 4.2.1) *Assume k is a perfect field. If $\mathscr{X}$ is a pointed $\mathbb{A}^1$–$(n-1)$–connected simplicial presheaf on $(\mathrm{Sm}_k)_{\mathrm{Nis}}$, with $n \geq 2$, then there is an exact sequence of $\mathbb{A}^1$–homotopy sheaves of the form*

$$\boldsymbol{\pi}^{\mathbb{A}^1}_{3n-2}(\mathscr{X}) \xrightarrow{\mathrm{E}} \boldsymbol{\pi}^{\mathbb{A}^1}_{3n-1}(\Sigma\mathscr{X}) \xrightarrow{\mathrm{H}} \boldsymbol{\pi}^{\mathbb{A}^1}_{3n-1}(\Sigma\mathscr{X}^{\wedge 2}) \xrightarrow{\mathrm{P}} \boldsymbol{\pi}^{\mathbb{A}^1}_{3n-3}(\mathscr{X}) \xrightarrow{\mathrm{E}} \cdots$$
$$\cdots \longrightarrow \boldsymbol{\pi}^{\mathbb{A}^1}_{q}(\mathscr{X}) \xrightarrow{\mathrm{E}} \boldsymbol{\pi}^{\mathbb{A}^1}_{q+1}(\Sigma\mathscr{X}) \xrightarrow{\mathrm{H}} \boldsymbol{\pi}^{\mathbb{A}^1}_{q+1}(\Sigma\mathscr{X}^{\wedge 2}) \xrightarrow{\mathrm{P}} \boldsymbol{\pi}^{\mathbb{A}^1}_{q-1}(\mathscr{X}) \longrightarrow \cdots,$$

where the map E is (simplicial) suspension, the map H is a James–Hopf invariant, and the map P is described, as above, in terms of Whitehead products.

We go on to discuss various consequences of the existence of this exact sequence. We analyze the low-degree portion of this sequence in Theorem 3.3.13 and give a more explicit description of the sequence in the first degree where the suspension map fails to be an isomorphism. When $\mathscr{X}$ is a motivic sphere, it is shown in Wickelgren and Williams [61] that the exact sequences displayed above can be extended to all degrees after localizing at 2. By suitably varying the input sphere, these sequences can be strung together to obtain the EHP spectral sequence converging to the 2–local S^1–stable $\mathbb{A}^1$–homotopy sheaves of spheres.

By construction, the E_1–differentials in this spectral sequence arise from the composite map HP, which in certain degrees we can analyze integrally. To state the result, recall that Morel [44, Corollary 6.43] computed the first nonvanishing $\mathbb{A}^1$–homotopy sheaf of a motivic sphere in terms of Milnor–Witt K-theory. Morel [44, Lemma 3.10] also showed that there is an isomorphism of rings $K_0^{\mathrm{MW}}(k) \cong \mathrm{GW}(k)$, ie the zeroth Milnor–Witt K-theory group of a field k is isomorphic to the Grothendieck–Witt ring of isomorphism classes of symmetric bilinear forms over k, defined to be the group completion of the monoid of isomorphism classes of nondegenerate symmetric bilinear forms. Given this terminology, the class of the composite HP can be seen to correspond with a symmetric bilinear form, which we can describe. More precisely, we establish the following result (see the statement in the body of the text and Remark 4.4.3 for a more conceptual explanation of the formula).

Theorem (see Theorem 4.4.1) *Assume k is a perfect field, and let p and q be integers with $p > 1$ and $q \geq 1$. The map*

$$\mathrm{HP}\colon \boldsymbol{K}_{2q}^{\mathrm{MW}} = \boldsymbol{\pi}_{2p+3}^{\mathbb{A}^1}(\Sigma(S^{p+1+q\alpha})^{\wedge 2}) \longrightarrow \boldsymbol{\pi}_{2p+1}(\Sigma(S^{p+q\alpha})^{\wedge 2}) = \boldsymbol{K}_{2q}^{\mathrm{MW}}$$

corresponds to the element $\langle 1\rangle + (-1)^{p+1+q}\langle -1\rangle^q \in \mathrm{GW}(k)$.

One consequence of this result is the following analog of the classical fact, due to Hopf, that $\pi_{4n-1}(S^{2n})$ is nontrivial.

Theorem (see Theorem 5.3.1) *Fix a base field k assumed to be perfect and to have characteristic unequal to 2. Let $n, q \geq 2$ be even integers, and let j be an integer. There is a surjection*

$$\boldsymbol{\pi}_{2n-1+j\alpha}^{\mathbb{A}^1} S^{n+q\alpha} \otimes \mathbb{Q} \xrightarrow{\mathrm{H}\otimes\mathbb{Q}} \boldsymbol{K}_{2q-j}^{\mathrm{MW}} \otimes \mathbb{Q},$$

and the sheaf $\boldsymbol{\pi}_{2n-1+j\alpha}^{\mathbb{A}^1} S^{n+q\alpha} \otimes \mathbb{Q}$ is nontrivial if either k is formally real or if $j \leq 2n-1$.

Relying on the computations of Asok and Fasel [5], we analyze the low-degree portion of the EHP sequence in great detail in the special case where $X = \mathbb{A}^3 \setminus 0$. In particular, we give a description of the next nonvanishing $\mathbb{A}^1$–homotopy sheaf (ie beyond that computed by Morel) of $\Sigma\mathbb{A}^3 \setminus 0 \cong \mathbb{P}^{1\wedge 3}$ in Theorem 5.2.5. The following statement is an easy-to-state special case of a more general result.

Theorem (see Theorem 5.2.5) *If k is a field of characteristic 0 containing an algebraically closed subfield, then, for any integer $i \geq 0$, there is an isomorphism of sheaves of the form*

$$\boldsymbol{\pi}_{4+i+5\alpha}^{\mathbb{A}^1}(S_s^{3+i} \wedge \mathbb{G}_m^{\wedge 3}) \cong \mathbb{Z}/24.$$

Remark Cohomology of homotopy sheaves of spheres such as those above appears in concrete applications to problems in algebra via techniques of obstruction theory; see eg Asok and Fasel [4; 5] for more details. Our description of the sheaf $\boldsymbol{\pi}_4^{\mathbb{A}^1}(S^{3+3\alpha})$ is wellsuited to such cohomology computations. Our computation allows us to state a precise conjecture (see Conjecture 5.2.10) regarding the structure of the sheaf $\boldsymbol{\pi}_n^{\mathbb{A}^1}(\mathbb{A}^n \setminus 0)$ for $n \geq 4$. An explicit description of the sheaf $\boldsymbol{\pi}_n^{\mathbb{A}^1}(\mathbb{A}^n \setminus 0)$ for $n = 2, 3$ was a key step in [4; 5] in the resolution of Murthy's conjecture regarding splitting of rank-n vector bundles on smooth affine $(n+1)$–folds over algebraically closed fields. A resolution of Conjecture 5.2.10 would, similarly, imply Murthy's conjecture in general.

We close this introduction with some general comments regarding prerequisites. When working with the (unstable) $\mathbb{A}^1$–homotopy category in general and Morel's $\mathbb{A}^1$–algebraic topology in particular, with the goal of making this paper as self-contained as possible, we have labored to present the material in an axiomatic framework involving the "unstable $\mathbb{A}^1$–connectivity property", which is introduced in Section 2.2. All of the results in Sections 2 and 3 are written from this axiomatic perspective. We hope this style of presentation makes the material accessible to people who have some familiarity with homotopy theory of simplicial presheaves and the constructions of Morel and Voevodsky [45], but not, for example, all of the technical results about strongly and strictly $\mathbb{A}^1$–invariant sheaves contained in the first five chapters of Morel [44]. Moreover, we hope that our presentation also makes [44] itself more accessible to the nonexpert.

For the most part, Section 4 is written in the same axiomatic framework. In contrast, Sections 4.4 and 5 require more background. In particular, this portion of the text requires familiarity with facts about strongly and strictly $\mathbb{A}^1$–invariant sheaves (see Section 5.1 for more precise statements), and known explicit computations of homotopy sheaves. In Section 5, we also appeal to structural results from the theory of quadratic forms and both the affirmation of the Milnor conjecture on quadratic forms and the Bloch–Kato conjecture.

Notation Throughout, the (undecorated) symbol S will be used to denote a base scheme assumed Noetherian and of finite Krull dimension. We write Sm_S for the category of schemes that are separated, smooth and of finite type over S. Script letters, eg, $\mathscr{X}$, $\mathscr{Y}$, will typically be used to denote "spaces", ie pointed simplicial presheaves on Sm_S (from Section 2.2 onward), while capital roman letters will typically be used to denote simplicial presheaves on more general sites. Typically, boldface letters will be used to denote strongly $\mathbb{A}^1$–invariant sheaves of groups (again, from Section 2.2 onward), with the exception of $\mathbf{C}$, which will always mean a category (often equipped with the structure of a site) and $\mathbf{R}$, which will be used to denote right derived functors.

Sheaf cohomology will always be taken with respect to the Nisnevich topology. See 2.2.3 for our conventions regarding motivic spheres; unfortunately the letter S appears in our notation for spheres, but since it will always be decorated with a superscript, we hope no confusion arises. See 2.2.4 for a summary of notation pertaining to homotopy sheaves, 2.2.5 for some discussion of our connectivity conventions, 2.2.9 for some recollections on our use of the term fiber sequence, 2.3.5 for conventions about relative connectivity, 3.3.2 for notation regarding $\mathbb{A}^1$–homology sheaves, and 3.3.6 for notation regarding the $\mathbb{A}^1$–tensor product. Finally, our conventions for loop spaces change in Sections 5.2 and 5.3; see 5.2.1 for more details.

Acknowledgements Asok would like to thank Frédéric Déglise for his hospitality during a stay at ENS Lyon in September 2012 when parts of this work were conceived and Fabien Morel for pointing out many years ago that a James-style model for (simplicial) loops on the suspension exists in $\mathbb{A}^1$–homotopy theory. This project was initially conceived as joint with Jean Fasel, and the authors would also like to sincerely thank him for his collaboration in the formative stages. Some of this work was done while Wickelgren was in residence at MSRI during the Spring 2014 semester supported by NSF grant 0932078 000. The authors are grateful to have been afforded the opportunity to work together in several places including the University of Duisburg–Essen during a special semester on motivic homotopy theory organized by Marc Levine (funded by the Alexander von Humboldt Foundation and the German Research Foundation) and the American Institute of Mathematics. Finally, we thank the referees for a very careful reading and a number of thoughtful suggestions that improved the clarity of the presentation.

Asok was partially supported by National Science Foundation Award DMS-1254892. Wickelgren was partially supported by National Science Foundation Award DMS-1406380.

2 The James construction revisited

The James construction on a CW complex was originally introduced by James in [29]. Milnor observed that the construction could be recast in the language of simplicial sets [39, page 120]. Using this translation, it is straightforward to develop a version of the James construction in the category of simplicial presheaves. Section 2.1 reviews the James construction in the category of simplicial sets and extends these constructions to simplicial presheaves; the main result in the context of simplicial presheaves is Proposition 2.1.6.

In this section, we aim to develop a version of the James model in the $\mathbb{A}^1$–homotopy category; this idea is due originally to Morel. Section 2.2 recalls a number of structural properties of the $\mathbb{A}^1$–homotopy category that will be used throughout the work: we point the reader to Definition 2.2.6, Lemma 2.2.11 and Theorem 2.2.12. Section 2.3 studies some aspects of $\mathbb{A}^1$–fiber sequences in the context of the axiomatic setup of

Section 2.2. Section 2.4 proves the main result, ie Theorem 2.4.2, which provides a James-style model for loops on the suspension of a space in $\mathbb{A}^1$–homotopy theory; this result depends on results about the Kan loop group in $\mathbb{A}^1$–homotopy theory, for which we refer the reader to Theorem 2.3.2.

2.1 The James construction in simplicial homotopy theory

Textbook treatments of the James construction can be found in [59, Chapter VII.2], for the category of CW complexes, and [62, Section 3.3.3] in the category of simplicial sets.

The James construction for simplicial sets Let K be a pointed simplicial set. An injection $\alpha\colon (1,2,\dots,m) \to (1,2,\dots,n)$ induces a map $\alpha_*\colon K^m \to K^n$. Let $\sim$ denote the equivalence relation on $\coprod_{n=0}^{\infty} K^n$ generated by $x \sim \alpha_*(x)$ for all *order-preserving* injections α. The James construction on K is defined by the formula

$$J(K) := \coprod_{n=0}^{\infty} K^n/{\sim},$$

ie $J(K)$ is the free (pointed) monoid on the pointed simplicial set K. The assignment $K \mapsto J(K)$ is functorial in K by definition. The James construction is filtered by pointed simplicial sets $J_n(X) \subset J(K)$, defined by

$$J_n(K) := \coprod_{m=0}^{n} K^m/{\sim}.$$

We consider also $F(K)$, the pointed free group functor as in [23, page 293] or [62, Section 3.2]. Because $J(K)$ is the free pointed monoid on K, there is an evident inclusion map $J(K) \hookrightarrow F(K)$. If ΣK denotes the Kan suspension [23, page 191], and $G(K)$ denotes the Kan loop group [23, page 276], then Milnor showed that there is a weak equivalence $F(K) \simeq G(\Sigma K)$ [23, Theorem V.6.15]. By [23, Corollary V.5.11], since ΣK is reduced, we conclude that $F(K)$ is a model for $\Omega\Sigma K$; here Ω is the derived loops (for a model, take naive loops on a fibrant model of the input). The following result details the main properties of $J(-)$.

Theorem 2.1.1 [62, Theorems 3.24 and 3.25] *Suppose K is a pointed simplicial set.*

(1) *If K is connected, there is a weak equivalence $J(K) \simeq F(K)$.*

(2) *For any integer $n \geq 1$ there is a cofiber sequence of the form*

$$J_{n-1}(K) \hookrightarrow J_n(K) \longrightarrow K^{\wedge n}.$$

(3) *The canonical map $\operatorname{colim}_n J_n(K) \to J(K)$ is an isomorphism.*

Remark 2.1.2 Consider the map $K = J_1(K) \to J(K)$. Under the weak equivalence $J(K) \simeq \Omega\Sigma K$ of point (1) of Theorem 2.1.1, this map corresponds to the unit map $K \to \Omega\Sigma K$ of the loops-suspension adjunction.

The James construction for simplicial presheaves Suppose $\mathbf{C}$ is a site. We will consider (pointed) simplicial presheaves on $\mathbf{C}$, though we do not introduce any special notation for this category. The category of (pointed) simplicial presheaves can be equipped with its injective local model structure [32]: cofibrations are given by monomorphisms, weak equivalences are defined locally with respect to the Grothendieck topology on $\mathbf{C}$ and fibrations are defined via the right-lifting property. Abusing terminology slightly, we will refer to the associated homotopy category as the (pointed) simplicial homotopy category. The category of pointed simplicial presheaves is a pointed category, ie the canonical map from the initial to the final object is an isomorphism. We will typically write $*$ for the final object. With this terminology, one can extend the definition of the James construction to (pointed) simplicial presheaves in a straightforward fashion by applying the constructions above sectionwise.

Definition 2.1.3 Assume X is a pointed simplicial presheaf on $\mathbf{C}$, and an $n \geq 0$ is an integer. Define pointed simplicial presheaves $G(X)$, $F(X)$, $J(X)$ and $J_n(X)$ by assigning to $U \in \mathbf{C}$ the following simplicial sets:

$$\begin{aligned} G(X)(U) &:= G(X(U)), \\ F(X)(U) &:= F(X(U)), \\ J(X)(U) &:= J(X(U)), \\ J_n(X)(U) &:= J_n(X(U)). \end{aligned}$$

We refer to the pointed simplicial presheaf $J(X)$ as the *James construction of* X, and $G(X)$ as the Kan loop group of X.

Remark 2.1.4 Since we have not assumed X to be reduced (ie having presheaf of 0–simplices the constant presheaf $*$) in the above, the simplicial presheaf $G(X)$ will not in general have the homotopy type of the loop space of X (eg, take $X = \Delta_1$ the simplicial interval, in which case $G(\Delta^1)$ is the constant simplicial group on $\mathbb{Z}$, which is not contractible).

The assignments $X \mapsto G(X)$, $X \mapsto F(X)$, $X \mapsto J(X)$, $X \mapsto J_n(X)$ are all evidently functorial in X. Moreover, there are morphisms $J_n(X) \to J(X)$ for any integer $n \geq 0$. We distinguish the morphism

$$\text{E}\colon X = J_1(X) \longrightarrow J(X); \tag{2.1.1}$$

we will refer to this morphism as *suspension* (see Remark 2.1.2).

Proposition 2.1.5 *Suppose X is a reduced pointed simplicial presheaf on $\mathbf{C}$, ie the presheaf of 0–simplices is $*$. There is a weak equivalence*

$$G(X) \simeq \Omega X.$$

Proof This follows by observing that the induced map on sections over $U \in \mathbf{C}$ is a weak equivalence by [23, Corollary V.5.11]. □

Proposition 2.1.6 *Suppose X is a pointed simplicial presheaf on $\mathbf{C}$.*

(1) *The map $J(X) \to F(X)$ is a sectionwise equivalence.*

(2) *The simplicial presheaf $J(X)$ is locally weakly equivalent to $\Omega\Sigma X$.*

Both of the weak equivalences just mentioned are functorial in X.

Proof Point (1) follows immediately from Theorem 2.1.1(1): the inclusion maps $J(X(U)) \to F(X(U))$ are weak equivalences; these maps are evidently functorial in U. For point (2), observe that [23, Theorem V.6.15] guarantees that for any object $U \in \mathbf{C}$ there is a functorial weak equivalence

$$G(\Sigma X(U)) \simeq F(X(U)),$$

where Σ is the Kan suspension. Since $\Sigma X(U)$ is by construction reduced, the result then follows from Proposition 2.1.5. □

The classifying space of the Kan loop group Suppose H is a simplicial presheaf of groups. If Y is a simplicial presheaf equipped with a right action $a\colon Y \times H \to Y$ of H, we will say that the action is *categorically free* if the morphism

$$Y \times H \xrightarrow{(a, p_Y)} Y \times Y$$

is a monomorphism. If Y carries a categorically free action of H, we write Y/H for the quotient, ie the colimit of the diagram $Y \leftarrow Y \times H \to Y$.

2.1.7 Following [37, Chapter 7], if H is a presheaf of groups, Y is a space with a right H–action and Z is a space with a left H–action, we may form the two-sided bar construction $B(Y, H, Z)$ as the "geometric realization" of a certain functorially constructed simplicial object $B(Y, H, Z)_\bullet$ having n–simplices of the form $Y \times H^{\times n} \times Z$. In the present context, $B(Y, H, Z)_\bullet$ is a simplicial object in the category of simplicial presheaves and the "geometric realization" is the homotopy colimit over Δ^{op}, ie

$$B(Y, H, X) := \operatorname*{hocolim}_{n \in \Delta^{\mathrm{op}}} B(Y, H, Z)_n.$$

When $Y = X = *$, then we use the following special terminology: by the *simplicial classifying space* BH we mean $B(*, H, *)$, by the *universal bundle* EH we mean $B(*, H, H)$, and by the *Borel construction* we mean $B(Y, H, *)$.

The next result, which is a (pre)sheaf-theoretic variant of a classical fact about the Kan loop group (see eg [14, point (5) on page 137]), follows immediately from Proposition 2.1.5; we use this result in the next section.

Proposition 2.1.8 *For X any reduced pointed simplicial presheaf on $\mathbf{C}$, there is a sectionwise weak equivalence $X \simeq BG(X)$.*

2.2 The unstable $\mathbb{A}^1$–connectivity property

Before discussing the James construction in $\mathbb{A}^1$–homotopy theory, we will recall some facts from $\mathbb{A}^1$–algebraic topology. We take $\mathbf{C} = \mathrm{Sm}_S$, ie the category of smooth schemes over S.

This category will be endowed throughout with the Nisnevich topology, as in [45, Section 3], and the category of simplicial presheaves on Sm_S may be equipped with a *simplicial* model structure, [32], local with respect to this topology. That is, the cofibrations are the monomorphisms of simplicial presheaves, and the weak equivalences may be detected on Nisnevich stalks. We warn the reader that in [44] and [45], contrary to our conventions, the motivic homotopy category is constructed using simplicial sheaves. In [33, Theorem 1.2], Jardine shows that the sheafification and the forgetful functor define an adjoint equivalent between the two theories.

By a *pointed space*, we will mean pointed simplicial presheaf on Sm_S. A model structure for the $\mathbb{A}^1$–homotopy category $\mathscr{H}(S)$ can be constructed by left Bousfield localization of the simplicial model structure of simplicial presheaves on Sm_S.

We will adopt the convention, at variance with that of [25], that homotopy limits will be calculated by first applying a functorial fibrant replacement objectwise to the diagram in question.

$\mathbb{A}^1$–localization Recall from [45, Section 3.2] the notion of an $\mathbb{A}^1$–local object. It will be useful to remember that the simplicial homotopy limit of a diagram of $\mathbb{A}^1$–local objects is again $\mathbb{A}^1$–local, this is the case because the fibrant, $\mathbb{A}^1$–local objects are the fibrant objects of a model category, and we may use [25, Theorem 18.5.2]. We begin by recalling the basic properties of "the" $\mathbb{A}^1$–localization functor.

Proposition 2.2.1 *There exists an endofunctor $\mathrm{L}_{\mathbb{A}^1}$ of the category of simplicial presheaves on $(\mathrm{Sm}_S)_{\mathrm{Nis}}$ and a natural transformation $\theta\colon \mathrm{id} \to \mathrm{L}_{\mathbb{A}^1}$ such that, for any space $\mathscr{X}$, the following statements hold:*

(i) *The space* $\mathrm{L}_{\mathbb{A}^1}\mathscr{X}$ *is fibrant and* $\mathbb{A}^1$*–local, and the map* $\mathscr{X} \to \mathrm{L}_{\mathbb{A}^1}\mathscr{X}$ *is an* $\mathbb{A}^1$*–weak equivalence.*

(ii) *If* $\mathscr{Y}$ *is any simplicially fibrant* $\mathbb{A}^1$*–local space, and* $f\colon \mathscr{X} \to \mathscr{Y}$ *is a morphism, then* f *factors as* $\mathscr{X} \to \mathrm{L}_{\mathbb{A}^1}\mathscr{X} \to \mathscr{Y}$*.*

(iii) *The functor* $\mathrm{L}_{\mathbb{A}^1}$ *commutes with the formation of finite limits.*

Comments on the proof Most of this statement is contained in [45]; we slightly modify the $\mathbb{A}^1$–localization functor given in [45, page 107]. The functor is constructed by repeated application of the singular construction and a fibrant replacement functor in the category of simplicial presheaves. The singular construction commutes with limits (see [45, page 87]). We use the Godement resolution functor of [45, Section 2, Theorem 1.66] as our functorial fibrant replacement for simplicial presheaves; this commutes with formation of finite limits by construction. □

In Morel's analysis, a distinguished role is played by Eilenberg–MacLane spaces [45, page 56] or classifying spaces of Nisnevich sheaves of groups [45, page 128] that are $\mathbb{A}^1$–local.

Definition 2.2.2 A sheaf of groups $\boldsymbol{G}$ is called *strongly* $\mathbb{A}^1$*–invariant* if $B\boldsymbol{G}$ is $\mathbb{A}^1$–local. A sheaf of abelian groups $\boldsymbol{A}$ is called strictly $\mathbb{A}^1$–invariant if $K(\boldsymbol{A}, i)$ is $\mathbb{A}^1$–local for every $i \geq 0$.

Homotopy sheaves and the unstable $\mathbb{A}^1$–connectivity property Suppose $\mathscr{X}$ and $\mathscr{Y}$ are pointed spaces. We write $[\mathscr{Y}, \mathscr{X}]_s$ for morphisms in the homotopy category of the injective local model structure (we will refer to this category as the *simplicial homotopy category*) and $[\mathscr{Y}, \mathscr{X}]_{\mathbb{A}^1}$ for morphisms in the $\mathbb{A}^1$–homotopy category. We now fix some conventions that will remain in force throughout the paper.

Notation 2.2.3 (spheres, suspension and looping) Write S^i_s to denote the simplicial i–sphere, and write $\mathbb{G}_m^{\wedge j}$ for the j–fold smash product of $\mathbb{G}_m$ (pointed by 1) with itself. Following conventions of $\mathbb{Z}/2$–equivariant homotopy theory, we write $S^{i+j\alpha}$ for the sphere $S^i_s \wedge \mathbb{G}_m^{\wedge j}$. If $j = 0$, the $\mathbb{G}_m$–term shall be dropped from the notation. The undecorated symbol Σ will be used for simplicial suspension. Likewise, we use Ω for the derived simplicial loops, a model for which is obtained by first taking a functorial fibrant replacement of the input and then applying naive loops.

Notation 2.2.4 (homotopy sheaves) We define homotopy sheaves $\boldsymbol{\pi}_i(\mathscr{X}, x)$ and $\mathbb{A}^1$–homotopy sheaves $\boldsymbol{\pi}_i^{\mathbb{A}^1}(\mathscr{X}, x)$ as Nisnevich sheaves associated with the presheaves

$$U \mapsto [S^i_s \wedge U_+, (\mathscr{X}, x)]_s \quad \text{and} \quad U \mapsto [S^i_s \wedge U_+, (\mathscr{X}, x)]_{\mathbb{A}^1} = [S^i_s \wedge U_+, (\mathrm{L}_{\mathbb{A}^1}\mathscr{X}, x)]_s.$$

Likewise, $\boldsymbol{\pi}^{\mathbb{A}^1}_{i+j\alpha}(\mathscr{X},x)$ is the sheafification for the Nisnevich topology of the presheaf

$$U \mapsto [S^{i+j\alpha} \wedge U_+, (\mathscr{X},x)]_{\mathbb{A}^1}.$$

For notational compactness, basepoints will typically be suppressed from notation.

Convention 2.2.5 (connectivity) We borrow various bits of terminology from classical homotopy theory: A pointed space $(\mathscr{X},x)$ is simplicially connected (resp. $\mathbb{A}^1$–connected) if the sheaf $\boldsymbol{\pi}_0(\mathscr{X})$ (resp. $\boldsymbol{\pi}^{\mathbb{A}^1}_0(\mathscr{X})$) is $*$. Similarly, if $n \geq 1$ is an integer, we will say that $(\mathscr{X},x)$ is simplicially n–connected (resp. $\mathbb{A}^1$–n–connected) if $(\mathscr{X},x)$ is simplicially (resp. $\mathbb{A}^1$–)connected and $\boldsymbol{\pi}_i(\mathscr{X},x)$ (resp. $\boldsymbol{\pi}^{\mathbb{A}^1}_i(\mathscr{X})$) is trivial for $i \leq n$.

Morel's approach to $\mathbb{A}^1$–algebraic topology in [44] consists in studying $\mathbb{A}^1$–local spaces via their Postnikov towers and, in doing this, it is important to understand the structural properties of $\mathbb{A}^1$–homotopy sheaves. Our discussion is inspired by Morel's axiomatic discussion of the so-called stable $\mathbb{A}^1$–connectivity property in [43, Section 6].

Definition 2.2.6 We will say that the *unstable $\mathbb{A}^1$–connectivity property holds for S* if the following two properties hold:

- $\boldsymbol{\pi}^{\mathbb{A}^1}_1(\mathscr{X})$ is strongly $\mathbb{A}^1$–invariant for any pointed space $(\mathscr{X},x)$.
- Any strongly $\mathbb{A}^1$–invariant sheaf of abelian groups $\boldsymbol{A}$ is strictly $\mathbb{A}^1$–invariant.

For the most part, the results in this text will be proven assuming that the unstable $\mathbb{A}^1$–connectivity holds over S. From this point of view, one of the key results of [44] is the following.

Theorem 2.2.7 (Morel) *If S is the spectrum of a perfect field, then the unstable $\mathbb{A}^1$–connectivity property holds for S.*

Proof See [44, Theorems 5.46 and 6.1]. Since the proof of [44, Lemma 1.15], a result due to Gabber, requires the restriction that k be infinite, use [26] to obtain the result in the stated generality. □

Remark 2.2.8 The unstable $\mathbb{A}^1$–connectivity property does not hold if S is a Noetherian scheme of Krull dimension ≥ 2; see Remark 3.3.5 for more details. Nevertheless, it is expected that the unstable $\mathbb{A}^1$–connectivity property holds for the spectrum of an arbitrary field. In fact, the perfection of the base field only intercedes in the verification of point (2) of Definition 2.2.6. There is some hope that it may hold for base schemes S that are regular of dimension ≤ 1.

Fiber sequences

Convention 2.2.9 (fiber and cofiber sequences) We use the terminology "(co)fiber sequence" as in [27, Definition 6.2.6]; we refer the reader there for more formal properties of (co)fiber sequences. We use the terminology *simplicial fiber sequence* for a fiber sequence in the injective local model structure on simplicial presheaves and $\mathbb{A}^1$*–fiber sequence* for a fiber sequence in the $\mathbb{A}^1$–local model structure. The theory of fiber sequences is simplified slightly by the fact that the injective local and $\mathbb{A}^1$–local model structures are right proper.

The following result, which is a version of [44, Lemma 6.51], studies the behavior of $\mathbb{A}^1$–local objects in simplicial fiber sequences (see [19, e.6, page 5] for a completely analogous result for localizations of the classical homotopy category). There appears to be a misprint in the statement of [44, Lemma 6.51]; the statement concerns $\mathbb{A}^1$–connectivity whereas the hypothesis used and the conclusion reached appear to concern $\mathbb{A}^1$–locality.

Lemma 2.2.10 *Suppose*

$$\mathscr{F} \longrightarrow \mathscr{E} \longrightarrow \mathscr{B}$$

is a simplicial fiber sequence of pointed spaces. If $\mathscr{B}$ and $\mathscr{F}$ are both $\mathbb{A}^1$–local and $\mathscr{B}$ is simplicially connected, then $\mathscr{E}$ is $\mathbb{A}^1$–local as well.

Proof The proof is that given in [44, Lemma 6.51], but we have added some details for the convenience of the reader and for the sake of completeness. By means of the existence of functorial factorizations, we may replace $\mathscr{E} \to \mathscr{B}$ by a simplicial fibration between simplicially fibrant pointed spaces, and we may assume $\mathscr{F}$ is the fiber over the basepoint of this map.

Write $\mathbf{R}\mathrm{Hom}(\mathbb{A}^1, \cdot)$ to denote the derived internal mapping object in the category of presheaves with the simplicial model structure. The model structure is closed monoidal, by [10, Section 4] for example, and the functor $\mathbf{R}\mathrm{Hom}(\mathbb{A}^1, \cdot)$ is a right Quillen functor, [27, Chapter 4]. In particular, this means that $\mathbf{R}\mathrm{Hom}(\mathbb{A}^1, \cdot)$ preserves simplicial fiber sequences.

Condition (2) of [45, Lemma 2.2.8], combined with the definition of internal mapping objects, allows us to say a space $\mathscr{X}$ is $\mathbb{A}^1$–local if and only if the map $\mathscr{X} \to \mathbf{R}\mathrm{Hom}(\mathbb{A}^1, \mathscr{X})$ induced by the projection map $\mathbb{A}^1 \to *$ is a simplicial weak equivalence. In the case of pointed spaces, we forget the basepoint and then apply this test.

Combining the above observations, one concludes that there is a morphism of simplicial fiber sequences of the following form:

$$\begin{tikzcd} \mathscr{F} \arrow[r] \arrow[d, "\sim"] & \mathscr{E} \arrow[r] \arrow[d] & \mathscr{B} \arrow[d, "\sim"] \\ \mathbf{R}\mathrm{Hom}(\mathbb{A}^1, \mathscr{F}) \arrow[r] & \mathbf{R}\mathrm{Hom}(\mathbb{A}^1, \mathscr{E}) \arrow[r] & \mathbf{R}\mathrm{Hom}(\mathbb{A}^1, \mathscr{B}) \end{tikzcd} \tag{2.2.1}$$

The two indicated arrows are simplicial weak equivalences.

It is now possible to test whether the map $\mathscr{E} \to \mathbf{R}\mathrm{Hom}(\mathbb{A}^1, \mathscr{E})$ a simplicial weak equivalence by arguing at points of the Nisnevich site. We refer to [34, Chapter 3 and Lemma 5.12] for the required local homotopy theory. If p^* is point, then from (2.2.1) we obtain a morphism of fiber sequences of Kan complexes of the following form:

$$\begin{tikzcd} p^*\mathscr{F} \arrow[r] \arrow[d, "\sim"] & p^*\mathscr{E} \arrow[r] \arrow[d] & p^*\mathscr{B} \arrow[d, "\sim"] \\ p^*\mathbf{R}\mathrm{Hom}(\mathbb{A}^1, \mathscr{F}) \arrow[r] & p^*\mathbf{R}\mathrm{Hom}(\mathbb{A}^1, \mathscr{E}) \arrow[r] & p^*\mathbf{R}\mathrm{Hom}(\mathbb{A}^1, \mathscr{B}) \end{tikzcd}$$

We wish to show that the map in the middle is a weak equivalence of (possibly disconnected) simplicial sets; that is, it induces an isomorphism on π_0 and on all homotopy groups for all choices of basepoint in $p^*\mathscr{E}$. It suffices, since $p^*\mathscr{E} \to p^*\mathscr{B}$ is a fibration, to consider basepoints lying in $p^*\mathscr{F}$, the fiber over the canonical basepoint of $p^*\mathscr{B}$.

The required isomorphism on homotopy groups and π_0 follows from a 5–lemma argument; the potentially problematic case of π_0 is handled by identifying $\pi_0(p^*\mathscr{E})$ with the orbit space of $\pi_0(p^*\mathscr{F})$ under the action of $\pi_1(p^*\mathscr{B}, b_0)$, and similarly for the derived mapping spaces. □

Basic consequences of the unstable $\mathbb{A}^1$–connectivity theorem We very briefly recall the Postnikov tower in the form we will use. For any pointed simplicially connected space $\mathscr{X}$, there is a tower of fibrations of the form

$$\begin{tikzcd} & & \mathscr{X} \arrow[dl] \arrow[d] \arrow[dr] & & \\ \cdots \arrow[r, "p_{i+2}"] & \mathscr{X}^{(i+1)} \arrow[r, "p_{i+1}"] & \mathscr{X}^{(i)} \arrow[r, "p_i"] & \mathscr{X}^{(i-1)} \arrow[r, "p_{i-1}"] & \cdots \end{tikzcd}$$

such that (a) $\mathrm{hofib}(p_j) = K(\boldsymbol{\pi}_j(\mathscr{X}), j)$ for any integer $j \geq 0$, and (b) $\mathscr{X} \simeq \mathrm{holim}_i\, \mathscr{X}^{(i)}$ [45, Definition 1.31 and Theorem 1.37]. We now deduce some consequences of the unstable $\mathbb{A}^1$–connectivity property.

Lemma 2.2.11 *Suppose the unstable $\mathbb{A}^1$–connectivity property holds for S.*

(1) *For any pointed space $(\mathscr{X}, x)$, and any integer $i \geq 2$, the sheaves $\boldsymbol{\pi}_i(\mathrm{L}_{\mathbb{A}^1}\mathscr{X}, x)$ are strictly $\mathbb{A}^1$–invariant.*

(2) *If $(\mathscr{X}, x)$ is a pointed, simplicially connected space, then $\mathscr{X}$ is $\mathbb{A}^1$–local if and only if $\boldsymbol{\pi}_1(\mathscr{X}, x)$ is strongly $\mathbb{A}^1$–invariant and $\boldsymbol{\pi}_i(\mathscr{X}, x)$ is strictly $\mathbb{A}^1$–invariant for any integer $i \geq 2$.*

Proof For (1), $\Omega^i \mathrm{L}_{\mathbb{A}^1}\mathscr{X}$ is $\mathbb{A}^1$–local and simplicially fibrant for every $i \geq 1$. In particular, the sheaf $\boldsymbol{\pi}_1(\Omega^{i-1}\mathrm{L}_{\mathbb{A}^1}\mathscr{X})$ is strongly $\mathbb{A}^1$–invariant. Since this is abelian when $i \geq 2$, we conclude by the assumption that the unstable $\mathbb{A}^1$–connectivity property holds that $\boldsymbol{\pi}_i(\mathrm{L}_{\mathbb{A}^1}\mathscr{X}, x)$ is strictly $\mathbb{A}^1$–invariant for $i \geq 2$.

For point (2), we use the existence and convergence of the Postnikov tower, together with an induction argument in combination with the results of point (1). Using this tower, it suffices to show that if $\mathscr{X}$ is a pointed, simplicially connected $\mathbb{A}^1$–local space, and we have a simplicial fiber sequence of the form

$$K(\boldsymbol{A}, n) \longrightarrow \mathscr{X}' \longrightarrow \mathscr{X},$$

where $\boldsymbol{A}$ is strongly $\mathbb{A}^1$–invariant if $n = 1$, and strictly $\mathbb{A}^1$–invariant if $n \geq 2$, then $\mathscr{X}'$ is $\mathbb{A}^1$–local as well. Either assumption guarantees that $K(\boldsymbol{A}, n)$ is $\mathbb{A}^1$–local and the result then follows by appeal to Lemma 2.2.10. □

The following result, which is called the unstable $\mathbb{A}^1$–connectivity theorem, justifies our terminology; this result is an axiomatic form of [44, Theorem 6.38].

Theorem 2.2.12 (unstable $\mathbb{A}^1$–connectivity theorem) *Suppose $n \geq 0$ is an integer, and $(\mathscr{X}, x)$ is a pointed, simplicially n–connected space. The space $\mathrm{L}_{\mathbb{A}^1}\mathscr{X}$ is simplicially connected, and if the unstable $\mathbb{A}^1$–connectivity property holds for S, then it is simplicially n–connected.*

Proof The case $n = 0$ of the theorem follows from [45, Section 2, Corollary 3.22] and does not require the unstable $\mathbb{A}^1$–connectivity property to hold. That result is presented without a proof in [45], but follows from the properties of the $\mathbb{A}^1$–localization functor. For a detailed proof, see eg [51, Theorem 1.2.20].

Now, we treat the case $n = 1$. We begin by establishing a general result. For any simplicially connected space $\mathscr{X}$, and any Nisnevich sheaf of groups $\boldsymbol{G}$, there is a functorial bijection

$$\mathrm{Hom}(\boldsymbol{\pi}_1(\mathscr{X}, x), \boldsymbol{G}) \cong [(\mathscr{X}, x), B\boldsymbol{G}]_s$$

by obstruction theory [44, Lemma B.7(1)]. If $\boldsymbol{G}$ is strongly $\mathbb{A}^1$–invariant, then $B\boldsymbol{G}$ is

$\mathbb{A}^1$–local, and there are functorial bijections of the form

$$\begin{aligned}[(\mathscr{X},x),\boldsymbol{BG}]_s &\cong [(\mathscr{X},x),\boldsymbol{BG}]_{\mathbb{A}^1}\\ &\cong [(\mathrm{L}_{\mathbb{A}^1}\mathscr{X},x),\boldsymbol{BG}]_{\mathbb{A}^1}\\ &\cong [(\mathrm{L}_{\mathbb{A}^1}\mathscr{X},x),\boldsymbol{BG}]_s.\end{aligned}$$

Since $\mathrm{L}_{\mathbb{A}^1}\mathscr{X}$ is simplicially connected by [45, Section 2, Corollary 3.22], we conclude that there is a functorial bijection of the form

$$\mathrm{Hom}(\boldsymbol{\pi}_1(\mathrm{L}_{\mathbb{A}^1}\mathscr{X},x),\boldsymbol{G}) \cong [(\mathrm{L}_{\mathbb{A}^1}\mathscr{X},x),\boldsymbol{BG}]_s.$$

Now, $\boldsymbol{\pi}_1(\mathrm{L}_{\mathbb{A}^1}\mathscr{X},x) = \boldsymbol{\pi}_1^{\mathbb{A}^1}(\mathscr{X})$ by definition, so combining all of the above isomorphisms, we conclude that if $\boldsymbol{G}$ is strongly $\mathbb{A}^1$–invariant, then

$$\mathrm{Hom}(\boldsymbol{\pi}_1(\mathscr{X},x),\boldsymbol{G}) \cong \mathrm{Hom}(\boldsymbol{\pi}_1^{\mathbb{A}^1}(\mathscr{X}),\boldsymbol{G}). \tag{2.2.2}$$

Having established this bijection, we can proceed to the proof of the main result.

If $\mathscr{X}$ is simplicially 1–connected, then $\mathrm{Hom}(\boldsymbol{\pi}_1(\mathscr{X},x),\boldsymbol{G}) = 0$ for any sheaf of groups $\boldsymbol{G}$. If $\boldsymbol{G}$ is furthermore strongly $\mathbb{A}^1$–invariant, we conclude by the isomorphism of (2.2.2) that $\mathrm{Hom}(\boldsymbol{\pi}_1^{\mathbb{A}^1}(\mathscr{X}),\boldsymbol{G}) = 0$. Since the unstable $\mathbb{A}^1$–connectivity property holds for k, we know that $\boldsymbol{\pi}_1^{\mathbb{A}^1}(\mathscr{X})$ is strongly $\mathbb{A}^1$–invariant. Therefore, by the Yoneda lemma, we know that $\boldsymbol{\pi}_1^{\mathbb{A}^1}(\mathscr{X})$ must be trivial.

For the general case, one proceeds by induction on n. If $\mathscr{X}$ is a simplicially $(n-1)$–connected space, $n \geq 2$, then for any sheaf of abelian groups $\boldsymbol{A}$,

$$\mathrm{Hom}(\boldsymbol{\pi}_n(\mathscr{X}),\boldsymbol{A}) \cong [(\mathscr{X},x),K(\boldsymbol{A},n)]_s;$$

this follows from [44, Lemma B.7(2)]. An argument completely analogous to the one above, this time using Lemma 2.2.11(1) to conclude that the higher $\mathbb{A}^1$–homotopy sheaves are strictly $\mathbb{A}^1$–invariant, shows that if $\boldsymbol{A}$ is any strictly $\mathbb{A}^1$–invariant sheaf, then

$$\mathrm{Hom}(\boldsymbol{\pi}_n(\mathscr{X}),\boldsymbol{A}) \cong \mathrm{Hom}(\boldsymbol{\pi}_n^{\mathbb{A}^1}(\mathscr{X}),\boldsymbol{A}).$$

If $\mathscr{X}$ is simplicially n–connected, then as before we can again conclude by appealing to the Yoneda lemma. □

Remark 2.2.13 We add one comment about the $n = 0$ case of the above theorem. In fact, [45, Section 2, Corollary 3.22] establishes a more general statement that we will frequently use below: if $\mathscr{X} \to \mathscr{X}'$ is an $\mathbb{A}^1$–weak equivalence with $\mathscr{X}'$ an $\mathbb{A}^1$–local space, then the induced morphism $\boldsymbol{\pi}_0(\mathscr{X}) \to \boldsymbol{\pi}_0(\mathscr{X}')$ is an epimorphism.

2.3 Further consequences of the unstable $\mathbb{A}^1$–connectivity property

In this section, we study further consequences of the unstable $\mathbb{A}^1$–connectivity property introduced in Section 2.2. In particular, we recast some results of Morel in our axiomatic framework. First, we provide an analog of Proposition 2.1.5 in the context of $\mathbb{A}^1$–homotopy theory (see Theorem 2.3.2); this result is a key technical tool in all that follows. In particular, it allows us to establish Theorem 2.3.3, which is a statement about preservation of simplicial fiber sequences under $\mathbb{A}^1$–localization. Consequences of this result include a relative version of the unstable connectivity theorem, which appears below as Corollary 2.3.6, and Theorem 2.3.8, which is a technical result about the interaction between Postnikov towers and $\mathbb{A}^1$–localization.

On $\mathbb{A}^1$–homotopy types of connected spaces We begin by establishing a result about the behavior of the classifying space of the Kan loop group under $\mathbb{A}^1$–localization; this result is culled from the proof of [44, Theorem 6.46]. Suppose G is a simplicial presheaf of groups. Since $\mathrm{L}_{\mathbb{A}^1}$ preserves finite products, $\mathrm{L}_{\mathbb{A}^1} G$ is again a simplicial presheaf of groups, the morphism $G \to \mathrm{L}_{\mathbb{A}^1} G$ is a homomorphism, and there is an induced morphism

$$BG \longrightarrow B\mathrm{L}_{\mathbb{A}^1} G. \tag{2.3.1}$$

Regarding this morphism, one has the following result.

Lemma 2.3.1 *If G is a simplicial presheaf of groups, then the functorial map $BG \to B\mathrm{L}_{\mathbb{A}^1} G$ is an $\mathbb{A}^1$–weak equivalence.*

Proof Recall from 2.1.7 that

$$BG = \operatorname*{hocolim}_{n \in \Delta^{\mathrm{op}}} B(*, G, *)_n.$$

Since the map $G^{\times n} \to (\mathrm{L}_{\mathbb{A}^1} G)^{\times n}$ is an $\mathbb{A}^1$–weak equivalence, and since hocolim preserves such equivalences [45, Section 2, Lemma 2.12], it follows that $BG \to B(\mathrm{L}_{\mathbb{A}^1} G)$ is an $\mathbb{A}^1$–weak equivalence. □

Theorem 2.3.2 *Suppose the unstable $\mathbb{A}^1$–connectivity property holds for S and $(\mathscr{X}, x)$ is a pointed reduced space. If $\pi_0(\mathrm{L}_{\mathbb{A}^1} G(\mathscr{X}))$ is strongly $\mathbb{A}^1$–invariant, then the objects $\mathrm{L}_{\mathbb{A}^1} \mathscr{X}$ and $B\mathrm{L}_{\mathbb{A}^1} G(\mathscr{X})$ are simplicially weakly equivalent.*

Proof By Proposition 2.1.8, since $\mathscr{X}$ is reduced, we know that there is a simplicial weak equivalence of the form $\mathscr{X} \simeq BG(\mathscr{X})$. In particular, $\mathrm{L}_{\mathbb{A}^1} \mathscr{X} \simeq \mathrm{L}_{\mathbb{A}^1} BG(\mathscr{X})$. It will be sufficient to prove that there is a simplicial weak equivalence $\mathrm{L}_{\mathbb{A}^1} BG(\mathscr{X}) \simeq B\mathrm{L}_{\mathbb{A}^1} G(\mathscr{X})$.

We show that $B\mathrm{L}_{\mathbb{A}^1}G(\mathscr{X})$ is $\mathbb{A}^1$–local. To this end, consider the sequence

$$\mathrm{L}_{\mathbb{A}^1}G(\mathscr{X}) \longrightarrow E\mathrm{L}_{\mathbb{A}^1}G(\mathscr{X}) \longrightarrow B\mathrm{L}_{\mathbb{A}^1}G(\mathscr{X}).$$

We know this is a simplicial fiber sequence by reference to [56] for instance. This fiber sequence yields a long exact sequence of homotopy sheaves. Since $E\mathrm{L}_{\mathbb{A}^1}G(\mathscr{X})$ is simplicially contractible, we conclude that

$$\boldsymbol{\pi}_{i+1}(B\mathrm{L}_{\mathbb{A}^1}G(\mathscr{X})) \cong \boldsymbol{\pi}_i(\mathrm{L}_{\mathbb{A}^1}G(\mathscr{X}))$$

for every integer $i \geq 0$. The space $B\mathrm{L}_{\mathbb{A}^1}G(\mathscr{X})$ is simplicially connected as it is the classifying space of a simplicial group. By assumption $\boldsymbol{\pi}_0(\mathrm{L}_{\mathbb{A}^1}G(\mathscr{X}))$ is strongly $\mathbb{A}^1$–invariant, so we conclude that $\boldsymbol{\pi}_1(B\mathrm{L}_{\mathbb{A}^1}G(\mathscr{X}))$ is strongly $\mathbb{A}^1$–invariant. Since the unstable $\mathbb{A}^1$–connectivity property holds for S, we conclude from Lemma 2.2.11(1) that $\boldsymbol{\pi}_j(B\mathrm{L}_{\mathbb{A}^1}G(\mathscr{X}))$ is strictly $\mathbb{A}^1$–invariant for $j \geq 2$. Therefore, by applying Lemma 2.2.11(2), we conclude that $B\mathrm{L}_{\mathbb{A}^1}G(\mathscr{X})$ is itself $\mathbb{A}^1$–local, and the map $B\mathrm{L}_{\mathbb{A}^1}G(\mathscr{X}) \to \mathrm{L}_{\mathbb{A}^1}B\mathrm{L}_{\mathbb{A}^1}G(\mathscr{X})$ is a simplicial weak equivalence. By Lemma 2.3.1, the map $\mathrm{L}_{\mathbb{A}^1}\,BG(\mathscr{X}) \to \mathrm{L}_{\mathbb{A}^1}B\mathrm{L}_{\mathbb{A}^1}G(\mathscr{X})$ is also a simplicial weak equivalence. It follows that there is a map $B\mathrm{L}_{\mathbb{A}^1}G(\mathscr{X}) \to \mathrm{L}_{\mathbb{A}^1}\,BG(\mathscr{X})$ that is also a simplicial weak equivalence, as required. □

On $\mathbb{A}^1$–fiber sequences The following result is a slight variant of [44, Theorem 6.53], which is presented there without proof.

Theorem 2.3.3 *Assume the unstable $\mathbb{A}^1$–connectivity property holds for S. Suppose*

$$\mathscr{F} \longrightarrow \mathscr{E} \xrightarrow{f} \mathscr{B}$$

is a simplicial fiber sequence of pointed spaces and assume that $\mathscr{B}$ is simplicially connected. If, in addition, $\boldsymbol{\pi}_0^{\mathbb{A}^1}(\Omega\mathscr{B})$ is strongly $\mathbb{A}^1$–invariant, then the canonical map

$$\mathrm{L}_{\mathbb{A}^1}\mathscr{F} = \mathrm{L}_{\mathbb{A}^1}\,\mathrm{hofib}(f) \longrightarrow \mathrm{hofib}(\mathrm{L}_{\mathbb{A}^1}(f))$$

is a simplicial weak equivalence. In particular, if $\boldsymbol{\pi}_1(\mathscr{B})$ is strongly $\mathbb{A}^1$–invariant (eg, trivial) then the canonical map of the previous display is a simplicial weak equivalence.

Proof The idea is to replace the simplicial fiber sequence in question by a "principal" fibration under the Kan loop group of the base and use Theorem 2.3.2. To this end, we begin with a reduction.

Step 1 For any simplicial presheaf $\mathscr{X}$, the map $\mathscr{X}(U) \to \mathrm{Ex}\,\mathscr{X}(U)$ is a (functorial) simplicial weak equivalence. Thus, without loss of generality we can assume that $\mathscr{B}$ is objectwise fibrant. Since $\mathscr{B}$ is connected, we can furthermore assume that $\mathscr{B}$

is also reduced. To see this, write $\mathscr{B}(0)$ for the zeroth level of the (sectionwise) Moore–Postnikov factorization of $\mathscr{B}$, and set $\mathscr{B}^{(0)}$ to be the pullback of the diagram:

$$* \longrightarrow \mathscr{B}(0) \longleftarrow \mathscr{B};$$

the space $\mathscr{B}^{(0)}$ is called the zeroth Eilenberg subcomplex of $\mathscr{B}$ (see [23, page 327, proof of Lemma VI.3.6]). By the existence of functorial factorizations, we may assume that, replacing $\mathscr{E}$ by a simplicially weakly equivalent space if necessary, that $\mathscr{E} \to \mathscr{B}$ is a simplicial fibration. In that case, we replace $\mathscr{E}$ by the pullback of the diagram $\mathscr{B}^{(0)} \to \mathscr{B} \leftarrow \mathscr{E}$; this does not change the simplicial homotopy type of the fiber.

Step 2 Since $\mathscr{B}$ is now assumed reduced, set $\mathscr{G} := G(\mathscr{B})$. Next, we claim that the simplicial fiber sequence is equivalent to the simplicial fiber sequence associated with a principal fibration under the Kan loop group. Indeed, since $\mathscr{B}$ is reduced, then $\mathscr{B} \simeq B\mathscr{G}$ by Proposition 2.1.8. Now, a priori there is an action of the h–group $\Omega\mathscr{B}$ on $\mathscr{F}$. Since $\mathscr{E} \to \mathscr{B}$ is a simplicial fibration by assumption, if we set $\mathscr{F}'$ to be the pullback of $E\mathscr{G} \to B\mathscr{G} \simeq \mathscr{B} \leftarrow \mathscr{E}$, then $\mathscr{F}'$ is simplicially weakly equivalent to $\mathscr{F}$ and carries an honest action of $\mathscr{G}$ (see 2.1.7 for our conventions regarding two-sided bar constructions). One then checks that there is an induced simplicial weak equivalence $B(\mathscr{F}', \mathscr{G}, *) \to \mathscr{E}$ making the simplicial fiber sequence $\mathscr{F}' \to B(\mathscr{F}', \mathscr{G}, *) \to B\mathscr{G}$ (see [37, Proposition 7.9]) weakly equivalent to the fiber sequence $\mathscr{F} \to \mathscr{E} \to \mathscr{B}$ with which we began.

Step 3a We now study what happens under $\mathbb{A}^1$–localization. First, since the hypotheses of Theorem 2.3.2 are satisfied by assumption, and the map $\mathscr{G} \to \mathrm{L}_{\mathbb{A}^1}\mathscr{G}$ induces a simplicial weak equivalence $\mathrm{L}_{\mathbb{A}^1}\mathscr{B} \simeq \mathrm{L}_{\mathbb{A}^1}B\mathscr{G} \to B\mathrm{L}_{\mathbb{A}^1}\mathscr{G}$. In particular, $B\mathrm{L}_{\mathbb{A}^1}\mathscr{G}$ is $\mathbb{A}^1$–local.

On the other hand, there is a simplicial fiber sequence of the form

$$\mathrm{L}_{\mathbb{A}^1}\mathscr{F}' \longrightarrow B(\mathrm{L}_{\mathbb{A}^1}\mathscr{F}', \mathrm{L}_{\mathbb{A}^1}\mathscr{G}, *) \longrightarrow B\mathrm{L}_{\mathbb{A}^1}\mathscr{G}.$$

Note that $\mathrm{L}_{\mathbb{A}^1}\mathscr{F}'$ is $\mathbb{A}^1$–local by assumption and $B\mathrm{L}_{\mathbb{A}^1}\mathscr{G}$ is $\mathbb{A}^1$–local and simplicially connected. Thus, by appeal to Lemma 2.2.10, we conclude that $B(\mathrm{L}_{\mathbb{A}^1}\mathscr{F}', \mathrm{L}_{\mathbb{A}^1}\mathscr{G}, *)$ is $\mathbb{A}^1$–local as well.

Next, the maps $\mathscr{F}' \to \mathrm{L}_{\mathbb{A}^1}\mathscr{F}'$ and $\mathscr{G} \to \mathrm{L}_{\mathbb{A}^1}\mathscr{G}$ induce $\mathbb{A}^1$–weak equivalences

$$\mathscr{F}' \times \mathscr{G}^{\times n} \to \mathrm{L}_{\mathbb{A}^1}\mathscr{F}' \times (\mathrm{L}_{\mathbb{A}^1}\mathscr{G})^{\times n}$$

for all n. Therefore, the induced morphism

$$B(\mathscr{F}', \mathscr{G}, *) \longrightarrow B(\mathrm{L}_{\mathbb{A}^1}\mathscr{F}', \mathrm{L}_{\mathbb{A}^1}\mathscr{G}, *)$$

is an $\mathbb{A}^1$–weak equivalence by [45, Section 2, Lemma 2.12].

If R_{Nis} denotes the simplicial fibrant replacement functor, then the map

$$B(\mathrm{L}_{\mathbb{A}^1}\mathscr{F}', \mathrm{L}_{\mathbb{A}^1}\mathscr{G}, *) \to R_{\mathrm{Nis}}B(\mathrm{L}_{\mathbb{A}^1}\mathscr{F}', \mathrm{L}_{\mathbb{A}^1}\mathscr{G}, *)$$

is a simplicial weak equivalence and $R_{\mathrm{Nis}}B(\mathrm{L}_{\mathbb{A}^1}\mathscr{F}', \mathrm{L}_{\mathbb{A}^1}\mathscr{G}, *)$ is $\mathbb{A}^1$–fibrant. Therefore, the composite map

$$B(\mathscr{F}', \mathscr{G}, *) \to R_{\mathrm{Nis}}B(\mathrm{L}_{\mathbb{A}^1}\mathscr{F}', \mathrm{L}_{\mathbb{A}^1}\mathscr{G}, *)$$

is also an $\mathbb{A}^1$–weak equivalence and factors through a simplicial weak equivalence of the form

$$\mathrm{L}_{\mathbb{A}^1}B(\mathscr{F}', \mathscr{G}, *) \longrightarrow R_{\mathrm{Nis}}B(\mathrm{L}_{\mathbb{A}^1}\mathscr{F}', \mathrm{L}_{\mathbb{A}^1}\mathscr{G}, *).$$

Step 3b The evident projections give morphisms

$$\mathrm{L}_{\mathbb{A}^1}B(\mathscr{F}', \mathscr{G}, *) \longrightarrow \mathrm{L}_{\mathbb{A}^1}B\mathscr{G} \quad \text{and} \quad R_{\mathrm{Nis}}B(\mathrm{L}_{\mathbb{A}^1}\mathscr{F}', \mathrm{L}_{\mathbb{A}^1}\mathscr{G}, *) \longrightarrow R_{\mathrm{Nis}}B\mathrm{L}_{\mathbb{A}^1}\mathscr{G}$$

that fit into the following commutative diagram:

$$\begin{tikzcd}
\mathrm{L}_{\mathbb{A}^1}B(\mathscr{F}', \mathscr{G}, *) \arrow[r] \arrow[d] & R_{\mathrm{Nis}}B(\mathrm{L}_{\mathbb{A}^1}\mathscr{F}', \mathrm{L}_{\mathbb{A}^1}\mathscr{G}, *) \arrow[d] \\
\mathrm{L}_{\mathbb{A}^1}B\mathscr{G} \arrow[r] & R_{\mathrm{Nis}}B\mathrm{L}_{\mathbb{A}^1}\mathscr{G}
\end{tikzcd}$$

The simplicial homotopy fiber of the first column is $\mathrm{hofib}(\mathrm{L}_{\mathbb{A}^1}f)$ by construction while the simplicial homotopy fiber of the second column is $\mathrm{L}_{\mathbb{A}^1}\,\mathrm{hofib}(f)$. Moreover, the diagram gives rise to a morphism of simplicial fiber sequences. Since the horizontal maps in the diagram are simplicial weak equivalences by the conclusions of the previous step, we conclude that the induced map of simplicial homotopy fibers is a simplicial weak equivalence.

The final statement is an immediate consequence of Lemma 2.3.4 below. □

Lemma 2.3.4 [44, Lemma 6.54] *If $\mathscr{X}$ is a pointed connected space, such that $\boldsymbol{\pi}_1(\mathscr{X}) = \boldsymbol{\pi}_0(\Omega\mathscr{X})$ is $\mathbb{A}^1$–invariant, then the morphism*

$$\boldsymbol{\pi}_0(\Omega\mathscr{X}) \longrightarrow \boldsymbol{\pi}_0(\mathrm{L}_{\mathbb{A}^1}\Omega\mathscr{X})$$

is an isomorphism. In particular, if $\boldsymbol{\pi}_1(\mathscr{X})$ is strongly $\mathbb{A}^1$–invariant, then so is $\boldsymbol{\pi}_0(\mathrm{L}_{\mathbb{A}^1}\Omega\mathscr{X})$.

Proof By [45, Section 2, Corollary 3.22], the morphism $\boldsymbol{\pi}_0(\Omega\mathscr{X}) \to \boldsymbol{\pi}_0(\mathrm{L}_{\mathbb{A}^1}\Omega\mathscr{X})$ is always an epimorphism (see Remark 2.2.13). Since $\boldsymbol{\pi}_0(\Omega\mathscr{X})$ is $\mathbb{A}^1$–invariant and has simplicial dimension 0 it is necessarily simplicially fibrant and therefore $\mathbb{A}^1$–local by [45, Section 2, Proposition 3.19]. Therefore, the morphism $\Omega\mathscr{X} \to \boldsymbol{\pi}_0(\Omega\mathscr{X})$ necessarily factors through $\mathrm{L}_{\mathbb{A}^1}\Omega\mathscr{X}$. Applying $\boldsymbol{\pi}_0$, we see that the identity map of

$\pi_0(\Omega\mathscr{X})$ factors through $\pi_0(\mathrm{L}_{\mathbb{A}^1}\Omega\mathscr{X})$. Using the epimorphism from the first sentence, we conclude that $\pi_0(\Omega\mathscr{X}) \to \pi_0(\mathrm{L}_{\mathbb{A}^1}\Omega\mathscr{X})$ is an isomorphism. □

The relative unstable $\mathbb{A}^1$–connectivity theorem Theorem 2.3.3 can be used to establish a relative version of the unstable connectivity theorem, Theorem 2.2.12. In the form below, this result is a variant of [44, Theorem 6.56] and will be used repeatedly in the sequel.

Convention 2.3.5 (relative connectivity) If $f\colon \mathscr{E} \to \mathscr{B}$ is a morphism of pointed spaces, we will say that f is simplicially i–connected or a simplicial i–equivalence if the simplicial homotopy fiber of f is $(i-1)$–connected. Likewise, we will say that f is $\mathbb{A}^1$–i–connected or an $\mathbb{A}^1$–i–equivalence if the $\mathbb{A}^1$–homotopy fiber of f is $\mathbb{A}^1$–$(i-1)$–connected.

Corollary 2.3.6 *Assume the unstable $\mathbb{A}^1$–connectivity property holds for S, and suppose $f\colon \mathscr{E} \to \mathscr{B}$ is a morphism of pointed spaces in which $\mathscr{B}$ is connected. If*

(i) $\pi_0^{\mathbb{A}^1}(\Omega\mathscr{B}) = \pi_0(\mathrm{L}_{\mathbb{A}^1}\Omega\mathscr{B})$ *is strongly $\mathbb{A}^1$–invariant, and*

(ii) $\mathrm{hofib}(f)$ *is n–connected for some integer $n \geq 2$,*

then the space $\mathrm{hofib}(\mathrm{L}_{\mathbb{A}^1}(f))$ *is $(n-1)$–connected as well. In particular, under hypothesis* (i), *if f is a simplicial n–equivalence, then f is also an $\mathbb{A}^1$–n–equivalence.*

Proof Since the unstable $\mathbb{A}^1$–connectivity property holds for S, and since by hypothesis (ii) the space $\mathrm{hofib}(f)$ is assumed $(n-1)$–connected, we may apply Theorem 2.2.12 to conclude that $\mathrm{L}_{\mathbb{A}^1}\,\mathrm{hofib}(f)$ is again n–connected. Again using the unstable $\mathbb{A}^1$–connectivity property for S, the fact that $\mathscr{B}$ is connected, and hypothesis (i), we may apply Theorem 2.3.3 to conclude that the canonical map $\mathrm{L}_{\mathbb{A}^1}\,\mathrm{hofib}(f) \to \mathrm{hofib}(\mathrm{L}_{\mathbb{A}^1}(f))$ is a simplicial weak equivalence. Combining these observations, we conclude that $\mathrm{hofib}(\mathrm{L}_{\mathbb{A}^1}(f))$ is n–connected as well. □

$\mathbb{A}^1$–localization of layers of Postnikov towers We can also deduce some stability properties for the layers of Postnikov towers under $\mathbb{A}^1$–localization. To this end, assume $(\mathscr{X}, x)$ is a pointed connected space. If $\mathscr{X}^{(n)}$ is the n^{th} layer of the Postnikov tower for $\mathscr{X}$, then we write $\mathscr{X}\langle n\rangle$ for the space fitting into a simplicial fibration sequence of the form

$$\mathscr{X}\langle n\rangle \longrightarrow \mathscr{X} \longrightarrow \mathscr{X}^{(n)}.$$

The space $\mathscr{X}\langle n\rangle$ is the n–fold connective cover of $\mathscr{X}$, in particular it is n–connected.

Lemma 2.3.7 *If $(\mathscr{X}, x)$ is a pointed simplicially connected space, and the unstable $\mathbb{A}^1$–connectivity property holds for our base S, then $(\mathrm{L}_{\mathbb{A}^1}\mathscr{X})\langle n\rangle$ is $\mathbb{A}^1$–local.*

Proof Assuming the unstable $\mathbb{A}^1$–connectivity property holds for our base S, we conclude from Lemma 2.2.11 that $(\mathrm{L}_{\mathbb{A}^1}\mathscr{X})^{(n)}$ is $\mathbb{A}^1$–local. Moreover, $(\mathrm{L}_{\mathbb{A}^1}\mathscr{X})^{(n)}$ is simplicially connected by Theorem 2.2.12 (though this does not require the unstable $\mathbb{A}^1$–connectivity property). It follows that, under these hypotheses, $(\mathrm{L}_{\mathbb{A}^1}\mathscr{X})\langle n\rangle$ is $\mathbb{A}^1$–local since it is the simplicial homotopy fiber of the map $\mathrm{L}_{\mathbb{A}^1}\mathscr{X} \to (\mathrm{L}_{\mathbb{A}^1}\mathscr{X})^{(n)}$, which has a connected base. □

By functoriality of the Postnikov tower, there is an induced morphism $\mathscr{X}\langle n\rangle \to (\mathrm{L}_{\mathbb{A}^1}\mathscr{X})\langle n\rangle$. Assuming the unstable $\mathbb{A}^1$–connectivity property holds for our base S, it follows from Lemma 2.3.7 that there is an induced morphism

$$\mathrm{L}_{\mathbb{A}^1}(\mathscr{X}\langle n\rangle) \longrightarrow (\mathrm{L}_{\mathbb{A}^1}\mathscr{X})\langle n\rangle.$$

Regarding this morphism, we have the following result, which is a variant of [44, Theorem 6.59 and Corollary 6.60].

Theorem 2.3.8 *Assume the unstable $\mathbb{A}^1$–connectivity property holds for S, and $\mathscr{X}$ is a pointed, connected space. Fix an integer $n \geq 1$. Suppose for each integer i with $1 \leq i \leq n$, the sheaf $\boldsymbol{\pi}_i(\mathscr{X})$ is strongly $\mathbb{A}^1$–invariant.*

(1) *The universal map $\boldsymbol{\pi}_i(\mathscr{X}) \to \boldsymbol{\pi}_i^{\mathbb{A}^1}(\mathscr{X})$ is an isomorphism if $i \leq n$.*

(2) *For each $i \leq n$, the morphism $\mathrm{L}_{\mathbb{A}^1}(\mathscr{X}\langle i\rangle) \to (\mathrm{L}_{\mathbb{A}^1}\mathscr{X})\langle i\rangle$ is a simplicial weak equivalence.*

(3) *The universal map $\boldsymbol{\pi}_{n+1}(\mathscr{X}) \to \boldsymbol{\pi}_{n+1}^{\mathbb{A}^1}(\mathscr{X})$ is the initial map from $\boldsymbol{\pi}_{n+1}(\mathscr{X})$ to a strictly $\mathbb{A}^1$–invariant sheaf of groups.*

Proof Since $\boldsymbol{\pi}_1(\mathscr{X})$ is assumed strongly $\mathbb{A}^1$–invariant, and for any integer $n \geq 1$ the map $\boldsymbol{\pi}_1(\mathscr{X}) \to \boldsymbol{\pi}_1(\mathscr{X}^{(n)})$ is an isomorphism, we conclude that $\boldsymbol{\pi}_1(\mathscr{X}^{(n)})$ is strongly $\mathbb{A}^1$–invariant for any $n \geq 1$. By assumption, the unstable $\mathbb{A}^1$–connectivity property holds for S. Thus, for $i \geq 2$ (or $i = 1$ if $\boldsymbol{\pi}_1^{\mathbb{A}^1}(\mathscr{X})$ is abelian) we conclude that $\boldsymbol{\pi}_i(\mathscr{X})$ is strictly $\mathbb{A}^1$–invariant. We may also apply Theorem 2.3.3 to conclude that the sequence

$$\mathrm{L}_{\mathbb{A}^1}(\mathscr{X}\langle n\rangle) \longrightarrow \mathrm{L}_{\mathbb{A}^1}\mathscr{X} \longrightarrow \mathrm{L}_{\mathbb{A}^1}(\mathscr{X}^{(n)})$$

is always a simplicial fiber sequence.

By Theorem 2.2.12, we know that $\mathrm{L}_{\mathbb{A}^1}(\mathscr{X}\langle n\rangle)$ is simplicially n–connected. Therefore, we conclude that

$$\begin{aligned} \pi_i(\mathrm{L}_{\mathbb{A}^1}\mathscr{X}) &\cong \boldsymbol{\pi}_i(\mathrm{L}_{\mathbb{A}^1}(\mathscr{X}^{(n)})) && \text{if } i \leq n, \\ \boldsymbol{\pi}_{n+1}(\mathrm{L}_{\mathbb{A}^1}\mathscr{X}) &\longrightarrow \boldsymbol{\pi}_{n+1}(\mathrm{L}_{\mathbb{A}^1}(\mathscr{X}^{(n)})) && \text{is an epimorphism.} \end{aligned} \tag{2.3.2}$$

There are also simplicial fiber sequences of the form

$$K(\boldsymbol{\pi}_i(\mathscr{X}), i) \longrightarrow \mathscr{X}^{(i)} \longrightarrow \mathscr{X}^{(i-1)}.$$

Since $\mathscr{X}$ is connected, point (2) of Lemma 2.2.11 and the assumptions about the sheaves $\boldsymbol{\pi}_i(\mathscr{X})$ guarantee that $\mathscr{X}^{(n)}$ is $\mathbb{A}^1$–local. Thus

$$\boldsymbol{\pi}_i(\mathscr{X}^{(n)}) \cong \boldsymbol{\pi}_i(\mathrm{L}_{\mathbb{A}^1}\mathscr{X}^{(n)}) \quad \text{if } i \leq n. \tag{2.3.3}$$

Now, we can put these facts together to prove the results.

For point (1), notice that by combining the isomorphisms of (2.3.2) and (2.3.3) we obtain for $i \leq n$ the series of isomorphisms

$$\begin{aligned} \boldsymbol{\pi}_i(\mathscr{X}) &\cong \boldsymbol{\pi}_i(\mathscr{X}^{(n)}) \\ &\cong \boldsymbol{\pi}_i(\mathrm{L}_{\mathbb{A}^1}\mathscr{X}^{(n)}) \\ &\cong \boldsymbol{\pi}_i(\mathrm{L}_{\mathbb{A}^1}\mathscr{X}) \\ &\cong \boldsymbol{\pi}_i^{\mathbb{A}^1}(\mathscr{X}), \end{aligned}$$

which is precisely what we wanted to show.

For point (2) we proceed as follows. From the isomorphisms established in point (1), we conclude that the map $\mathscr{X}^{(i)} \to (\mathrm{L}_{\mathbb{A}^1}\mathscr{X})^{(i)}$ is a simplicial weak equivalence. On the other hand, we already saw that $\mathscr{X}^{(i)}$ is $\mathbb{A}^1$–local for $i \leq n$. Thus, the map $\mathrm{L}_{\mathbb{A}^1}\mathscr{X}^{(i)} \to (\mathrm{L}_{\mathbb{A}^1}\mathscr{X})^{(i)}$ is a simplicial weak equivalence for $i \leq n$. Since $(\mathrm{L}_{\mathbb{A}^1}\mathscr{X})\langle i\rangle$ is by definition the simplicial homotopy fiber of $\mathrm{L}_{\mathbb{A}^1}\mathscr{X} \to (\mathrm{L}_{\mathbb{A}^1}\mathscr{X})^{(i)}$, it follows from the fiber sequence in the previous paragraph that the induced map $\mathrm{L}_{\mathbb{A}^1}(\mathscr{X}\langle i\rangle) \to (\mathrm{L}_{\mathbb{A}^1}\mathscr{X})\langle i\rangle$ is a simplicial weak equivalence for $i \leq n$.

Finally, for point (3), begin by observing that if $\boldsymbol{A}$ is a strictly $\mathbb{A}^1$–invariant sheaf, then since $\mathscr{X}\langle n\rangle$ is n–connected, obstruction theory (see [44, Lemma B.7]) gives a bijection

$$\mathrm{Hom}(\boldsymbol{\pi}_{n+1}(\mathscr{X}), \boldsymbol{A}) \cong [\mathscr{X}\langle n\rangle, K(\boldsymbol{A}, n+1)]_s.$$

Since $K(\boldsymbol{A}, n+1)$ is $\mathbb{A}^1$–local, any map $\mathscr{X}\langle n\rangle \to K(\boldsymbol{A}, n+1)$ factors through $\mathrm{L}_{\mathbb{A}^1}(\mathscr{X}\langle n\rangle)$. However, as $\mathrm{L}_{\mathbb{A}^1}(\mathscr{X}\langle n\rangle) \to (\mathrm{L}_{\mathbb{A}^1}\mathscr{X})\langle n\rangle$ is a simplicial weak equivalence by point (2), the result follows from the fact that $\boldsymbol{\pi}_{n+1}((\mathrm{L}_{\mathbb{A}^1}\mathscr{X})\langle n\rangle) = \boldsymbol{\pi}_{n+1}^{\mathbb{A}^1}(\mathscr{X})$. □

2.4 James-style models for loop spaces in $\mathbb{A}^1$–homotopy theory

In this section, we discuss the James model for loop spaces in $\mathbb{A}^1$–homotopy theory. The construction involves comparing the James model and the Kan loop group model, as was the case in the setting of simplicial homotopy theory. If $(\mathscr{X}, x)$ is a pointed simplicially connected space, then $\Omega \mathrm{L}_{\mathbb{A}^1}\mathscr{X}$ is a model for the $\mathbb{A}^1$–derived loop space of $\mathscr{X}$. The map $\Omega\mathscr{X} \to \Omega\mathrm{L}_{\mathbb{A}^1}\mathscr{X}$ factors through a morphism

$$\mathrm{L}_{\mathbb{A}^1}\Omega\mathscr{X} \longrightarrow \Omega\mathrm{L}_{\mathbb{A}^1}\mathscr{X},$$

which need not be a simplicial weak equivalence. The following result, which is a variant of [44, Theorem 6.46], gives a necessary and sufficient condition for the above morphism to be a simplicial weak equivalence.

Theorem 2.4.1 *Assume the unstable $\mathbb{A}^1$–connectivity property holds for S and suppose $(\mathscr{X}, x)$ is a pointed space. If $\boldsymbol{\pi}_0(\mathrm{L}_{\mathbb{A}^1}\Omega\mathscr{X})$ is strongly $\mathbb{A}^1$–invariant, then*

$$\mathrm{L}_{\mathbb{A}^1}\Omega\mathscr{X} \longrightarrow \Omega\mathrm{L}_{\mathbb{A}^1}\mathscr{X}$$

is a simplicial weak equivalence.

Proof Since $\Omega\mathscr{X}$ only depends on the simplicial connected component of the basepoint x, without loss of generality we can assume that $\mathscr{X}$ is simplicially connected. In that case, the result follows immediately from Theorem 2.3.3 applied to the simplicial fiber sequence $\Omega\mathscr{X} \to * \to \mathscr{X}$. □

We now use Proposition 2.1.6, the James construction in the category of simplicial presheaves, together with the result just established about models for $\mathbb{A}^1$–derived loop spaces to produce a James-style model for loops on the suspension in the $\mathbb{A}^1$–homotopy category.

Theorem 2.4.2 *Suppose that the unstable $\mathbb{A}^1$–connectivity property holds for S and $f\colon \mathscr{X} \to \mathscr{Y}$ is a morphism of pointed simplicially connected spaces.*

(1) *There is a functorial simplicial weak equivalence $\mathrm{L}_{\mathbb{A}^1}J(\mathscr{X}) \simeq \Omega\mathrm{L}_{\mathbb{A}^1}\Sigma\mathscr{X}$.*

(2) *If f is an $\mathbb{A}^1$–weak equivalence, the map $J(f)$ is an $\mathbb{A}^1$–weak equivalence.*

Proof By Proposition 2.1.6, there is a simplicial weak equivalence $J(\mathscr{X}) \simeq \Omega\Sigma\mathscr{X}$. Thus, there is a simplicial weak equivalence

$$\mathrm{L}_{\mathbb{A}^1}J(\mathscr{X}) \simeq \mathrm{L}_{\mathbb{A}^1}\Omega\Sigma\mathscr{X}.$$

Since $\mathscr{X}$ is connected, $\Sigma\mathscr{X}$ is necessarily 1–connected (this follows by checking on stalks). Therefore, by Lemma 2.3.4, we conclude that $\boldsymbol{\pi}_0(\mathrm{L}_{\mathbb{A}^1}\Omega\Sigma\mathscr{X})$ is strongly $\mathbb{A}^1$–invariant. Thus, we can apply Theorem 2.4.1 to conclude that $\mathrm{L}_{\mathbb{A}^1}\Omega\Sigma\mathscr{X} \simeq \Omega\mathrm{L}_{\mathbb{A}^1}\Sigma\mathscr{X}$.

For (2), it suffices to observe that if $f\colon \mathscr{X} \to \mathscr{Y}$ is an $\mathbb{A}^1$–weak equivalence, then by [45, Section 3, Lemma 2.13] the map $\Sigma\mathscr{X} \to \Sigma\mathscr{Y}$ is an $\mathbb{A}^1$–weak equivalence. It follows immediately that the induced morphism $\Omega\Sigma\mathscr{X} \to \Omega\Sigma\mathscr{Y}$ is an $\mathbb{A}^1$–weak equivalence. By part (1), we conclude that $J(X) \to J(Y)$ is an $\mathbb{A}^1$–weak equivalence. □

3 The EHP sequence in $\mathbb{A}^1$–homotopy theory

In this section, we study the analog in $\mathbb{A}^1$–homotopy theory of Whitehead's EHP exact sequence from the introduction. We begin by recasting this exact sequence in the homotopy theory of simplicial sets (see Proposition 3.1.2), and then explaining how to extend this result to simplicial presheaves on a site (see Proposition 3.1.4). For convenience, we will assume our site has enough points. In Section 3.2, we construct a version of Whitehead's exact sequence in $\mathbb{A}^1$–homotopy theory (see Theorem 3.2.1). In Section 3.3, we study the low-degree portion of the exact sequence of Theorem 3.2.1 and study very explicitly the first degree in which the suspension fails to be an isomorphism. The main result is Theorem 3.3.13, which depends on various facts about $\mathbb{A}^1$–homology.

3.1 The EHP sequence in simplicial homotopy theory

In this section, we recall Whitehead's refinement of the Freudenthal suspension theorem and adapt this result to the context of simplicial presheaves. This result appears as [59, Chapter XII, Theorem 2.2] and the main novelty of this section is that we give a different derivation of the exact sequence that we learned from Mike Hopkins; this version allows more precise control at the end of the sequence. The translation to the setting of simplicial presheaves is then straightforward.

The classical EHP sequence We begin by recalling the combinatorial construction of James–Hopf maps. We refer the reader to [62, page 169] for more details.

Definition 3.1.1 Suppose K is a pointed simplicial set and $r \geq 1$ is an integer. Define a morphism of simplicial sets

$$\mathrm{H}_r\colon J(K) \to J(K^{\wedge r})$$

that in each simplicial degree is given by the formula

$$\mathrm{H}_r(x_1 \dots x_q) = \prod_{1 \leq i_1 < \cdots < i_r \leq q} x_{i_1} \wedge \cdots \wedge x_{i_r},$$

where the product on the right-hand side is taken in (left-to-right) lexicographic order. We refer to H_r as a *simplicial James–Hopf invariant*.

Note that H_r is, by definition, functorial in the input simplicial set K. Directly from the definition of H_r it follows that if $r \geq 2$, then the composite $K \xrightarrow{\mathrm{E}} J(K) \xrightarrow{\mathrm{H}_r} J(K^{\wedge r})$

is trivial. We fix $r = 2$, and write H for H_2. There is a commutative diagram:

(3.1.1)
$$\begin{tikzcd} K \arrow[d, "\phi"] \arrow[dr, "\mathrm{E}"] & & \\ \operatorname{hofib}\mathrm{H} \arrow[r] & J(K) \arrow[r, "\mathrm{H}"] & J(K^{\wedge 2}) \end{tikzcd}$$

Proposition 3.1.2 *Suppose K is $(n-1)$–connected where $n \geq 2$. Then the morphism $\phi\colon K \to \operatorname{hofib}\mathrm{H}$ is $(3n-2)$–connected. In particular, we obtain a long exact sequence of homotopy groups:*

(3.1.2)
$$\begin{tikzcd}[column sep=small] \pi_{3n-2}(K) \arrow[d, two heads] & \pi_{3n-1}(\Sigma K) \arrow[d, no head, "\cong"] & \pi_{3n-1}(\Sigma(K^{\wedge 2})) \arrow[d, no head, "\cong"] & & \\ \pi_{3n-2}(\operatorname{hofib}\mathrm{H}) \arrow[r] & \pi_{3n-2}(J(K)) \arrow[r, "\mathrm{H}"] & \pi_{3n-2}(J(K^{\wedge 2})) \arrow[r, "\mathrm{P}"] & \pi_{3n-3}(K) \arrow[r, "\mathrm{E}"] & \cdots \end{tikzcd}$$

$$\begin{tikzcd}[column sep=small] \cdots \arrow[r] & \pi_q(K) \arrow[r, "\mathrm{E}"] & \pi_q(J(K)) \arrow[r, "\mathrm{H}"] \arrow[d, no head, "\cong"] & \pi_q(J(K^{\wedge 2})) \arrow[r, "\mathrm{P}"] \arrow[d, no head, "\cong"] & \pi_{q-1}(K) \arrow[r] & \cdots \\ & & \pi_{q+1}(\Sigma K) & \pi_{q+1}(\Sigma(K^{\wedge 2})) & & \end{tikzcd}$$

Proof Since K is $(n-1)$–connected, we conclude that $K^{\wedge 2}$ is $(2n-1)$–connected. Therefore, $J(K) \simeq \Omega\Sigma K$ is $(n-1)$–connected, and $J(K^{\wedge 2}) \simeq \Omega\Sigma K^{\wedge 2}$ is $(2n-1)$–connected.

We consider the Serre spectral sequence in homology $\mathrm{H}_*(\,\cdot\,,\mathbb{Z})$ associated with the simplicial fiber sequence

$$\operatorname{hofib}\mathrm{H} \longrightarrow J(K) \longrightarrow J(K^{\wedge 2}).$$

Since $n \geq 1$ by assumption, $J(K^{\wedge 2})$ is simply connected.

By use of the Hilton–Milnor splitting [59, Chapter VII, Theorem 2.10] there are isomorphisms

$$\widetilde{\mathrm{H}}_*(J(K),\mathbb{Z}) \cong \bigoplus_{i=1}^{\infty} \widetilde{\mathrm{H}}_*(K^{\wedge i},\mathbb{Z}) \quad \text{and} \quad \widetilde{\mathrm{H}}_*(J(K^{\wedge 2}),\mathbb{Z}) \cong \bigoplus_{i=1}^{\infty} \widetilde{\mathrm{H}}_*(K^{\wedge 2i},\mathbb{Z}).$$

We remark in passing that the map $\mathrm{E}\colon K \to J(K)$ induces an isomorphism of $\widetilde{\mathrm{H}}_*(K,\mathbb{Z})$ with the first summand of $\widetilde{\mathrm{H}}_*(J(K),\mathbb{Z}) \cong \bigoplus_{i=1}^{\infty} \widetilde{\mathrm{H}}_*(K^{\wedge i},\mathbb{Z})$; this appears in the proof of [59, Chapter VII, Theorem 2.10].

In the range where $p+q < 3n$, the E^2 page of the spectral sequence takes a particularly simple form: $\mathrm{E}^2_{0,q} = \mathrm{H}_q(\operatorname{hofib} H,\mathbb{Z})$ and $\mathrm{E}^2_{p,0} = \mathrm{H}_p(J(K^{\wedge 2}),\mathbb{Z}) = \mathrm{H}_p(K^{\wedge 2},\mathbb{Z})$, and all other groups are necessarily 0.

From [36, Theorem 6.2], we know that the composite map

$$\mathrm{H}_*(K^{\wedge 2}, \mathbb{Z}) \longrightarrow \mathrm{H}_*(J(K), \mathbb{Z}) \xrightarrow{\mathrm{H}} \mathrm{H}_*(J(K^{\wedge 2}), \mathbb{Z}) \longrightarrow \mathrm{H}_*(K^{\wedge 2}, \mathbb{Z})$$

is the identity. This observation implies that there are no nonzero differentials in our spectral sequence having source $\mathrm{E}^*_{p,0}$ with $p < 4n$. For degree reasons, therefore, there can be no nonzero differentials the targets of which are the groups $\mathrm{E}^*_{0,q}$ with $q < 3n-1$ either, and in the range where $p+q \leq 3n-2$, the sequence collapses at the E^2 page. We obtain $\widetilde{\mathrm{H}}_{\leq 3n-2}(J(K), \mathbb{Z}) = \widetilde{\mathrm{H}}_{\leq 3n-2}(K^{\wedge 2}, \mathbb{Z}) \oplus \widetilde{\mathrm{H}}_{\leq 3n-2}(\operatorname{hofib} \mathrm{H}, \mathbb{Z})$.

In the given range, therefore, we have a commutative diagram of homology groups (with $\mathbb{Z}$ coefficients)

$$\begin{tikzcd}
0 \arrow[r] & \widetilde{\mathrm{H}}_{\leq 3n-2}(\operatorname{hofib}\mathrm{H}) \arrow[r] & \widetilde{\mathrm{H}}_{\leq 3n-2}(J(K)) \arrow[r] & \widetilde{\mathrm{H}}_{\leq 3n-2}(J(K^{\wedge 2})) \arrow[r] & 0 \\
0 \arrow[r] & \widetilde{\mathrm{H}}_{\leq 3n-2}(K) \arrow[r] \arrow[u, "\phi_*"] & \widetilde{\mathrm{H}}_{\leq 3n-2}(K) \oplus \widetilde{\mathrm{H}}_{\leq 3n-2}(K^{\wedge 2}) \arrow[r] \arrow[u, equal] & \widetilde{\mathrm{H}}_{\leq 3n-2}(K^{\wedge 2}) \arrow[r] \arrow[u, equal] & 0
\end{tikzcd}$$

from which it follows that the map ϕ_* is a homology isomorphism in the stated range. In particular, the map ϕ is $(3n-2)$–connected. The long exact sequence (3.1.2) now follows from the long exact sequence in homotopy associated with the simplicial fiber sequence

$$\operatorname{hofib}\mathrm{H} \longrightarrow J(K) \longrightarrow J(K^{\wedge 2}). \qquad \square$$

The EHP sequence for simplicial presheaves Using the results of the previous section, we can generalize Proposition 3.1.2 to the situation of pointed simplicial presheaves on a site $\mathbf{C}$ equipped with a local model structure; for simplicity, we assume that $\mathbf{C}$ has enough points. Functoriality of the simplicial James–Hopf invariants allows Definition 3.1.1 to be extended to simplicial presheaves.

Definition 3.1.3 If X is a pointed simplicial presheaf on $\mathbf{C}$, define morphisms

$$\mathrm{H}_r\colon J(X) \longrightarrow J(X^{\wedge r})$$

by $\mathrm{H}_r\colon J(X)(U) \to J(X^{\wedge r})(U)$. Set $\mathrm{H} := \mathrm{H}_2$.

As before, the composite map $X \xrightarrow{\mathrm{E}} J(X) \xrightarrow{\mathrm{H}_r} J(X^{\wedge r})$ is null. The next result extends Proposition 3.1.2 to simplicial presheaves.

Proposition 3.1.4 *Suppose* $\mathbf{C}$ *is a site that has enough points. Suppose,* $n \geq 1$ *is an integer, and* X *is a pointed* $(n-1)$*–connected simplicial presheaf. Let* $\mathrm{E}\colon X \to J(X)$ *be as in* (2.1.1), $\mathrm{H}\colon J(X) \to J(X^{\wedge 2})$ *as in* Definition 3.1.3, *and let* ϕ *be a lift of the* *map* E: $X \to J(X)$ *to a map* ϕ: $X \to \operatorname{hofib} \mathrm{H}$. *The map* ϕ *is* $(3n-2)$*–connected and there is a long exact sequence of homotopy sheaves of the form*

$$\begin{aligned}(3.1.3)\quad &\boldsymbol{\pi}_{3n-2}(X) \xrightarrow{\mathrm{E}} \boldsymbol{\pi}_{3n-2}(J(X)) \xrightarrow{\mathrm{H}} \boldsymbol{\pi}_{3n-2}(J(X^{\wedge 2})) \xrightarrow{\mathrm{P}} \boldsymbol{\pi}_{3n-3}(X) \xrightarrow{\mathrm{E}} \cdots \\ &\cdots \longrightarrow \boldsymbol{\pi}_q(X) \xrightarrow{\mathrm{E}} \boldsymbol{\pi}_q(J(X)) \xrightarrow{\mathrm{H}} \boldsymbol{\pi}_q(J(X^{\wedge 2})) \xrightarrow{\mathrm{P}} \boldsymbol{\pi}_{q-1}(X) \longrightarrow \cdots.\end{aligned}$$

Remark 3.1.5 Proposition 2.1.6 guarantees the existence of isomorphisms of homotopy sheaves of the form $\boldsymbol{\pi}_q(J(X)) \cong \boldsymbol{\pi}_{q+1}(X)$ and $\boldsymbol{\pi}_q(J(X^{\wedge 2})) \cong \boldsymbol{\pi}_{q+1}(X^{\wedge 2})$.

Proof In outline, we argue at points to reduce to the classical EHP sequence. In more detail, let F denote the homotopy fiber of the map H: $J(X) \to J(X^{\wedge 2})$ in the local model structure. Since the composite $\mathrm{H} \circ \mathrm{E}$: $X \to J(X) \to J(X^{\wedge 2})$ is null, there is a lift of E: $X \to J(X)$ to a map ϕ: $X \to F$ as follows:

$$(3.1.4)\qquad \begin{tikzcd} X \arrow[d, dashed, "\phi"'] \arrow[dr, "\mathrm{E}"] & & \\ F \arrow[r] & J(X) \arrow[r, "\mathrm{H}"] & J(X^{\wedge 2}) \end{tikzcd}$$

If q^* is a point of the site $\mathbf{C}$, then q^* preserves fiber sequences, and commutes with the formation of $J(\cdot)$ and E, H. In particular, applying q^* throughout, we see using Proposition 3.1.2 that $q^*\phi$ is $(3n-2)$–connected. Since this holds for all such q^*, we deduce that the map ϕ is itself $(3n-2)$–connected. The long exact sequence follows. □

3.2 The construction of the EHP sequence in $\mathbb{A}^1$–homotopy theory

We now transport the EHP sequence studied in the previous section to $\mathbb{A}^1$–homotopy theory. The basic idea is to appeal to Proposition 3.1.4 and use facts about when $\mathbb{A}^1$–localization preserves simplicial fiber sequences from Section 2.3. If we $\mathbb{A}^1$–localize the simplicial James–Hopf map H of Definition 3.1.3 (we abuse notation and write H for the resulting map), then we can consider the following sequence of morphisms

$$(3.2.1)\qquad \mathrm{L}_{\mathbb{A}^1}\mathscr{X} \longrightarrow \mathrm{L}_{\mathbb{A}^1} J(\mathscr{X}) \xrightarrow{\mathrm{H}} \mathrm{L}_{\mathbb{A}^1} J(\mathscr{X}^{\wedge 2}).$$

The next result gives an analog of Whitehead's classical exact sequence in $\mathbb{A}^1$–homotopy theory.

Theorem 3.2.1 *Assume the unstable $\mathbb{A}^1$–connectivity property holds for S and suppose $\mathscr{X}$ is a pointed $\mathbb{A}^1$–$(n-1)$–connected space, with $n \geq 2$. There is an exact sequence of homotopy sheaves of the form*

$$\begin{aligned}(3.2.2)\quad &\boldsymbol{\pi}^{\mathbb{A}^1}_{3n-2}(\mathscr{X}) \xrightarrow{\mathrm{E}} \boldsymbol{\pi}^{\mathbb{A}^1}_{3n-2}(J(\mathscr{X})) \xrightarrow{\mathrm{H}} \boldsymbol{\pi}^{\mathbb{A}^1}_{3n-2}(J(\mathscr{X}^{\wedge 2})) \xrightarrow{\mathrm{P}} \boldsymbol{\pi}^{\mathbb{A}^1}_{3n-3}(\mathscr{X}) \xrightarrow{\mathrm{E}} \cdots \\ &\cdots \longrightarrow \boldsymbol{\pi}^{\mathbb{A}^1}_q(\mathscr{X}) \xrightarrow{\mathrm{E}} \boldsymbol{\pi}^{\mathbb{A}^1}_q(J(\mathscr{X})) \xrightarrow{\mathrm{H}} \boldsymbol{\pi}^{\mathbb{A}^1}_q(J(\mathscr{X}^{\wedge 2})) \xrightarrow{\mathrm{P}} \boldsymbol{\pi}^{\mathbb{A}^1}_{q-1}(\mathscr{X}) \longrightarrow \cdots.\end{aligned}$$

Remark 3.2.2 Theorem 2.4.2 guarantees the existence of isomorphisms of sheaves $\boldsymbol{\pi}_q^{\mathbb{A}^1}(J(\mathscr{X})) \cong \boldsymbol{\pi}_{q+1}^{\mathbb{A}^1}(\Sigma\mathscr{X})$ and $\boldsymbol{\pi}_q(J(\mathscr{X}^{\wedge 2})) \cong \boldsymbol{\pi}_{q+1}^{\mathbb{A}^1}(\Sigma\mathscr{X}^{\wedge 2})$.

Proof We proceed as in the proof of Proposition 3.1.4, with the part of $\mathscr{X}$ played by $\mathrm{L}_{\mathbb{A}^1}\mathscr{X}$. By hypothesis, $\mathrm{L}_{\mathbb{A}^1}\mathscr{X}$ is simplicially $(n-1)$–connected. As before, set up the diagram:

(3.2.3)
$$\begin{tikzcd} \mathrm{L}_{\mathbb{A}^1}\mathscr{X} \arrow[d, dashed, "\phi"] \arrow[dr, "\mathrm{E}"] & & \\ \mathscr{F} \arrow[r] & J(\mathrm{L}_{\mathbb{A}^1}\mathscr{X}) \arrow[r, "\mathrm{H}"] & J((\mathrm{L}_{\mathbb{A}^1}\mathscr{X})^{\wedge 2}) \end{tikzcd}$$

Since $n \geq 1$, the space $J((\mathrm{L}_{\mathbb{A}^1}\mathscr{X})^{\wedge 2})$ is simplicially 1–connected. Then, using the unstable $\mathbb{A}^1$–connectivity property, we may apply Theorem 2.3.3 to conclude that applying $\mathrm{L}_{\mathbb{A}^1}$ to the simplicial fiber sequence in (3.1.4) results in a simplicial fiber sequence of the form

$$\mathrm{L}_{\mathbb{A}^1}\mathscr{F} \longrightarrow \mathrm{L}_{\mathbb{A}^1} J(\mathrm{L}_{\mathbb{A}^1}\mathscr{X}) \longrightarrow \mathrm{L}_{\mathbb{A}^1} J((\mathrm{L}_{\mathbb{A}^1}\mathscr{X})^{\wedge 2}). \tag{3.2.4}$$

Since the map $\mathscr{X} \to \mathrm{L}_{\mathbb{A}^1}\mathscr{X}$ is an $\mathbb{A}^1$–weak equivalence, Theorem 2.4.2(2) implies that there are weak equivalences of the form $\mathrm{L}_{\mathbb{A}^1} J(\mathrm{L}_{\mathbb{A}^1}\mathscr{X}) \simeq \mathrm{L}_{\mathbb{A}^1} J(\mathscr{X})$ and $\mathrm{L}_{\mathbb{A}^1} J((\mathrm{L}_{\mathbb{A}^1}\mathscr{X})^{\wedge 2}) \simeq \mathrm{L}_{\mathbb{A}^1} J(\mathscr{X}^{\wedge 2})$.

Since the unstable $\mathbb{A}^1$–connectivity property holds for S, the sheaves $\boldsymbol{\pi}_i^{\mathbb{A}^1}(\mathscr{X})$ are strictly $\mathbb{A}^1$–invariant by Lemma 2.2.11(1). Then Theorem 2.3.8 implies that

$$\boldsymbol{\pi}_i(\mathscr{F}) \cong \boldsymbol{\pi}_i^{\mathbb{A}^1}(\mathscr{F}) \cong \boldsymbol{\pi}_i(\mathrm{L}_{\mathbb{A}^1}\mathscr{F}) \cong \boldsymbol{\pi}_i(\mathrm{L}_{\mathbb{A}^1}\mathscr{X}) = \boldsymbol{\pi}_i^{\mathbb{A}^1}(\mathscr{X}) \quad \text{for } 1 \leq i \leq 3n-3.$$

These observations suffice to establish exactness everywhere except the leftmost part of the long exact sequence.

The map $\phi\colon \mathrm{L}_{\mathbb{A}^1}\mathscr{X} \to \mathscr{F}$ is simplicially $(3n-2)$–connected, and since $n \geq 2$, the connectivity of $\mathscr{X}$ implies that $\boldsymbol{\pi}_0^{\mathbb{A}^1}(\Omega\mathscr{F}) \simeq *$, by means of Theorem 2.2.12 for example. Thus, we can apply Corollary 2.3.6 to conclude that $\mathrm{L}_{\mathbb{A}^1}\phi\colon \mathrm{L}_{\mathbb{A}^1}\mathscr{X} \to \mathrm{L}_{\mathbb{A}^1}\mathscr{F}$ is also simplicially $(3n-2)$–connected. Therefore, there is a surjective map $\phi_*\colon \boldsymbol{\pi}_{3n-2}^{\mathbb{A}^1}(\mathscr{X}) \twoheadrightarrow \boldsymbol{\pi}_{3n-2}^{\mathbb{A}^1}(\mathscr{F})$ factoring $\mathrm{E}\colon \boldsymbol{\pi}_{3n-2}^{\mathbb{A}^1}(\mathscr{X}) \twoheadrightarrow \boldsymbol{\pi}_{3n-2}^{\mathbb{A}^1}(J(\mathscr{X}))$, yielding the exactness of the long exact sequence at the left as well. □

Remark 3.2.3 Assume the unstable $\mathbb{A}^1$–connectivity property holds for S. If $\mathscr{X}$ is a simplicially $(n-1)$–connected space, then $J(\mathrm{L}_{\mathbb{A}^1}\mathscr{X}^{\wedge 2})$ is at least $\mathbb{A}^1$–$(2n-1)$–connected by Theorem 2.2.12. Theorem 3.2.1 is therefore a refinement of Morel's suspension theorem, [44, Theorem 6.61].

3.3 Analyzing the $\mathbb{A}^1$–EHP sequence in low degrees

The goal of this section is to study the low-degree portion of the EHP sequence in $\mathbb{A}^1$–algebraic topology. To do this, given an $\mathbb{A}^1$–$(n-1)$–connected space $\mathscr{X}$, we will show that $\mathscr{X}^{\wedge 2}$ is at least $\mathbb{A}^1$–$(2n-1)$–connected, identify the first nonvanishing $\mathbb{A}^1$–homotopy sheaf of $\mathscr{X}^{\wedge 2}$ and use this to give a more explicit form of the EHP sequence in the first degree in which the suspension map is not an isomorphism. Granted the results of previous sections, and some results about $\mathbb{A}^1$–homology recalled below, the argument is a straightforward translation of a classical argument due to J H C Whitehead [60, Theorem 2] in the case of spheres and more generally by P Hilton [24, Theorem 2.1].

Some connectivity estimates Suppose $(\mathscr{X}, x)$ and $(\mathscr{Y}, y)$ are two pointed spaces. We will assume that $\mathscr{X}$ is $\mathbb{A}^1$–$(m-1)$–connected, and $\mathscr{Y}$ is $\mathbb{A}^1$–$(n-1)$–connected. Without loss of generality, we will assume that $m \leq n$.

Lemma 3.3.1 *Assume the unstable $\mathbb{A}^1$–connectivity property holds over S. The wedge sum $\mathscr{X} \vee \mathscr{Y}$ is at least $\mathbb{A}^1$–$(m-1)$–connected, and the smash product $\mathscr{X} \wedge \mathscr{Y}$ is at least $\mathbb{A}^1$–$(m+n-1)$–connected.*

Proof For the first statement, observe that the map $\mathscr{X} \vee \mathscr{Y} \to \mathrm{L}_{\mathbb{A}^1}\mathscr{X} \vee \mathrm{L}_{\mathbb{A}^1}\mathscr{Y}$ is an $\mathbb{A}^1$–weak equivalence by [45, Section 2, Lemma 2.11]. Since taking stalks commutes with coproducts, we conclude that the stalks of $\mathrm{L}_{\mathbb{A}^1}\mathscr{X} \vee \mathrm{L}_{\mathbb{A}^1}\mathscr{Y}$ are at least $(m-1)$–connected. Under the hypotheses, Theorem 2.2.12 implies that $\mathscr{X} \vee \mathscr{Y}$ is at least $(m-1)$–connected.

The second statement is established similarly. By two applications of [45, Section 3, Lemma 2.13] we can conclude that the map $\mathscr{X} \wedge \mathscr{Y} \to \mathrm{L}_{\mathbb{A}^1}\mathscr{X} \wedge \mathrm{L}_{\mathbb{A}^1}\mathscr{Y}$ is an $\mathbb{A}^1$–weak equivalence. Again by checking on stalks, and using the unstable $\mathbb{A}^1$–connectivity theorem one concludes that $\mathscr{X} \wedge \mathscr{Y}$ is at least $\mathbb{A}^1$–$(m+n-1)$–connected. □

$\mathbb{A}^1$–homology The $\mathbb{A}^1$–derived category may be constructed as a left Bousfield localization of the derived category of presheaves of abelian groups on Sm_S with respect to a notion of $\mathbb{A}^1$–quasi-isomorphism [44, Section 6.2].[1]

Morel gives a construction of an $\mathbb{A}^1$–localization functor $\mathrm{L}^{\mathrm{ab}}_{\mathbb{A}^1}$ [44, Lemma 6.18]; this functor is an endofunctor of the category of chain complexes of Nisnevich sheaves of abelian groups, and there is a natural transformation $\theta\colon \mathrm{id} \to \mathrm{L}^{\mathrm{ab}}_{\mathbb{A}^1}$ such that for any complex A, there is a quasi-isomorphism $A \to \mathrm{L}^{\mathrm{ab}}_{\mathbb{A}^1}(A)$ with target that is fibrant and $\mathbb{A}^1$–local.

[1] Morel works with the derived category of Nisnevich sheaves of abelian groups, but the exact functor of Nisnevich sheafification induces a Quillen equivalence between the model we use and Morel's model.

Notation 3.3.2 If $\mathscr{X}$ is a space, then we consider $C_*(\mathscr{X})$, the normalized chain complex associated with the simplicial presheaf of free abelian groups $\mathbb{Z}(\mathscr{X})$. The $\mathbb{A}^1$–singular chain complex of $\mathscr{X}$ is the complex $\mathrm{L}^{\mathrm{ab}}_{\mathbb{A}^1} C_*(\mathscr{X})$, which we may also denote $C^{\mathbb{A}^1}_*(\mathscr{X})$. The structure morphism $\mathscr{X} \to S$ induces a morphism $C^{\mathbb{A}^1}_*(\mathscr{X}) \to C^{\mathbb{A}^1}_*(S)$, and we define $\widetilde{C}^{\mathbb{A}^1}_*(\mathscr{X})$ as the kernel of this morphism.

The $\mathbb{A}^1$–homology sheaves of $\mathscr{X}$, denoted $\boldsymbol{H}^{\mathbb{A}^1}_i(\mathscr{X})$, are defined as the Nisnevich sheafifications of the homology presheaves $H_i(\mathrm{L}^{\mathrm{ab}}_{\mathbb{A}^1} C_*(\mathscr{X}))$. If A is a complex of presheaves of abelian groups, we will abuse notation and define $\boldsymbol{H}^{\mathbb{A}^1}_i(A)$ to be the Nisnevich sheafification of the homology presheaf $H_i(\mathrm{L}^{\mathrm{ab}}_{\mathbb{A}^1} A)$. We define $\widetilde{\boldsymbol{H}}^{\mathbb{A}^1}_i(\mathscr{X})$ as $\ker(\boldsymbol{H}^{\mathbb{A}^1}_i(\mathscr{X}) \to \boldsymbol{H}^{\mathbb{A}^1}_i(S))$.

The Dold–Kan adjunction shows that the Eilenberg–MacLane space associated with an $\mathbb{A}^1$–local complex is an $\mathbb{A}^1$–local space [15, Proposition 4 and (3.5)]. Note, however, that the ordinary singular chain complex of an $\mathbb{A}^1$–local space can *fail* to be $\mathbb{A}^1$–local (the standard counterexample is $\mathbb{G}_m$). The following property is the analog of the unstable $\mathbb{A}^1$–connectivity property of Definition 2.2.6 and was studied in [43, Section 6.2] in the closely related context of S^1–spectra.

Definition 3.3.3 The *stable $\mathbb{A}^1$–connectivity property holds for S* if $\mathrm{L}^{\mathrm{ab}}_{\mathbb{A}^1}$ preserves (-1)–connected complexes.

Theorem 3.3.4 [43, Theorem 6.1.8] *The stable $\mathbb{A}^1$–connectivity property holds for the spectrum of a field.*[2]

Remark 3.3.5 Ayoub [8] has shown that if S is a Noetherian scheme of Krull dimension $d \geq 2$, then the stable $\mathbb{A}^1$–connectivity property may fail for S in a very strong sense. Ayoub's counterexample is constructed in Voevodsky's derived category of motives. As noted above, if a complex of sheaves of abelian groups is $\mathbb{A}^1$–local, then the associated Eilenberg–MacLane space is $\mathbb{A}^1$–local as well. Therefore, Ayoub's counterexample can be transported to yield a counterexample to the unstable $\mathbb{A}^1$–connectivity property over S.

If the stable $\mathbb{A}^1$–connectivity property holds, the $\mathbb{A}^1$–derived category has a number of very nice properties. Write Ab_S for the category of Nisnevich sheaves of abelian groups on Sm_S, and $\mathrm{Ab}^{\mathbb{A}^1}_S$ for the full subcategory of strictly $\mathbb{A}^1$–invariant sheaves. Before proceeding, we introduce the following notation.

[2] For k finite, use the result of [26] for the same reason as in the proof of Theorem 2.2.7.

Notation 3.3.6 Given two strictly $\mathbb{A}^1$–invariant sheaves, set

$$\boldsymbol{A} \otimes^{\mathbb{A}^1} \boldsymbol{B} := \boldsymbol{H}_0^{\mathbb{A}^1}(\boldsymbol{A} \otimes^{\mathrm{L}} \boldsymbol{B}).$$

Remark 3.3.7 The unit object for the $\mathbb{A}^1$–tensor product is the strictly $\mathbb{A}^1$–invariant sheaf $\mathbb{Z}$.

With this notation, the following result holds.

Lemma 3.3.8 [43, Lemma 6.2.13] *If the stable $\mathbb{A}^1$–connectivity property holds over S, then $\mathrm{Ab}_S^{\mathbb{A}^1}$ is an abelian category and the inclusion functor $\mathrm{Ab}_S^{\mathbb{A}^1} \to \mathrm{Ab}_S$ is an exact embedding. Moreover, the bifunctor $(\boldsymbol{A}, \boldsymbol{B}) \mapsto \boldsymbol{A} \otimes^{\mathbb{A}^1} \boldsymbol{B}$ equips the category $\mathrm{Ab}_S^{\mathbb{A}^1}$ with a symmetric monoidal structure.*

The next result is closely related to [43, Remark 6.2.6] (apply that remark to shifted suspension spectra of suitably highly connected pointed spaces).

Proposition 3.3.9 *Assume the unstable and stable $\mathbb{A}^1$–connectivity properties hold for S, and suppose m, n are integers ≥ 1. If $\mathscr{X}$ is $\mathbb{A}^1$–$(m-1)$–connected, and $\mathscr{Y}$ is $\mathbb{A}^1$–$(n-1)$–connected, there are canonical isomorphisms*

$$\begin{aligned}&\tilde{\boldsymbol{H}}_i^{\mathbb{A}^1}(\mathscr{X} \times \mathscr{Y}) \\ &\quad \xrightarrow{\sim} \begin{cases} \tilde{\boldsymbol{H}}_i^{\mathbb{A}^1}(\mathscr{X}) \oplus \tilde{\boldsymbol{H}}_i^{\mathbb{A}^1}(\mathscr{Y}) & \text{if } 0 \leq i \leq m+n-1, \\ (\tilde{\boldsymbol{H}}_m^{\mathbb{A}^1}(\mathscr{X}) \otimes^{\mathbb{A}^1} \tilde{\boldsymbol{H}}_n^{\mathbb{A}^1}(\mathscr{Y})) \oplus \tilde{\boldsymbol{H}}_{m+n}^{\mathbb{A}^1}(\mathscr{X}) \oplus \tilde{\boldsymbol{H}}_{m+n}^{\mathbb{A}^1}(\mathscr{Y}) & \text{if } i = m+n. \end{cases}\end{aligned}$$

Proof Consider the inclusion map

$$\mathscr{X} \vee \mathscr{Y} \longrightarrow \mathscr{X} \times \mathscr{Y}.$$

The cone of this inclusion map is $\mathscr{X} \wedge \mathscr{Y}$. Note also that after a single suspension, the inclusion map is split by the projection. As a consequence, there are direct sum decompositions of the form

$$\tilde{\boldsymbol{H}}_i^{\mathbb{A}^1}(\mathscr{X} \times \mathscr{Y}) \cong \tilde{\boldsymbol{H}}_i^{\mathbb{A}^1}(\mathscr{X}) \oplus \tilde{\boldsymbol{H}}_i^{\mathbb{A}^1}(\mathscr{Y}) \oplus \tilde{\boldsymbol{H}}_i^{\mathbb{A}^1}(\mathscr{X} \wedge \mathscr{Y}).$$

Under the assumption that the unstable and stable $\mathbb{A}^1$–connectivity theorems hold for S, Lemma 3.3.1 together with the usual $\mathbb{A}^1$–Hurewicz theorem [44, Theorem 6.57] immediately imply the result for $i \leq m+n-1$ (note that Morel's Hurewicz theorem holds in this context under the assumptions given: simply replace the appeal to [44, Theorem 6.56] in Morel's proof by an appeal to Corollary 2.3.6).

It remains to treat the case $i = m+n$. In that case, let $\tilde{C}_*^{\mathbb{A}^1}(\mathscr{X})$ and $\tilde{C}_*^{\mathbb{A}^1}(\mathscr{Y})$ be the $\mathbb{A}^1$–chain complexes of $\mathscr{X}$ and $\mathscr{Y}$; recall that these are obtained by taking the chain complex associated with the free abelian group on $\mathscr{X}$ and then $\mathbb{A}^1$–localizing the result. By replacing $\widetilde{C}_*^{\mathbb{A}^1}(\mathscr{X})$ by a shift, we may assume $n=0$ and similarly for $\mathscr{Y}$ and $m=0$. The complex $\widetilde{C}_*^{\mathbb{A}^1}(\mathscr{X}) \otimes^{\mathrm{L}} \widetilde{C}_*^{\mathbb{A}^1}(\mathscr{Y})$ is also concentrated in degrees ≥ 0, and since the stable $\mathbb{A}^1$–connectivity property holds for S it follows that $\mathrm{L}^{\mathrm{ab}}_{\mathbb{A}^1}(\widetilde{C}_*^{\mathbb{A}^1}(\mathscr{X}) \otimes^{\mathrm{L}} \widetilde{C}_*^{\mathbb{A}^1}(\mathscr{Y}))$ is concentrated in degrees ≥ 0. Since the zeroth homology is obtained by truncation with respect to the homotopy t–structure, it follows that

$$\boldsymbol{H}_0^{\mathbb{A}^1}(\widetilde{C}_*^{\mathbb{A}^1}(\mathscr{X}) \otimes^{\mathrm{L}} \widetilde{C}_*^{\mathbb{A}^1}(\mathscr{Y})) \cong \boldsymbol{H}_0^{\mathbb{A}^1}(\mathscr{X}) \otimes^{\mathbb{A}^1} \boldsymbol{H}_0^{\mathbb{A}^1}(\mathscr{Y}).$$

To conclude it remains to identify the left-hand side in terms of the smash product $\mathscr{X} \wedge \mathscr{Y}$. For this, it suffices to observe that the Eilenberg–Zilber theorem implies the existence of an isomorphism of the form $\widetilde{C}_*^{\mathbb{A}^1}(\mathscr{X}) \otimes^{\mathrm{L}} \widetilde{C}_*^{\mathbb{A}^1}(\mathscr{Y}) \cong \widetilde{C}_*^{\mathbb{A}^1}(\mathscr{X} \wedge \mathscr{Y})$. □

Corollary 3.3.10 *Assume the unstable and stable $\mathbb{A}^1$–connectivity properties hold for S and suppose m, n are integers ≥ 2. If $(\mathscr{X}, x)$ is a pointed $\mathbb{A}^1$–$(m-1)$–connected space and $(\mathscr{Y}, y)$ is a pointed $\mathbb{A}^1$–$(n-1)$–connected space, then there is a canonical isomorphism*

$$\boldsymbol{\pi}_{n+m}^{\mathbb{A}^1}(\mathscr{X} \wedge \mathscr{Y}) \xrightarrow{\sim} \boldsymbol{\pi}_m^{\mathbb{A}^1}(\mathscr{X}) \otimes^{\mathbb{A}^1} \boldsymbol{\pi}_n^{\mathbb{A}^1}(\mathscr{Y}).$$

Proof By Lemma 3.3.1, we know $\mathscr{X} \wedge \mathscr{Y}$ is $\mathbb{A}^1$–$(m+n-1)$–connected. By the $\mathbb{A}^1$–Hurewicz theorem [44, Theorem 6.57], it suffices to prove the result in $\mathbb{A}^1$–homology (again, as in the proof of Proposition 3.3.9, this holds under our assumption that the unstable and stable $\mathbb{A}^1$–connectivity properties hold). In that case, it follows immediately from the proof of Proposition 3.3.9. □

The connection with the results of Hilton and Whitehead mentioned at the beginning of this section is contained in the next result, which describes the first "nonlinear" $\mathbb{A}^1$–homotopy sheaf of a wedge sum. Given the above results, the proof is a direct consequence of the $\mathbb{A}^1$–homotopy excision theorem (aka Blakers–Massey theorem;[3] see eg [6, Theorem 3.1], [51, Theorem 2.3.8] or [61, Proposition 2.20]) and is left to the reader.

Corollary 3.3.11 *Assume the unstable and stable $\mathbb{A}^1$–connectivity properties hold for S and suppose $m, n \geq 2$ are integers. Suppose $\mathscr{X}$ is a pointed $\mathbb{A}^1$–$(m-1)$–connected space and $\mathscr{Y}$ is a pointed $\mathbb{A}^1$–$(n-1)$–connected space. There are canonical*

[3]By inspecting the proof, one sees that this result is a direct consequence of the relative connectivity theorem (Corollary 2.3.6) and therefore holds over any base scheme S for which the unstable $\mathbb{A}^1$–connectivity property holds.

isomorphisms

$$\boldsymbol{\pi}_i^{\mathbb{A}^1}(\mathscr{X} \vee \mathscr{Y}) \xrightarrow{\sim} \begin{cases} \boldsymbol{\pi}_i^{\mathbb{A}^1}(\mathscr{X}) \oplus \boldsymbol{\pi}_i^{\mathbb{A}^1}(\mathscr{X}) & \text{if } 1 \le i \le m+n-2, \\ \boldsymbol{\pi}_{m+n-1}^{\mathbb{A}^1}(\mathscr{X}) \oplus \boldsymbol{\pi}_{m+n-1}^{\mathbb{A}^1}(\mathscr{X}) \oplus \boldsymbol{\pi}_m^{\mathbb{A}^1}(\mathscr{X}) \otimes^{\mathbb{A}^1} \boldsymbol{\pi}_n^{\mathbb{A}^1}(\mathscr{Y}) & \text{when } i = m+n-1. \end{cases}$$

Remark 3.3.12 As in classical homotopy theory, the computation of the homotopy of a wedge sum allows one to study homotopy operations. The first "nonlinear" summand in the homotopy of a wedge sum is closely related to the Whitehead product studied in Section 4.1, though we have not attempted to establish equivalence of the definitions.

The EHP sequence in low degrees

Theorem 3.3.13 *Assume the unstable and stable $\mathbb{A}^1$–connectivity properties hold for S. Let $n \geq 2$ be an integer. If $\mathscr{X}$ is an $\mathbb{A}^1$–$(n-1)$–connected space, then there is an exact sequence of the form*

$$\boldsymbol{\pi}_{2n+1}^{\mathbb{A}^1}(\Sigma\mathscr{X}) \xrightarrow{\mathrm{H}} \boldsymbol{\pi}_n^{\mathbb{A}^1}(\mathscr{X}) \otimes^{\mathbb{A}^1} \boldsymbol{\pi}_n^{\mathbb{A}^1}(\mathscr{X}) \xrightarrow{\mathrm{P}} \boldsymbol{\pi}_{2n-1}^{\mathbb{A}^1}(\mathscr{X}) \xrightarrow{\mathrm{E}} \boldsymbol{\pi}_{2n}^{\mathbb{A}^1}(\Sigma\mathscr{X}) \longrightarrow 0.$$

In particular, one has an exact sequence as above if S is the spectrum of an (infinite) perfect field.

Proof Consider the exact sequence of Theorem 3.2.1. Lemma 3.3.1 implies that $\Sigma\mathscr{X} \wedge \mathscr{X}$ is at least $\mathbb{A}^1$–$2n$–connected, and thus $J(\mathscr{X} \wedge \mathscr{X})$ is at least $\mathbb{A}^1$–$(2n-1)$–connected. This immediately yields the surjectivity in the statement. Corollary 3.3.10 then yields the identification of $\boldsymbol{\pi}_{2n}^{\mathbb{A}^1}(J(\mathscr{X} \wedge \mathscr{X}))$ with the $\mathbb{A}^1$–tensor product term. The final statement is a consequence of Theorems 2.2.7 and 3.3.4. □

Remark 3.3.14 The exact sequence of Theorem 3.3.13 when $\mathscr{X} = \mathbb{A}^3 \setminus 0$ is precisely the one described in [2, Theorem 4]. One notational benefit of the statement of Theorem 3.3.13 is that the quadratic nature of the James–Hopf invariants is apparent.

4 Some E_1 differentials in the $\mathbb{A}^1$–EHP sequence

The goal of this section is to analyze the morphisms in the $\mathbb{A}^1$–EHP sequence. As mentioned in the introduction, classically, the morphism P can be described in terms of Whitehead products. In Section 4.1, we extend the definition of Whitehead product to the theory of simplicial presheaves (see Definition 4.1.2) to make it available in the $\mathbb{A}^1$–homotopy category as well. These results are written in the generality of simplicial presheaves on a site with enough points. We then use the results of Section 4.1 to show that the map P in Theorem 3.2.1 can indeed be expressed in terms of the Whitehead product (see Theorem 4.2.1); this requires that the unstable $\mathbb{A}^1$–connectivity property holds for S.

The $\mathbb{A}^1$–EHP spectral sequence is created by combining $\mathbb{A}^1$–EHP sequences into an exact couple. However, since the $\mathbb{A}^1$–EHP sequences of Theorem 3.2.1 are truncated, some algebraic manipulation is required to form an exact couple (for example extending the sequences to the left by a kernel and then zeros), and the resulting spectral sequence will differ from the $\mathbb{A}^1$–EHP spectral sequence. Nevertheless, it is shown in [61] that after localizing at 2, the exact sequences of Theorem 3.2.1 are "low-degree portions" of suitable long exact sequences, and these long exact sequences yield the 2–primary $\mathbb{A}^1$–EHP sequence (with the expected convergence properties).

The analysis of morphisms in the $\mathbb{A}^1$–EHP sequence described above can be used to describe some differentials on the E_1 page in the $\mathbb{A}^1$–EHP spectral sequence. The desired E_1–differentials (given by the composite HP linking the EHP sequences of different spheres) are then determined by the James–Hopf invariant of a Whitehead product. The axiomatic approach to Hopf invariants of Boardman and Steer [12] determines these James–Hopf invariants. In Section 4.3, we recast some results of Boardman and Steer in the context of simplicial presheaves. The main result is Proposition 4.3.5, which holds in the generality of simplicial presheaves on a site with enough points. In contrast, the remaining Section 4.4 is more specific to $\mathbb{A}^1$–homotopy theory; Theorem 4.4.1 identifies an E_1–differential in the $\mathbb{A}^1$–EHP sequence with multiplication by a given element of $\mathrm{GW}(k)$.

4.1 Whitehead products for simplicial presheaves

In this section, we give a construction of Whitehead products in $\mathbb{A}^1$–homotopy theory, the construction generalizes classical results of [13; 1] to the context of simplicial presheaves. Suppose $\mathbf{C}$ is a site and as in Section 2.1 consider the category of pointed simplicial presheaves on $\mathbf{C}$ with its injective local model structure. If X and Y are pointed simplicial presheaves, we write $[X, Y]$ for morphisms in the associated (pointed) homotopy category. Unfortunately, it is also standard to use the notation $[-,-]$ for Whitehead products, but we hope that context will ensure that no confusion arises.

Recall that the constant simplicial presheaf S^1 is an H–cogroup object in the category of pointed simplicial presheaves on $\mathbf{C}$, and there is an induced H–cogroup structure on ΣW for any pointed space W. In particular, for any space Z, the space $\mathrm{Map}(\Sigma W, Z)$ has the structure of an H–group, functorially in both Z and W. We will write $\cdot$ for the product in $\mathrm{Map}(\Sigma W, Z)$ and $(-)^{-1}$ for the inversion map; the constant map to the basepoint $*$ serves as a unit.

Suppose given pointed simplicial presheaves X, Y and Z. The product projections $p_X\colon X \times Y \to X$ and $p_Y\colon X \times Y \to Y$ are pointed, and they induce morphisms $\Sigma p_X\colon \Sigma(X \times Y) \to \Sigma X$ and $\Sigma p_Y\colon \Sigma(X \times Y) \to \Sigma Y$. In addition, the canonical map $X \vee Y \to X \times Y$ induces a morphism $\Sigma(X \vee Y) \simeq \Sigma X \vee \Sigma Y \to \Sigma(X \times Y)$ that fits into a cofiber sequence with cofiber $\Sigma(X \wedge Y)$.

Construction 4.1.1 (Whitehead product) Given maps $\alpha\colon \Sigma X \to Z$ and $\beta\colon \Sigma Y \to Z$, composition with the projections yields morphisms $a := \alpha \circ \Sigma p_X$ and $b := \beta \circ \Sigma p_Y$. With respect to the product structure on $\mathrm{Map}(\Sigma(X \times Y), Z)$, we may consider the map

$$(a^{-1} \cdot b^{-1}) \cdot (a \cdot b)\colon \Sigma(X \times Y) \longrightarrow Z.$$

We embed the map $(a^{-1} \cdot b^{-1}) \cdot (a \cdot b)$ into the following diagram:

$$\begin{tikzcd}
\Sigma X \vee \Sigma Y \arrow[r] \arrow[dr] & \Sigma(X \times Y) \arrow[r] \arrow[d, "(a^{-1}\cdot b^{-1})\cdot(a\cdot b)"] & \Sigma(X \wedge Y) \\
& Z &
\end{tikzcd}$$

The pullback of $(a^{-1} \cdot b^{-1}) \cdot (a \cdot b)$ to $\Sigma(X \vee Y)$ has a prescribed null-homotopy described as follows. The composition of the inclusion $\Sigma(X \times *) \to \Sigma(X \times Y)$ with π_Y is the constant map. Thus, if we pull back $(a^{-1} \cdot b^{-1}) \cdot (a \cdot b)$ to $\Sigma(X \times *)$ the result coincides with the pullback of $(a^{-1} \cdot *^{-1}) \cdot (a \cdot *)$. There is a canonical homotopy between $(a^{-1} \cdot *^{-1}) \cdot (a \cdot *)$ and the constant map $*$. Switching the roles of X and Y and a and b, the pullback of $(a^{-1} \cdot b^{-1}) \cdot (a \cdot b)$ to $\Sigma(* \times Y)$ also admits a specified null-homotopy. Thus, the pullback of $(a^{-1} \cdot b^{-1}) \cdot (a \cdot b)$ to $\Sigma(X \vee Y)$ comes equipped with a specified null-homotopy. By means of this null-homotopy, the map $(a^{-1} \cdot b^{-1}) \cdot (a \cdot b)$ passes to a well-defined homotopy class of maps $\Sigma(X \wedge Y) \to Z$; we write $[\alpha, \beta]$ for any representative of this class.

Since the sequence $1 \to [\Sigma(X \wedge Y), Z] \to [\Sigma(X \times Y), Z] \to [\Sigma(X \vee Y)] \to 1$ is exact, the choice of null-homotopy does not affect the homotopy class of $[\alpha, \beta]$.

Definition 4.1.2 Given maps $\alpha\colon \Sigma X \to Z$ and $\beta\colon \Sigma Y \to Z$, a representative for the homotopy class of maps

$$[\alpha, \beta]\colon \Sigma(X \wedge Y) \longrightarrow Z.$$

in Construction 4.1.1 (or the homotopy class itself) is a *Whitehead product* of α and β.

Following classical conventions, we write ι_X for the identity map on a (pointed) simplicial presheaf X, and by a slight abuse of notation, we also let $\iota_{\Sigma X}$ denote the inclusion $\Sigma X \to \Sigma X \vee \Sigma Y$. The construction above with $\alpha = \iota_{\Sigma X}$ and $\beta = \iota_{\Sigma Y}$ also yields a canonical map

$$[\iota_{\Sigma X}, \iota_{\Sigma Y}]\colon \Sigma(X \wedge Y) \longrightarrow \Sigma X \vee \Sigma Y$$

that can be thought of as a universal Whitehead product in the sense that the Whitehead product of Definition 4.1.2 can be obtained by composing $[\iota_{\Sigma X}, \iota_{\Sigma Y}]$ with the map $\alpha \vee \beta\colon \Sigma X \vee \Sigma Y \to Z$. Regarding this product, we have the following result, which is a straightforward consequence of [1, Theorem 4.2].

To state this result, introduce the following notation. For X and Y pointed spaces, let $\tau\colon \Sigma\Sigma(X \wedge Y) \xrightarrow{\sim} \Sigma\Sigma(X \wedge Y)$ denote the map which switches the two suspensions and is the identity on $X \wedge Y$. Let $\wp\colon \Sigma\Sigma(X \wedge Y) \xrightarrow{\sim} \Sigma X \wedge \Sigma Y$ be the map which does not change the order of the suspensions and is the identity on X and Y.

Lemma 4.1.3 *If X and Y are pointed connected spaces, then there is a cofiber sequence of the form*

$$\Sigma(X \wedge Y) \xrightarrow{[\iota_{\Sigma X}, \iota_{\Sigma Y}]} \Sigma X \vee \Sigma Y \longrightarrow \Sigma X \times \Sigma Y,$$

where the second map is the usual map from the sum to the product, and such that the induced weak equivalence

$$\Sigma^2(X \wedge Y) \xrightarrow{\sim} \Sigma X \wedge \Sigma Y$$

is homotopic to $\wp\tau$.

Proof Recall that if X and Y are two pointed spaces, their join, typically denoted $X * Y$, is the homotopy pushout of the diagram $X \leftarrow X \times Y \to Y$. There is a functorial sectionwise weak equivalence $X * Y \simeq \Sigma X \wedge Y$. Using this identification, the universal Whitehead product can be thought of as a map with source $X * Y \to \Sigma X \vee \Sigma Y$ (see [1, Definition 2.3]).

We first treat the case of simplicial sets. Thus, suppose A and B are connected simplicial sets. Let $C\big(A * B \to \Sigma A \vee \Sigma B\big)$ be the reduced mapping cone of $[\iota_{\Sigma A}, \iota_{\Sigma B}]$. In the proof of [1, Theorem 4.2], one finds a natural map

$$\Theta\colon C\big(A * B \to \Sigma A \vee \Sigma B\big) \longrightarrow \Sigma A \times \Sigma B$$

(this map is called G in [1] and is originally due to D E Cohen [13]). We claim the map Θ is a homology isomorphism. In [1, Theorem 4.2], this is fact is established for A and B polyhedra with one of A or B compact. As a consequence, it is true for finite simplicial sets. Since homology commutes with filtered direct limits and since the map Θ is evidently compatible with passing to sub-simplicial-sets by inspection, it follows that Θ is a homology isomorphism for A and B arbitrary simplicial sets.

Now, we treat the general case of simplicial presheaves. Write $C([\iota_{\Sigma X}, \iota_{\Sigma Y}])$ for the cofiber of $[\iota_{\Sigma X}, \iota_{\Sigma Y}]\colon X * Y \to \Sigma X \vee \Sigma Y$. It follows that there is a map $\Theta\colon C([\iota_{\Sigma X}, \iota_{\Sigma Y}]) \to \Sigma X \times \Sigma Y$ defined sectionwise, ie for each object $U \in \mathbf{C}$ define $\Theta(U)\colon C([\iota_{\Sigma X}, \iota_{\Sigma Y}])(U) \to \Sigma X(U) \times \Sigma Y(U)$ to be the map above. For each such U, the map $\Theta(U)$ is a homology equivalence between simply connected simplicial sets, and therefore a sectionwise weak equivalence. Combining with the sectionwise weak equivalence $X * Y \to \Sigma(X \wedge Y)$ and using the compatibility of the definitions of the Whitehead product we have established the claimed cofiber sequence.

The cofiber sequence identifies the suspension of $\Sigma(X \wedge Y)$ with the homotopy cofiber of $\Sigma X \vee \Sigma Y \to \Sigma X \times \Sigma Y$. To prove that the homotopy class of the resulting map

$$\Sigma^2(X \wedge Y) \xrightarrow{\sim} \Sigma X \wedge \Sigma Y \tag{4.1.1}$$

is $\wp\tau$, by working sectionwise it suffices to establish the analogous claim in the context of simplicial sets, which in turn can be reduced to checking the claim in the context of CW complexes, as considered in [1].

We recall the following constructions from [1]. Let S denote the unreduced suspension, T denote the unreduced cone, and C denote the reduced cone. Let A and B be pointed, locally finite, connected, CW complexes. The map Θ induces a map

$$C\big(A * B \to \Sigma A \vee \Sigma B\big)/(\Sigma A \vee \Sigma B) \longrightarrow (\Sigma A \times \Sigma B)/(\Sigma A \vee \Sigma B). \tag{4.1.2}$$

By construction [1, Theorem 4.2, Lemma 4.1], (4.1.2) is induced from a map of pairs

$$(T(A * B), A * B) \longrightarrow (\Sigma A \times \Sigma B, \Sigma A \vee \Sigma B), \tag{4.1.3}$$

constructed as follows. Define

$$N\colon (T(A * B), A * B) \longrightarrow (TA \times TB, TA \times B \cup A \times TB)$$

by

$$N(u,(t,a,b)) = \begin{cases} (u,a) \times \big(1-2t(1-u), b\big) & \text{if } 0 \le t \le \frac{1}{2}, \\ \big(1-2(1-t)(1-u), a\big) \times (u,b) & \text{if } \frac{1}{2} \le t \le 1. \end{cases}$$

For a space W, let $t_W\colon TW \to SW$ and $s_W\colon SW \to \Sigma W$ be the quotient maps. Then

(4.1.3) is defined to be the composite:

$$\begin{tikzcd}
(T(A*B), A*B) \arrow[rr, "N"] & & (TA\times TB, TA\times B\cup A\times TB) \arrow[d, "t_A\times t_B"] \\
(\Sigma A\times \Sigma B, \Sigma A\vee \Sigma B) & & (SA\times SB, SA\vee SB) \arrow[ll, "S_A\times S_B"']
\end{tikzcd}$$

There is a weak equivalence $\mu'\colon A*B\to \Sigma A\wedge B$ given by quotienting by points of the forms $(a,*,t)$ and $(*,b,t)$, where $*$ denotes the base points, and t denotes the coordinate of the interval in the standard representation of the join. Let μ be a homotopy inverse. The map (4.1.1) in the homotopy category is $\overline{M}\circ\Sigma(\mu)$, where $\overline{M}$ denotes the map on quotient spaces associated to the map of pairs $(S_X\times S_Y)\circ(t_X\times t_Y)\circ N$. It therefore suffices to show that

$$(u,t)\longmapsto \begin{cases} \bigl(u, 1-2t(1-u)\bigr) & \text{if } 0\le t\le \frac{1}{2}, \\ \bigl(1-2(1-t)(1-u), u\bigr) & \text{if } \frac{1}{2}\le t\le 1 \end{cases}$$

determines an endomorphism of $S^1\wedge S^1$ of degree -1. This is easily checked: for example, the fiber over $\bigl(\frac{1}{4},\frac{3}{4}\bigr)$ consists of the single point $(u,t)=\bigl(\frac{1}{4},\frac{1}{6}\bigr)$ and the Jacobian determinant at the point $\bigl(\frac{1}{4},\frac{1}{6}\bigr)$ is negative. □

Proposition 4.1.4 *If Z is an h–space in the category of pointed simplicial presheaves on $\mathbf{C}$, then $[\alpha,\beta]=0$ for all $\alpha\in[\Sigma X,Z]$ and $\beta\in[\Sigma Y,Z]$.*

Proof If Z is an h–space, then the group $[\Sigma(X\times Y),Z]$ is an abelian group by the Eckmann–Hilton argument. Therefore, the commutator must be zero. □

4.2 On the map P in the $\mathbb{A}^1$–EHP sequence

In this section, we return to the setting of $\mathbb{A}^1$–homotopy theory and we analyze the map

$$\mathrm{P}\colon \boldsymbol{\pi}_{i+1}^{\mathbb{A}^1}(J(\mathscr{X}^{\wedge 2}))\longrightarrow \boldsymbol{\pi}_i^{\mathbb{A}^1}(\mathscr{X})$$

in the exact sequence of Theorem 3.2.1 under the additional assumption that the pointed space $\mathscr{X}$ is itself a suspension $\mathscr{X}=\Sigma\mathscr{Z}$ (see the beginning of Section 2.2 for a reminder regarding conventions). In direct analogy with the results mentioned in the introduction, the map P can be described in terms of the Whitehead product introduced in Definition 4.1.2; the main result is Theorem 4.2.1.

The map P was defined as the connecting homomorphism in the exact sequence of Theorem 3.2.1. This exact sequence does not arise directly from a fiber sequence, however. If $\mathscr{X}$ is as above, then we can recast the sequence of (3.2.1) as the homotopy commutative diagram:

$$\begin{tikzcd} \mathrm{L}_{\mathbb{A}^1}\mathscr{X} \arrow[r] \arrow[d, "\mathrm{E}"] & * \arrow[d] \\ \mathrm{L}_{\mathbb{A}^1} J(\mathscr{X}) \arrow[r, "\mathrm{H}"] & \mathrm{L}_{\mathbb{A}^1} J(\mathscr{X}^{\wedge 2}) \end{tikzcd}$$

By functoriality of homotopy fibers, there is an induced morphism

$$\mathrm{hofib}\big(\mathrm{L}_{\mathbb{A}^1}\mathscr{X} \to \mathrm{L}_{\mathbb{A}^1} J(\mathscr{X})\big) \longrightarrow \Omega\mathrm{L}_{\mathbb{A}^1} J(\mathscr{X}^{\wedge 2})$$

and the connectivity of this map is what allows us to define the map P in Theorem 3.2.1.

In the range where the map above is connected, it makes sense to consider the composite map

$$\boldsymbol{\pi}_i^{\mathbb{A}^1}(\Sigma\mathscr{Z}\wedge\mathscr{Z}) \longrightarrow \boldsymbol{\pi}_{i+1}^{\mathbb{A}^1}(J(\Sigma\mathscr{Z}\wedge\Sigma\mathscr{Z})) \xrightarrow{\mathrm{P}} \boldsymbol{\pi}_i^{\mathbb{A}^1}(\Sigma\mathscr{Z}). \tag{4.2.1}$$

Precisely, if $\mathscr{Z}$ is $(n-2)$–connected, then this composite is defined for $i \le 3n-3$. We shall furthermore see that, provided the $\mathbb{A}^1$–connectivity property holds for S and $n \ge 2$, the first map in (4.2.1) is in fact an isomorphism in the range being considered.

On the other hand, we saw in Section 4.1 that the Whitehead square of the identity $[\iota_{\Sigma\mathscr{Z}}, \iota_{\Sigma\mathscr{Z}}]$ gives a morphism $\Sigma\mathscr{Z}\wedge\mathscr{Z} \to \Sigma\mathscr{Z}$. We will abuse notation and write $[\iota_{\Sigma\mathscr{Z}}, \iota_{\Sigma\mathscr{Z}}]$ for the map

$$[\iota_{\Sigma\mathscr{Z}}, \iota_{\Sigma\mathscr{Z}}]\colon \mathrm{L}_{\mathbb{A}^1}\Sigma\mathscr{Z}\wedge\mathscr{Z} \longrightarrow \mathrm{L}_{\mathbb{A}^1}\Sigma\mathscr{Z}.$$

This morphism induces a pushforward map on homotopy sheaves

$$[\iota_{\Sigma\mathscr{Z}}, \iota_{\Sigma\mathscr{Z}}]_*\colon \boldsymbol{\pi}_i^{\mathbb{A}^1}(\Sigma\mathscr{Z}\wedge\mathscr{Z}) \longrightarrow \boldsymbol{\pi}_i^{\mathbb{A}^1}(\Sigma\mathscr{Z})$$

that we would like to compare with the map (4.2.1). The next result, which gives precisely such a comparison, provides an analog of [59, Theorem XII.2.4] or, rather, its extension to general $(n-1)$–connected spaces in the spirit of [20, Theorem 3.1 and page 231], in the context of unstable $\mathbb{A}^1$–homotopy theory.

Theorem 4.2.1 *Assume the unstable $\mathbb{A}^1$–connectivity property holds for S and suppose $n \ge 2$ is an integer. If $\mathscr{Z}$ is an $\mathbb{A}^1$–$(n-2)$–connected pointed space, and $\mathscr{X} = \Sigma\mathscr{Z}$, then for any positive integer $i \le 3n-4$ the composite morphism of* (4.2.1) *fits into a commutative diagram of the following form:*

$$\begin{tikzcd} \boldsymbol{\pi}_i^{\mathbb{A}^1}(\Sigma(\mathscr{Z}\wedge\mathscr{Z})) \arrow[r, "\sim"] \arrow[rr, bend right, "{[\iota_{\Sigma\mathscr{Z}},\iota_{\Sigma\mathscr{Z}}]_*}"'] & \boldsymbol{\pi}_{i+1}^{\mathbb{A}^1}(J(\mathscr{X}^{\wedge 2})) \arrow[r, "\mathrm{P}"] & \boldsymbol{\pi}_i^{\mathbb{A}^1}(\mathscr{X}) \end{tikzcd}$$

The isomorphism of homotopy sheaves in the diagram is induced by 2–fold suspension.

Proof Without loss of generality, assume $\mathscr{X}$ is $\mathbb{A}^1$–fibrant.

We begin with the cofiber sequence

$$\Sigma(\mathscr{X} \wedge \mathscr{X}) \xrightarrow{u} \Sigma\mathscr{X} \vee \Sigma\mathscr{X} \longrightarrow \Sigma\mathscr{X} \times \Sigma\mathscr{X}$$

of Lemma 4.1.3. Here u denotes the "universal" Whitehead product; the map

$$[\iota_{\Sigma\mathscr{X}}, \iota_{\Sigma\mathscr{X}}]\colon \Sigma(\mathscr{X} \wedge \mathscr{X}) \to \Sigma\mathscr{X}$$

is obtained by composing u with a fold map.

By construction, the map u is represented by $u\colon \mathscr{C} \to \Sigma\mathscr{X} \vee \Sigma\mathscr{X}$, where $\mathscr{C}$ denotes the reduced mapping cone of $\Sigma(\mathscr{X} \vee \mathscr{X}) \to \Sigma(\mathscr{X} \times \mathscr{X})$. We now consider the following commutative diagram:

$$\begin{tikzcd}
\mathscr{C} \arrow[r, "u"] \arrow[d, equal] & \Sigma\mathscr{X} \vee \Sigma\mathscr{X} \arrow[r] \arrow[d] & \Sigma\mathscr{X} \times \Sigma\mathscr{X} \arrow[r] \arrow[d] & \Sigma\mathscr{X} \wedge \Sigma\mathscr{X} \arrow[d, equal] \\
\mathscr{C} \arrow[r, "{[\iota_{\Sigma\mathscr{X}}, \iota_{\Sigma\mathscr{X}}]}"] & \Sigma\mathscr{X} \arrow[r] & J_2(\Sigma\mathscr{X}) \arrow[r] & \Sigma\mathscr{X} \wedge \Sigma\mathscr{X}
\end{tikzcd} \tag{4.2.2}$$

The vertical maps are, reading left to right, the identity, the fold map, the canonical map from the product to J_2 and the identity. The horizontal arrows $\Sigma\mathscr{X} \to J_2(\Sigma\mathscr{X})$ in the center are the canonical inclusions. We observe that the center square is a pushout, being the definition of $J_2(\Sigma\mathscr{X})$, and therefore the rightmost two horizontal maps are quotient maps. Since $\Sigma\mathscr{X} \vee \Sigma\mathscr{X} \to \Sigma\mathscr{X} \times \Sigma\mathscr{X}$ is a cofibration, this square is in fact a homotopy pushout square. The two rows are homotopy cofiber sequences.

We form the following diagram, where the upper row is a homotopy cofiber sequence and the lower row a fiber sequence:

$$\begin{tikzcd}
\Sigma\mathscr{X} \wedge \mathscr{X} \arrow[r, "{[\iota_{\Sigma\mathscr{X}}, \iota_{\Sigma\mathscr{X}}]}"] \arrow[d, dashed, "g"] & \Sigma\mathscr{X} \arrow[r] \arrow[d, "t"] & J_2(\Sigma\mathscr{X}) \arrow[r] \arrow[d] & \Sigma\mathscr{X} \wedge \Sigma\mathscr{X} \arrow[d, dashed, "f"] \\
\Omega J((\Sigma\mathscr{X})^{\wedge 2}) \arrow[r, "s"] & \operatorname{hofib} \mathrm{H} \arrow[r] & J(\Sigma\mathscr{X}) \arrow[r, "\mathrm{H}"] & J((\Sigma\mathscr{X})^{\wedge 2})
\end{tikzcd} \tag{4.2.3}$$

The maps indicated by the dashed arrows are adjoint to one another: by [48, Chapter I.3, proof of Proposition 6, 3.13] the map $\Sigma^2\mathscr{X} \wedge \mathscr{X} \to \Sigma\mathscr{X} \wedge \Sigma\mathscr{X} \xrightarrow{f} J((\Sigma\mathscr{X})^{\wedge 2})$ is inverse adjoint to g. By Lemma 4.1.3, the map $\Sigma^2\mathscr{X} \wedge \mathscr{X} \to \Sigma\mathscr{X} \wedge \Sigma\mathscr{X}$ is simplicially homotopic to $\wp\tau$, which reverses the order of the two suspensions, inducing -1 in the homotopy category.

The map $f\colon \Sigma\mathscr{X} \wedge \Sigma\mathscr{X} \to J((\Sigma\mathscr{X})^{\wedge 2})$ is given by the canonical map $\Sigma\mathscr{X} \times \Sigma\mathscr{X} \to J(\Sigma\mathscr{X})$ followed by H; an analysis of H (Definition 3.1.1) now shows that f is the suspension map E as applied to $\Sigma\mathscr{X} \wedge \Sigma\mathscr{X}$. Therefore, g is the 2–fold suspension map.

We may $\mathbb{A}^1$–localize throughout. We do not draw (4.2.3) a second time with $\mathrm{L}_{\mathbb{A}^1}$ prepended to all terms. Since all spaces appearing above are at least 1–connected, Corollary 2.3.6 applies throughout and maps that are n–connected remain n–connected after $\mathbb{A}^1$–localization. We apply $\pi_i^{\mathbb{A}^1}$ to the leftmost square to obtain a commuting square:

$$\begin{tikzcd}
\pi_i^{\mathbb{A}^1}(\Sigma\mathscr{X}\wedge\mathscr{X}) \arrow[r, "{[\iota_{\Sigma\mathscr{X}},\iota_{\Sigma\mathscr{X}}]}"] \arrow[d, "g_*"] & \pi_i^{\mathbb{A}^1}(\Sigma\mathscr{X}) \arrow[d, "t_*"] \\
\pi_{i+1}^{\mathbb{A}^1}(J((\Sigma\mathscr{X})^{\wedge 2})) \arrow[r, "s_*"] & \pi_i^{\mathbb{A}^1}(\mathrm{hofib}\,\mathrm{H})
\end{tikzcd}$$

The map g_*, the 2–fold suspension map, is an isomorphism when $i \le 4n-3$. When $i \le 3n-2$, the map t_* is an isomorphism, and by definition $\mathrm{P} = t_*^{-1} \circ s_*$; the result follows. □

4.3 The James–Hopf invariant of a Whitehead product

In this section, we collect some properties of the James–Hopf invariant defined in Section 3.1. Probably due to proliferation of different definitions of "Hopf invariants" made at the time, Boardman and Steer [12] made an axiomatic study of such invariants. We adapt some of their results to the context under consideration.

Let $\mathbf{C}$ be a site with enough points and consider simplicial presheaves on $\mathbf{C}$ equipped with the injective local model structure. For pointed simplicial presheaves X and Y, we continue to use the notation $[X, Y]$ for morphisms in the associated (pointed) homotopy category, and rely on context to distinguish this notation from that for the Whitehead product.

Cup products Before passing to the main results, it will be necessary to recall some constructions from [12].

Construction 4.3.1 (cup product [12, Definition 1.3]) Let X, Y and Z be pointed simplicial presheaves. The reduced diagonal map $X \to X \wedge X$ is the composite of the diagonal $X \to X \times X$ and the map $X \times X \to X \wedge X$; this map is null-homotopic if X is a suspension. The smash product induces a map

$$[\Sigma^m X, Y] \times [\Sigma^n X, Z] \longrightarrow [\Sigma^m X \wedge \Sigma^n X, Y \wedge Z].$$

The reduced diagonal then induces a morphism

$$\Sigma^{m+n} X \longrightarrow \Sigma^{m+n} X \wedge X \simeq \Sigma^m X \wedge \Sigma^n X;$$

the isomorphism on the right does not permute the suspension factors. The composite of these two morphisms defines the cup-product pairing

$$\smile\colon [\Sigma^m X, Y] \times [\Sigma^n X, Z] \longrightarrow [\Sigma^{n+m} X, Y \wedge Z].$$

The following result, which is identical in form to [12, Lemma 1.4] summarizes the properties of cup products we use.

Lemma 4.3.2 *Suppose X, Y and Z are pointed simplicial presheaves on $\mathbf{C}$.*

(1) *The cup-product pairing is bilinear and associative.*

(2) *If X is itself a suspension, the cup-product pairing is trivial.*

Proof For point (1), note that the smash product is bilinear and associative, and the pullback is a homomorphism. Thus, the bilinearity and associativity of the cup product follow immediately. As regards point (2), if there exists a pointed simplicial presheaf W such that $X = \Sigma W$, then the reduced diagonal map $\Sigma W \to \Sigma W \wedge \Sigma W$ is simplicially null-homotopic, which means the cup-product pairing is trivial. □

Hopf invariants after Boardman and Steer Now, fix a pointed space Z and consider the James–Hopf invariant $\mathrm{H}\colon J(Z) \to J(Z^{\wedge 2})$ from Definition 3.1.3. Applying $[X, -]$ to this morphism, the identification of $J(-)$ with $\Omega\Sigma(-)$ of Proposition 2.1.6 and the loops suspension adjunction, H determines a map

$$\mathrm{H}\colon [\Sigma X, \Sigma Z] \longrightarrow [\Sigma X, \Sigma Z \wedge Z];$$

in an abuse of notation, we denote this map also by H, but the context should make clear which version we mean.

Following Boardman and Steer [12, Definition 2.1], define

$$\lambda_2\colon [\Sigma X, \Sigma Z] \to [\Sigma^2 X, \Sigma^2 (Z \wedge Z)]$$

by $\lambda_2 = \Sigma \circ \mathrm{H}$. Note that λ_2 vanishes on suspensions. The classical interaction between λ_2 and the group operation $\cdot$ induced by the h–cogroup structure of a suspension generalizes to the context under consideration; this is a special case of the Cartan formula [12, Definition 2.1(c) and Theorem 3.15], which uses the notion of cup products just introduced. To state the result, let $\wp\colon \Sigma Z \wedge \Sigma Z \xrightarrow{\sim} \Sigma^2 Z \wedge Z$ be the map which does not change the order of the suspensions or the order of the copies of Z. The next result provides a direct analog of [12, Formula 3.14], but we include the proof for the convenience of the reader.

Lemma 4.3.3 *Suppose that X and Z are pointed simplicial presheaves on $\mathbf{C}$. Given $\alpha_1, \dots, \alpha_n$ in $[X, Z]$, consider $\Sigma\alpha_i \in [\Sigma X, \Sigma Z]$. The following formula holds in $[\Sigma^2 X, \Sigma^2 Z \wedge Z]$:*

$$\lambda_2(\Sigma\alpha_1 \cdots \Sigma\alpha_n) = \prod_{1 \le i < j \le n} \wp(\Sigma\alpha_i \smile \Sigma\alpha_j),$$

where the product $\Sigma\alpha_1 \cdots \Sigma\alpha_n$ is taken with respect to the group structure on $[\Sigma X, \Sigma Z]$ and $\prod$ denotes the group operation $\cdot$ ordered lexicographically from left to right.

Proof We abuse notation slightly and write α_i for a specified representative of the homotopy class of maps $\alpha_i \in [X, Z]$. In that case, $\Sigma\alpha_1 \cdots \Sigma\alpha_n$ is adjoint to the element of $[X, J(Z)]$ determined by $x \mapsto \alpha_1(x)\alpha_2(x) \cdots \alpha_n(x)$ for $x \in X(U)$.

Then $\mathrm{H}(\Sigma\alpha_1 \cdots \Sigma\alpha_n)$ is adjoint to the element of $[X, J(Z \wedge Z)]$ represented by

$$x \mapsto \prod_{1 \le i < j \le n} \alpha_i(x) \wedge \alpha_j(x) = \prod_{1 \le i < j \le n} (\alpha_i \wedge \alpha_j)(x \wedge x),$$

where $\prod$ denotes the group operation on $J(Z \wedge Z)$ ordered lexicographically from left to right. Thus, $\lambda_2(\Sigma\alpha_1 \cdots \Sigma\alpha_n) = \prod_{1 \le i < j \le n} \wp(\Sigma\alpha_i \smile \Sigma\alpha_j)$, as claimed. $\square$

Let

$$s\colon \Sigma(X \times Y) \longrightarrow \Sigma(X \wedge Y) \quad \text{and} \quad \mathfrak{S}\colon \Sigma^2(X \times Y) \longrightarrow \Sigma^2(X \wedge Y)$$

be the suspension and 2–fold suspension of the usual map from the product to the smash. The group operation on $[\Sigma^2(X \times Y), \Sigma^2 Z \wedge Z]$ is abelian, so we use the symbol $+$ instead of $\cdot$ when writing this operation.

Proposition 4.3.4 *Let X and Y be pointed simplicial presheaves on $\mathbf{C}$ and assume both are suspensions. If $\alpha \in [\Sigma X, \Sigma Z]$ and $\beta \in [\Sigma Y, \Sigma Z]$ are suspensions, then*

$$\mathfrak{S}^* \lambda_2[\alpha, \beta] = -\wp(\pi_Y^* \beta \smile \pi_X^* \alpha) + \wp(\pi_X^* \alpha \smile \pi_Y^* \beta).$$

We remind the reader that $[\alpha, \beta]$ denotes the Whitehead product of α and β. Here $\pi_X^*\colon [\Sigma X, \Sigma Z] \to [\Sigma(X \times Y), \Sigma Z]$ is the map induced by projection.

Proof All cup products will be followed by $\wp$, so we suppress $\wp$ from the notation. By definition, $s^*[\alpha, \beta] = \pi_X^* \alpha^{-1} \cdot \pi_Y^* \beta^{-1} \cdot \pi_X^* \alpha \cdot \pi_Y^* \beta$. By the naturality of λ_2 and Lemma 4.3.3,

$$\begin{aligned} \mathfrak{S}^* \lambda_2[\alpha, \beta] = {} & \pi_X^* \alpha^{-1} \smile \pi_Y^* \beta^{-1} + \pi_X^* \alpha^{-1} \smile \pi_X^* \alpha + \pi_X^* \alpha^{-1} \smile \pi_Y^* \beta \\ & + \pi_Y^* \beta^{-1} \smile \pi_X^* \alpha + \pi_Y^* \beta^{-1} \smile \pi_Y^* \beta + \pi_X^* \alpha \smile \pi_Y^* \beta. \end{aligned}$$

Since X is a suspension, the cup product $\alpha^{-1} \smile \alpha$ vanishes by Lemma 4.3.2(2), whence $\pi_X^*\alpha^{-1} \smile \pi_X^*\alpha = 0$. The same reasoning shows that $\pi_Y^*\beta^{-1} \smile \pi_Y^*\beta = 0$. By Lemma 4.3.2(1), the cup product is bilinear and associative, and so we obtain the following formula:

$$\begin{aligned}\mathfrak{S}^*\lambda_2[\alpha,\beta] &= \pi_X^*\alpha \smile \pi_Y^*\beta - \pi_X^*\alpha \smile \pi_Y^*\beta - \pi_Y^*\beta \smile \pi_X^*\alpha + \pi_X^*\alpha \smile \pi_Y^*\beta \\ &= -\pi_Y^*\beta \smile \pi_X^*\alpha + \pi_X^*\alpha \smile \pi_Y^*\beta. \qquad \square\end{aligned}$$

Hopf invariants of Whitehead products Let Z be a pointed simplicial presheaf on $\mathbf{C}$. Consider the maps

$$[\iota_{\Sigma Z}, \iota_{\Sigma Z}]\colon \Sigma Z \wedge Z \longrightarrow \Sigma Z \quad \text{and} \quad \Sigma \mathrm{H}[\iota_{\Sigma Z}, \iota_{\Sigma Z}]\colon \Sigma^2 Z \wedge Z \longrightarrow \Sigma^2 Z \wedge Z.$$

Let $e\colon Z \wedge Z \to Z \wedge Z$ denote the exchange map, ie the map that permutes the two factors.

Proposition 4.3.5 *Let Z be a pointed simplicial presheaf on $\mathbf{C}$ that is a suspension. In the homotopy category, there is an equality*

$$\Sigma \mathrm{H}[\iota_{\Sigma Z}, \iota_{\Sigma Z}] = -\Sigma^2 e + \Sigma^2 \iota_{Z\wedge Z}.$$

Proof Let $\pi_i\colon \Sigma(Z\times Z) \to \Sigma Z$ denote the suspension of the i^{th} projection for $i = 1, 2$. Let $\iota\colon \Sigma Z \to \Sigma Z$ denote the identity. Let $\wp\colon \Sigma Z \wedge \Sigma Z \to \Sigma^2(Z \wedge Z)$ be the permutation that does not change the order of suspensions, and let $\wp^{-1}\colon \Sigma^2(Z\wedge Z) \to \Sigma Z \wedge \Sigma Z$ be its inverse.

Consider the map $\mathfrak{S}\colon \Sigma^2(Z \times Z) \to \Sigma^2(Z \wedge Z)$ as introduced above. In that case, Proposition 4.3.4 allows us to conclude the following equality holds:

$$\mathfrak{S}^* \Sigma \mathrm{H}[\iota,\iota] = -\wp(\pi_2^*\iota \smile \pi_1^*\iota) + \wp(\pi_1^*\iota \smile \pi_2^*\iota).$$

Write $\Delta\colon Z \times Z \to (Z \times Z) \wedge (Z \times Z)$ for the reduced diagonal map. Let

$$\wp'\colon \Sigma^2(Z \times Z) \wedge (Z \times Z) \longrightarrow \Sigma(Z \times Z) \wedge \Sigma(Z \times Z)$$

denote the permutation which does not swap the order of the two suspensions in $\Sigma^2(Z \times Z) \wedge (Z \times Z)$. Note that

$$(\pi_1^*\iota \wedge \pi_2^*\iota) \circ \wp'\Sigma^2\Delta\colon \Sigma^2(Z \times Z) \longrightarrow \Sigma(Z \times Z) \wedge \Sigma(Z \times Z) \longrightarrow \Sigma(Z) \wedge \Sigma(Z)$$

is homotopic to $\mathfrak{S}$ followed by the permutation $\wp^{-1}$. Therefore, up to homotopy, $(\pi_1^*\iota \wedge \pi_2^*\iota) \circ \wp'\Sigma^2\Delta = \wp^{-1}\mathfrak{S}$.

By definition, $\pi_1^*\iota \smile \pi_2^*\iota = (\pi_1^*\iota \wedge \pi_2^*\iota) \circ \wp'\Sigma^2\Delta$. Thus $\pi_1^*\iota \smile \pi_2^*\iota = \wp^{-1}\mathfrak{S}$ in the homotopy category. Applying $\wp$ to both sides, we conclude that

$$\wp(\pi_1^*\iota \smile \pi_2^*\iota) = \mathfrak{S}.$$

Note that

$$(\pi_2^*\iota \wedge \pi_1^*\iota) \circ \wp' \Sigma^2 \Delta\colon \Sigma^2(Z \times Z) \longrightarrow \Sigma(Z \times Z) \wedge \Sigma(Z \times Z) \longrightarrow \Sigma(Z) \wedge \Sigma(Z)$$

is homotopic to $\wp^{-1} \circ \Sigma^2 e \circ \mathfrak{S}$, whence

$$\wp(\pi_2^*\iota \smile \pi_1^*\iota) = \mathfrak{S}^* \Sigma^2 e.$$

It follows that

$$\mathfrak{S}^* \Sigma \mathrm{H}[\iota_{\Sigma Z}, \iota_{\Sigma Z}] = \mathfrak{S}^*(-\Sigma^2 e + \Sigma^2 \iota_{Z \wedge Z}).$$

To conclude, we simply observe that $\mathfrak{S}^*$ is injective. Indeed, this follows from the standard fact that for simplicial presheaves X and Y, after a single suspension, the cofiber sequence $X \vee Y \to X \times Y \to X \wedge Y$ is split by the sum of the projections Σp_X and Σp_Y. In that case, the long exact sequence in homotopy obtained by mapping any suspension of the above cofiber sequence into the space Z splits into a collection of short exact sequences. □

4.4 The composite HP for a sphere as an element of $\boldsymbol{K}_0^{\mathrm{MW}}(k)$

In this section, we analyze the composite map HP for a sphere. Up to this point in the paper, we have worked either in the context of simplicial presheaves on a site having enough points or in the unstable $\mathbb{A}^1$–homotopy theory over a base for which the unstable $\mathbb{A}^1$–connectivity property holds. The results in this section differ from those earlier in the paper because they will use finer structure of the $\mathbb{A}^1$–homotopy category over a base field k assumed to be perfect (and infinite for those being especially careful). We will try to be clear about exactly which ingredients do not follow from the "axiomatic" point of view.

Morel shows the sheaf $\boldsymbol{\pi}_p^{\mathbb{A}^1}(S^{p+q\alpha})$ (see Notation 2.2.4) is isomorphic to the Milnor–Witt K-theory sheaf $\boldsymbol{K}_q^{\mathrm{MW}}$ for $p \geq 2$ (or, somewhat exceptionally, for $p = 1$ and $q = 2$) [44, Theorem 1.23]. Stringing the EHP exact sequences of Theorem 3.2.1 for different spheres together, one obtains the following diagram:

$$\begin{tikzcd}
 & \boldsymbol{\pi}_{i+3}^{\mathbb{A}^1}(S^{2p+3+2q\alpha}) \arrow[d, "\mathrm{P}"] & & \\
\boldsymbol{\pi}_i^{\mathbb{A}^1}(S^{p+q\alpha}) \arrow[r, "\mathrm{E}"] & \boldsymbol{\pi}_{i+1}^{\mathbb{A}^1}(S^{p+1+q\alpha}) \arrow[r, "\mathrm{H}"] \arrow[d, "\mathrm{E}"] & \boldsymbol{\pi}_{i+1}^{\mathbb{A}^1}(S^{2p+1+2q\alpha}) \arrow[r, "\mathrm{P}"] & \boldsymbol{\pi}_{i-1}^{\mathbb{A}^1}(S^{p+q\alpha}) \\
 & \boldsymbol{\pi}_{i+2}^{\mathbb{A}^1}(S^{p+2+q\alpha}) & &
\end{tikzcd}$$

The composite map HP becomes the E_1–differential in the EHP spectral sequence.

When $i=2p$, the composite map HP is, by means of Morel's computations, a morphism

$$\mathrm{HP}\colon \boldsymbol{K}^{\mathrm{MW}}_{2q} \longrightarrow \boldsymbol{K}^{\mathrm{MW}}_{2q}.$$

Note that by definition of Milnor–Witt K-theory sheaves, there is a ring homomorphism $\boldsymbol{K}^{\mathrm{MW}}_0(k) \to \mathrm{Hom}(\boldsymbol{K}^{\mathrm{MW}}_{2q}, \boldsymbol{K}^{\mathrm{MW}}_{2q})$ induced on sections by multiplication; moreover this homomorphism is necessarily injective. Lemma 5.1.3, combined with the computation of contractions of Milnor–Witt K-theory sheaves (see the discussion after Lemma 5.1.1), implies this morphism is an isomorphism if k has characteristic unequal to 2; that this map is an isomorphism is also true if k has characteristic 2 but a different proof is required; see the discussion after Lemma 5.1.1 for more details. In any case, the map HP corresponds to an element of $\mathrm{Hom}(\boldsymbol{K}^{\mathrm{MW}}_{2q}, \boldsymbol{K}^{\mathrm{MW}}_{2q})$; we will see below that it always lies in the subring $\boldsymbol{K}^{\mathrm{MW}}_0(k)$.

In order to state the result, we need some more precise information about the structure of the Milnor–Witt K-theory ring. Recall that $\boldsymbol{K}^{\mathrm{MW}}_*(k)$ is generated by elements $[a] \in k^*$ of degree $+1$ and an element η of degree -1 subject to various relations [44, Definition 3.1]. For a unit $a \in k^*$, set $\langle a\rangle := 1+\eta[a]$; the identification $\boldsymbol{K}^{\mathrm{MW}}_0(k) \cong \mathrm{GW}(k)$ sends the element $\langle a\rangle$ to the class of the 1–dimensional symmetric bilinear form of the same name [44, Lemma 3.10]. Following Morel [44, page 51] or [40, Section 6.1], we set $\epsilon := -\langle -1\rangle$. The class ϵ is related to the map $\mathbb{G}_m \wedge \mathbb{G}_m \to \mathbb{G}_m \wedge \mathbb{G}_m$ that exchanges the two factors: see [40, Lemma 6.1.1(2)] for a "stable" statement or [44, Lemma 3.43] for an "unstable" statement. Note that $1-\epsilon$ is the hyperbolic form h, which intercedes in the definition of Milnor–Witt K-theory.

Theorem 4.4.1 *Assume k is a perfect field and suppose p and q are integers with $p>1$ and $q \geq 1$. The map*

$$\mathrm{HP}\colon \boldsymbol{K}^{\mathrm{MW}}_{2q} = \boldsymbol{\pi}^{\mathbb{A}^1}_{2p+2} J((S^{p+1+q\alpha})^{\wedge 2}) \longrightarrow \boldsymbol{\pi}_{2p} J((S^{p+q\alpha})^{\wedge 2}) = \boldsymbol{K}^{\mathrm{MW}}_{2q}$$

is given by $1-(-1)^p\epsilon^q \in \boldsymbol{K}^{\mathrm{MW}}_0(k)$. Equivalently,

$$\mathrm{HP} = \begin{cases} 0 & \text{if } p \text{ and } q \text{ are even,} \\ 2 & \text{if } p \text{ is odd and } q \text{ is even,} \\ h & \text{if } p \text{ is even and } q \text{ is odd,} \\ 1+\epsilon & \text{if } p \text{ and } q \text{ are odd.} \end{cases}$$

Remark 4.4.2 See Remark 5.1.2 for more details on the situation when k has characteristic 2. There is a corresponding statement if $q=0$ as well, but in that case, the composite in question gets identified with an element of $\mathbb{Z}$, not $\boldsymbol{K}^{\mathrm{MW}}_0(k)$ and the result is the classical computation of HP.

Proof Under the hypothesis on k, the unstable $\mathbb{A}^1$–connectivity property holds by Theorem 2.2.7. The hypotheses on p and q are simply those that need to be imposed to appeal to Morel's computations of homotopy sheaves.

We begin by applying Theorem 4.2.1 with $\mathscr{X} = S^{p+1+q\alpha}$, $\mathscr{X} = S^{p+q\alpha}$, $n = p+1$ and $i = 2p+1$ to interpret P as a Whitehead product. More precisely, since $1 \le p$, it follows that $i = (2p+1) \le 3(p+1)-3$, and so we know that P is induced by the map $[\iota_{\Sigma\mathscr{X}}, \iota_{\Sigma\mathscr{X}}]_*$ in degree $2p+1$. It follows that the composite HP is isomorphic to the map obtained by applying $\boldsymbol{\pi}^{\mathbb{A}^1}_{2p+1}$ to $\mathrm{H}[\iota_{\Sigma\mathscr{X}}, \iota_{\Sigma\mathscr{X}}]$.

Next, we appeal to our results about James–Hopf invariants of Whitehead products (Proposition 4.3.5) to produce an explicit formula for HP in terms of the exchange map $e\colon \mathscr{X} \wedge \mathscr{X} \to \mathscr{X} \wedge \mathscr{X}$. Indeed, Proposition 4.3.5 yields an equality of the form

$$\Sigma\mathrm{H}[\iota_{\Sigma\mathscr{X}}, \iota_{\Sigma\mathscr{X}}] = -\Sigma^2 e + \Sigma^2 \iota_{\mathscr{X}\wedge\mathscr{X}}.$$

However, recall that by Theorem 3.2.1 (combined with Remark 3.2.3), the suspension map

$$\Sigma\colon [\Sigma(\mathscr{X}\wedge\mathscr{X}), \Sigma(\mathscr{X}\wedge\mathscr{X})] \longrightarrow [\Sigma^2(\mathscr{X}\wedge\mathscr{X}), \Sigma^2(\mathscr{X}\wedge\mathscr{X})]$$

is an isomorphism and we conclude that the following equality holds:

$$\mathrm{H}[\iota_{\Sigma\mathscr{X}}, \iota_{\Sigma\mathscr{X}}] = -\Sigma e + \Sigma \iota_{\mathscr{X}\wedge\mathscr{X}}.$$

Therefore, we see that HP is isomorphic to the map induced by applying $\boldsymbol{\pi}^{\mathbb{A}^1}_{2p+1}$ to $-\Sigma e + \Sigma\iota_{\mathscr{X}\wedge\mathscr{X}}$.

Now we identify the homotopy class of the exchange map, e. Since e can be effected by pairwise exchanging copies of S^1, and $\mathbb{G}_m$, it suffices to understand the effect of each such exchange on a homotopy class. By [44, Lemma 3.43], in the presence of a single suspension, the exchange of copies of $\mathbb{G}_m$ contributes a factor of ϵ. It is well known that the permutation map on $S^1 \wedge S^1$ has degree -1. Combining these two observations, a straightforward induction argument allows us to conclude that e has degree $(-1)^p\epsilon^q$. Thus, the map induced by applying $\boldsymbol{\pi}^{\mathbb{A}^1}_{2p+1}(-)$ to $-\Sigma e + \Sigma\iota_{\mathscr{X}\wedge\mathscr{X}}$ is multiplication by $1-(-1)^p\epsilon^q$. Since $\epsilon^2 = 1$ in $\boldsymbol{K}^{\mathrm{MW}}_0(k)$ by [44, Lemma 3.5], the statement of the theorem follows by simply listing the possible cases. □

Remark 4.4.3 Classically, the composite HP computed above is either 2 or 0 depending on the parity of the dimension of the sphere in question (this follows immediately from the definition of the James–Hopf invariant and symmetry properties of the Whitehead square of the identity). If one invokes real realization [45, page 121], Theorem 4.4.1 can be viewed as a direct analog of this classical result. First, observe that $\boldsymbol{K}^{\mathrm{MW}}_0(\mathbb{R}) \cong \mathrm{GW}(\mathbb{R}) \cong \mathbb{Z} \oplus \mathbb{Z}$; this identification sends a symmetric bilinear form over the real numbers to its rank and signature. Under real realization, the sphere $S^{p+q\alpha}$ is sent to the $\mathbb{Z}/2$–equivariant sphere of the same name. In particular, the $\mathbb{Z}/2$–fixed-point locus of $S^{p+q\alpha}$ is simply S^p, while the fixed-point locus under the trivial group is the sphere S^{p+q}. From this point of view, the formula for HP from Theorem 4.4.1 simply reflects the relative parities of p and $p+q$: the signature keeps track of the degree on fixed-point loci for $\mathbb{Z}/2$, while the rank keeps track of the degree on fixed-point loci for the trivial subgroup. For example, when p is even and q is odd, HP is multiplication by h, which has rank 2 and signature 0.

5 Applications

Here we collect some computational applications of Theorems 3.2.1 and 3.3.13. Section 5.1 is of a preliminary nature and contains a number of results about Milnor–Witt K-theory sheaves that are used elsewhere in the text; some of these facts are certainly well known, but we could not find good references. Section 5.2 contains new computations of a family of unstable $\mathbb{A}^1$–homotopy sheaves of motivic spheres: it contains the first computation since Morel's of an S^1–stable $\mathbb{A}^1$–homotopy sheaf (see Theorems 5.2.5 and 5.2.9). Finally, Section 5.3 contains results regarding unstable rationalized motivic homotopy sheaves, and S^1–stable homotopy sheaves of Voevodsky's mod m motivic Eilenberg–MacLane spaces (see Theorems 5.3.1 and 5.3.3). While it should be clear from the referencing, essentially all of the results of this section require finer properties of the unstable $\mathbb{A}^1$–homotopy category than merely the unstable $\mathbb{A}^1$–connectivity property.

5.1 On Milnor–Witt K-theory sheaves

In this section we study some properties of the Milnor–Witt K-theory sheaves $\boldsymbol{K}_n^{\mathrm{MW}}$ [44, Section 3]. By [44, Theorem 3.37], the sheaves $\boldsymbol{K}_n^{\mathrm{MW}}$ are strictly $\mathbb{A}^1$–invariant sheaves for any integer n. In fact, for any integer $n \geq 1$, the sheaf $\boldsymbol{K}_n^{\mathrm{MW}}$ is the free strictly $\mathbb{A}^1$–invariant sheaf on the sheaf of pointed sets $\mathbb{G}_m^{\wedge n}$, and $\boldsymbol{K}_0^{\mathrm{MW}}$ is the free strictly $\mathbb{A}^1$–invariant sheaf on $\mathbb{G}_m/\mathbb{G}_m^{\times 2}$ by [44, Theorem 3.46] (not pointed in this case).

Basic properties of Milnor–Witt K-theory sheaves If $\boldsymbol{M}$ is a presheaf of groups (actually, pointed sets suffices), then its contraction $\boldsymbol{M}_{-1}$ is the presheaf of groups on Sm_k defined by

$$\boldsymbol{M}_{-1}(U) := \ker\big(\boldsymbol{M}(\mathbb{G}_m \times U) \xrightarrow{(1\times\mathrm{id})^*} \boldsymbol{M}(U)\big),$$

where $1\colon \operatorname{Spec} k \to \mathbb{G}_m$ is the unit map. The next result summarizes the properties of contractions we will use.

Lemma 5.1.1 [44, Lemmas 2.32 and 7.33] *The assignment $\boldsymbol{M} \mapsto \boldsymbol{M}_{-1}$ defines an endofunctor of the category of strictly (or strongly) $\mathbb{A}^1$–invariant sheaves which preserves exact sequences.*

We will freely use the fact [3, Lemma 2.9][4] that $(\boldsymbol{K}_n^{\mathrm{MW}})_{-j} = \boldsymbol{K}_{n-j}^{\mathrm{MW}}$ for any pair of integers n, j. We write $\boldsymbol{W}$ for the sheaf of unramified Witt groups, and $\boldsymbol{I}^n \subset \boldsymbol{W}$ for the subsheaves of unramified powers of the fundamental ideal in the Witt ring [42, Section 2.1]. For any integer m, we write $\boldsymbol{K}_n^{\mathrm{M}}/m$ for the mod m Milnor K-theory sheaf. The contractions of $\boldsymbol{K}_m^{\mathrm{M}}$ are summarized in [3, Lemma 2.7]. There is a canonical morphism $\boldsymbol{K}_n^{\mathrm{M}}/2 \to \boldsymbol{I}^n/\boldsymbol{I}^{n+1}$; the Milnor conjecture on quadratic forms, now a theorem, asserts that this morphism is an isomorphism [46; 42].

Suppose k is a base field of characteristic unequal to 2. Morel established [41, Théorème 5.3] under this hypothesis that there is a fiber product presentation of $\boldsymbol{K}_n^{\mathrm{MW}}$ relating the various sheaves described in the previous paragraph. For any integer n, there is a fiber product diagram of the following form:[5]

$$\begin{array}{ccc} \boldsymbol{K}_n^{\mathrm{MW}} & \longrightarrow & \boldsymbol{I}^n \\ \downarrow & & \downarrow \\ \boldsymbol{K}_n^{\mathrm{M}} & \longrightarrow & \boldsymbol{K}_n^{\mathrm{M}}/2 \end{array}$$

By convention $\boldsymbol{K}_n^{\mathrm{M}} = \boldsymbol{K}_n^{\mathrm{M}}/2 = 0$ for $n < 0$, whereas $\boldsymbol{I}^n \cong \boldsymbol{W}$ for $n < 0$.

This fiber product presentation yields two fundamental exact sequences

$$\begin{aligned} 0 \longrightarrow \boldsymbol{I}^{n+1} \longrightarrow \boldsymbol{K}_n^{\mathrm{MW}} \longrightarrow \boldsymbol{K}_n^{\mathrm{M}} \longrightarrow 0, \\ 0 \longrightarrow 2\boldsymbol{K}_n^{\mathrm{M}} \longrightarrow \boldsymbol{K}_n^{\mathrm{MW}} \longrightarrow \boldsymbol{I}^n \longrightarrow 0. \end{aligned} \tag{5.1.1}$$

We use these sequences repeatedly in the sequel.

Remark 5.1.2 The assumption that k has characteristic unequal to 2 is inessential above: the fiber product presentation exists without this condition as one can see by inspecting the proofs, and appealing to the results of Kato [35] on the characteristic 2 version of Milnor's conjecture involving symmetric bilinear forms instead of [46]. However, a detailed proof of this generalization does not appear in the literature, and

[4] This identification is due to Morel and appears in several places in [44] but without a proof. The proof given in [3] requires k to have characteristic unequal to 2 since it depends on the Gersten conjecture for the sheaves $\boldsymbol{I}^j$. The result can also be demonstrated when k has characteristic 2 if one appeals to Morel's Gersten–Schmid resolution of $\boldsymbol{K}_n^{\mathrm{MW}}$. Nevertheless, since we will momentarily restrict to the case where k has characteristic different from 2 for other reasons, the result from [3] is sufficient for our purposes.

[5] See [22] for some corrections to [41].

since in later applications we will be restricted to the characteristic unequal to 2 case anyway, we have not pursued this generalization.

On the structure of contracted sheaves

Lemma 5.1.3 *Suppose $\boldsymbol{M}$ is a strictly $\mathbb{A}^1$–invariant sheaf.*

(1) *For any integer $n \geq 1$, there are isomorphisms*

$$\mathrm{Hom}(\boldsymbol{K}_n^{\mathrm{MW}}, \boldsymbol{M}) \cong \boldsymbol{M}_{-n}(k).$$

(2) *If $n \geq 2$, the evident map $\mathrm{Hom}(\boldsymbol{K}_n^{\mathrm{MW}}, \boldsymbol{M}) \to \mathrm{Hom}(\boldsymbol{K}_{n-1}^{\mathrm{MW}}, \boldsymbol{M}_{-1})$ induced by contraction is an isomorphism compatible with the identification of point (1).*

Proof Write $\underline{\mathrm{Hom}}_*$ for the internal Hom in the category of presheaves of pointed sets on Sm_k. In that case, unwinding the definitions, there is an identification $M_{-1} = \mathrm{Hom}_*(\mathbb{G}_m, M)$. A straightforward induction argument combined with the adjunction between $\wedge$ and Hom_* then shows $M_{-n} = \underline{\mathrm{Hom}}_*(\mathbb{G}_m^{\wedge n}, M)$.

For $n \geq 1$, [44, Theorem 3.37] shows that $\boldsymbol{K}_n^{\mathrm{MW}}$ is the free strictly $\mathbb{A}^1$–invariant sheaf of groups on the sheaf of pointed sets $\mathbb{G}_m^{\wedge n}$. As a consequence, there are functorial identifications

$$\underline{\mathrm{Hom}}_*(\mathbb{G}_m^{\wedge n}, M) \xrightarrow{\sim} \underline{\mathrm{Hom}}(\boldsymbol{K}_n^{\mathrm{MW}}, \boldsymbol{M}),$$

where $\underline{\mathrm{Hom}}$ on the right-hand side is the internal Hom in the category of presheaves of abelian groups. To complete the verification of point (1), simply take sections over k.

In light of the discussion of the previous paragraphs, to establish point (2) one simply observes that, as long as $n \geq 2$, the map in question arises via the following sequence of identifications:

$$\begin{aligned}\underline{\mathrm{Hom}}(\boldsymbol{K}_n^{\mathrm{MW}}, \boldsymbol{M}) &\cong \underline{\mathrm{Hom}}_*(\mathbb{G}_m^{\wedge n}, \boldsymbol{M}) \\ &\cong \underline{\mathrm{Hom}}_*(\mathbb{G}_m^{\wedge (n-1)}, \underline{\mathrm{Hom}}_*(\mathbb{G}_m, \boldsymbol{M})) \\ &\cong \underline{\mathrm{Hom}}(\boldsymbol{K}_{n-1}^{\mathrm{MW}}, \boldsymbol{M}_{-1}). \end{aligned}$$ □

Lemma 5.1.4 *Suppose $\boldsymbol{M}$ is a strictly $\mathbb{A}^1$–invariant sheaf.*

(1) *There is an isomorphism*

$$\mathrm{Hom}(\boldsymbol{K}_0^{\mathrm{MW}}, \boldsymbol{M}) \cong \boldsymbol{M}(k) \times {}_h\boldsymbol{M}_{-1}(k),$$

where ${}_h\boldsymbol{M}_{-1}(k)$ is the h–torsion subgroup. The first map is induced by the projection $\boldsymbol{K}_0^{\mathrm{MW}} \to \mathbb{Z}$, while the second is induced by a splitting of the map $\boldsymbol{I} \to \boldsymbol{K}_0^{\mathrm{MW}}$.

(2) *For any integer $n \geq 1$, the map $\mathrm{Hom}(\boldsymbol{K}_n^{\mathrm{MW}}, \boldsymbol{M}) \to \mathrm{Hom}(\boldsymbol{K}_0^{\mathrm{MW}}, \boldsymbol{M}_{-n})$ induced by contraction has image the factor $\boldsymbol{M}_{-n}(k)$ of the product described in point (1).*

Proof Note that $\boldsymbol{K}_0^{\mathrm{MW}} = \boldsymbol{H}_0^{\mathbb{A}^1}(\mathbb{G}_m/\mathbb{G}_m^{\times 2}) = \widetilde{\boldsymbol{H}}_0^{\mathbb{A}^1}(\mathbb{G}_m/\mathbb{G}_m^{\times 2}{}_+)$ by [44, Theorem 3.46] and what follows amounts to unwinding the proof of this result. Note that there is a split cofiber sequence $S_k^0 \to \mathbb{G}_m/\mathbb{G}_m^{\times 2}{}_+ \to \mathbb{G}_m/\mathbb{G}_m^{\times 2}$.

By adjunction, one then obtains identifications of the form

$$\begin{aligned}\underline{\mathrm{Hom}}(\boldsymbol{K}_0^{\mathrm{MW}}, \boldsymbol{M}) &\cong \underline{\mathrm{Hom}}_*(\mathbb{G}_m/\mathbb{G}_m^{\times 2}{}_+, \boldsymbol{M})\\ &\cong \underline{\mathrm{Hom}}_*(S_k^0, \boldsymbol{M}) \times \underline{\mathrm{Hom}}_*(\mathbb{G}_m/\mathbb{G}_m^{\times 2}, \boldsymbol{M})\\ &\cong \underline{\mathrm{Hom}}(\mathbb{Z}, \boldsymbol{M}) \times \underline{\mathrm{Hom}}(\widetilde{\boldsymbol{H}}_0^{\mathbb{A}^1}(\mathbb{G}_m/\mathbb{G}_m^{\times 2}), \boldsymbol{M})\\ &\cong \boldsymbol{M} \times \underline{\mathrm{Hom}}(\widetilde{\boldsymbol{H}}_0^{\mathbb{A}^1}(\mathbb{G}_m/\mathbb{G}_m^{\times 2}), \boldsymbol{M}).\end{aligned}$$

Under this decomposition, the projection map $\boldsymbol{K}_0^{\mathrm{MW}} \to \mathbb{Z}$ is precisely the rank map. On the other hand, the splitting $\mathbb{G}_m/\mathbb{G}_m^{\times 2}{}_+ \cong S_k^0 \vee \mathbb{G}_m/\mathbb{G}_m^{\times 2}$ corresponds to the splitting $\boldsymbol{I} \to \boldsymbol{K}_0^{\mathrm{MW}}$ as described before [44, Corollary 3.47].

Next, there is an exact sequence of Nisnevich sheaves of abelian groups of the form

$$\mathbb{G}_m \xrightarrow{x \mapsto x^2} \mathbb{G}_m \longrightarrow \mathbb{G}_m/\mathbb{G}_m^{\times 2} \longrightarrow 0.$$

Taking reduced $\mathbb{A}^1$–homology (as the composite of taking the (based) free abelian group functor and the exact functor $\mathrm{L}_{\mathbb{A}^1}^{\mathrm{ab}}$) yields an exact sequence of the form

$$\widetilde{\boldsymbol{H}}_0^{\mathbb{A}^1}(\mathbb{G}_m) \longrightarrow \widetilde{\boldsymbol{H}}_0^{\mathbb{A}^1}(\mathbb{G}_m) \longrightarrow \widetilde{\boldsymbol{H}}_0^{\mathbb{A}^1}(\mathbb{G}_m/\mathbb{G}_m^{\times 2}) \longrightarrow 0.$$

The map $\widetilde{\boldsymbol{H}}_0^{\mathbb{A}^1}(\mathbb{G}_m) = \boldsymbol{K}_1^{\mathrm{MW}} \to \boldsymbol{K}_1^{\mathrm{MW}}$ induced by the squaring map on $\mathbb{G}_m$ is multiplication by $h = \langle 1 \rangle + \langle -1 \rangle$ by [44, Lemma 3.14]. Thus, we conclude that

$$\mathrm{Hom}(\widetilde{\boldsymbol{H}}_0^{\mathbb{A}^1}(\mathbb{G}_m/\mathbb{G}_m^{\times 2}), \boldsymbol{M}) \cong {}_h\boldsymbol{M}_{-1}(k),$$

which is what we wanted to show.

For point (2), we appeal to Lemma 5.1.3. Indeed, it suffices by Lemma 5.1.3(2) and induction to treat the case where $n = 1$. As in the proof of Lemma 5.1.3, adjunction yields identifications of the form

$$\underline{\mathrm{Hom}}(\boldsymbol{K}_1^{\mathrm{MW}}, \boldsymbol{M}) \cong \underline{\mathrm{Hom}}_*(\mathbb{G}_m, \boldsymbol{M}) \cong \underline{\mathrm{Hom}}_*(S_k^0, \underline{\mathrm{Hom}}_*(\mathbb{G}_m, \boldsymbol{M})) \cong \underline{\mathrm{Hom}}(\mathbb{Z}, \boldsymbol{M}_{-1}).$$

In particular, the map

$$\mathrm{Hom}(\boldsymbol{K}_1^{\mathrm{MW}}, \boldsymbol{M}) \longrightarrow \mathrm{Hom}(\boldsymbol{K}_0^{\mathrm{MW}}, \boldsymbol{M}_{-1})$$

induced by contraction factors through the isomorphism

$$\mathrm{Hom}(\boldsymbol{K}_1^{\mathrm{MW}}, \boldsymbol{M}) \cong \mathrm{Hom}(\mathbb{Z}, \boldsymbol{M}_{-1}).$$

We treat the universal case: taking $\boldsymbol{M} = \boldsymbol{K}_1^{\mathrm{MW}}$, we see that the map induced by contraction factors through $\mathrm{Hom}(\mathbb{Z}, \boldsymbol{K}_0^{\mathrm{MW}})$. Such homomorphisms correspond to elements of $\boldsymbol{K}_0^{\mathrm{MW}}(k)$ via the image of $1 \in \mathbb{Z}$ and under the identification, the identity map $\boldsymbol{K}_1^{\mathrm{MW}} \to \boldsymbol{K}_1^{\mathrm{MW}}$ is sent to the class of $\langle 1 \rangle$.

The identification $(\boldsymbol{K}_1^{\mathrm{MW}})_{-1} \cong \boldsymbol{K}_0^{\mathrm{MW}}$ can be seen in terms of the fiber product presentations $\boldsymbol{K}_1^{\mathrm{MW}} \xrightarrow{\sim} \boldsymbol{K}_1^{\mathrm{M}} \times_{\boldsymbol{K}_1^{M}/2} \boldsymbol{I}$ and $\boldsymbol{K}_0^{\mathrm{MW}} \xrightarrow{\sim} \mathbb{Z} \times_{\mathbb{Z}/2} \boldsymbol{W}$. The symbol map $\mathbb{G}_m \to \boldsymbol{K}_1^{\mathrm{MW}}$ can be thought of as a set-theoretic splitting of the projection map $\boldsymbol{K}_1^{\mathrm{MW}} \to \boldsymbol{K}_1^{M}$. After contraction, this projection is sent to the rank map $\boldsymbol{K}_0^{\mathrm{MW}} \to \mathbb{Z}$. The factorization produced in the previous paragraph thus corresponds to a splitting of the rank map $\boldsymbol{K}_0^{\mathrm{MW}} \to \mathbb{Z}$. On the other hand, the decomposition in point (1) corresponded with a decomposition $\boldsymbol{K}_0^{\mathrm{MW}} \cong \mathbb{Z} \oplus \boldsymbol{I}$ and under this identification the unit $\langle 1 \rangle$ is sent to $(1, 0)$, thus we conclude that the projection onto the other factor is the zero map. □

Lemma 5.1.5 *Fix a base field k. If $\phi\colon \boldsymbol{K}_n^{\mathrm{MW}} \to \boldsymbol{M}$ is a morphism of sheaves such that $\phi_{-j} = 0$, then*

(1) *if $n \geq j \geq 0$, the morphism ϕ is trivial;*

(2) *if $0 \leq n < j$, the morphism ϕ factors through a morphism $\boldsymbol{K}_n^{\mathrm{MW}}/\boldsymbol{I}^j \to \boldsymbol{M}$.*

Proof Factor $\phi\colon \boldsymbol{K}_n^{\mathrm{MW}} \twoheadrightarrow \mathrm{Im}(\phi) \hookrightarrow \boldsymbol{M}$. Since the inclusion of the abelian category of strictly $\mathbb{A}^1$–invariant sheaves into the abelian category of abelian sheaves is exact (see Lemma 3.3.8) we can assume without loss of generality that $\mathrm{Im}(\phi)$ is strictly $\mathbb{A}^1$–invariant. Thus, it suffices to consider the case where $\phi\colon \boldsymbol{K}_n^{\mathrm{MW}} \twoheadrightarrow \boldsymbol{M}$ is an epimorphism and $\boldsymbol{M}$ is strictly $\mathbb{A}^1$–invariant.

If $\boldsymbol{M}_{-0} = 0$, then $\boldsymbol{M} = 0$, and there is nothing to check. Therefore, we can assume without loss of generality that $j \geq 1$. By Lemma 5.1.3, for any integer $r \geq 1$, $\boldsymbol{M}_{-r}(k) \cong \mathrm{Hom}(\boldsymbol{K}_r^{\mathrm{MW}}, \boldsymbol{M})$. For (1) we simply observe that if $n \geq j$, then the morphism $\boldsymbol{K}_n^{\mathrm{MW}} \to \boldsymbol{M}$ is the trivial map since $\boldsymbol{M}_{-n}(k) = 0$ as well.

For (2), begin by observing that, since $0 \leq n < j$, we can consider the following diagram:

$$\boldsymbol{K}_j^{\mathrm{MW}} \twoheadrightarrow \boldsymbol{I}^j \hookrightarrow \boldsymbol{I}^{n+1} \hookrightarrow \boldsymbol{K}_n^{\mathrm{MW}} \twoheadrightarrow \boldsymbol{M}$$

Reading from the left, the first and third maps are those in the exact sequences in (5.1.1), the map $\boldsymbol{I}^j \hookrightarrow \boldsymbol{I}^{n+1}$ is the standard inclusion (since $j \geq n+1$, this makes sense), and the final epimorphism is the one given by the assumptions. Since $\boldsymbol{M}_{-j} = 0$ (and $j \geq 1$ by assumption), this composite is trivial, which means the map $\boldsymbol{K}_n^{\mathrm{MW}} \to \boldsymbol{M}$ factors through the quotient $\boldsymbol{K}_n^{\mathrm{MW}}/\boldsymbol{I}^j$. □

Remark 5.1.6 This result will be applied below with $\boldsymbol{K}_n^{\mathrm{MW}} \to \boldsymbol{M}$ a map to a strictly $\mathbb{A}^1$–invariant sheaf with strictly $\mathbb{A}^1$–invariant cokernel.

Some result on $\mathbb{A}^1$–tensor products

Proposition 5.1.7 *There is an isomorphism $\boldsymbol{K}^{\mathrm{MW}}_m \otimes^{\mathbb{A}^1} \boldsymbol{K}^{\mathrm{MW}}_n \xrightarrow{\sim} \boldsymbol{K}^{\mathrm{MW}}_{m+n}$, for any integers $m, n \geq 1$.*

Proof Morel computed $\widetilde{\boldsymbol{H}}^{\mathbb{A}^1}_{n-1}(\mathbb{A}^i \setminus 0) \cong \boldsymbol{K}^{\mathrm{MW}}_n$ [44, Theorem 6.40] (strictly speaking, this result is stated for $n \geq 2$, but the result is true for $n = 1$ as well by unwinding the definitions and appealing to [44, Theorem 3.37]). There are identifications

$$\Sigma \mathbb{A}^m \setminus 0 \wedge \mathbb{A}^n \setminus 0 \;\simeq\; \mathbb{A}^m \setminus 0 * \mathbb{A}^n \setminus 0 \;\simeq\; \mathbb{A}^{m+n} \setminus 0$$

(here $*$ means join), for any $m, n \geq 1$. Proposition 3.3.9 then yields

$$\boldsymbol{K}^{\mathrm{MW}}_m \otimes^{\mathbb{A}^1} \boldsymbol{K}^{\mathrm{MW}}_n \cong \widetilde{\boldsymbol{H}}^{\mathbb{A}^1}_{n+m-2}(\mathbb{A}^m \setminus 0 \wedge \mathbb{A}^n \setminus 0),$$

which when combined with the suspension isomorphism $\widetilde{\boldsymbol{H}}^{\mathbb{A}^1}_{n+m-2}(\mathbb{A}^m \setminus 0 \wedge \mathbb{A}^n \setminus 0) \cong \widetilde{\boldsymbol{H}}^{\mathbb{A}^1}_{n+m-1}(\mathbb{A}^m \setminus 0 * \mathbb{A}^n \setminus 0) \cong \boldsymbol{K}^{\mathrm{MW}}_{n+m}$ yields the result. □

Lemma 5.1.8 *For any integers $m, n \geq 1$, and any integer $r \geq 0$, there are canonical isomorphisms $\boldsymbol{K}^{\mathrm{MW}}_m \otimes^{\mathbb{A}^1} \boldsymbol{K}^{\mathrm{M}}_n / r \cong \boldsymbol{K}^{\mathrm{M}}_{m+n}/r$. There are also canonical isomorphisms $\boldsymbol{K}^{\mathrm{M}}_m / r \otimes^{\mathbb{A}^1} \boldsymbol{K}^{\mathrm{M}}_n / r \cong \boldsymbol{K}^{\mathrm{M}}_{m+n}/r$.*

Proof By Lemma 5.1.3, we can identify $\mathrm{Hom}(\boldsymbol{K}^{\mathrm{MW}}_{n+1}, \boldsymbol{K}^{\mathrm{MW}}_n) \cong \boldsymbol{K}^{\mathrm{MW}}_{-1}(k) \cong \boldsymbol{W}(k)$ for $n \geq 2$ (the final identification by [44, Lemma 3.10]). The group $\boldsymbol{K}^{\mathrm{MW}}_{-1}(k)$ contains the element η and we refer to the corresponding map $\boldsymbol{K}^{\mathrm{MW}}_{n+1} \to \boldsymbol{K}^{\mathrm{MW}}_n$ using the same notation. Unwinding the definitions, this map corresponds to the composite of the $\boldsymbol{K}^{\mathrm{MW}}_{n+1} \to \boldsymbol{I}^{n+1}$ defined on sections by multiplication by η and the inclusion map $\boldsymbol{I}^{n+1} \hookrightarrow \boldsymbol{K}^{\mathrm{MW}}_n$.

By the discussion of the previous paragraph, the first exact sequence of (5.1.1) yields the exact sequence

$$\boldsymbol{K}^{\mathrm{MW}}_{n+1} \xrightarrow{\eta} \boldsymbol{K}^{\mathrm{MW}}_n \longrightarrow \boldsymbol{K}^{\mathrm{M}}_n \longrightarrow 0.$$

Tensoring this exact sequence with $\boldsymbol{K}^{\mathrm{MW}}_m$ and applying Proposition 5.1.7 we conclude that there is an exact sequence

$$\boldsymbol{K}^{\mathrm{MW}}_{m+n+1} \xrightarrow{\eta} \boldsymbol{K}^{\mathrm{MW}}_{m+n} \longrightarrow \boldsymbol{K}^{\mathrm{MW}}_m \otimes^{\mathbb{A}^1} \boldsymbol{K}^{\mathrm{M}}_n \longrightarrow 0.$$

However, this sequence identifies $\boldsymbol{K}^{\mathrm{MW}}_m \otimes^{\mathbb{A}^1} \boldsymbol{K}^{\mathrm{M}}_n \cong \boldsymbol{K}^{\mathrm{M}}_{m+n}$. Repeating this discussion using the exact sequence $\boldsymbol{K}^{\mathrm{M}}_n \to \boldsymbol{K}^{\mathrm{M}}_n \to \boldsymbol{K}^{\mathrm{M}}_n / r$ we obtain the isomorphism of the statement. Repeating this discussion in the other factor allows us to obtain the final statement. □

Rationalized Milnor–Witt K-theory sheaves Now, we turn our attention to rationalized Milnor–Witt sheaves $\boldsymbol{K}_n^{\mathrm{MW}} \otimes \mathbb{Q}$, which will reappear in Section 5.3.

Lemma 5.1.9 *Fix a base field k, assumed to have characteristic unequal to 2.*

(1) *There is a canonical isomorphism $\boldsymbol{K}_n^{\mathrm{MW}} \otimes \mathbb{Q} \xrightarrow{\sim} \boldsymbol{K}_n^{\mathrm{M}} \otimes \mathbb{Q} \times \boldsymbol{I}^n \otimes \mathbb{Q}$ for every integer n.*

(2) *If k is not formally real, then $\boldsymbol{I}^n \otimes \mathbb{Q}$ is trivial, ie $\boldsymbol{K}_n^{\mathrm{MW}} \otimes \mathbb{Q} \xrightarrow{\sim} \boldsymbol{K}_n^{\mathrm{M}} \otimes \mathbb{Q}$.*

Proof The first statement follows from the fiber product presentation of $\boldsymbol{K}_n^{\mathrm{MW}}$ together with the fact that $\boldsymbol{K}_n^{\mathrm{M}}/2 \otimes \mathbb{Q} = 0$. For the second statement, since $\boldsymbol{I}^n$ is an unramified sheaf, it suffices to show that $\boldsymbol{I}^n(L) \otimes \mathbb{Q} = 0$ for L a finitely generated extension of k. If k is not formally real, then any extension field has the same property, and the result follows immediately from the fact that $\boldsymbol{I}^n(L)$ is a 2–torsion sheaf if L is not formally real [18, Proposition 31.4]. □

Corollary 5.1.10 *Fix a base field k, assumed to have characteristic unequal to 2.*

(1) *$\boldsymbol{K}_n^{\mathrm{MW}} \otimes \mathbb{Q}$ is nontrivial for any integer $n \geq 0$.*

(2) *If k is formally real, then $\boldsymbol{K}_n^{\mathrm{MW}} \otimes \mathbb{Q}$ is nontrivial for any integer n.*

Proof Both tensoring with $\mathbb{Q}$ and contraction are exact endofunctors of the category of strictly $\mathbb{A}^1$–invariant sheaves (see Lemma 5.1.1) and it follows immediately from the definitions that the two constructions commute, ie if $\boldsymbol{M}$ is strictly $\mathbb{A}^1$–invariant, then $(\boldsymbol{M} \otimes \mathbb{Q})_{-1} \cong \boldsymbol{M}_{-1} \otimes \mathbb{Q}$. Thus, to show $\boldsymbol{K}_n^{\mathrm{MW}} \otimes \mathbb{Q}$ is nontrivial, it suffices to show that $(\boldsymbol{K}_n^{\mathrm{MW}} \otimes \mathbb{Q})_{-m} = (\boldsymbol{K}_n^{\mathrm{MW}})_{-m} \otimes \mathbb{Q} \cong \boldsymbol{K}_{n-m}^{\mathrm{MW}} \otimes \mathbb{Q}$ is nontrivial for some $m > 0$. There is a canonical identification $\boldsymbol{K}_{n-m}^{\mathrm{MW}} \otimes \mathbb{Q} \cong \boldsymbol{K}_{n-m}^{\mathrm{M}} \otimes \mathbb{Q} \oplus \boldsymbol{I}^{n-m} \otimes \mathbb{Q}$, by Lemma 5.1.9(1).

For (1), take $m = n$, and observe that $\boldsymbol{K}_0^{\mathrm{M}} \cong \mathbb{Z}$. For (2) observe that if $m > n$, then $\boldsymbol{K}_{n-m}^{\mathrm{MW}} \otimes \mathbb{Q} \cong \boldsymbol{W} \otimes \mathbb{Q}$. Since k is assumed formally real, we can choose an ordering of k [18, Proposition 31.20], and thus find a real closed field k' containing k. In that case, observe that $W(k') \cong \mathbb{Z}$ by Sylvester's law of inertia [18, Proposition 31.5]. Thus, $\boldsymbol{W}(k') \otimes \mathbb{Q}$ is nonzero, so the sheaf $\boldsymbol{W} \otimes \mathbb{Q}$ is nontrivial. □

Remark 5.1.11 Once again, the assumption that k has characteristic unequal to 2 is inessential. This assumption only appears by way of our appeal to Morel's fiber square presentation of $\boldsymbol{K}_n^{\mathrm{MW}}$ (see Remark 5.1.2).

Remark 5.1.12 Rationalized Milnor K-theory sheaves can be quite large. If L is an infinite field, write L^{alg} for an algebraic closure. The Bloch–Kato conjecture [54; 55] implies that for $n \geq 2$, the groups $K_n^M(L^{\mathrm{alg}})$ are (nontrivial) uniquely divisible (these groups are evidently divisible for $n = 1$). By using transfers in Milnor K-theory [11, Section I.5] it is easy to see that the restriction map $K^M_n(L) \to K^M_n(L^{\mathrm{alg}})$ is injective modulo torsion. Any element $\alpha \in K^M_n(L)$ that goes to zero in $K^M_n(L^{\mathrm{alg}})$ necessarily goes to zero in a finite extension L'/L. In that case, the composite $K^M_n(L) \to K^M_n(L') \to K^M_n(L)$ of restriction with transfer is multiplication by the degree. Thus, $[L' : L]\alpha = 0$, ie α is torsion. Equivalently, one can use the identification of Milnor K-theory with motivic cohomology [38, Theorem 5.1] and transfers there.

5.2 On $\mathbb{A}^1$–homotopy sheaves of spheres

The goal of this section is to establish Conjecture 5 of [2]. The results below depends rather heavily on the results of [5] and thus we assume throughout this section that k is an infinite perfect field of characteristic unequal to 2.

On the computation of $\boldsymbol{\pi}^{\mathbb{A}^1}_{3+j\alpha}(S^{2+3\alpha})$ To begin, we recall some results from [5] where we used the notation $\boldsymbol{\pi}_{3,j}(\mathbb{A}^3 \setminus 0)$ for the sheaf in the title. One begins by considering the fiber sequence

$$\mathrm{SL}_4 / \mathrm{Sp}_4 \longrightarrow \mathrm{SL}_6 / \mathrm{Sp}_6 \longrightarrow \mathbb{A}^5 \setminus 0. \tag{5.2.1}$$

A stable range was described for the homotopy sheaves of $\mathrm{SL}_{2n} / \mathrm{Sp}_{2n}$ in [5, Proposition 4.2.2] in terms of Grothendieck–Witt sheaves (see [5, Sections 3.1 and 3.3] and the references there for explication of the notation). Also obtained there was a short exact sequence of sheaves of the form

$$\boldsymbol{GW}^3_5 \longrightarrow \boldsymbol{K}^{\mathrm{MW}}_5 \longrightarrow \boldsymbol{\pi}^{\mathbb{A}^1}_3(S^{2+3\alpha}) \longrightarrow \boldsymbol{GW}^3_4 \longrightarrow 0. \tag{5.2.2}$$

The cokernel of morphism $\boldsymbol{GW}^3_5 \to \boldsymbol{K}^{\mathrm{MW}}_5$ was called $\boldsymbol{F}_5$ and a description of $\boldsymbol{F}_5$ was given in [5, Theorem 4.4.1]. Before discussing the structure of this morphism, we introduce a further convention to simplify the notation.

Convention 5.2.1 Write $\Omega(-)$ for the $\mathbb{A}^1$–derived loop functor, ie $\Omega \mathrm{L}_{\mathbb{A}^1}(-)$.

The map $\boldsymbol{K}^{\mathrm{MW}}_5 \to \boldsymbol{\pi}^{\mathbb{A}^1}_3(S^{2+3\alpha})$ is by construction induced by a morphism $\Omega S^{4+5\alpha} \to S^{2+3\alpha}$. The composite map

$$\delta\colon S^{3+5\alpha} \longrightarrow \Omega S^{4+5\alpha} \longrightarrow S^{2+3\alpha}$$

was shown to be a generator of $\boldsymbol{\pi}_{3+5\alpha}(S^{2+3\alpha})$ in [5, Proposition 5.2.1]. We deduce a few simple consequences of these results now.

Lemma 5.2.2 *Suppose $j \geq 6$ is an integer.*

(1) *There is a canonical isomorphism $\boldsymbol{\pi}^{\mathbb{A}^1}_{3+j\alpha}(S^{2+3\alpha}) \cong \boldsymbol{W}$.*

(2) *Any $\mathbb{A}^1$–homotopy class of maps $S^{3+j\alpha} \to S^{2+3\alpha}$ lifts uniquely along δ to a map $S^{3+j\alpha} \to S^{3+5\alpha}$.*

Proof For point (1), begin by applying [44, Theorem 6.13] to the exact sequence of (5.2.2). By [5, Proposition 3.4.3], we observe that $(\boldsymbol{GW}^3_4)_{-j} = 0$ for $j \geq 5$. By [5, Lemma 3.4.1], if $j = 5$, we conclude that $(\boldsymbol{GW}^3_5)_{-5} = \boldsymbol{GW}^2_0 \cong \mathbb{Z}$. Therefore, $(\boldsymbol{GW}^3_5)_{-j} = 0$ for $j \geq 6$. Thus, we conclude that there is a sequence of isomorphisms

$$\boldsymbol{\pi}^{\mathbb{A}^1}_{3+j\alpha}(S^{2+3\alpha}) \cong \boldsymbol{\pi}^{\mathbb{A}^1}_{3+j\alpha}(\Omega S^{4+5\alpha}) \cong \boldsymbol{\pi}^{\mathbb{A}^1}_{4+j\alpha}(S^{4+5\alpha}) \cong (\boldsymbol{K}^{\mathrm{MW}}_5)_{-j} \cong \boldsymbol{W}$$

if $j \geq 6$.

For point (2), take a map $\phi\colon S^{3+j\alpha} \to S^{2+3\alpha}$ as in the statement. Mapping $S^{3+j\alpha}$ into the fiber sequence of (5.2.2), the argument of point (1) shows that, for $j \geq 6$, such a map lifts uniquely to a map $S^{3+j\alpha} \to \Omega S^{4+5\alpha}$. Since $S^{3+5\alpha}$ is $\mathbb{A}^1$–2–connected, the unit of the loop-suspension adjunction $S^{3+5\alpha} \to \Omega S^{4+5\alpha}$ induces an isomorphism on $\mathbb{A}^1$–homotopy sheaves in degrees ≤ 4 by, eg, Theorem 3.2.1 and Remark 3.2.3. Therefore, ϕ lifts uniquely along δ. □

On the computation of $\pi^{\mathbb{A}^1}_{j+1}(S^{j+3\alpha})$ for $j \geq 3$

Proposition 5.2.3 *If k is a field of characteristic 0 and containing a quadratically closed subfield, then $\pi^{\mathbb{A}^1}_{4+6\alpha}(S^{3+3\alpha}) = 0$.*

Proof With $\mathscr{X}$ equal to $S^{2+3\alpha}$, which is $\mathbb{A}^1$–1–connected, Theorem 3.3.13, the exactness of contraction and [44, Theorem 6.13] yield the following exact sequence:

$$\boldsymbol{\pi}^{\mathbb{A}^1}_{5+6\alpha}(S^{3+3\alpha}) \to \boldsymbol{\pi}^{\mathbb{A}^1}_{5+6\alpha}(S^{5+6\alpha}) \xrightarrow{\mathrm{P}} \boldsymbol{\pi}^{\mathbb{A}^1}_{3+6\alpha}(S^{2+3\alpha}) \to \boldsymbol{\pi}^{\mathbb{A}^1}_{4+6\alpha}(S^{3+3\alpha}) \to 0. \tag{5.2.3}$$

We have $\boldsymbol{\pi}^{\mathbb{A}^1}_{5+6\alpha}(S^{5+6\alpha}) \cong \boldsymbol{K}^{\mathrm{MW}}_0$ again by Morel's computations [44, Theorem 1.23]. By Theorem 4.2.1, the morphism P is induced by composition with

$$[\iota_{S^{2+3\alpha}}, \iota_{S^{2+3\alpha}}]\colon S^{3+6\alpha} \longrightarrow S^{2+3\alpha};$$

we will refer to this map as composition with the Whitehead square of the identity. By Lemma 5.2.2(2), this map lifts uniquely through δ to a map

$$[\iota_{S^{2+3\alpha}}, \iota_{S^{2+3\alpha}}]\colon S^{3+6\alpha} \longrightarrow S^{3+5\alpha},$$

which (by [44, Corollary 6.43]) can be viewed as an element of $\boldsymbol{W}(k)$. The exact sequence (5.2.3) becomes, by Lemma 5.2.2(1),

$$\boldsymbol{K}^{\mathrm{MW}}_0 \xrightarrow{\mathrm{P}} \boldsymbol{W} \longrightarrow \boldsymbol{\pi}^{\mathbb{A}^1}_{4+6\alpha}(S^{3+3\alpha}) \longrightarrow 0.$$

We claim each of the sheaves in the above exact sequence are sheaves of $\boldsymbol{K}_0^{\mathrm{MW}}$–modules in a natural way, and that the morphisms are morphisms of sheaves of $\boldsymbol{K}_0^{\mathrm{MW}}$–modules. To see this, it suffices to observe that the portion of the $\mathbb{A}^1$–EHP sequence under consideration takes the form $\boldsymbol{\pi}_{3+6\alpha}(\Omega^2 S^{5+6\alpha}) \to \boldsymbol{\pi}^{\mathbb{A}^1}_{3+6\alpha}(S^{2+3\alpha}) \to \boldsymbol{\pi}^{\mathbb{A}^1}_{3+6\alpha}(\Omega S^{3+3\alpha})$, and the $\boldsymbol{K}_0^{\mathrm{MW}}$–module structure is induced by precomposition with $\boldsymbol{\pi}_{3+6\alpha}(S^{3+6\alpha})$. From these observations it follows that the map P is determined by an element of $\mathrm{Hom}_{\boldsymbol{K}_0^{\mathrm{MW}}}(\boldsymbol{K}_0^{\mathrm{MW}}, \boldsymbol{W}) = \boldsymbol{W}(k)$.

Assume first that k is a quadratically closed field of characteristic 0. In that case $\boldsymbol{W}(k) = \mathbb{Z}/2$, and, to establish the claim, it suffices to prove that our morphism is nontrivial. To see this, fix an embedding $k \hookrightarrow \mathbb{C}$. Using complex realization (see [45, Section 3, Lemma 3.4] or [16]), and the fact that complex realization takes spheres to spheres, it suffices to prove that composition with the Whitehead square of the identity is nontrivial after taking $\mathbb{C}$–points. Serre showed that $\pi_9(S^5) = \mathbb{Z}/2$ and that the Whitehead square of the identity on S^5 is a generator, [50, Section 41]. Consequently, we conclude that the our morphism P also corresponds to the nontrivial element of $\boldsymbol{W}(k)$ and is therefore an epimorphism.

If L/k is an extension field, then the morphism P in our sequence viewed over the base field L is pulled back from the morphism P over k. Thus, by appeal to the conclusion of the previous paragraph, we conclude in this case as well that the morphism $\boldsymbol{K}_0^{\mathrm{MW}} \to \boldsymbol{W}$ is necessarily the standard epimorphism, and therefore that $\boldsymbol{\pi}^{\mathbb{A}^1}_{4+6\alpha}(S^{3+3\alpha}) = 0$. □

Remark 5.2.4 In the preceding proof, the assumption that k has characteristic 0 can likely be weakened to the assumption that k has characteristic unequal to 2 via appeal to étale realization [28]. As a consequence, the same remark applies to all statements below appealing to Proposition 5.2.3. Removing the assumption that k contains a quadratically closed subfield will probably require different techniques. Nevertheless, it seems likely that the "lifted" map $[\iota_{S^{2+3\alpha}}, \iota_{S^{2+3\alpha}}]\colon S^{3+6\alpha} \to S^{3+5\alpha}$ is simply a suspension of η and the above result can be established without reference to realization of any sort.

The above vanishing statement has a number of useful consequences.

Theorem 5.2.5 *If k is a field of characteristic 0 and containing a quadratically closed subfield, then for every integer $j \geq 3$, there is an exact sequence of the form*

$$0 \longrightarrow \boldsymbol{F}_5' \longrightarrow \boldsymbol{\pi}^{\mathbb{A}^1}_{j+1}(S^{j+3\alpha}) \longrightarrow \boldsymbol{GW}^3_4 \longrightarrow 0, \tag{5.2.4}$$

together with an epimorphism $\boldsymbol{K}_5^{\mathrm{M}}/24 \to \boldsymbol{F}_5'$ that becomes an isomorphism after 4–fold contraction. Moreover, the composite map $\boldsymbol{K}_5^{\mathrm{M}}/24 \to \boldsymbol{\pi}_4^{\mathbb{A}^1}(S^{3+3\alpha})$ determines an isomorphism $\mathbb{Z}/24 \cong \boldsymbol{\pi}^{\mathbb{A}^1}_{4+5\alpha}(S^{3+3\alpha})$.

Proof We treat the case where $j = 3$, building upon the analysis in the proof of Proposition 5.2.3. Since $S^{3+3\alpha}$ is $\mathbb{A}^1$–2–connected, the case $j \geq 4$ will follow immediately from this case and the $\mathbb{A}^1$–simplicial suspension theorem (Theorem 3.2.1 and Remark 3.2.3).

Take $\mathscr{X} = S^{2+3\alpha}$ in Theorem 3.3.13 and consider the map P: $\boldsymbol{K}_6^{\mathrm{MW}} = \boldsymbol{\pi}_5^{\mathbb{A}^1}(S^{5+6\alpha}) \to \boldsymbol{\pi}_3^{\mathbb{A}^1}(S^{2+3\alpha})$. Recall the exact sequence of (5.2.2), which appears as the horizontal line of (5.2.5):

$$
\begin{tikzcd}
 & & \boldsymbol{\pi}_5^{\mathbb{A}^1}(S^{5+6\alpha}) = \boldsymbol{K}_6^{\mathrm{MW}} \arrow[d, "\mathrm{P}"] \arrow[dr, dotted] & & \\
\boldsymbol{GW}^3_5 \arrow[r] & \boldsymbol{K}_5^{\mathrm{MW}} \arrow[r] & \boldsymbol{\pi}_3^{\mathbb{A}^1}(S^{2+3\alpha}) \arrow[r] \arrow[d, "\mathrm{E}"] & \boldsymbol{GW}^3_4 \arrow[r] & 0 \\
 & & \boldsymbol{\pi}_4^{\mathbb{A}^1}(S^{3+3\alpha}) \arrow[ur, dashed] \arrow[d] & & \\
 & & 0 & &
\end{tikzcd}
\tag{5.2.5}
$$

The vertical sequence is the EHP sequence applied to $S^{2+3\alpha}$. The dotted diagonal map is an element of $\mathrm{Hom}(\boldsymbol{K}_6^{\mathrm{MW}}, \boldsymbol{GW}^3_4) \cong (\boldsymbol{GW}^3_4)_{-6}(k)$ by Lemma 5.1.3. On the other hand, [5, Proposition 3.4.3] allows us to conclude that $(\boldsymbol{GW}^3_4)_{-6} = 0$ so this diagonal map vanishes, and therefore there is an induced epimorphism, denoted by the dashed diagonal arrow in (5.2.5), $\boldsymbol{\pi}_4^{\mathbb{A}^1}(S^{3+3\alpha}) \to \boldsymbol{GW}^3_4$, as required by the theorem. By combining a portion of diagram (5.2.5) with the exact sequence $0 \to \boldsymbol{I}^6 \to \boldsymbol{K}_5^{\mathrm{MW}} \to \boldsymbol{K}_5^{\mathrm{M}} \to 0$ we obtain diagram (5.2.6), ignoring the dotted arrow for the moment:

$$
\begin{tikzcd}
 & 0 \arrow[d] & & & \\
 & \boldsymbol{I}^6 \arrow[d] & & & \\
\boldsymbol{GW}^3_5 \arrow[r] & \boldsymbol{K}_5^{\mathrm{MW}} \arrow[r] \arrow[d] & \boldsymbol{\pi}_3^{\mathbb{A}^1}(S^{2+3\alpha}) \arrow[r] \arrow[d, "\mathrm{E}"] & \boldsymbol{GW}^3_4 \arrow[r] \arrow[d, equal] & 0 \\
 & \boldsymbol{K}_5^{\mathrm{M}} \arrow[r, dotted] \arrow[d] & \boldsymbol{\pi}_4^{\mathbb{A}^1}(S^{3+3\alpha}) \arrow[r] \arrow[d] & \boldsymbol{GW}^3_4 \arrow[r] & 0 \\
 & 0 & 0 & &
\end{tikzcd}
\tag{5.2.6}
$$

As we know that $\boldsymbol{\pi}_4^{\mathbb{A}^1}(S^{3+3\alpha})_{-6} = 0$, it follows from Lemma 5.1.5 that the composite map $\boldsymbol{K}_5^{\mathrm{MW}} \to \boldsymbol{\pi}_4^{\mathbb{A}^1}(S^{3+3\alpha})$ factors through $\boldsymbol{K}_5^{\mathrm{MW}}/\boldsymbol{I}^6 = \boldsymbol{K}_5^{\mathrm{M}}$, giving the dotted arrow in diagram (5.2.6).

We define $\boldsymbol{F}_5$, as in [5], to be the cokernel of the map $\boldsymbol{GW}_5^3 \to \boldsymbol{K}_5^{\mathrm{MW}}$, and define $\boldsymbol{F}_5'$ to be the image of $\boldsymbol{F}_5$ in $\boldsymbol{\pi}_4^{\mathbb{A}^1}(S^{3+3\alpha})$. The exact sequence

$$0 \longrightarrow \boldsymbol{F}_5' \longrightarrow \boldsymbol{\pi}_4^{\mathbb{A}^1}(S^{3+3\alpha}) \longrightarrow \boldsymbol{GW}_4^3 \longrightarrow 0$$

is an immediate consequence of this definition. Furthermore, there is a diagram of exact sequences:

$$\begin{tikzcd}
0 \arrow[r] & \boldsymbol{I}^6 \arrow[r] \arrow[d, equal] & \boldsymbol{K}_5^{\mathrm{MW}} \arrow[r] \arrow[d, two heads] & \boldsymbol{K}_5^{\mathrm{M}} \arrow[r] \arrow[d, two heads] & 0 \\
& \boldsymbol{I}^6 \arrow[r] & \boldsymbol{F}_5 \arrow[r] & \boldsymbol{F}_5' \arrow[r] & 0
\end{tikzcd} \tag{5.2.7}$$

To determine the behavior of $\boldsymbol{F}_5'$, we need finer information regarding the sheaf $\boldsymbol{F}_5$ as described in [5, Theorems 4.3.1 and 4.4.1]. We provide a brief recapitulation of that description here. The sheaf $\boldsymbol{F}_5$ is identified there as a quotient of a fibered product as follows. One defines a sheaf $\boldsymbol{S}_5$, the cokernel of a "Chern class" map $\boldsymbol{K}_5^{Q} \to \boldsymbol{K}_5^{\mathrm{M}}$ [3, Definition 3.6]. The sheaf $\boldsymbol{S}_5$ is equipped with a canonical surjection onto $\boldsymbol{K}_5^{\mathrm{M}}/2$ (see [3, Lemma 3.13] and the subsequent discussion). One then defines a sheaf $\boldsymbol{T}_5$ to be the fiber product of $\boldsymbol{S}_5$ and $\boldsymbol{I}^5$ over $\boldsymbol{K}_5^{\mathrm{M}}/2$ [3, page 911]; the maps $\boldsymbol{I}^5 \to \boldsymbol{K}_5^{\mathrm{M}}/2$ and $\boldsymbol{K}_5^{\mathrm{M}} \to \boldsymbol{S}_5 \to \boldsymbol{K}_5^{\mathrm{M}}/2$ coincide with the defining maps in the fiber product presentation in $\boldsymbol{K}_5^{\mathrm{MW}}$. By [5, Theorem 4.3.1] (see [5, Theorem 4.4.1]), there is an epimorphism $\boldsymbol{T}_5 \to \boldsymbol{F}_5$ and this epimorphism becomes an isomorphism after 4–fold contraction by [5, Lemma 5.1.1]. Assembling all the above, there is a diagram of short exact sequences, enlarging (5.2.7),

$$\begin{tikzcd}
0 \arrow[r] & \boldsymbol{I}^6 \arrow[r] \arrow[d, equal] & \boldsymbol{K}_5^{\mathrm{MW}} \arrow[r] \arrow[d, two heads] & \boldsymbol{K}_5^{\mathrm{M}} \arrow[r] \arrow[d, two heads] & 0 \\
0 \arrow[r] & \boldsymbol{I}^6 \arrow[r] \arrow[d, equal] & \boldsymbol{T}_5 \arrow[r] \arrow[d, two heads, "\phi"] & \boldsymbol{S}_5 \arrow[r] \arrow[d, two heads, "\phi'"] & 0 \\
& \boldsymbol{I}^6 \arrow[r] & \boldsymbol{F}_5 \arrow[r] & \boldsymbol{F}_5' \arrow[r] & 0
\end{tikzcd} \tag{5.2.8}$$

where ϕ, and therefore ϕ', becomes an isomorphism after 4–fold contraction.

There is an epimorphism $\boldsymbol{K}^{\mathrm{M}}_5/24 \twoheadrightarrow \boldsymbol{S}_5$ that becomes an isomorphism after 4–fold contraction, [3, Corollary 3.11]. It follows there is such an epimorphism $\boldsymbol{K}^{\mathrm{M}}_5/24 \twoheadrightarrow \boldsymbol{F}'_5$ as well.

Since $(\boldsymbol{K}^{\mathrm{M}}_5)_{-5} \cong \mathbb{Z}$, and $(\boldsymbol{GW}^3_4)_{-5} = 0$, the latter by [5, Proposition 3.4.3], the 5–fold contraction of the sequence

$$0 \longrightarrow \boldsymbol{F}'_5 \longrightarrow \boldsymbol{\pi}^{\mathbb{A}^1}_4(S^{3+3\alpha}) \longrightarrow \boldsymbol{GW}^3_4 \longrightarrow 0$$

reduces to an isomorphism $\mathbb{Z}/24 \cong \boldsymbol{\pi}^{\mathbb{A}^1}_{4+5\alpha}(S^{3+3\alpha})$. □

Remark 5.2.6 Because of the observation of [5, Remark 5.1.2], we do not know whether the map $\boldsymbol{K}^{\mathrm{M}}_5/24 \to \boldsymbol{F}'_5$ of Theorem 5.2.5 is an isomorphism after 3–fold contraction. Nevertheless, it seems likely that this is the case.

Consider the motivic Hopf map $\nu\colon S^{3+4\alpha} \to S^{2+2\alpha}$. The standard construction of this map is via the Hopf construction [44, page 190] on the multiplication map $\mathrm{SL}_2 \times \mathrm{SL}_2 \to \mathrm{SL}_2$.

Corollary 5.2.7 *If k is a field having characteristic 0 and containing a quadratically closed subfield, then for every integer $j \geq 3$, the group $\boldsymbol{\pi}^{\mathbb{A}^1}_{j+1+5\alpha}(S^{j+3\alpha}) \cong \mathbb{Z}/24$ is generated by $\Sigma^{j-2+\alpha}\nu$.*

Proof This follows by combining Theorem 5.2.5 and [5, Corollary 5.3.1]. □

Remark 5.2.8 Since $\eta\colon S^{1+2\alpha} \to S^{1+\alpha}$ we can consider $\nu \wedge \eta\colon S^{4+6\alpha} \to S^{3+3\alpha}$. Proposition 5.2.3 then guarantees that $\nu \wedge \eta$ and $\eta \wedge \nu$ are null-homotopic. Since they remain null-homotopic after suspension, we obtain a purely unstable proof of one of the motivic null-Hopf relations [17, Proposition 5.4].

Similarly, if $\eta_s\colon S^3_s \to S^2_s$ is the simplicial Hopf map, then we can consider the composite map

$$\Sigma^{2+2\alpha}\eta \circ \Sigma^{2+3\alpha}\eta \circ \Sigma^{2+4\alpha}\eta \circ \Sigma^{1+6\alpha}\eta_s\colon S^{4+6\alpha} \longrightarrow S^{3+3\alpha}.$$

Once again, Proposition 5.2.3 implies this composition is null-homotopic. Stabilizing with respect to $\mathbb{P}^1$–suspension, this implies the relation $\eta^3\eta_s = 0$ in the motivic stable homotopy ring. This relation is an incarnation of the fact that the topological Hopf map η_{top} satisfies $\eta^4_{\mathrm{top}} = 0$.

The existence of such null-homotopies allows us to construct new elements in unstable homotopy sheaves of motivic spheres using Toda brackets [52]. It would be interesting to study such constructions more systematically.

On the structure of $\pi^{\mathbb{A}^1}_{n+1}(S^{n-1+n\alpha})$ for $n \geq 4$ Finally, we are able to establish [2, Conjecture 5] under the additional hypothesis that our base field contains a quadratically closed field having characteristic 0.

Theorem 5.2.9 *Suppose k is a field that contains a quadratically closed field having characteristic 0. For every integer $n \geq 4$, the map $\nu_n := \Sigma^{(n-2)+(n-2)\alpha}\nu$ induces a nontrivial morphism*

$$(\nu_n)_*\colon \boldsymbol{K}^{\mathrm{M}}_{n+2}/24 \longrightarrow \boldsymbol{\pi}^{\mathbb{A}^1}_{n+1}(S^{n+n\alpha}).$$

Proof This follows essentially from Corollary 5.2.7. In more detail, the map $(\nu_n)_*$ determines a morphism $\boldsymbol{K}^{\mathrm{MW}}_{n+2} \to \boldsymbol{\pi}^{\mathbb{A}^1}_{n+1}(S^{n+n\alpha})$, but by construction, this morphism factors through $\mathbb{P}^1$–suspension. In particular, since the map $\boldsymbol{K}^{\mathrm{MW}}_5 \to \boldsymbol{\pi}^{\mathbb{A}^1}_4(S^{3+3\alpha})$ factors through a morphism $\boldsymbol{K}^{\mathrm{M}}_5/24 \to \boldsymbol{\pi}^{\mathbb{A}^1}_4(S^{3+3\alpha})$, we conclude that for any integer $n \geq 4$, the morphism $\boldsymbol{K}^{\mathrm{MW}}_{n+2} \to \boldsymbol{\pi}^{\mathbb{A}^1}_{n+1}(S^{n+n\alpha})$ factors through a map $\boldsymbol{K}^{\mathrm{MW}}_{n-3} \otimes^{\mathbb{A}^1} \boldsymbol{K}^{\mathrm{M}}_5/24 \to \boldsymbol{\pi}^{\mathbb{A}^1}_{n+1}(S^{n+n\alpha})$. Lemma 5.1.8 allows us to conclude $\boldsymbol{K}^{\mathrm{MW}}_{n-3} \otimes^{\mathbb{A}^1} \boldsymbol{K}^{\mathrm{M}}_5/24 \cong \boldsymbol{K}^{\mathrm{M}}_{n+2}/24$, which is precisely what we wanted to show. □

Recall that in [7, Theorem 5], a morphism

$$\boldsymbol{\pi}^{\mathbb{A}^1}_{n+1}(S^{n+n\alpha}) \longrightarrow \boldsymbol{GW}^n_{n+1}$$

is constructed using "Suslin matrices". The composite map

$$\boldsymbol{K}^{\mathrm{MW}}_{n+2} \longrightarrow \boldsymbol{\pi}^{\mathbb{A}^1}_{n+1}(S^{n+n\alpha}) \longrightarrow \boldsymbol{GW}^n_{n+1}$$

is, by means of Lemma 5.1.3, determined by an element of $(\boldsymbol{GW}^n_{n+1})_{-n-2}(k)$; since the latter group is trivial by [5, Proposition 3.4.3], we conclude that this composite is trivial. Combining these observations with Theorem 5.2.9 and the connectivity estimate from the $\mathbb{A}^1$–simplicial suspension theorem (see Theorem 3.2.1 and Remark 3.2.3), we now refine [2, Conjecture 7].

Conjecture 5.2.10 *For any pair of integers $n \geq 4$ and $i \geq 0$, there is an exact sequence of the form*

$$\boldsymbol{K}^{\mathrm{M}}_{n+2}/24 \longrightarrow \boldsymbol{\pi}^{\mathbb{A}^1}_n(S^{(n-1+i)+n\alpha}) \longrightarrow \boldsymbol{GW}^n_{n+1};$$

the right-hand map becomes an epimorphism after $(n-3)$–fold contraction, and the sequence becomes a short exact sequence after n–fold contraction.

Remark 5.2.11 In private communication from 2005, Morel stated a conjecture about the stable $\boldsymbol{\pi}_1$ sheaf of the motivic sphere spectrum. Conjecture 5.2.10 can be thought of as an unstable refinement of Morel's conjecture. Morel's conjecture has been verified in various situations. K Ormsby and P-A Østvær verified Morel's conjecture after taking sections over fields of small cohomological dimension [47]. Much more generally, work of Østvær, O Röndigs and M Spitzweck has verified Morel's conjecture over fields having characteristic 0 [49] (or, more generally, after inverting the characteristic exponent of the base field). While these results provide evidence for Conjecture 5.2.10, without a version of the suspension theorem for $\mathbb{P}^1$–suspension these stable results do not imply our conjecture.

5.3 Other computations

In this section, we establish nontriviality of unstable rationalized $\mathbb{A}^1$–homotopy sheaves of motivic spheres. We then go on to compute the first S^1–stable $\mathbb{A}^1$–homotopy sheaf of a mod m motivic Eilenberg–MacLane space.

Rationalized $\mathbb{A}^1$–homotopy sheaves of spheres The computations of Morel of $\mathbb{A}^1$–homotopy sheaves of spheres yield isomorphisms $\boldsymbol{\pi}^{\mathbb{A}^1}_{2n-1} S^{2n-1+2q\alpha} \cong \boldsymbol{K}^{\mathrm{MW}}_{2q}$ for $2n-1 \geq 2$ [44, Theorem 6.40]. By [44, Theorem 6.13] (see also [44, Corollary 6.43]), for any integer j there are induced isomorphisms $\pi^{\mathbb{A}^1}_{2n-1+j\alpha} S^{2n-1+2q\alpha} \cong (\boldsymbol{K}^{\mathrm{MW}}_{2q})_{-j}$. In these degrees, the James–Hopf invariant map H of Section 3.2 yields a morphism

$$\mathrm{H}\colon \boldsymbol{\pi}^{\mathbb{A}^1}_{2n-1+j\alpha} S^{n+q\alpha} \longrightarrow \boldsymbol{\pi}^{\mathbb{A}^1}_{2n-1+j\alpha} S^{2n-1+2q\alpha} \cong \boldsymbol{K}^{\mathrm{MW}}_{2q-j}.$$

We now study the rationalized version of this map. The next result provides an analog of the fact, due to Hopf, that there is a surjection $\pi_{4n-1}(S^{2n}) \to \mathbb{Z}$.

Theorem 5.3.1 *Fix a base field k, assumed to be perfect and to have characteristic unequal to 2. Let $n > 2$ and $q \geq 2$ be even integers.*

(1) *For any integer $j \geq 0$, the sequence of sheaves*

$$\boldsymbol{\pi}^{\mathbb{A}^1}_{2n-2+j\alpha} S^{n-1+q\alpha} \otimes \mathbb{Q} \xrightarrow{\mathrm{E}\otimes\mathbb{Q}} \boldsymbol{\pi}^{\mathbb{A}^1}_{2n-1+j\alpha} S^{n+q\alpha} \otimes \mathbb{Q} \xrightarrow{\mathrm{H}\otimes\mathbb{Q}} \boldsymbol{K}^{\mathrm{MW}}_{2q-j} \otimes \mathbb{Q} \longrightarrow 0$$

is exact.

(2) *If k is not formally real, then for any integer j satisfying $0 \leq j \leq 2q$, the sheaf $\boldsymbol{\pi}^{\mathbb{A}^1}_{2n-1+j\alpha} S^{n+q\alpha} \otimes \mathbb{Q}$ is nontrivial.*

(3) *If k is formally real, then for any integer $j \geq 0$, the sheaf $\boldsymbol{\pi}^{\mathbb{A}^1}_{2n-1+j\alpha} S^{n+q\alpha} \otimes \mathbb{Q}$ is nontrivial.*

Proof Tensoring with $\mathbb{Q}$ and contraction are exact functors on the category of strictly $\mathbb{A}^1$–invariant sheaves of abelian groups (see Lemma 5.1.1). Combining [44, Theorem 6.13] with the exact sequence of Theorem 3.2.1 (which applies since $n \geq 3$ by assumption) and then tensoring with $\mathbb{Q}$, we obtain exactness of the above sequence at $\pi_{2n-1}^{\mathbb{A}^1} S^{n+q\alpha} \otimes \mathbb{Q}$. Since n and q are even by assumption, the class $1+(-1)^{n+q}\langle -1\rangle^q$ equals 2. Surjectivity of $\mathrm{H}\otimes\mathbb{Q}$ follows from Theorem 4.4.1.

Points (2) and (3) follow immediately from Corollary 5.1.10. □

Remark 5.3.2 A corresponding statement holds for $q=0$ as well, but that result follows immediately from the classical computation of nonzero rational homotopy groups of spheres.

Some S^1–stable $\mathbb{A}^1$–homotopy sheaves of motivic Eilenberg–MacLane spaces Set $K_n := K(\mathbb{Z}(n), 2n)$ and $K_n/m := K(\mathbb{Z}/m(n), 2n)$ where for an abelian group A, the space $K(A(n), 2n)$ is a motivic Eilenberg–MacLane space in the sense of Voevodsky; see, for example, [53, Section 2]. We write $\boldsymbol{H}^i_{\text{ét}}(\mu_m^{\otimes n})$ for the Nisnevich sheafification of the presheaf $U \mapsto H^i_{\text{ét}}(U, \mu_m^{\otimes m})$. In the next result, which is an analog of a result appearing in [9, Example 5.11], we adhere to Convention 5.2.1.

Theorem 5.3.3 *Assume k is a field having characteristic exponent p. Fix integers $i \geq 1$ and $m, n \geq 2$ and assume m is coprime to p.*

1. *The space $\Sigma^i K_n/m$ is $\mathbb{A}^1$–$(n+i-1)$–connected.*
2. *If j is an integer satisfying $0 \leq j \leq n-1$, then there are isomorphisms of the form*
$$\boldsymbol{\pi}^{\mathbb{A}^1}_{n+j+i}(\Sigma^i K_n/m) \xrightarrow{\sim} \boldsymbol{H}^{n-j}_{\text{ét}}(\mu_m^{\otimes n}).$$
3. *There is an exact sequence of the form*
$$\boldsymbol{H}^0_{\text{ét}}(\mu_m^{\otimes n}) \longrightarrow \boldsymbol{\pi}^{\mathbb{A}^1}_{2n+i}(\Sigma^i K_n/m) \longrightarrow \boldsymbol{K}^{\mathrm{M}}_{2n}/m \longrightarrow 0,$$
and $\boldsymbol{H}^0_{\text{ét}}(\mu_m^{\otimes n})$ is killed by a single contraction.

Proof The space K_n/m is $\mathbb{A}^1$–$(n-1)$–connected and it is possible to describe all its higher $\mathbb{A}^1$–homotopy sheaves. If m is prime to p by the Bloch–Kato conjecture (in Beilinson–Lichtenbaum form) together with $\mathbb{A}^1$–representability of mod m motivic cohomology [54; 55], there are isomorphisms of the form $\boldsymbol{\pi}^{\mathbb{A}^1}_{n+r}(K_n/m) \cong \boldsymbol{H}^{n-r}_{\text{ét}}(\mu_m^{\otimes n})$. In particular, $\boldsymbol{H}^{n-r}_{\text{ét}}(\mu_m^{\otimes n})$ is isomorphic to $\boldsymbol{K}^{\mathrm{M}}_n/m$ for $r=0$ and vanishes for $r>n$.

We begin by investigating what occurs after a single suspension. By the $\mathbb{A}^1$–Freudenthal suspension theorem, the map $\boldsymbol{\pi}^{\mathbb{A}^1}_r(K_n/m) \to \boldsymbol{\pi}^{\mathbb{A}^1}_{r+1}(\Sigma K_n/m)$ is an isomorphism for $r \leq 2n-2$. We now show that this map is an isomorphism for $r = 2n-1$ as well.

Theorem 3.3.13 applied with $\mathscr{X} = K_n/m$ yields an exact sequence of the form
$$\boldsymbol{\pi}^{\mathbb{A}^1}_n(K_n/m) \otimes^{\mathbb{A}^1} \boldsymbol{\pi}^{\mathbb{A}^1}_n(K_n/m) \longrightarrow \boldsymbol{\pi}^{\mathbb{A}^1}_{2n-1}(K_n/m) \longrightarrow \boldsymbol{\pi}^{\mathbb{A}^1}_{2n}(\Sigma K_n/m) \longrightarrow 0.$$

By the discussion of the previous paragraph combined with Lemma 5.1.8, we conclude $\boldsymbol{\pi}_n^{\mathbb{A}^1}(K_n/m) \otimes^{\mathbb{A}^1} \boldsymbol{\pi}_n^{\mathbb{A}^1}(K_n/m) \cong \boldsymbol{K}_n^{\mathrm{M}}/m \otimes^{\mathbb{A}^1} \boldsymbol{K}_n^{\mathrm{M}}/m \cong \boldsymbol{K}_{2n}^{\mathrm{M}}/m$. Thus, in the above exact sequence the left-hand map is a map $\boldsymbol{K}_{2n}^{\mathrm{M}}/m \cong \boldsymbol{\pi}_n^{\mathbb{A}^1}(K_n/m) \otimes^{\mathbb{A}^1} \boldsymbol{\pi}_n^{\mathbb{A}^1}(K_n/m) \to \boldsymbol{\pi}_{2n-1}^{\mathbb{A}^1}(K_n/m)$.

One knows $(\boldsymbol{H}_{\text{ét}}^i(\mu_m^{\otimes n}))_{-s} \cong \boldsymbol{H}_{\text{ét}}^{i-s}(\mu_m^{\otimes n-s})$ (appeal to [38, Example 23.3] and sheafify). Since étale cohomology vanishes in negative degrees, we conclude that $\boldsymbol{H}_{\text{ét}}^1(\mu_m^{\otimes n})$ is killed by 2–contractions. Since there is an epimorphism $\boldsymbol{K}_{2n}^{\mathrm{MW}} \to \boldsymbol{K}_{2n}^{\mathrm{M}}/m$, and $(\boldsymbol{K}_1^{\mathrm{M}}/m)_{-2n} = 0$, by appealing to Lemma 5.1.3, we may conclude that the left-hand morphism in the exact sequence displayed in the previous paragraph is the trivial map. Therefore, $\boldsymbol{H}_{\text{ét}}^1(\mu_m^{\otimes n}) \cong \boldsymbol{\pi}_{2n-1}^{\mathbb{A}^1}(K_n/m) \to \boldsymbol{\pi}_{2n}^{\mathbb{A}^1}(\Sigma K_n/m)$ is an isomorphism.

In light of the discussion above, by reading the exact sequence of Theorem 3.3.13 farther to the left, we conclude that there is a short exact sequence of the form

$$\boldsymbol{\pi}_{2n}^{\mathbb{A}^1}(K_n/m) \longrightarrow \boldsymbol{\pi}_{2n+1}^{\mathbb{A}^1}(\Sigma K_n/m) \longrightarrow \boldsymbol{K}_{2n}^{\mathrm{M}}/m \longrightarrow 0.$$

Then $\boldsymbol{\pi}_{2n}^{\mathbb{A}^1}(K_n/m) \cong \boldsymbol{H}_{\text{ét}}^0(\mu_m^{\otimes n})$ and this sheaf is killed by a single contraction as discussed in the previous paragraph.

For $i \geq 1$ and $0 \leq j \leq n$, the map $\boldsymbol{\pi}_{n+j+i}(\Sigma^i K_n/m) \to \boldsymbol{\pi}_{n+j+i+1}^{\mathbb{A}^1}(\Sigma^{i+1} K_n/m)$ is an isomorphism by the $\mathbb{A}^1$–Freudenthal suspension theorem. Combining these observations establishes the points listed above. □

Remark 5.3.4 It is possible to treat the case where m is a power of p as well, but the answer is simpler. If m is a power of p, then Geisser and Levine computed the homotopy sheaves of K_n/m: by [21, Theorem 8.3], $\boldsymbol{\pi}_i^{\mathbb{A}^1}(K_n/m)$ is nonvanishing if and only if $i = n$, in which case $\boldsymbol{\pi}_n^{\mathbb{A}^1}(K_n/m)$ may be described as the unramified Milnor K-theory sheaf $\boldsymbol{K}_n^M/m$ (see [21, Theorem 8.1]). In this case, we conclude that $\pi_{n+j+i}^{\mathbb{A}^1}(\Sigma^i K_n/m)$ simply vanishes for $1 < i \leq n-1$.

References

[1] **M Arkowitz**, *The generalized Whitehead product*, Pacific J. Math. 12 (1962) 7–23 MR

[2] **A Asok**, **J Fasel**, *Toward a meta-stable range in $\mathbb{A}^1$–homotopy theory of punctured affine spaces*, Oberwolfach Rep. 10 (2013) 1892–1895 MR

[3] **A Asok**, **J Fasel**, *Algebraic vector bundles on spheres*, J. Topol. 7 (2014) 894–926 MR

[4] **A Asok**, **J Fasel**, *A cohomological classification of vector bundles on smooth affine threefolds*, Duke Math. J. 163 (2014) 2561–2601 MR

[5] **A Asok**, **J Fasel**, *Splitting vector bundles outside the stable range and $\mathbb{A}^1$–homotopy sheaves of punctured affine spaces*, J. Amer. Math. Soc. 28 (2015) 1031–1062 MR

[6] **A Asok**, **J Fasel**, *Comparing Euler Classes*, The Quarterly Journal of Mathematics 67 (2016) 603–635

[7] **A Asok**, **J Fasel**, *An explicit KO–degree map and applications*, Journal of Topology 10 (2017) 268–300

[8] **J Ayoub**, *Un contre-exemple à la conjecture de $\mathbb{A}^1$–connexité de F Morel*, C. R. Math. Acad. Sci. Paris 342 (2006) 943–948 MR

[9] **W D Barcus**, *The stable suspension of an Eilenberg–MacLane space*, Trans. Amer. Math. Soc. 96 (1960) 101–114 MR

[10] **C Barwick**, *On left and right model categories and left and right Bousfield localizations*, Homology Homotopy Appl. 12 (2010) 245–320 MR

[11] **H Bass**, **J Tate**, *The Milnor ring of a global field*, from "Algebraic K–theory, II: 'Classical' algebraic K–theory and connections with arithmetic" (H Bass, editor), Springer, Berlin (1973) 349–446. Lecture Notes in Math., Vol. 342 MR

[12] **J M Boardman**, **B Steer**, *On Hopf invariants*, Comment. Math. Helv. 42 (1967) 180–221 MR

[13] **D E Cohen**, *Products and carrier theory*, Proc. London Math. Soc. 7 (1957) 219–248 MR

[14] **E B Curtis**, *Simplicial homotopy theory*, Advances in Math. 6 (1971) 107–209 MR

[15] **P Deligne**, *Voevodsky's lectures on motivic cohomology 2000/2001*, from "Algebraic topology" (N A Baas, E M Friedlander, B Jahren, P A Østvær, editors), Abel Symp. 4, Springer, Berlin (2009) 355–409 MR

[16] **D Dugger**, **D C Isaksen**, *Topological hypercovers and $\mathbb{A}^1$–realizations*, Math. Z. 246 (2004) 667–689 MR

[17] **D Dugger**, **D C Isaksen**, *Motivic Hopf elements and relations*, New York J. Math. 19 (2013) 823–871 MR

[18] **R Elman**, **N Karpenko**, **A Merkurjev**, *The algebraic and geometric theory of quadratic forms*, American Mathematical Society Colloquium Publications 56, Amer. Math. Soc., Providence, RI (2008) MR

[19] **E D Farjoun**, *Cellular spaces, null spaces and homotopy localization*, Lecture Notes in Mathematics 1622, Springer, Berlin (1996) MR

[20] **T Ganea**, *On the homotopy suspension*, Comment. Math. Helv. 43 (1968) 225–234 MR

[21] **T Geisser**, **M Levine**, *The K–theory of fields in characteristic p*, Invent. Math. 139 (2000) 459–493 MR

[22] **S Gille**, **S Scully**, **C Zhong**, *Milnor–Witt K–groups of local rings*, Advances in Mathematics 286 (2016) 729–753

[23] **P G Goerss**, **J F Jardine**, *Simplicial homotopy theory*, Progress in Mathematics 174, Birkhäuser, Basel (1999) MR

[24] **P J Hilton**, *On the homotopy groups of unions of spaces*, Comment. Math. Helv. 29 (1955) 59–92 MR

[25] **P S Hirschhorn**, *Model categories and their localizations*, Mathematical Surveys and Monographs 99, Amer. Math. Soc., Providence, RI (2003) MR

[26] **A Hogadi**, **G Kulkarni**, *Gabber's presentation lemma for finite fields*, preprint (2016) arXiv

[27] **M Hovey**, *Model categories*, Mathematical Surveys and Monographs 63, Amer. Math. Soc., Providence, RI (1999) MR

[28] **D C Isaksen**, *Etale realization on the $\mathbb{A}^1$–homotopy theory of schemes*, Adv. Math. 184 (2004) 37–63 MR

[29] **I M James**, *Reduced product spaces*, Ann. of Math. 62 (1955) 170–197 MR

[30] **I M James**, *On the suspension triad*, Ann. of Math. 63 (1956) 191–247 MR

[31] **I M James**, *On the suspension sequence*, Ann. of Math. 65 (1957) 74–107 MR

[32] **J F Jardine**, *Simplicial presheaves*, J. Pure Appl. Algebra 47 (1987) 35–87 MR

[33] **J F Jardine**, *Motivic symmetric spectra*, Doc. Math. 5 (2000) 445–552 MR

[34] **J F Jardine**, *Local homotopy theory*, Springer, New York (2015) MR

[35] **K Kato**, *Symmetric bilinear forms, quadratic forms and Milnor K–theory in characteristic two*, Invent. Math. 66 (1982) 493–510 MR

[36] **N J Kuhn**, *The transfer and James–Hopf invariants*, Math. Z. 196 (1987) 391–405 MR

[37] **J P May**, *Classifying spaces and fibrations*, Mem. Amer. Math. Soc. 155, Amer. Math. Soc., Providence, RI (1975) MR

[38] **C Mazza**, **V Voevodsky**, **C Weibel**, *Lecture notes on motivic cohomology*, Clay Mathematics Monographs 2, Amer. Math. Soc., Providence, RI (2006) MR

[39] **J Milnor**, *On the construction FK*, from "Algebraic topology—a student's guide" (J F Adams, editor), London Mathematical Society Lecture Note Series 4, Cambridge University Press (1972) 119–136 MR

[40] **F Morel**, *An introduction to $\mathbb{A}^1$–homotopy theory*, from "Contemporary developments in algebraic K–theory" (M Karoubi, A O Kuku, C Pedrini, editors), ICTP Lect. Notes XV, Abdus Salam Int. Cent. Theoret. Phys., Trieste (2004) 357–441 MR

[41] **F Morel**, *Sur les puissances de l'idéal fondamental de l'anneau de Witt*, Comment. Math. Helv. 79 (2004) 689–703 MR

[42] **F Morel**, *Milnor's conjecture on quadratic forms and mod 2 motivic complexes*, Rend. Sem. Mat. Univ. Padova 114 (2005) 63–101 MR

[43] **F Morel**, *The stable $\mathbb{A}^1$–connectivity theorems*, K-Theory 35 (2005) 1–68 MR

[44] **F Morel**, *$\mathbb{A}^1$–algebraic topology over a field*, Lecture Notes in Mathematics 2052, Springer, Heidelberg (2012) MR

[45] **F Morel**, **V Voevodsky**, *A^1–homotopy theory of schemes*, Inst. Hautes Études Sci. Publ. Math. 90 (1999) 45–143 MR

[46] **D Orlov**, **A Vishik**, **V Voevodsky**, *An exact sequence for $K_*^M/2$ with applications to quadratic forms*, Ann. of Math. 165 (2007) 1–13 MR

[47] **K M Ormsby**, **P A Østvær**, *Stable motivic π_1 of low-dimensional fields*, Adv. Math. 265 (2014) 97–131 MR

[48] **D G Quillen**, *Homotopical algebra*, Lecture Notes in Mathematics 43, Springer, Berlin (1967) MR

[49] **O Röndigs**, **M Spitzweck**, **P A Østvær**, *The first stable homotopy groups of motivic spheres*, preprint (2016) arXiv

[50] **J-P Serre**, *Cohomologie modulo 2 des complexes d'Eilenberg–MacLane*, Comment. Math. Helv. 27 (1953) 198–232 MR

[51] **F Strunk**, *On motivic spherical bundles*, PhD thesis, Institut für Mathematik, Universität Osnabrück (2012) Available at `https://tinyurl.com/strunkthesis`

[52] **H Toda**, *Composition methods in homotopy groups of spheres*, Annals of Mathematics Studies 49, Princeton University Press (1962) MR

[53] **V Voevodsky**, *Reduced power operations in motivic cohomology*, Publ. Math. Inst. Hautes Études Sci. 98 (2003) 1–57 MR

[54] **V Voevodsky**, *Motivic cohomology with $\mathbf{Z}/2$–coefficients*, Publ. Math. Inst. Hautes Études Sci. 98 (2003) 59–104 MR

[55] **V Voevodsky**, *On motivic cohomology with $\mathbf{Z}/l$–coefficients*, Ann. of Math. 174 (2011) 401–438 MR

[56] **M Wendt**, *Classifying spaces and fibrations of simplicial sheaves*, J. Homotopy Relat. Struct. 6 (2011) 1–38 MR

[57] **G W Whitehead**, *A generalization of the Hopf invariant*, Ann. of Math. 51 (1950) 192–237 MR

[58] **G W Whitehead**, *On the Freudenthal theorems*, Ann. of Math. 57 (1953) 209–228 MR

[59] **G W Whitehead**, *Elements of homotopy theory*, Graduate Texts in Mathematics 61, Springer, New York (1978) MR

[60] **J H C Whitehead**, *On adding relations to homotopy groups*, Ann. of Math. 42 (1941) 409–428 MR

[61] **K Wickelgren**, **B Williams**, *The simplicial EHP Sequence in $\mathbb{A}^1$–algebraic topology*, preprint (2014) arXiv

[62] **J Wu**, *Simplicial objects and homotopy groups*, from "Braids" (A J Berrick, F R Cohen, E Hanbury, Y-L Wong, J Wu, editors), Lect. Notes Ser. Inst. Math. Sci. Natl. Univ. Singap. 19, World Sci. Publ., Hackensack, NJ (2010) 31–181 MR

Department of Mathematics, University of Southern California
3620 S Vermont Ave, Los Angeles, CA 90089-2532, United States

School of Mathematics, Georgia Institute of Technology
686 Cherry Street, Atlanta, GA 30308, United States

Department of Mathematics, University of British Columbia
1984 Mathematics Road, Vancouver BC V6T 1Z2, Canada

asok@usc.edu, kwickelgren3@math.gatech.edu, tbjw@math.ubc.ca

Proposed: Haynes Miller Received: 6 August 2015
Seconded: Peter Teichner, Mark Behrens Revised: 7 July 2016

Geometry & Topology 21 (2017) 2161–2230

Symplectic and contact differential graded algebras

TOBIAS EKHOLM
ALEXANDRU OANCEA

We define Hamiltonian simplex differential graded algebras (DGA) with differentials that deform the high-energy symplectic homology differential and wrapped Floer homology differential in the cases of closed and open strings in a Liouville manifold of finite type, respectively. The order-m term in the differential is induced by varying natural degree-m coproducts over an $(m-1)$–simplex, where the operations near the boundary of the simplex are trivial. We show that the Hamiltonian simplex DGA is quasi-isomorphic to the (nonequivariant) contact homology algebra and to the Legendrian homology algebra of the ideal boundary in the closed and open string cases, respectively.

53D40, 53D42; 16E45, 18G55

1 Introduction

Let X be a Liouville manifold, and let $L \subset X$ be an exact Lagrangian submanifold. (We use the terminology of Cieliebak and Eliashberg [15] for Liouville manifolds, cobordisms etc throughout the paper.) Assume that (X, L) is cylindrical at infinity, meaning that outside a compact set, (X, L) looks like $([0,\infty) \times Y, [0,\infty) \times \Lambda)$, where Y is a contact manifold, $\Lambda \subset Y$ a Legendrian submanifold, and the Liouville form on $[0,\infty) \times Y$ is the symplectization form $e^t\alpha$ for α a contact form on Y and t the standard coordinate in $[0,\infty)$.

There are a number of Floer homological theories associated to this geometric situation. For example, there is *symplectic homology* $\mathrm{SH}(X)$ which can be defined (see Bourgeois and Oancea [11], Seidel [39] and Viterbo [42]) using a time-dependent Hamiltonian $H\colon X \times I \to \mathbb{R}$, $I = [0,1]$, which is a small perturbation of a time-independent Hamiltonian that equals a small positive constant in the compact part of X and is linearly increasing of certain slope in the coordinate $r = e^t$ in the cylindrical end at infinity, and then taking a certain limit over increasing slopes. The chain complex underlying $\mathrm{SH}(X)$ is denoted by $\mathrm{SC}(X)$ and is generated by the 1–periodic orbits of the Hamiltonian vector field X_H of H, graded by their Conley–Zehnder indices. These fall into two classes: low-energy orbits in the compact part of X and (reparametrizations of) Reeb orbits of α in the region in the end where H increases from a function that is

Published: 19 May 2017 DOI: 10.2140/gt.2017.21.2161

close to zero to a function of linear growth. The differential counts Floer holomorphic cylinders interpolating between the orbits. These are solutions $u\colon \mathbb{R}\times S^1 \to X$, $S^1 = I/\partial I$, of the Floer equation

$$(du - X_H \otimes dt)^{0,1} = 0, \tag{1-1}$$

where $s+it \in \mathbb{R}\times S^1$ is a standard complex coordinate and the complex antilinear part is taken with respect to a chosen adapted almost complex structure J on X. The 1–periodic orbits of H are closed loops that are critical points of an action functional, and cylinders solving (1-1) are similar to instantons that capture the effect of tunneling between critical points. Because of this and analogies with (topological) string theory, we say that symplectic homology is a *theory of closed strings*.

The open string analogue of $\mathrm{SH}(X)$ is a corresponding theory for paths with endpoints in the Lagrangian submanifold $L\subset X$. It is called the *wrapped Floer homology* of L and here denoted by $\mathrm{SH}(L)$. Its underlying chain complex $\mathrm{SC}(L)$ is generated by Hamiltonian time-1 chords that begin and end on L, graded by a Maslov index. Again these fall into two classes: high-energy chords that correspond to Reeb chords of the ideal Legendrian boundary Λ of L and low-energy chords that correspond to critical points of H restricted to L. The differential on $\mathrm{SC}(L)$ counts Floer holomorphic strips with boundary on L interpolating between Hamiltonian chords, ie solutions

$$u\colon (\mathbb{R}\times I, \partial(\mathbb{R}\times I)) \to (X, L)$$

of (1-1).

We will also consider a mixed version of open and closed strings. The graded vector space underlying the chain complex is simply $\mathrm{SC}(X,L)=\mathrm{SC}(X)\oplus\mathrm{SC}(L)$, and the differential $d_1\colon \mathrm{SC}(X,L)\to\mathrm{SC}(X,L)$ has the following matrix form with respect to this decomposition (subscripts "c" and "o" refer to *closed* and *open*, respectively):

$$d_1 = \begin{pmatrix} d_{\mathrm{cc}} & d_{\mathrm{oc}} \\ 0 & d_{\mathrm{oo}} \end{pmatrix}.$$

Here d_{cc} and d_{oo} are the differentials on $\mathrm{SC}(X)$ and $\mathrm{SC}(L)$, respectively, and $d_{\mathrm{oc}}\colon \mathrm{SC}(L)\to\mathrm{SC}(X)$ is a chain map of degree -1. (There is also a closed-open map $d_{\mathrm{co}}\colon \mathrm{SC}(X)\to\mathrm{SC}(L)$, but we will not use it here.) Each of these three maps counts solutions of (1-1) on a Riemann surface with two punctures, one positive regarded as input, and one negative regarded as output. For d_{cc} the underlying Riemann surface is the cylinder, for d_{oo} the underlying Riemann surface is the strip, and for d_{oc} the underlying Riemann surface is the cylinder $\mathbb{R}\times S^1$ with a slit at $[0,\infty)\times\{1\}$ (or equivalently, a disk with two boundary punctures, a sphere with two interior punctures, and a disk with positive boundary puncture and negative interior puncture). We will denote the corresponding homology by $\mathrm{SH}(X,L)$.

In order to count the curves in the differential over integers, we use index bundles to orient solution spaces, and for that we assume that the pair (X, L) is relatively spin; see Fukaya, Oh, Ohta and Ono [26]. As the differential counts Floer-holomorphic curves, it respects the energy filtration, and the subspace generated by the low-energy chords and orbits is a subcomplex. We denote the corresponding high-energy quotient by $\mathrm{SC}^+(X, L)$ and its homology by $\mathrm{SH}^+(X, L)$. We define similarly $\mathrm{SC}^+(X)$, $\mathrm{SC}^+(L)$, $\mathrm{SH}^+(X)$ and $\mathrm{SH}^+(L)$.

In the context of Floer homology, the cylinders and strips above are the most basic Riemann surfaces, and it is well known that more complicated Riemann surfaces Σ can be included in the theory as follows; see Ritter [36] and Seidel [39]. Pick a family of 1–forms B with values in Hamiltonian vector fields on X over the appropriate Deligne–Mumford space of domains and count rigid solutions of the Floer equation

$$(du - B)^{0,1} = 0, \tag{1-2}$$

where $B(s + it) = X_{H_t} \otimes dt$ in cylindrical coordinates $s + it$ near the punctures of Σ. The resulting operation descends to homology as a consequence of gluing and Gromov–Floer compactness. A key condition for solutions of (1-2) to have relevant compactness properties is that B is required to be nonpositive in the following sense. For each $x \in X$, we get a 1–form $B(x) = X_{H_z}(x) \otimes \beta$ on Σ with values in $T_x X$, where $H_z\colon X \to \mathbb{R}$ is a family of Hamiltonian functions parametrized by $z \in \Sigma$ and β is a 1–form on Σ. The nonpositivity condition is then that the 2–form $d(H_z(x)\,\beta)$ associated to B is a nonpositive multiple of the area form on Σ for each $x \in X$.

The most important such operations on $\mathrm{SH}(X)$ are the BV-operator and the pair-of-pants product. The BV-operator corresponds to solutions of a parametrized Floer equation analogous to (1-1) which twists the cylinder one full turn. The pair-of-pants product corresponds to a sphere with two positive and one negative puncture and restricts to the cup product on the ordinary cohomology of X, which here appears as the low-energy part of $\mathrm{SH}(X)$. Analogously on $\mathrm{SH}(L)$, the product corresponding to the disk with two positive and one negative boundary puncture restricts to the cup product on the cohomology of L, and the disk with one positive interior puncture and two boundary punctures of opposite signs expresses $\mathrm{SH}(L)$ as a module over $\mathrm{SH}(X)$.

The BV-operator and the pair-of-pants product are generally nontrivial operations. In contrast, arguing along the lines of Seidel [39, Section 8a] and Ritter [36, Theorem 6.10], one shows that the operations determined by Riemann surfaces with at least two negative punctures are often trivial on $\mathrm{SH}^+(X, L)$. Basic examples of this phenomenon are the operations D_m given by disks and spheres with one positive and $m \geq 2$ negative punctures. By pinching the 1–form B in (1-2) in the cylindrical end at one of the m

negative punctures, it follows that, up to homotopy, D_m factors through the low-energy part of the complex $\mathrm{SC}(X, L)$. In particular, on the high-energy quotient $\mathrm{SC}^+(X, L)$, the operation is trivial if the 1–form is pinched near at least one negative puncture.

The starting point for this paper is to study operations d_m that are associated to natural families of forms B that interpolate between all ways of pinching near negative punctures. More precisely, for disks and spheres with one positive and m negative punctures, we take B in (1-2) to have the form $B = X_H \otimes w_j\, dt$ in the cylindrical end, with coordinate $s + it$ in $[0, \infty) \times I$ for open strings and in $[0, \infty) \times S^1$ for closed strings, near the j^{th} puncture. Here w_j is a positive function with a minimal value called *weight*. By Stokes' theorem, in order for B to satisfy the nonpositivity condition, the sum of weights at the negative ends must be greater than the weight at the positive end. Thus the choice of 1–form is effectively parametrized by an $(m-1)$–simplex and the equation (1-2) associated to a form which lies in a small neighborhood of the boundary of the simplex, where at least one weight is very small, has no solutions with all negative punctures at high-energy chords or orbits. The operation d_m is then defined by counting rigid solutions of (1-2) where B varies over the simplex bundle. Equivalently, we count solutions with only high-energy asymptotes in the class dual to the fundamental class of the sphere bundle over Deligne–Mumford space obtained as the quotient space after fiberwise identification of the boundary of the simplex to a point. In particular, curves contributing to d_m have formal dimension $-(m-1)$.

Our first result says that the operations d_m combine to give a DGA differential. The *Hamiltonian simplex DGA* $\mathcal{SC}^+(X, L)$ is the unital algebra generated by the generators of $\mathrm{SC}^+(X, L)$ with grading shifted down by 1, where orbits sign-commute with orbits and chords but where chords do not commute. Let $d\colon \mathcal{SC}^+(X, L) \to \mathcal{SC}^+(X, L)$ be the map defined on generators b by

$$d\, b = d_1 b + d_2 b + \cdots + d_m b + \cdots,$$

and extend it by the Leibniz rule.

Theorem 1.1 *The map d is a differential, $d \circ d = 0$, and the homotopy type of the Hamiltonian simplex DGA $\mathcal{SC}^+(X, L)$ depends only on (X, L). Furthermore, $\mathcal{SC}^+(X, L)$ is functorial in the following sense. If $(X_0, L_0) = (X, L)$, if (X_{10}, L_{10}) is a Liouville cobordism with negative end $(\partial X_0, \partial L_0)$, and if (X_1, L_1) denotes the Liouville manifold obtained by gluing (X_{10}, L_{10}) to (X_0, L_0), then there is a DGA map*

$$\Phi_{X_{10}}\colon \mathcal{SC}^+(X_1, L_1) \to \mathcal{SC}^+(X_0, L_0),$$

and the homotopy class of this map is an invariant of (X_{10}, L_{10}) up to Liouville homotopy.

If $L = \varnothing$ in Theorem 1.1, then we get a Hamiltonian simplex DGA $\mathcal{SC}^+(X)$ generated by high-energy Hamiltonian orbits. This DGA is (graded) commutative. Also, the quotient $\mathcal{SC}^+(L)$ of $\mathcal{SC}^+(X,L)$ by the ideal generated by orbits is a Hamiltonian simplex DGA generated by high-energy chords of L. We write $\mathcal{SH}^+(X,L)$ for the homology DGA of $\mathcal{SC}^+(X,L)$ and use the notation $\mathcal{SH}^+(X)$ and $\mathcal{SH}^+(L)$ with a similar meaning. If X is the cotangent bundle of a manifold $X = T^*M$, then $\mathrm{SH}(X)$ is isomorphic to the homology of the free loop space of M (see Abbondandolo and Schwarz [2], Abouzaid [3], Salamon and Weber [38] and Viterbo [41]), and the counterpart of d_2 in string topology is nontrivial (see Goresky and Hingston [27]). Also, if b is a generator of $\mathcal{SC}^+(X_1,L_1)$, then with $\Phi = \Phi_{X_{10}}$ the DGA map in Theorem 1.1, $\Phi(b)$ can be expanded as $\Phi(b) = \Phi_1(b) + \Phi_2(b) + \cdots$, where $\Phi_m(b)$ represents the homogeneous component of monomials of degree m. The linear component Φ_1 in this expansion induces the Viterbo functoriality map $\mathrm{SC}^+(X_1,L_1) \to \mathrm{SC}^+(X_0,L_0)$; see Cieliebak and Oancea [17] and Viterbo [42].

Our second result expresses $\mathcal{SC}^+(X,L)$ in terms of the ideal boundary $(Y,\Lambda) = (\partial X, \partial L)$. Recall that the usual contact homology DGA $\widetilde{\mathcal{A}}(Y,\Lambda)$ is generated by closed Reeb orbits in Y and by Reeb chords with endpoints on Λ; see Eliashberg, Givental and Hofer [25]. Here we use the differential that is naturally augmented by rigid once-punctured spheres in X and by rigid once-boundary punctured disks in X with boundary in L. (In the terminology of Bourgeois, Ekholm and Eliashberg [7], the differential counts anchored spheres and disks). In Bourgeois and Oancea [10], a nonequivariant version of linearized orbit contact homology was introduced. In Section 6, we extend this construction and define a nonequivariant DGA that we call $\mathcal{A}(Y,\Lambda)$, which is generated by decorated Reeb orbits and by Reeb chords. We give two definitions of the differential on $\mathcal{A}(Y,\Lambda)$, one using Morse–Bott curves and one using curves holomorphic with respect to a domain dependent almost complex structure. In analogy with the algebras considered above, we write $\mathcal{A}(Y)$ for the subalgebra generated by decorated orbits and $\mathcal{A}(\Lambda)$ for the quotient by the ideal generated by decorated orbits.

In Sections 2.6 and 6.1, we introduce a continuous 1–parameter deformation of the simplex family of 1–forms B that turns off the Hamiltonian term in (1-2) by sliding its support to the negative end in the domains of the curves and that leads to the following result.

Theorem 1.2 *The deformation that turns the Hamiltonian term off gives rise to a DGA map*

$$\Phi\colon \mathcal{A}(Y,\Lambda) \to \mathcal{SC}^+(X,L).$$

The map Φ is a quasi-isomorphism that takes the orbit subalgebra $\mathcal{A}(Y)$ quasi-isomorphically to the orbit subalgebra $\mathcal{SC}^+(X)$. Furthermore, it descends to the quotient $\mathcal{A}(\Lambda)$ and maps it to $\mathcal{SC}^+(L)$ as a quasi-isomorphism.

The usual (equivariant) contact homology DGA $\widetilde{\mathcal{A}}(Y,\Lambda)$ is also quasi-isomorphic to a Hamiltonian simplex DGA that corresponds to a version of symplectic homology defined by a time-independent Hamiltonian; see Theorem 6.5. For the corresponding result on the linear level see Bourgeois and Oancea [12].

Remark 1.3 As is well known, the constructions of the DGAs $\widetilde{\mathcal{A}}(Y,\Lambda)$ and $\mathcal{A}(Y,\Lambda)$, of the orbit augmentation induced by X, and of symplectic homology for time-independent Hamiltonians with time-independent almost complex structures, require the use of abstract perturbations for the pseudoholomorphic curve equation in a manifold with cylindrical end. This is an area where much current research is being done and there are several approaches, some of an analytical character (see eg Hofer, Wysocki and Zehnder [29; 30]), others of more algebraic topological flavor (see eg Pardon [35]), and others of more geometric flavor (see eg Fukaya, Oh, Ohta and Ono [26]). Here we will not enter into the details of this problem but merely assume such a perturbation scheme has been fixed. More precisely, the proofs that the differential in the definition of the Hamiltonian simplex DGA squares to zero and that the maps induced by cobordisms are chain maps of DGAs do not require the use of any abstract perturbation scheme; standard transversality arguments suffice. On the other hand, our proof of invariance of the Hamiltonian simplex DGA in Section 5.4 does use an abstract perturbation scheme (in its simplest version: to count rigid curves over the rationals). Also, it gives equivalences of DGAs under deformations as in the original version of symplectic field theory; see Eliashberg, Givental and Hofer [25] and compare the discussion in Pardon [34, Remark 1.3].

Theorem 1.2 relates symplectic field theory (SFT) and Hamiltonian Floer theory. On the linear level the relation is rather direct (see Bourgeois and Oancea [10]), but not for the SFT DGA. The first candidate for a counterpart on the Hamiltonian Floer side collects the standard coproducts to a DGA differential, but that DGA is trivial by pinching. To see that, recall the sphere bundle over Deligne–Mumford space obtained by identifying the boundary points in each fiber of the simplex bundle. The coproduct DGA then corresponds to counting curves lying over the homology class of a point in each fiber, but that point can be chosen as the base point where all operations are trivial. The object that is actually isomorphic to the SFT DGA is the Hamiltonian simplex DGA related to the fundamental class of the spherization of the simplex bundle.

In light of this, the following picture of the relation between Hamiltonian Floer theory and SFT emerges. The Hamiltonian Floer theory holomorphic curves solve a Cauchy–Riemann equation with Hamiltonian 0–order term chosen consistently over Deligne–Mumford space. These curves are less symmetric than their counterparts in SFT, which are defined without additional 0–order term. Accordingly, the moduli spaces of Hamiltonian Floer theory have more structure and admit natural deformations and actions, eg parametrized by simplices which control deformations of the weights at the negative punctures and an action of the framed little disk operad; see Section 7. The SFT moduli spaces are, in a sense, homotopic to certain essential strata inside the Hamiltonian Floer theory moduli spaces (see also Remark 6.4), and the structure and operations that they carry are intimately related to the natural actions mentioned. From this perspective, this paper studies the most basic operations, ie the higher coproducts, determined by simplices parametrizing weights at the negative punctures; see Section 2.3.

We end the introduction by a comparison between our constructions and other well-known constructions in Floer theory. In the case of open strings, the differential $d = \sum_{j=1}^{\infty} d_j$ can be thought of as a sequence of operations $(d_1, d_2, \dots, d_m, \dots)$ on the vector space $\mathrm{SC}^+(L)$. These operations define the structure of an ∞–coalgebra on $\mathrm{SC}^+(L)$ (with grading shifted down by one) and $\mathcal{SC}(L)^+$ is the cobar construction for this ∞–coalgebra. This point of view is dual to that of the Fukaya category, in which the primary objects of interest are ∞–algebras. In the Fukaya category setting, algebraic invariants are obtained by applying (variants of) the Hochschild homology functor. In the DGA setting, invariants are obtained more directly as the homology of the Hamiltonian simplex DGA.

Acknowledgements Ekholm was partially funded by the Knut and Alice Wallenberg Foundation as a Wallenberg scholar and by the Swedish Research Council, 2012-2365. Oancea was partially funded by the European Research Council, StG-259118-STEIN.

We thank the organizers of the Gökova 20th Geometry and Topology Conference held in May 2013 for an inspiring meeting, where this project started. We also thank Mohammed Abouzaid for valuable discussions and an anonymous referee for careful reading and for helping us improve the paper. Part of this work was carried out while Oancea visited the Simons Center for Geometry and Physics at Stony Brook in July 2014.

2 Simplex bundles over Deligne–Mumford space, splitting compatibility and 1–forms

The Floer theories we study use holomorphic maps of disks and spheres with one positive and several negative punctures. Configuration spaces for such maps naturally fiber over the corresponding Deligne–Mumford space that parametrizes their domains. In this section we endow the Deligne–Mumford space with additional structure needed to define the relevant solution spaces. More precisely, we parametrize 1–forms with nonpositive exterior derivative by a simplex bundle over Deligne–Mumford space that respects certain restriction maps at several level curves in the boundary. We then combine these forms with a certain type of Hamiltonian to get nonpositive forms with values in Hamiltonian vector fields, suitable as 0–order perturbations in the Floer equation.

2.1 Asymptotic markers and cylindrical ends

We will use punctured disks and spheres with a fixed choice of cylindrical end at each puncture. Here, a *cylindrical end* at a puncture is defined to be a biholomorphic identification of a neighborhood of that puncture with one of the following punctured model Riemann surfaces:

- Negative interior puncture:
$$Z^- = (-\infty, 0) \times S^1 \approx D^2 \setminus \{0\},$$
where $D^2 \subset \mathbb{C}$ is the unit disk in the complex plane.
- Positive interior puncture:
$$Z^+ = (0, \infty) \times S^1 \approx \mathbb{C} \setminus \bar{D}^2.$$
- Negative boundary puncture:
$$\Sigma^- = (-\infty, 0) \times [0, 1] \approx (D^2 \setminus \{0\}) \cap H,$$
where $H \subset \mathbb{C}$ denotes the closed upper half plane.
- Positive boundary puncture:
$$\Sigma^+ = (0, \infty) \times [0, 1] \approx (\mathbb{C} \setminus \bar{D}^2) \cap H.$$

Each of the above model surfaces has a canonical complex coordinate of the form $z = s + it$. Here $s \in \mathbb{R}$ at all punctures, with $s > 0$ or $s < 0$ according to whether the puncture is positive or negative. At interior punctures, $t \in S^1$, and at boundary punctures, $t \in [0, 1]$.

The automorphism group of the cylindrical end at a boundary puncture is $\mathbb{R}$ and the end is thus well defined up to a contractible choice of automorphisms. For a positive or negative interior puncture, the corresponding automorphism group is $\mathbb{R} \times S^1$. Thus the cylindrical end is well defined up to a choice of automorphism in a space homotopy equivalent to S^1. To remove the S^1–ambiguity, we fix an *asymptotic marker* at the puncture, ie a tangent half-line at the puncture, and require that it corresponds to $(0, \infty) \times \{1\}$ or to $(-\infty, 0) \times \{1\}$, $1 \in S^1$, at positive or negative punctures, respectively. The cylindrical end at an interior puncture *with asymptotic marker* is then well defined up to contractible choice.

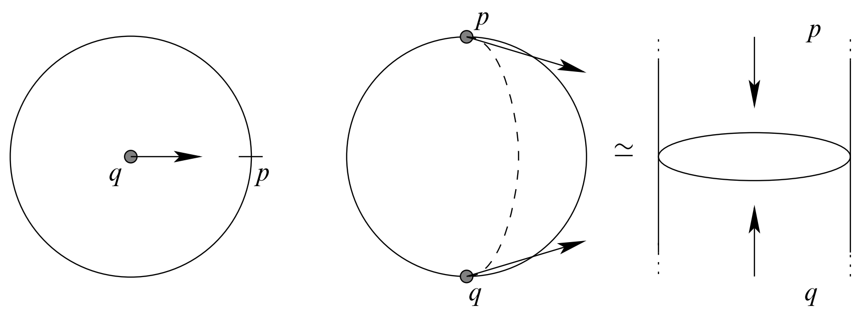


Figure 1: Inducing markers at negative interior punctures

We next consider various ways to induce asymptotic markers at interior punctures that we will eventually assemble into a coherent choice of asymptotic markers over the space of punctured spheres and disks. Consider first a disk D with interior punctures and with a distinguished boundary puncture p. Then p determines an asymptotic marker at any interior puncture q as follows. There is a unique holomorphic diffeomorphism $\psi\colon D \to D^2 \subset \mathbb{C}$ with $\psi(q) = 0$ and $\psi(p) = 1$. Define the asymptotic marker at q in D to correspond to the direction of the real line at $0 \in D^2$, ie the direction given by the vector $d\psi^{-1}(0)\cdot 1$. See Figure 1.

Similarly, on a sphere S, a distinguished interior puncture p with asymptotic marker determines an asymptotic marker at any other interior puncture q as follows. There is a holomorphic map $\psi\colon S \to \mathbb{R}\times S^1$ taking p to ∞, q to $-\infty$ and the asymptotic marker to the tangent vector of $\mathbb{R}\times\{1\}$. We take the asymptotic marker at q to correspond to the tangent vector of $\mathbb{R}\times\{1\}$ at $-\infty$ under ψ. See Figure 1.

For a more unified notation below we use the following somewhat involved convention for our spaces of disks and spheres. Let $h \in \{0,1\}$. For $h = 1$ and $m,k \geq 0$, let $\mathcal{D}'_{h;hm,k} = \mathcal{D}'_{1;m,k}$ denote the moduli space of disks with one positive boundary puncture, $m \geq 0$ negative boundary punctures and k negative interior punctures. For $h = 0$ and $k \geq 0$, let $\mathcal{D}'_{h;hm,k} = \mathcal{D}'_{0;0,k}$ denote the moduli space of spheres with one positive interior puncture with asymptotic marker and k negative interior punctures.

As explained above there are then, for both $h = 0$ and $h = 1$, induced asymptotic markers at all the interior negative punctures of any element in $\mathcal{D}'_{h;hm,k}$. The space $\mathcal{D}'_{h;hm,k}$ admits a natural compactification that consists of several level disks and spheres; see [8, Section 4] and also [31]. We introduce the following notation to describe the boundary. Consider a several-level curve. We associate to it a downwards oriented rooted tree Γ with one vertex for the positive puncture of each component of the several-level curve and one edge for each one of the negative punctures of the

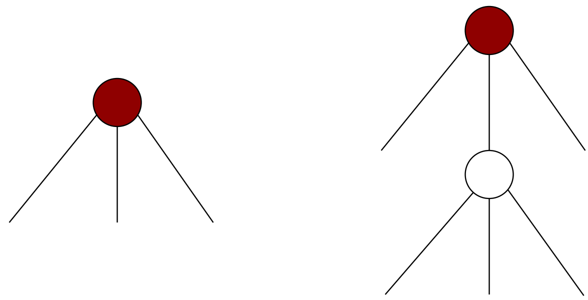


Figure 2: A curve in the main stratum of $\mathcal{D}'_{h;hm,k}$ with $hm+k=3$ (left) and a 2–level curve in the boundary of $\mathcal{D}'_{h;hm,k}$ with $hm+k=5$ (right)

components of the several-level curves. See Figure 2 for examples. Here the root of the tree is the positive puncture of the top-level curve and the edges attached to it are the edges of the negative punctures in the top level oriented away from the root. The definition of Γ is inductive: the vertex of the positive puncture of a curve C in the j^{th} level is attached to the edge of the negative puncture of a curve in the $(j-1)^{\text{st}}$ level where it is attached. All edges of negative punctures of C are attached to the vertex of the positive puncture of C and oriented away from it. Then the boundary strata of $\mathcal{D}'_{h;hm,k}$ are in one-to-one correspondence with such graphs Γ and the components of the several-level curve are in one-to-one correspondence with downwards oriented subtrees consisting of one vertex and all edges emanating from it. For example the graph of a curve lying in the interior of $\mathcal{D}'_{h;hm;k}$ is simply a vertex with $hm+k$ edges attached and oriented away from the vertex. To distinguish the edges of such graphs Γ, we call an edge a *gluing edge* if it is attached to two vertices and *free* if it is attached only to one vertex.

Note next that the induced asymptotic markers are compatible with the level structure in the boundary of $\mathcal{D}'_{h;hm,k}$ in the sense that they vary continuously with the domain inside the compactification. To see this, note that in a boundary stratum corresponding to a graph Γ, it is sufficient to study neck stretching for cylinders corresponding to linear subgraphs of Γ, and here the compatibility of asymptotic markers with the level structure is obvious.

Consider the bundle $\mathcal{C}'_{h;hm,k} \to \mathcal{D}'_{h;hm,k}$, with $h \in \{0,1\}$ and $m,k \geq 0$, of disks or spheres with punctures with cylindrical ends compatible with the markers. The fiber of this bundle is contractible so there exists a section. We next show that there is also a section over the compactification of $\mathcal{D}'_{h;hm,k}$. The proof is by induction on $hm+k \geq 3$. We first choose cylindrical ends for disks and spheres with three punctures. Gluing these we get cylindrical ends in a neighborhood of the boundary of the moduli space of disks and spheres with four punctures. Since the fiber of $\mathcal{C}'_{h;hm;k}$ is contractible, this choice can be extended continuously over the whole space of disks and spheres with four punctures. Assume by induction that cylindrical ends for disks and spheres with less than $hm+k$ negative punctures have been chosen to be *splitting compatible*; ie in such a way that near the boundary of any moduli space of disks and spheres with $hm'+k' < hm+k$ negative punctures, the cylindrical ends are induced via gluing from the moduli spaces of disks and spheres with less than $hm'+k'$ negative punctures. We claim that such a choice that is splitting compatible determines a well-defined splitting compatible section of the bundle $\mathcal{C}'_{h;hm,k} \to \mathcal{D}'_{h;hm,k}$ near its boundary via gluing. Indeed, given a stratum in the boundary corresponding to a graph Γ as above, the gluing construction determines a section on the intersection between $\mathcal{D}'_{h;hm,k}$ and some open neighborhood of that stratum in the compactification of $\mathcal{D}'_{h;hm,k}$. Splitting compatibility ensures that local sections determined by different strata in the boundary coincide on overlaps; see [40, Lemma 9.3]. Finally, to complete the induction, note that the resulting section defined in a neighborhood of the boundary extends to a global section because the fiber of the bundle $\mathcal{C}'_{h;hm,k} \to \mathcal{D}'_{h;hm,k}$ is contractible.

Let $\{\mathcal{D}_{h;hm,k}\}_{h\in\{0,1\},\,k,m\geq 0}$, $\mathcal{D}_{h;hm,k}\colon \mathcal{D}'_{h;hm,k} \to \mathcal{C}'_{h;hm,k}$ denote a system of sections as in the inductive construction above, with $\mathcal{D}_{h;hm,k}$ defined over the compactification of $\mathcal{D}'_{h;hm,k}$. We say that

$$\mathcal{D} = \bigcup_{h\in\{0,1\};\,m,k\geq 0} \mathcal{D}_{h;hm,k}$$

is a *system of cylindrical ends that is compatible with breaking*.

We identify $\mathcal{D}_{h;hm,k}$ with its graph and think of it as a subset of $\mathcal{C}'_{h;hm,k}$. The projection of $\mathcal{D}_{h;hm,k}$ onto $\mathcal{D}'_{h;hm,k}$ is a homeomorphism and, after using smooth approximation, a diffeomorphism with respect to the natural stratification of the space determined by several-level curves. Via this projection we endow $\mathcal{D}_{h;hm,k}$ with the structure of a set consisting of (several-level) curves with additional data corresponding to a choice of a cylindrical end neighborhood at each puncture.

A neighborhood of a several-level curve $S \in \mathcal{D}_{h;hm,k}$ can then be described as follows. Consider the graph Γ determined by S. Let $V(\Gamma) = \{v_0, v_1, \dots, v_r\}$ denote the vertices of Γ with v_0 the top vertex, and let $E_g(\Gamma) = \{e_1, \dots, e_s\}$ denote the gluing edges of Γ. Let U_j be neighborhoods in $\mathcal{D}_{h_j;h_j m_j,k_j}$ of the component corresponding to v_j. Then a neighborhood U of S is given by

$$U = \Bigl(\prod_{v_j\in V(\Gamma)} U_j\Bigr) \times \Bigl(\prod_{e_l\in E_g(\Gamma)} (\rho_{0;l}, \infty)\Bigr), \tag{2-1}$$

where $\rho_{0;l} \geq 0$ for $1 \leq j \leq s$. Here the *gluing parameters* $\rho_l \in (\rho_{0;l}, \infty)$ measure the length of the breaking cylinder or strip corresponding to the gluing edge e_l. More

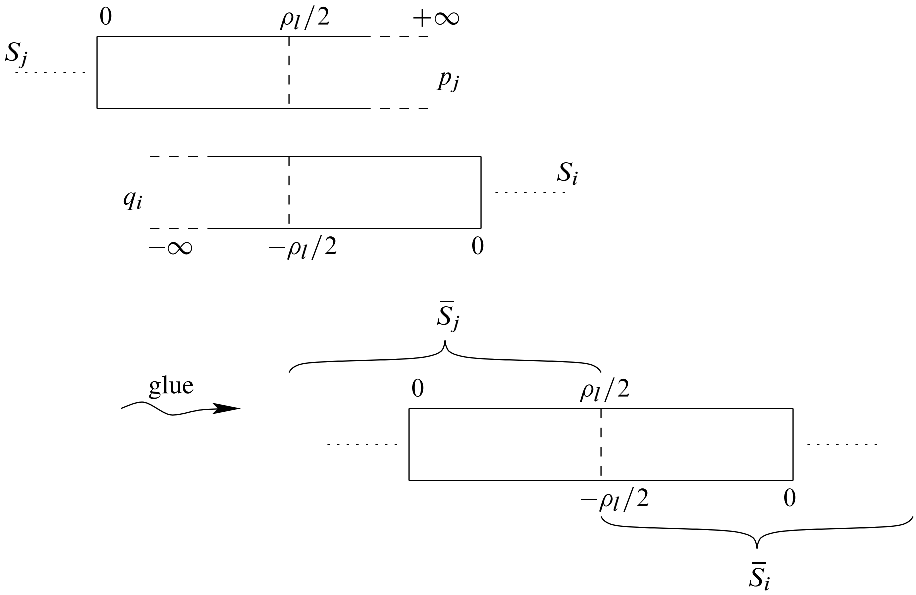


Figure 3: Gluing of a nodal curve in cylindrical coordinates

precisely, assume that e_l connects v_i and v_j and corresponds to the curve S_j of v_j attached at its positive puncture p_j to a negative puncture q_i of the curve S_i of v_i. Then, given the cylindrical ends $(-\infty, 0] \times S^1$ (interior case) or $(-\infty, 0] \times [0, 1]$ (boundary case) for q_i, respectively $[0, \infty) \times S^1$ (interior case) or $[0, \infty) \times [0, 1]$ (boundary case) for p_j, the glued curve corresponding to the parameter $\rho_l \in (\rho_{0;l}, \infty)$ is obtained via the gluing operation on these cylindrical ends defined by cutting out $\bigl(-\infty, -\frac{1}{2}\rho_l\bigr) \times S^1$ or $\bigl(-\infty, -\frac{1}{2}\rho_l\bigr) \times [0, 1]$ from the cylindrical end of q_i, cutting out $\bigl(\frac{1}{2}\rho_l, \infty\bigr) \times S^1$ or $\bigl(\frac{1}{2}\rho_l, \infty\bigr) \times [0, 1]$ from the cylindrical end of p_j, and gluing the remaining compact domains in the cylindrical ends by identifying $\bigl\{-\frac{1}{2}\rho_l\bigr\} \times S^1$ with $\bigl\{\frac{1}{2}\rho_l\bigr\} \times S^1$, respectively $\bigl\{-\frac{1}{2}\rho_l\bigr\} \times [0, 1]$ with $\bigl\{\frac{1}{2}\rho_l\bigr\} \times [0, 1]$. We refer to the resulting compact domain as the *breaking cylinder or strip*, and we refer to $\bigl\{-\frac{1}{2}\rho_l\bigr\} \times S^1 \equiv \bigl\{\frac{1}{2}\rho_l\bigr\} \times S^1$ or $\bigl\{-\frac{1}{2}\rho_l\bigr\} \times [0, 1] \equiv \bigl\{\frac{1}{2}\rho_l\bigr\} \times [0, 1]$ as its *middle circle or segment*. Given a several-level curve S in this neighborhood we write $\overline{S}_j$ for the closures of the components that remain if the middle circle or segment in each breaking cylinder or strip is removed, and that correspond to subsets of the levels S_j of the broken curve. See Figure 3.

2.2 Almost complex structures

We next introduce splitting compatible families of almost complex structures over $\mathcal{D}$. Let $\mathcal{J}(X)$ denote the space of almost complex structures on X compatible with ω

and adapted to the contact form α in the cylindrical end; ie if $J \in \mathcal{J}$ then in the cylindrical end J preserves the contact planes and takes the vertical direction to the Reeb direction. Our construction of a family of almost complex structures is inductive. We start with strips, cylinders and cylinders with slits with coordinates $s + it$. Here we require that $J = J_t$ depends only on the I or $I/\partial I$ coordinate. Assume that we have defined a family of almost complex structures J_z for all curves $D_{h;hm,k}$, $hm + k \leq p$ which have the form above in every cylindrical end and which commute with restriction to components for several-level curves. By gluing we then have a field of almost complex structures in a neighborhood of the boundary of $\mathcal{D}_{h;hm,k}$ for $hm + k = p + 1$. Since $\mathcal{J}$ is contractible, it is easy to see that we can extend this family to all of $\mathcal{D}_{h;hm,k}$. We call the resulting family of almost complex structures over the universal curve corresponding to $\mathcal{D}$ *splitting compatible*.

2.3 A simplex bundle

Consider the trivial bundle

$$\mathcal{E}^{hm+k-1} = \mathcal{D}_{h;hm,k} \times \Delta^{hm+k-1} \to \mathcal{D}_{h;hm,k}$$

over $\mathcal{D}_{h;hm,k}$, with fiber the open $(hm+k-1)$–simplex

$$\Delta^{hm+k-1} = \bigl\{(s_1, \dots, s_{hm+k}) : \textstyle\sum_i s_i = 1,\ s_i > 0\bigr\}.$$

Since the bundle is trivial, it extends as such over the compactification of $\mathcal{D}_{h;hm,k}$. We think of the coordinates of a point $(s_1, \dots, s_{hm+k}) \in \Delta^{hm+k-1}$ over a disk or sphere $D_{h;hm,k} \in \mathcal{D}_{h;hm,k}$ as representing weights at its negative punctures, and we think of the positive puncture as carrying the weight 1.

We next define restriction maps for $\mathcal{E}^{hm+k-1}$ over the boundary of $\mathcal{D}_{h;hm,k}$. Let $s = (s_1, \dots, s_{hm+k}) \in \Delta^{hm+k-1}$ denote the weights of a several-level curve S in the boundary of $\mathcal{D}_{h;hm,k}$ with graph Γ. Let S_j be a component of this building corresponding to the vertex v_j of Γ, with positive puncture q_0 and negative punctures $q_1, \dots, q_n$. Define the weight $w(q_l)$ at q_l for $l = 0, \dots, n$ as follows. For $l = 0$, $w(q_0)$ equals the sum of all weights at negative punctures q of the total several-level curve for which there exists a level-increasing path in Γ from v_j to q. For $l \geq 1$, if the edge of the negative puncture q_l is free then $w(q_l)$ equals the weight of the puncture q_l as a puncture of the total several-level curve, and if the edge is a gluing edge connecting v_j and v_t, then $w(q_l)$ equals the sum of all weights at negative punctures q of the total several-level curve for which there exists a level-increasing path in Γ from v_t to q. Note that $w(q_0) = w(q_1) + \cdots + w(q_n)$ by construction.

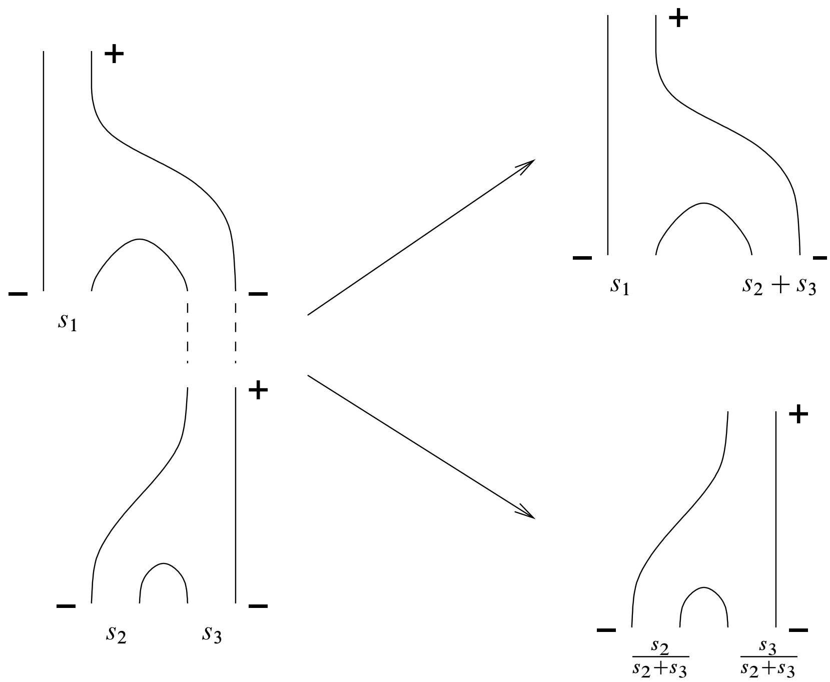


Figure 4: Component restriction maps

The *component restriction map* r_j then takes the point $s \in \Delta^{hm+k-1}$ over S to the point

$$r_j(s) = \frac{1}{w(q_0)}(w(q_1), \dots, w(q_n)) \in \Delta^{n-1}$$

over S_j in $\mathcal{E}^{n-1}$. The component restriction map r_j is defined on the restriction of $\mathcal{E}^{hm+k-1}$ to the stratum that corresponds to Γ in the boundary of $\mathcal{D}_{h;hm,k}$.

2.4 Superharmonic functions and nonpositive 1–forms

Our main Floer homological constructions involve studying Floer holomorphic curves parametrized by finite-dimensional families of 1–forms with values in Hamiltonian vector fields. As discussed in Section 1, it is important that the 1–forms are nonpositive; ie the associated 2–forms are nonpositive multiples of the area form. Furthermore, in order to derive basic homological algebra equations, the 1–forms must be gluing/breaking compatible on the boundary of Deligne–Mumford space. In this section we construct a family of superharmonic functions parametrized by $\mathcal{E}$ that is compatible with the component restriction maps at several-level curves. The differentials of these functions multiplied by the complex unit i then give a family of 1–forms with nonpositive exterior derivative that constitutes the basis for our construction of the 0–order term in the Floer equation.

Fix a smooth decreasing function $\kappa\colon (0,1]\to[0,\infty)$ such that $\kappa(1)=0$ and

$$\lim_{s\to 0+}\kappa(s)=+\infty. \tag{2-2}$$

We will refer to κ as a *stretching profile*.

We will construct a family of functions over curves in $\mathcal{D}$ parametrized by the bundle $\mathcal{E}$ in the following sense. If $e\in\mathcal{E}$ belongs to the fiber over a one-level curve $D_{h;hm,k}\in\mathcal{D}_{h;hm,k}$, then $g_e\colon D_{h;hm,k}\to\mathbb{R}$. If $D_{h;hm,k}$ is a several-level curve with graph Γ and components S_j corresponding to its vertices v_j for $j=0,\dots,s$, then g_e is the collection of functions $g_{r_0(e)},\dots,g_{r_s(e)}$ on $S_0,\dots,S_s$, where r_j denotes the component restriction map to S_j. Our construction uses induction on the number of negative punctures and on the number of levels.

In the first case, $hm+k=1$, and the domain is the strip $\mathbb{R}\times[0,1]$, the cylinder $\mathbb{R}\times S^1$ or the cylinder with a slit (which we view as a subset of $\mathbb{R}\times S^1$). Over these domains, the fiber of $\mathcal{E}$ is a point e, and we take the function g_e to be the projection to the $\mathbb{R}$–factor.

For $hm+k>1$, we specify properties of the functions separately for one-level curves in the interior of $\mathcal{D}_{h;hm,k}$ and for a neighborhood of several-level curves near the boundary. We start with one-level curves. Let e be a section of $\mathcal{E}$ over one-level curves in the interior $\mathring{\mathcal{D}}_{h;hm,k}$. Let $D_{h;hm,k}\in\mathring{\mathcal{D}}_{h;hm,k}$ and write $e=(w_1,\dots,w_{hm+k})\in\Delta^{hm+k-1}$.

We say that a smooth family of functions g_e over the interior satisfies the *one-level conditions* if the following hold (we write $\pi\colon\mathcal{E}\to\mathcal{D}$ for the projection):

(I) There is a constant $c_0=c_0(\pi(e))$ such that in a neighborhood of infinity in the cylindrical end at the positive puncture

$$g_e(s+it)=c_0+s, \tag{2-3}$$

where $s+it$ is the complex coordinate in the cylindrical end, ie in $[0,\infty)\times S^1$ for an interior puncture and in $[0,\infty)\times[0,1]$ for a boundary puncture; see Section 2.1.

(II_1) There are constants $\sigma=\sigma(\pi(e))\in[1,2)$, $R=R(\pi(e))>0$, $c_j=c_j(e)$ and $c_j'=c_j'(e)$ for $j=1,\dots,hm+k$, such that in a neighborhood of infinity in the cylindrical end of the j^{th} negative puncture of the form $(-\infty,0]\times S^1$ for interior punctures or $(-\infty,0]\times[0,1]$ for boundary punctures, we have $g_e(s+it)=g_e(s)$, where

$$g_e(s)=\begin{cases} c_j'+\sigma w_j s & \text{for } -R\ge s\ge -R-\kappa(w_j),\\ c_j+s & \text{for } -R-\kappa(w_j)-1\ge s>-\infty,\end{cases} \tag{2-4}$$

is a concave function, $g_e''(s) \leq 0$, and where κ is the stretching profile (2-2). In particular, for each weight w_j at a negative puncture there is a cylinder or strip region of length at least $\kappa(w_j)$ along which $g_e(s+it) = \epsilon s + C$, with $0 < \epsilon \leq 2w_j$.

(III) The function is superharmonic: $\Delta g_e \leq 0$ everywhere.

(IV) When $h = 1$ so that $D_{h;hm,k}$ is a disk, the derivative of g_e in the direction of the normal ν of the boundary $\partial D_{h;hm,k}$ vanishes everywhere:

$$\frac{\partial g_e}{\partial \nu} = 0 \quad \text{along } \partial D_{h;hm,k}.$$

Remark 2.1 The reason for having $g_e(s) = c_j + s$ rather than $g_e(s) = c_j + \sigma w_j s$ near infinity in (2-4) is to make the functions compatible with splitting. Indeed, the weight equals 1 at the positive puncture of any domain.

Remark 2.2 For the boundary condition IV, note that for the cylinder with a slit, in local coordinates $u + iv$, $v \geq 0$, at the end of the slit, the standard function looks like $g_e(u+iv) = u^2 - v^2$, and $\partial g_e / \partial v = 0$.

Remark 2.3 The appearance of the "extra factor" σ in (2-4) is to allow for a certain interpolation below; see the proof of Lemma 2.4. As we shall see, we can take σ arbitrarily close to 1 on compact sets of $\mathring{\mathcal{D}}_{h;hm,k}$. As mentioned in Section 1, one of the main uses of weights is to force solutions to degenerate for small weights, and for desired degenerations it is enough that σ be uniformly bounded. At the opposite end we find the following restriction on σ: superharmonicity in the cylindrical end near a negative puncture where the weight is w_j implies that $\sigma w_j \leq 1$, and in particular $\sigma \to 1$ if $w_j \to 1$. In general, superharmonicity of the function g_e is equivalent to the differential $d(-i^* dg_e)$ being nonpositive with respect to the conformal area form on the domain $D_{h;hm,k}$. This is compatible with Stokes' theorem, which gives

$$\int_{D_{h;hm,k}} -d(i^* dg_e) = 1 - (hm + k) \leq 0.$$

We will next construct families of functions satisfying the one-level condition over any compact subset of the interior of $\mathcal{D}_{h;hm,k}$. Later we will cover all of $\mathcal{D}_{h;hm,k}$ with a system of neighborhoods of the boundary where condition II_1 above is somewhat weakened but still strong enough to ensure degeneration for small weights.

Lemma 2.4 *If $e\colon \mathring{\mathcal{D}}_{h;hm,k} \to \mathcal{E}$ is a constant section, then over any compact subset $\mathcal{K} \subset \mathring{\mathcal{D}}_{h;hm,k}$, there is a family of functions g_e that satisfies the one-level conditions. Moreover, we can take σ in II_1 arbitrarily close to 1.*

Proof For simpler notation, let $D = D_{h;hm,k}$. Consider first the case when the positive puncture p and all the negative punctures $q_1, \dots, q_k$ are interior. Fix an additional marked point in the domain. For each q_j, fix a conformal map to $\mathbb{R} \times S^1$ which takes the positive puncture to ∞, the marked point to some point in $\{0\} \times S^1$, and the negative puncture to $-\infty$. Fix $\sigma \in (1,2)$ and let $g'_j \colon D \to \mathbb{R}$ be the function $g'_j = \frac{1}{2}(1+\sigma)w_j s_j + c_j$ with s_j the $\mathbb{R}$–coordinate on $\mathbb{R} \times S^1$. Let g_j be a concave approximation of this function with second derivative nonzero only on two intervals of finite length located near $\pm\infty$, linear of slope w_j near $+\infty$ and linear of slope σw_j near $-\infty$; see Figure 5. Note in particular that since $\sigma > 1$ the derivative of g_j will be strictly negative in both intervals. We will use these regions below. Consider the function

$$g = \sum_{j=1}^{k} g_j.$$

Then g is superharmonic but it does not quite have the right behavior at the punctures. Here however, the leading terms are correct and the errors are exponentially small. To see this consider a negative puncture q_j as a point in the cylinder $\mathbb{R} \times S^1$ used to define g_m for $j \neq m$. Let $s + it \in (-\infty, 0) \times S^1$ be the coordinates of the cylindrical end near q_j. The change of variables $z = e^{2\pi(s+it)}$ defines a complex coordinate centered at q_j, with respect to which g_m has a Taylor expansion $g_m(z) = a_{m,0} + a_{m,1}z + a_{m,2}z^2 + \cdots$ around 0. We thus find $g_m(s+it) = a_{m,0} + a_{m,1}e^{2\pi(s+it)} + a_{m,2}e^{4\pi(s+it)} + \cdots$, so that in the cylindrical end near q_j,

$$g(s+it) = g_j(s+it) + \sum_{m \neq j} a_{m,0} + \mathcal{O}(e^{-2\pi|s|}).$$

Thus the error

$$g(s+it) - g_j(s+it) - \text{const} = \mathcal{O}(e^{-2\pi|s|})$$

is exponentially small. We turn off these exponentially small errors in a neighborhood of q_j in the region of support of the second derivative of g_j so that $g(s+it) = g_j(s+it) + \text{const}$ in a neighborhood of infinity as desired.

We can arrange the parameters so that the resulting function satisfies (2-3) near the positive puncture, and it satisfies the top equation in the right-hand side of (2-4) in some neighborhood of q_j. In order to achieve the bottom equation in a neighborhood of q_j we use $\sigma w_j \leq 1$ and simply replace the linear function of slope σw_j by a concave function that interpolates between it and the linear function of slope 1. The fact that we can take σ arbitrarily close to 1 follows from the construction.

The case of boundary punctures can be treated in exactly the same way. In case of a positive boundary puncture and a negative interior puncture we replace the cylinder

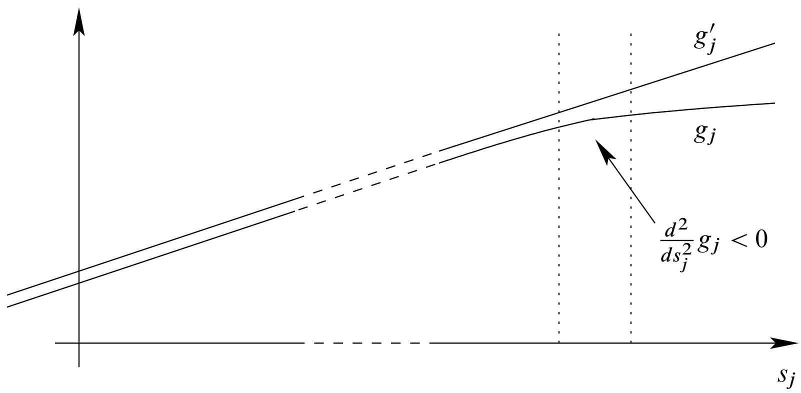


Figure 5: A function g_j that is strictly concave on the region of concavity near $+\infty$

above with the cylinder with a slit along $[0,\infty)\times\{1\}$ and in case of both positive and negative boundary punctures we use the cylinder with a slit all along $\mathbb{R}\times\{1\}$. $\square$

Remark 2.5 For future reference we call the regions in the cylindrical ends where $\Delta g_e < 0$ *regions of concavity*.

We next want to define a corresponding notion for several-level curves. To this end we consider nested neighborhoods

$$\cdots \subset \mathcal{N}^\ell \subset \mathcal{N}^{\ell-1} \subset \mathcal{N}^{\ell-2} \subset \cdots \subset \mathcal{N}^2,$$

where $\mathcal{N}^j$ is a neighborhood of the subset $\mathcal{D}^j \subset \mathcal{D}$ of j–level curves. Consider constant sections e of $\mathcal{E}^{hm+k-1}$ over $\mathring{\mathcal{D}}_{h;hm,k}$ and let g_e be a family of functions. The *ℓ–level conditions* are the same as the one-level conditions I, III and IV, and also the following new condition:

(II_ℓ) For curves in $\mathcal{N}^\ell - \mathcal{N}^{\ell-1}$ with $e=(w_1,\dots,w_{hm+k})$ and any j, there is a strip or cylinder region of length at least $\kappa((w_j)^{1/\ell})$, where $g_e(s+it)=\epsilon s + C$ for $0<\epsilon\le 2(w_j)^{1/\ell}$.

Our next lemma shows that there is a family of functions g_e that satisfies the ℓ–level condition and that is also compatible with splittings into several-level curves in the following sense.

We say that a family of functions g_e as above is *splitting compatible* if the following holds. If $S_\nu \in \mathring{\mathcal{D}}_{h;hm,k}$, $\nu = 1,2,3,\dots$, is a family of curves that converges as $\nu\to\infty$ to an ℓ–level curve with components $S_0,\dots,S_m$ and if $K_\nu \subset S_\nu$ is any compact subset that converges to a compact subset K_j of S_j, then there is a sequence of constants c_ν such that the restriction $g_e|_{K_\nu} + c_\nu$ converges to $g_{r_j(e_j)}|_{K_j}$, where $r_j(e)$ is the component restriction of e to S_j.

Lemma 2.6 *There exists a system of neighborhoods*

$$\cdots \subset \mathcal{N}^\ell \subset \mathcal{N}^{\ell-1} \subset \mathcal{N}^{\ell-2} \subset \cdots \subset \mathcal{N}^2,$$

and a splitting compatible family of functions g_e parametrized by constant sections of $\mathcal{E}$ that satisfies the ℓ–level condition for all $\ell \geq 1$.

Proof The proof is inductive. In the first case $hm + k = 2$ there are only one-level curves and we use the canonical functions g_e discussed above. Consider next a gluing compatible section e over $\mathcal{D}_{h;hm,k}$ with $hm + k = 3$. This space is an interval and the boundary points correspond to two-level curves S with both levels S_0 and S_1 in $\mathcal{D}_{h;hm,k}$, $hm + k = 2$. Consider a neighborhood of such a two-level curve in $\mathcal{D}_{h;hm,k}$ parametrized by a gluing parameter $\rho \in [0, \infty)$; see (2-1). Assume that the positive puncture of S_1 is attached at a negative puncture of S_0. Write $S(\rho) \in \mathcal{D}_{h;hm,k}$, $hm + k = 3$ for the resulting domain, and write $S_j(\rho)$ for the part of the curve $S(\rho)$ that is naturally a subset of S_j. Note that $\rho \to \infty$ as we approach the boundary; see the discussion in Section 2.1. Let $g_{r_0(e)}$ and $g_{r_1(e)}$ denote the functions of the component restrictions of e to S_0 and S_1. If we are sufficiently close to the boundary so that ρ is sufficiently large, then there is a constant $c(\rho)$ such that

$$c(\rho) = g_{r_0(e)}|_{\partial S_0(\rho)} - g_{r_1(e)}|_{\partial S_1(\rho)}. \tag{2-5}$$

We then define the function $g_e(\rho)\colon S(\rho) \to \mathbb{R}$ as

$$g_e(\rho) = \begin{cases} g_{r_0(e)} & \text{on } S_0(\rho), \\ c(\rho) + g_{r_1(e)} & \text{on } S_1(\rho). \end{cases}$$

Then $g_e(\rho)$ is smooth, satisfies I, III and IV, and has the required properties for restrictions to levels. Furthermore, the restriction of $g_e(\rho)$ to $S_0(\rho)$ satisfies (2-4) with σw_j replaced by $\sigma w(q_0)$, where $w(q_0)$ is the weight of $r_0(e)$ at the negative puncture q_0 of S_0 where S_1 is attached (except that the interval in the second equation is not infinite but finite) and the restriction of $g_e(\rho)$ to $S_1(\rho)$ satisfies (2-4) with the weights of $r_1(e)$ at the negative ends of S_1. Let $w_j(r_1(e))$ denote the weights at the negative punctures q_j of S which are negative punctures of S_1, seen as negative punctures of S_1. Then by definition,

$$w_j = w(q_0) w_j(r_1(e)).$$

Since

$$(w_j)^{1/2} = \big(w(q_0) w_j(r_1(e))\big)^{1/2} \geq \min\big(w(q_0), w_j(r_1(e))\big)$$

and

$$\kappa((w_j)^{1/2}) \le \kappa(\min(w(q_0), w_j(r_1(e)))) = \max\bigl(\kappa(w(q_0)), \kappa(w_j(r_1(e)))\bigr),$$

we find that there exists a strip or cylinder region of length at least $\kappa((w_j)^{1/2})$ where $g_e(s+it) = \epsilon s + C$, with $0 < \epsilon \le 2(w_j)^{1/2}$. Thus the two-level condition II_2 holds.

We next want to extend the family of functions over all of $\mathcal{D}_{h;hm,k}$, $hm+k=3$, respecting condition II_2. To this end we consider a neighborhood $\mathcal{N}^{2'}$ of the broken curves in the boundary where the glued functions described above are defined. Using the gluing parameter this neighborhood can be identified with a half infinite interval (ρ_0, ∞), where ∞ corresponds to the broken curve. In some neighborhood (ρ_1, ∞) of ∞ we use the glued functions above. As the gluing parameter decreases in (ρ_0, ρ_1) we deform the derivative of the function as follows: we decrease it uniformly below the gluing region and increase the length of the region near the negative puncture where it is small, until we reach the one-level function. See Figure 6. For this family g_e, conditions I, II_2, III and IV hold everywhere, and II_1 holds in the compact subset of $\mathring{\mathcal{D}}_{h;hm,k}$ which is the complement of a suitable subset $\mathcal{N}^2 \subset \mathcal{N}^{2'}$.

For more general two-level curves with $hm+k>3$ lying in $\mathcal{N}_2 - \mathcal{N}_3$, we argue in exactly the same way using the gluing parameter to interpolate between the natural gluing of the functions of the component restrictions of e and the function of e (see Lemma 2.4) satisfying the one-level condition.

Consider next the general case. Assume that we have found a splitting compatible family of functions g_e, associated to a constant section e defined over the subset $\mathcal{D}^\ell$ consisting of all curves in $\mathcal{D}$ with at most ℓ levels, that satisfies conditions I, III and IV everywhere, and assume that there are nested neighborhoods

$$\mathcal{N}^\ell \subset \mathcal{N}^{\ell-1} \subset \mathcal{N}^{\ell-2} \subset \cdots \subset \mathcal{N}^2,$$

where $\mathcal{N}^j$ is a neighborhood of $\mathcal{D}^j$ in $\mathcal{D}^\ell$ such that condition II_j holds in $\mathcal{N}^j - \mathcal{N}^{j-1}$.

Consider a curve S in the boundary of $\mathcal{D}_{h;hm,k}$ with $\ell+1$ levels. Assume that the top-level curve S_0 of S has r negative punctures at which there are curves $S_1, \dots, S_r$ of levels $\le \ell$ attached. Let $r_j(e)$ denote the component restriction to S_j for $j = 0, 1, \dots, r$. Our inductive assumption gives a smooth splitting compatible family of superharmonic functions with properties I, III and IV for curves in a neighborhood of these broken configurations depending smoothly on $r_j(e)$. Denote the corresponding functions by $g_{r(e_j)}\colon S_j \to \mathbb{R}$. Consider now a coordinate neighborhood U of the form (2-1) around S:

$$U = U^0 \times \prod_{j=1}^{r} (\rho_0^j, \infty) \times U^j.$$

Let $\rho = (\rho_1, \dots, \rho_r)$. For curves $S_j \in U^j$, write $S(\rho)$ for the curve that results from gluing these according to ρ and in analogy with the two-level case, write $S_j(\rho)$ for the part of $S(\rho)$ that is naturally a subset of S_j. Our inductive assumption then shows that there are constants $c_j(r_0(e), r_j(e), \rho_j)$ for $j = 1, \dots, r$ such that

$$c_j(r_0(e), r_j(e), \rho_j) = g^0_{r_0(e)}|_{\partial_j S_0(\rho)} - g_{r_j(e)}|_{\partial S_j(\rho)}, \tag{2-6}$$

where $\partial_j S_0(\rho)$ is the boundary component of $S_0(\rho)$ where $S_j(\rho)$ is attached. Define the function $g_e(\rho)\colon S(\rho) \to \mathbb{R}$ as

$$g_e(\rho) = \begin{cases} g_{r_0(e)} & \text{on } S_0(\rho), \\ g_{r_j(e)} + c_j(r_0(e), r_j(e), \rho_j) & \text{on } S_j(\rho),\ j = 1, \dots, r. \end{cases}$$

The splitting compatibility of the cylindrical ends (see Section 2.1) guarantees that the cylindrical ends on the curves in a neighborhood of the $(\ell{+}1)$–level curve are independent of breaking, ie independent of the way in which the curve is obtained by gluing from some other curve with more levels. Since the shifting constants above are defined in terms of gluing parameters in cylindrical ends, this splitting compatibility then implies the splitting compatibility of the family of functions. It is immediate that the function $g_e(\rho)$ satisfies I, III and IV. We show that condition $\mathrm{II}_{\ell+1}$ holds. Let q be a negative puncture in some S_j for $j = 1, \dots, r$. Let

$$w^{j\prime}_q = w^0_j w^j_q,$$

where w^0_j is the weight of $r_0(e)$ at the negative puncture of S_0 where S_j is attached and where w^j_q is the weight of $r_j(e)$ at the negative puncture q. Then $w^{j\prime}_q w^{j\prime}_q$ is the weight of the puncture q seen as a negative puncture of S. Since

$$(w^{j\prime}_q)^{1/(\ell+1)} = (w^0_j w^j_q)^{1/(\ell+1)} \geq \min(w^0_j, (w^j_q)^{1/\ell})$$

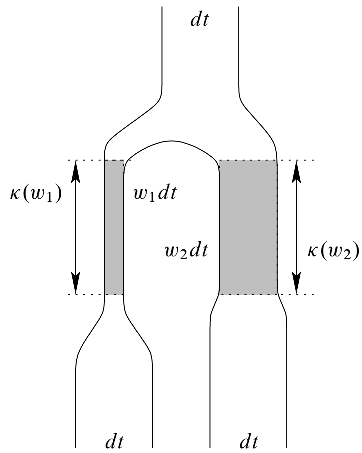
and

$$\kappa((w^{j\prime}_q)^{1/(\ell+1)}) \leq \kappa(\min(w^0_j, (w^j_q)^{1/\ell})) = \max(\kappa(w^0_j), \kappa((w^j_q)^{1/\ell})),$$

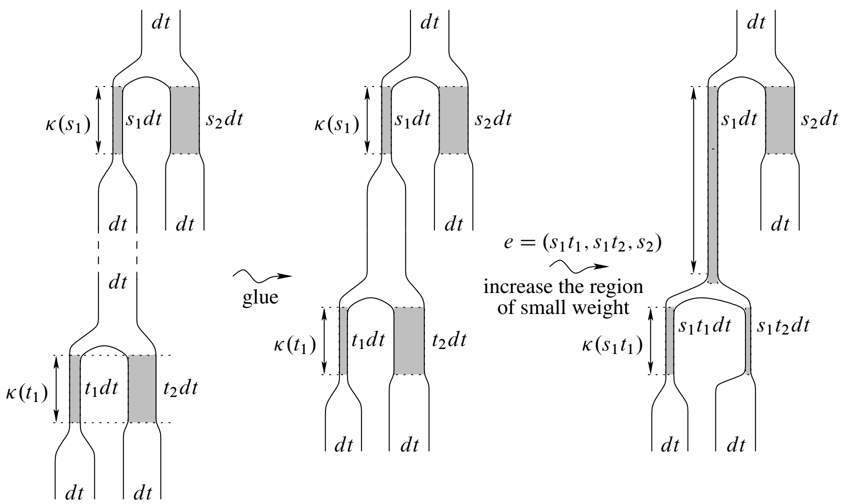
we deduce that condition $\mathrm{II}_{\ell+1}$ holds.

This defines $g_e(\rho)$ in a collar neighborhood of the boundary of $\mathcal{D}^{\ell+1}$. As in the two-level case above we get a family g'_e on the complement of half the collar neighborhood, and then by interpolation we obtain a gluing compatible family over all of $\mathcal{D}^{\ell+1}$ that satisfies conditions II_ℓ and $\mathrm{II}_{\ell+1}$ with respect to an appropriate neighborhood $\mathcal{N}^{\ell+1}$, as required. □

Using the splitting compatible family of subharmonic functions parametrized by $\mathcal{E}$, we define a family of nonpositive 1–forms on the domains in $\mathcal{D}$, likewise parametrized



(I)



(II)

Figure 6: The top picture shows a function that meets the one-level condition. The size of the derivative of the function is indicated by the width of the strip. The lower picture shows how the function changes in a neighborhood of the boundary: very near to the broken curve, we simply glue the functions of the pieces. Moving away from the boundary, we increase the region where the slope is small as indicated and increase the length of the thin regions near the negative puncture until they meet the one-level condition. Further in (not shown in the picture), we interpolate to the one-level function.

by $\mathcal{E}$, as follows. Let i denote the complex structure on the domain $D_{h;hm,k}$ and define

$$\beta_e = -i^* dg_e = \frac{\partial g_e}{\partial \sigma} d\tau - \frac{\partial g_e}{\partial \tau} d\sigma, \tag{2-7}$$

where $\sigma + i\tau$ is a complex coordinate on $D_{h;hm,k}$. Then we find that

$$d\beta_e = (\Delta g_e) d\sigma \wedge d\tau \leq 0,$$

with strict inequality in regions of concavity.

2.5 Hamiltonians

We consider two types of Hamiltonians: one for defining the Hamiltonian simplex DGA that we call *one-step Hamiltonian* and one for defining cobordism maps between DGAs that we call *two-step Hamiltonian*. We use the following convention: if $H\colon X \to \mathbb{R}$ is a Hamiltonian function then we define the corresponding Hamiltonian vector field X_H by

$$\omega(X_H, \cdot) = -dH.$$

Let (X, L) be a Liouville pair with end $[0,\infty) \times (Y, \Lambda)$, and recall our notation $r = e^t$, where t is the coordinate on the factor $[0,\infty)$. We first consider time-independent *one-step* Hamiltonians $H\colon X \to \mathbb{R}$. Such a function has the following properties:

- For small $\epsilon > 0$, $\frac{\epsilon}{2} \leq H \leq \epsilon$ and H is a Morse function on the compact manifold with boundary $X \setminus (0,\infty) \times Y$.
- On $[0,\infty) \times Y$, we have that $H(r, y) = h(r)$ is a function of r only with $h'(r) > 0$ and $h''(r) \geq 0$ such that $H(r) = ar + b$ for $r \geq 1$, where $a > 0$ and b are real constants. We require that a is distinct from the length of any closed Reeb orbit or of any Reeb chord with endpoints on Λ.

Note that in the symplectization part, where $H = h(r)$, the Hamiltonian vector field is proportional to the Reeb vector field R of the contact form α on Y:

$$X_H = h'(r) R.$$

Consider the time-1 flow of the Hamiltonian vector field X_H of H. Hamiltonian chords and orbits then come in two classes. Low-energy orbits that correspond to critical points of H that we take to lie off of L and low-energy chords that correspond to critical points of $H|_L$. The low-energy chords and orbits are generically transverse. High-energy orbits and chords are reparametrizations of Reeb chords and orbits. The chords are generically transverse but the orbits are generically transverse only in the directions transverse to the orbit but not along the orbit. Following [16], we pick a small positive time-dependent perturbation of H near each orbit based on a Morse function on the orbit that gives two orbits of the time-dependent Hamiltonian corresponding to H. We call the resulting Hamiltonian a time-dependent one-step Hamiltonian.

Let (X_0, L_0) be a Liouville pair with end $[0,\infty)\times(Y_0,\Lambda_0)$ and consider a symplectic (Liouville) cobordism (X_{10}, L_{10}) with negative end (Y_0,Λ_0) and positive end (Y_1,Λ_1). Gluing (X_{10}, L_{10}) to (X_0, L_0), we build a new Liouville pair (X_1, L_1) which contains the compact part of (X_0, L_0), connected via $[-R,0]\times(Y_0,\Lambda_0)$ to a compact version (X'_{10}, L'_{10}) of the cobordism, and finally its cylindrical end. Consider time-independent *two-step* Hamiltonians $H\colon X_1\to\mathbb{R}$. Such functions have the following properties:

- For small $\epsilon>0$, we have that $\frac{\epsilon}{2}\le H\le\epsilon$ and that H is a Morse function on X'_0, the complement of $[-R,0]\times Y_0$ in the compact part of X_0.
- On $[-R,-1]\times Y_0$, we have that $H(r,y)=h(r)$ is a function of r only with $h'(r)>0$ and $h''(r)\ge 0$ such that $H(r)=ar+b$ for $r\ge -R+1$, where $a>0$ and b are real constants. We require that a is distinct from the length of any closed Reeb orbit or Reeb chord with endpoints on Λ_0 in Y_0.
- $h'(r)\le 0$ on $[-1,0]\times Y_0$, and the function becomes constant near $0\times Y_0$.
- Over X'_{10}, the function is an approximately constant Morse function.
- Finally, in the positive end, the function has the standard affine form of a one-step Hamiltonian.

Let H_1 be a time-dependent one-step Hamiltonian on X_1 and let H_0 be a two-step Hamiltonian on X_1 with respect to the cobordism X_{01} such that $H_0\ge H_1$.

We consider chords and orbits of both Hamiltonians. The action of a chord or orbit $\gamma\colon[0,1]\to X$ of H_j is

$$\mathfrak{a}(\gamma)=\int_0^1\gamma^*\lambda-\int_0^1 H_j(\gamma(t))\,dt.$$

The nonpositivity of our 1–forms implies that, if we have $D_{1;m,k}\in\mathcal{D}_{1;m,k}$ and $u\colon(D_{1;m,k},\partial D_{1;m,k})\to(X_1,L_1)$ lies in the space of solutions of the Floer equation $\mathcal{F}(a;\boldsymbol{b},\boldsymbol{\eta})$ as defined in Section 4.1 below, with a a chord, $\boldsymbol{b}=b_1\cdots b_m$ a word in chords, and $\boldsymbol{\eta}=\eta_1\cdots\eta_k$ a word in periodic orbits, then

$$\mathfrak{a}(a)-(\mathfrak{a}(b_1)+\cdots+\mathfrak{a}(b_m))-(\mathfrak{a}(\eta_1)+\cdots+\mathfrak{a}(\eta_k))\ge 0.$$

Likewise if $u\in\mathcal{F}(\gamma,\boldsymbol{\eta})$ as defined in Section 4.1 below, with γ a periodic orbit and $\boldsymbol{\eta}=\eta_1\cdots\eta_k$ a word in periodic orbits, then

$$\mathfrak{a}(\gamma)-(\mathfrak{a}(\eta_1)+\cdots+\mathfrak{a}(\eta_k))\ge 0.$$

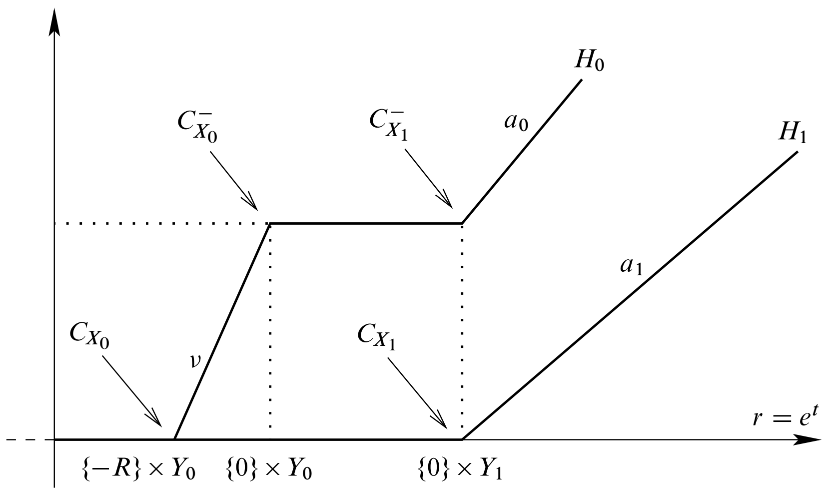


Figure 7: Hamiltonians for the cobordism map

Lemma 2.7 *The Hamiltonian chords and orbits of H_1 decompose into the following subsets:*

- O_{X_1}*: the chords and orbits that correspond to critical points of H_1 in X_1. If $\gamma \in O_{X_1}$, then $\mathfrak{a}(\gamma) \approx 0$.*
- C_{X_1}*: Hamiltonian chords and orbits located near $\{0\} \times Y_1$, and corresponding to Reeb chord and orbits in (Y_1, α_1). If $\gamma \in C_{X_1}$, then $\mathfrak{a}(\gamma) > 0$.*

The Hamiltonian chords and orbits of H_0 decompose into the following subsets:

- O_{X_0}*: the chords and orbits that correspond to critical points of H_0 in X_0. If $\gamma \in O_{X_0}$, then $\mathfrak{a}(\gamma) \approx 0$.*
- C_{X_0}*: Hamiltonian chords and orbits located near $\{-R\} \times Y_0$. If $\gamma \in C_{X_0}$, then $\mathfrak{a}(\gamma) > 0$.*
- $C^-_{X_0}$*: Hamiltonian chords and orbits located near $\{0\} \times Y_0$. Given $\nu > 0$, if R is chosen small enough, then every $\gamma \in C^-_{X_0}$ has $\mathfrak{a}(\gamma) < 0$.*
- $O_{X_{01}}$*: the chords and orbits that correspond to critical points of H_0 in X_{01}. If $\gamma \in O_{X_{01}}$, then $\mathfrak{a}(\gamma) < 0$.*
- $C^-_{X_1}$*: Hamiltonian chords or orbits located near $\{0\} \times Y_1$. If $a_0 < \nu(1 - e^{-R})$, then for any chord or orbit $\gamma^- \in C^-_{X_1}$, we have $\mathfrak{a}(\gamma^-) < 0$.*

Proof Straightforward calculation. □

2.6 Nonpositive 1–forms of Hamiltonian vector fields

Let H be a one-step time-independent Hamiltonian and H_t, $t \in [0,1]$, an associated time-dependent one-step Hamiltonian on X. We will define nonpositive 1–forms with values in Hamiltonian vector fields parametrized by splitting compatible constant sections of $\mathcal{E}$. As before, our construction is inductive. Before we enter the actual construction, recall the notion of nonpositivity for 1–forms B on a Riemann surface Σ with values in Hamiltonian vector fields on X; see Section 1. Each $x \in X$ gives a 1–form on Σ with values in T_xX, $B(x) = X_{H_z}(x) \otimes \beta$, where $H_z\colon X \to \mathbb{R}$, is a family of Hamiltonian functions parametrized by $z \in \Sigma$ and β is a 1–form on Σ. The nonpositivity condition is then that the 2–form associated to B, $d(H_z(x)\,\beta)$, is a nonpositive multiple of the area form on Σ for each $x \in X$.

Let $I = [0,1]$ and $S^1 = I/\partial I$. For cylinders, strips, and cylinders with a slit with coordinates $s + it$, $s \in \mathbb{R}$, $t \in I$ or $t \in I/\partial I$ we use the time-dependent Hamiltonian throughout and define

$$B = X_{H_t} \otimes dt.$$

For $x \in X$, the associated 2–form is $d(H_t(x)\,dt) = 0$ and B is nonpositive.

Consider next disks and spheres in $D_{h;hm,k}$ with $hm + k = 2$. Fix a cut-off function $\psi\colon D_{h;hm,k} \to [0,1]$ which equals 0 outside the cylindrical ends, which equals 1 in a neighborhood of each cylindrical end, and such that $d\psi$ has support in the regions of concavity only. Furthermore, we take the cut-off function to depend on the first coordinate only in the cylindrical end $[0,\infty)\times S^1$ or $(-\infty,0]\times S^1$ at interior punctures and $[0,\infty)\times[0,1]$ or $(-\infty,0]\times[0,1]$ at boundary punctures. Let H_t, $t \in I$, denote the time-dependent one-step Hamiltonian and H the time-independent one, chosen such that $H_t(x) \geq H(x)$ for all $(x,t) \in X \times I$. Let $H_t^{\psi} = (1-\psi)H + \psi H_t$. Define

$$B = X_{H_t^{\psi}} \otimes \beta.$$

For $x \in X$ the associated 2–form is as follows: in the complements of cylindrical ends near the punctures it is given by

$$d(H_t^{\psi}(x)\beta) = H(x)d\beta \leq 0,$$

since $H(x) \geq 0$, and in the cylindrical ends near the punctures, with coordinates $s + it$, by

$$d(H_t^{\psi}(x)\beta) = \psi'(s)\big(H_t(x) - H(x)\big)ds \wedge \beta + \big((1-\psi)H + \psi H_t\big)d\beta \leq 0,$$

where the last inequality holds provided H_t is sufficiently close to H, so that the second term dominates when the first is nonvanishing. (Here we used that $dt \wedge \beta = 0$

in the cylindrical end.) We now extend this field of 1–forms with values in Hamiltonian vector fields over all of $\mathcal{D}$ using induction. For one-level curves in the interior of $\mathcal{D}_{h;hm,k}$, a straightforward extension of the above including more than two ends gives a nonpositive form. For several-level curves, gluing the 1–forms of the components define forms with desired properties in a neighborhood of the boundary of $\mathcal{D}_{h;hm,k}$. Finally, we interpolate between the two fields of forms over a collar region near the boundary using the interpolation of the form part β; see Lemma 2.6. We denote the resulting form with nonpositive differential by B.

Consider next the case of two-step Hamiltonians. As for the one-level Hamiltonians we insert a small time-dependent perturbation near all Reeb orbits of positive action and we get a 0–order term B exactly as above, just replace the one-step Hamiltonian with the two-step Hamiltonian everywhere. Note that, with this definition, the Hamiltonian is time dependent near each puncture. We use this when relating the nonequivariant contact homology to the Morse–Bott version of symplectic homology; see Section 6.

We will consider one further type of 1–form with values in Hamiltonian vector fields that we use to interpolate between one-step and two-step Hamiltonians. Let $H_0 = H$ be the two-step Hamiltonian above and let H_1 be a one-step Hamiltonian on X_1 with $H_1 \le H_0$ everywhere. Let $\phi\colon \mathbb{R} \to \mathbb{R}$ be a smooth function with nonpositive derivative supported in $[-1,1]$ such that $\phi = 1$ in $(-\infty,-1]$ and $\phi = 0$ in $[1,\infty)$. Let $\phi_T = \phi(\cdot - T)$ for $T \in \mathbb{R}$ so that ϕ_T has nonpositive derivative supported in $[T-1, T+1]$, $\phi_T = 1$ in $(-\infty, T-1]$, and $\phi_T = 0$ in $[T+1,\infty)$. Recall the superharmonic field of functions $g = g_e$, $e \in \mathcal{E}$ and let B_0 and B_1 be the fields of 1–forms parametrized by $\mathcal{E}$ associated to H_0 and H_1, respectively, constructed above. Fix an orientation-reversing diffeomorphism $T\colon (0,1) \to \mathbb{R}$. Then the *interpolation form*

$$B_\tau = (1-\phi_{T(\tau)} \circ g)B_1 + (\phi_{T(\tau)} \circ g)B_0 \tag{2-8}$$

is a 1–form with values in Hamiltonian vector fields of the Hamiltonian

$$(1-\phi_{T(\tau)} \circ g)H_1 + (\phi_{T(\tau)} \circ g)H_0.$$

We check that it is nonpositive. For fixed $x \in X_1$, the associated 2–form is

$$\begin{aligned} dB_\tau &= d\big((1-\phi_T \circ g)H_1(x)\beta + (\phi_T \circ g)H_0(x)\beta\big) \\ &= (1-\phi_T \circ g)d(H_1\beta) + (\phi_T \circ g)d(H_0\beta) + \phi_T'(g)\big(H_0(x) - H_1(x)\big)dg \wedge \beta \\ &\le 0, \end{aligned}$$

where the inequality follows since the first term is a convex combination of nonpositive forms and the second is nonpositive as well since $\beta = -i^* dg$ and hence $dg \wedge \beta \ge 0$. Note also that for $\tau = 0$ and $\tau = 1$, we have $B_\tau = B_0$ and $B_\tau = B_1$, respectively.

2.7 Determinant bundles and orientations

We use the field of 1–forms B parametrized by constant splitting compatible sections of $\mathcal{E}$ and almost complex structures over $\mathcal{D}$ to define the Floer equation

$$\overline{\partial}_F u = (du - B)^{0,1} = 0$$

for $u\colon (D_{h;hm,k}, \partial D_{h;hm,k}) \to (X, L)$. In order to study properties of the solution space we will consider the corresponding linearized operator $L\overline{\partial}_F$ which maps vector fields v with one derivative in L^p into complex antilinear maps $L\overline{\partial}_F(v)\colon T_z D_{h;hm,k} \to T_{u(z)} X$, in case of nonempty boundary the vector fields are tangent to L along the boundary. The linearized operator is elliptic and it defines an index bundle over the space of maps. This index bundle is orientable provided the Lagrangian L is relatively spin as was shown in [26]. In this paper we will not use specifics of the index bundle beyond it being orientable. We will however use it to orient solution spaces of the Floer equation. For that purpose we fix capping operators for each Hamiltonian chord and orbit and use linear gluing results to find a system of coherent orientations of the index bundle. The main requirement here is that the positive and negative capping operator at each chord or orbit glues to the operator on a disk or sphere which has a fixed orientation of the index bundle over domains without punctures. The details of this linear analysis are similar to [21] for chords and [25] for orbits. There is however one point where the situation in this paper differs. Namely, our main equation depends on extra parameters corresponding to the simplex and the orientations we use depend on this. In order to get the right graded sign behavior for our Hamiltonian simplex DGA we will use the following conventions.

The index bundle corresponding to the parametrized problem is naturally identified with the index bundle for the unparametrized problem stabilized by the tangent space of the simplex. Here we use the following orientation convention for the simplex. The simplex is given by the equation

$$w_1 + \cdots + w_m = 1,$$

and we think of its tangent space stably as the kernel-cokernel pair $(\mathbb{R}^m, \mathbb{R})$. We use the standard oriented basis $\partial_1, \dots, \partial_m$ of $\mathbb{R}^m$ and ∂_0 of $\mathbb{R}$. We then think of the direction ∂_j as a stabilization of the capping operator of the j^{th} negative puncture and of ∂_0 as a stabilization of that at the positive puncture and get the induced orientation of the index bundle over $\mathcal{E}$ by gluing these stabilized operators. Then the index bundle orientations reflect Conley–Zehnder/Maslov grading in the DGA as usual. We give a more detailed discussion of index bundles and sign rules in the DGA in the appendix.

3 Properties of Floer solutions

In this section we establish two basic results about Floer holomorphic curves. First we prove that the $\mathbb{R}$–factor of any Floer holomorphic curve in the cylindrical end of a Liouville manifold satisfies a maximum principle. This result allows us to establish the correct form of Gromov–Floer compactness for our theories. Second we establish an elementary energy bound that ensures our Floer equations do not have any solutions with only high-energy asymptotes near the boundary of the parametrizing simplex.

3.1 A maximum principle for solutions of Floer equations

Consider a 1–parameter family of fields of splitting compatible 1–forms $B = B_\tau$, $\tau \in [0, 1]$, parametrized by constant sections of $\mathcal{E}$ and constructed from one-step and two-step Hamiltonians as in Section 2.6. (Fields of forms constructed only from a one-step Hamiltonian appear here as special cases corresponding to constant $\tau = 1$.) Let J be a splitting compatible field of almost complex structures over $\mathcal{D}$. Recall that this means in particular that if $S = D_{h;hm,k} \in \mathcal{D}_{h;hm,k}$, then J_z is an almost complex structure on X for each $z \in S$ such that in any cylindrical end with coordinate $s + it$, $J_{s+it} = J_t$; see Section 2.2.

We make the following *nondegeneracy* assumption. The one and two-step Hamiltonians are both linear at infinity $H(r, y) = h(r) = ar + b$ for real constants $a > 0$ and b. We assume that the length ℓ of any Reeb orbit or Reeb chord satisfies

$$\ell \neq a. \tag{3-1}$$

Note that the set of Reeb chord and orbit lengths is discrete and hence the condition on the Hamiltonians holds generically.

Consider now a solution $u\colon S \to X$ of the Floer equation

$$(du - B)^{0,1} = 0,$$

where the complex antilinear component of the map $(du - B)\colon T_zS \to T_{u(z)}X$ is taken with respect to the almost complex structures J_z on X and j on S.

Lemma 3.1 *If the nondegeneracy condition at punctures* (3-1) *is satisfied, then* $u(S)$ *is contained in the compact subset* $\{r \leq 1\}$.

Proof Assume that there exists $z \in S$ such that

$$r(z) = r(u(z)) > 1. \tag{3-2}$$

Following [4, Section 7], we show that (3-2) leads to a contradiction. Fix a regular value $r' > 1$ of the smooth function $r \circ u$ such that

$$S' = \{z \in S : r \circ u(z) \geq r'\} \neq \varnothing.$$

Then S' is a Riemann surface with boundary with corners and its boundary can be decomposed as $\partial S' = \partial_{r'} S' \cup \partial_L S'$, where $u(\partial_{r'} S') \subset \{r = r'\}$ and $u(\partial_L S') \subset L$. Here both $\partial_{r'} S'$ and $\partial_L S'$ are finite unions of circles and closed intervals. The intervals in $\partial_{r'} S'$ and $\partial_L S'$ intersect at their endpoints that are the corners of $\partial S'$.

Define the energy of $u\colon S \to X$ to be

$$E(u) = \frac{1}{2} \int_S \|du - B\|^2,$$

where we measure the norm with respect to the metric $\omega(\cdot\,, J_z\,\cdot)$. A straightforward computation shows that

$$E(u) = \int_S u^*\omega - u^* dH_z \wedge \beta.$$

Recall the family of 2–forms $\theta(x) = d(H_z(x)\beta)$ associated to the one form B with values in Hamiltonian vector fields parametrized by $x \in X$, and recall the nonpositivity condition for B which says that $\theta(x)$ is a nonpositive 2–form for each $x \in X$; see Section 1. Consider the energy of S':

$$\begin{aligned} E(u|_{S'}) &= \int_{S'} u^*\omega - u^* dH_z \wedge \beta \\ &\le \int_{S'} u^*\omega - u^* dH_z \wedge \beta - \theta(u(z)) \\ &= \int_{\partial S'} u^*\, r\alpha - H_z(u(z))\beta. \end{aligned}$$

Since $\alpha|_L = 0$ and $\beta|_{\partial S} = 0$, and since $H_z(r, y) = ar + b$ in the region $\{r \ge 1\}$ where $b < 0$ for $a > 0$ sufficiently large, the last integral satisfies

$$\begin{aligned} \int_{\partial S'} u^*\, r\alpha - H_z(u(z))\beta &= \int_{\partial_{r'} S'} u^*\, r\alpha - a\, u^* r\, \beta - b\beta \\ &= \int_{\partial_{r'} S'} u^*\, r\alpha - a\, u^* r\, \beta - b \int_{S'} d\beta \\ &\le \int_{\partial_{r'} S'} u^*\, r\alpha - a\, u^* r\, \beta \\ &= r' \int_{\partial_{r'} S'} \alpha \circ (du - X_H \otimes \beta) \\ &= r' \int_{\partial_{r'} S'} \alpha \circ J_z \circ (du - X_H \otimes \beta) \circ (-i) \le 0, \end{aligned}$$

where i is the complex structure on S. Here we use the identities $\alpha \circ J_z = dr$ and $dr(X_{H_z}) = 0$. The last inequality uses that $u(S')$ is contained in $\{r \ge r'\}$. Indeed, if v is a positively oriented tangent vector to $\partial_{r'}S'$, then $-iv$ points outwards, and therefore $d(r\circ u)(-iv)\le 0$.

We find that $E(u|_{S'})\le 0$, which implies that u satisfies $du-X_H\otimes\beta=0$ on S'. Since u intersects the level $r=r'$, it then follows that the image of any connected component of S' under u is contained in the image of a Reeb orbit or chord in this level set. Note that this conclusion is independent of the choice of regular level set $r'>1$ such that $S'=u^{-1}(\{r\ge r'\})\neq\varnothing$. Since such regular level sets exist (and are actually dense) in the interior of the original interval $(1,r')$, we get a contradiction. □

3.2 An action bound

In this section we establish an elementary action bound that we will use to show that our $\mathcal{E}$–families of Floer equations have no solutions with only high-energy asymptotes near the boundary of the fiber simplex. Consider a Liouville manifold with an exact Lagrangian submanifold (X,L) and let H_t be a one or two-step time-dependent Hamiltonian as above and let $\epsilon_0>0$ denote the smallest value of the action

$$\mathfrak{a}(\gamma)=\int_\gamma \lambda-H_t\,dt \tag{3-3}$$

of a Hamiltonian chord or orbit γ corresponding to a Reeb chord or orbit. Then any high-energy chord or orbit has action at least ϵ_0.

Let $u\colon S\to X$ for $S=D_{h;hm,k}$ be a solution of the Floer equation

$$(du-B)^{0,1}=0,$$

asymptotic at the positive puncture to a periodic orbit or chord γ.

Lemma 3.2 *There are constants $L,\epsilon>0$ such that the following holds for any $L'\ge L$ and $0<\epsilon'\le\epsilon$. If there is a strip region $V=[0,L']\times I$ or cylinder region $V=[0,L']\times S^1$ in S of length L' that separates a negative puncture q from the positive puncture p and such that $B=X_H\otimes\epsilon\,dt$, in standard coordinates $s+it$ in V, where H does not depend on $s+it$, then q maps to a low-energy chord or orbit.*

Remark 3.3 As will be seen from the proof, the constant $L>0$ depends only on ϵ_0, the action E_+ of the Hamiltonian chord or orbit at p, $M=\max_{\{r\le 1\}}H$ and $C=\max_{\{r\le 1\}}\|\lambda\|$, where λ is the Liouville form, while the constant $\epsilon>0$ depends also on $F=\max_{\{r\le 1\}}\|X_H\|$.

Proof We use notation as in the proof of Lemma 3.1 and Remark 3.3. Consider the energy

$$
\begin{aligned}
E = \int_S \|du - B\|^2 &= \int_S u^*\omega - u^* dH_z \wedge \beta \\
&\le \int_S u^*\omega - u^* dH_z \wedge \beta - \theta(u(z)) \\
&= E_+ - \sum_{i=1}^{hm+k} E_{i,-},
\end{aligned}
$$

where E_+ is the action at the positive puncture and $E_{i,-}$, $i = 1, \dots, hm + k$ are the actions at the negative punctures. In particular, the action $E_{i,-}$ at any of the negative punctures satisfies

$$E_{i,-} \le E_+. \tag{3-4}$$

Also, because each of the actions $E_{i,-}$ is positive, we have

$$E \le E_+.$$

Consider now the contribution to the energy from the strip or cylinder region V. Fix $\eta > 0$ and note that, in the strip case, the measure of the set of points $s \in [0, L]$ such that

$$\int_{\{s\}\times[0,1]} \|\partial_t u - \epsilon X_H\|^2 \, dt \ge \eta$$

is bounded by E/η (similarly for the integral over $\{s\} \times S^1$ in the cylinder case). In particular if $L > E/\eta$ we have that there are slices $\gamma' = \{s_0\} \times [0, 1]$ in the strip case or $\gamma' = \{s_0\} \times S^1$ in the cylinder case for which $\|\partial_t u(s_0, \cdot) - \epsilon X_H\|^2_{L^2} \le \eta$, which implies

$$\|\partial_t u(s_0, \cdot)\|_{L^2} \le \sqrt{\eta} + \epsilon F.$$

We obtain for the action of γ' the estimate

$$
\begin{aligned}
\left| \int_{\gamma'} \lambda - \epsilon H \, dt \right| &\le C \|\partial_t u(s_0, \cdot)\|_{L^1} + \epsilon M \\
&\le C\sqrt{\eta} + \epsilon(CF + M).
\end{aligned}
$$

Applying Stokes' theorem to the energy integral of the part S' of S containing γ' and the negative puncture q then shows as in (3-4) that the energy of the chord or orbit at q is $< \epsilon_0$, provided $\eta = E/L$ and ϵ are sufficiently small. □

4 Properties of spaces of Floer solutions

Let (X, L) be a Liouville pair as before. Consider an $\mathcal{E}$–family B_τ, $\tau \in [0,1]$, of interpolation splitting compatible 1–forms over $\mathcal{D}$ with values in Hamiltonian vector fields (see Section 2.6), and a field of domain dependent almost complex structures (see Section 2.2) where the Hamiltonians satisfy the nondegeneracy condition at infinity. Here we think of $(X, L) = (X_1, L_1)$ constructed from a cobordism if $\tau \in [0,1)$, and if $\tau = 1$ we also allow standard Liouville pairs.

This data allows us to study the Floer equation

$$(du - B_\tau)^{0,1} = 0 \tag{4-1}$$

for $u\colon (D_{h;hm,k}, \partial D_{h;hm,k}) \to (X, L)$. We will refer to solutions of (4-1) as *Floer holomorphic curves*.

4.1 Transversality and dimension

In order to express the dimensions of moduli spaces of Floer holomorphic curves, we use Conley–Zehnder indices for chords and orbits (with conventions as in [13, Appendix A.1]). They are defined as follows. If γ is a Hamiltonian orbit, then fix a disk D_γ (recall that we assume $\pi_1(X) = 1$) that bounds γ and a trivialization of the tangent bundle TX over D_γ. The Conley–Zehnder index $\mathrm{CZ}(\gamma) \in \mathbb{Z}$ of a Hamiltonian orbit is then defined using the path of linear symplectic matrices that arises as the linearization of the Hamiltonian flow along γ in this trivialization; see [37]. Then $\mathrm{CZ}(\gamma)$ is independent of the choice of trivialization since $c_1(X) = 0$.

If c is a Hamiltonian chord, we pick a capping disk D_c mapping the unit disk into X as follows. Pick a base point in each component of the Lagrangian L. Fix paths connecting base points in different components and along these paths fix paths of Lagrangian tangent planes connecting the tangent planes of the Lagrangian L at the base points. (We use the constant path with the constant tangent plane at the base point connecting the base point in a given component to itself.) In the disk D_c we map the boundary arc ∂D_c^- between -1 and 1 to the Hamiltonian chord, and we map the boundary arc ∂D_c^+ between 1 and -1 as follows: the boundary arc between 1 and $e^{\pi i/4}$ is mapped to the component of L that contains the Hamiltonian chord endpoint, and connects the latter to the base point; the arc between $e^{\pi i/4}$ and $e^{3\pi i/4}$ follows the path between base points; finally, the arc between $e^{3\pi i/4}$ and -1 is mapped to the connected component of L that contains the Hamiltonian chord start point, and connects the base point to the Hamiltonian chord start point. This then gives the following loop Γ_c of Lagrangian planes: along ∂D_c^+ we follow first the tangent planes of L starting at the endpoint of the chord and ending at the base point, then the planes along the path connecting base points, then again planes tangent to L from the base point to the start point of the chord; along ∂D_c^- we transport the tangent plane of the Lagrangian at the chord start point by the linearization of the Hamiltonian flow along the chord, and finally we close up by a rotation along the complex angle in the positive direction connecting the transported Lagrangian plane to the tangent plane at the endpoint of the chord. We define

$$\mathrm{CZ}(c) = \mu(\Gamma_c),$$

where μ denotes the Maslov index of Γ_c read in a trivialization of TX over ∂D_c that extends over D_c. This is then well defined since $c_1(X) = 0$ and since the Maslov class of L vanishes.

Remark 4.1 The Conley–Zehnder index $\mathrm{CZ}(c)$ of a Reeb chord c with both endpoints in one component of the Lagrangian submanifold is independent of all choices. For chords with endpoints in distinct components CZ is independent up to an over all shift that depends on the choice of tangent planes along the path connecting base points.

We also define positive and negative capping operators. For chords c these operators $o_\pm(c)$ are defined using capping disks. This capping operator is a linearized Floer-operator on a once boundary-punctured disk, with Lagrangian boundary condition given by the tangent planes along the capping path oriented from the endpoint of the chord to the start point for the positive capping operator $o_+(c)$ and with the reverse path for the negative capping operator $o_-(c)$. We assume (as is true for generic data) that the image of the Lagrangian tangent plane at the start point of the chord under the linearized flow is transverse to the tangent plane at the endpoint. For orbits, the capping operators $o_\pm(\gamma)$ are operators on punctured spheres with positive or negative puncture with asymptotic behavior determined by the linearized Hamiltonian flow along the orbit γ. More precisely, the capping operators are then $\bar\partial$–operators perturbed by a 0–order term acting on the Sobolev space of vector fields v on the punctured sphere S or disk D that in the latter case are tangent to the Lagrangian along ∂D with one derivative in L^p, $p > 2$.

We find that the chord capping operators are Fredholm and their index is given by the formula [32; 22]

$$\mathrm{index}(o_+(c)) = n + (\mathrm{CZ}(c) - n) = \mathrm{CZ}(c), \quad \mathrm{index}(o_-(c)) = n - \mathrm{CZ}(c).$$

The orbit capping operators have index [3; 13]

$$\mathrm{index}(o_+(\gamma)) = n + \mathrm{CZ}(\gamma), \quad \mathrm{index}(o_-(\gamma)) = n - \mathrm{CZ}(\gamma).$$

Let a be a Hamiltonian chord, γ a Hamiltonian orbit, $\boldsymbol{b} = b_1 \cdots b_m$ a word of Hamiltonian chords and $\boldsymbol{\eta} = \eta_1 \cdots \eta_k$ a word of Hamiltonian orbits. Let $\sigma \in \mathcal{E}$ be a splitting compatible constant section over $\mathcal{D}$, which takes values in the simplex Δ^{hm+k-1} over the interior of $\mathcal{D}_{h;hm,k}$.

When the number of boundary components of the source curve is $h = 1$, we consider the moduli space $\mathcal{F}^{\sigma}_{\tau}(a; \boldsymbol{b}, \boldsymbol{\eta})$ of solutions

$$u\colon (D_{1;m,k}, \partial D_{1;m,k}) \to (X, L), \quad D_{1;m,k} \in \mathcal{D}_{1;m,k},$$

of the Floer equation

$$(du - B^{\sigma}_{\tau})^{0,1} = 0.$$

Here B^{σ}_{τ} is the 1–form with values in Hamiltonian vector fields determined by $\sigma \in \mathcal{E}$. The map u converges at the positive puncture to a, and at the negative punctures to $b_1, \dots, b_m, \eta_1, \dots, \eta_k$. Note that if $\tau \neq 1$, then a is a chord of the Hamiltonian H_0 and b_j and η_l chords and orbits of H_1. The interior negative punctures are endowed with asymptotic markers induced from the positive boundary puncture as in Section 2.1. We write

$$\mathcal{F}_{\tau}(a; \boldsymbol{b}, \boldsymbol{\eta}) = \bigcup_{\sigma \in \Delta^{hm+k-1}} \mathcal{F}^{\sigma}_{\tau}(a; \boldsymbol{b}, \boldsymbol{\eta}).$$

(Recall that the family B^{σ} depends smoothly on σ.) We also write

$$\mathcal{F}_{\mathbb{R}}(a; \boldsymbol{b}, \boldsymbol{\eta}) = \bigcup_{\tau \in (0,1)} \mathcal{F}_{\tau}(a; \boldsymbol{b}, \boldsymbol{\eta}), \quad \mathcal{F}^{\sigma}_{\mathbb{R}}(a; \boldsymbol{b}, \boldsymbol{\eta}) = \bigcup_{\tau \in (0,1)} \mathcal{F}^{\sigma}_{\tau}(a; \boldsymbol{b}, \boldsymbol{\eta}).$$

In the case that the domain is a cylinder or strip we will discuss the definition of the moduli space $\mathcal{F}_{\mathbb{R}}$ with more details in Remark 4.3. (Intuitively, the parameter τ moves the interpolation region along the domain, but in these domains we also divide by the natural $\mathbb{R}$–translation. This then together has the same effect as simply fixing the location of the interpolation region.)

When the number of boundary components of the source curve is $h = 0$, we similarly consider the moduli space $\mathcal{F}^{\sigma}_{\tau}(\gamma; \boldsymbol{\eta})$ of solutions

$$u\colon D_{0;0,k} \to X, \quad D_{0;0,k} \in \mathcal{D}_{0;0,k},$$

of the Floer equation

$$(du - B^{\sigma}_{\tau})^{0,1} = 0,$$

converging at the positive puncture to γ, and at the negative punctures to $\eta_1, \dots, \eta_k$. Again, if $\tau \neq 1$, then γ is an orbit of the Hamiltonian H_0 and the η_l orbits of H_1. Here the positive puncture has a varying asymptotic marker, which induces asymptotic markers at all the negative punctures as described in Section 2.1. We write

$$\mathcal{F}_\tau(\gamma;\boldsymbol{\eta}) = \bigcup_{\sigma\in\Delta^{k-1}} \mathcal{F}_\tau^\sigma(\gamma;\boldsymbol{\eta})$$

and

$$\mathcal{F}_\mathbb{R}(\gamma;\boldsymbol{\eta}) = \bigcup_{\tau\in(0,1)} \mathcal{F}_\tau(\gamma;\boldsymbol{\eta}), \quad \mathcal{F}_\mathbb{R}^\sigma(\gamma;\boldsymbol{\eta}) = \bigcup_{\tau\in(0,1)} \mathcal{F}_\tau^\sigma(\gamma;\boldsymbol{\eta}).$$

Remark 4.2 Floer equations corresponding to one-step Hamiltonians are a special case of the above, corresponding to $\tau = 1$. We sometimes use a simpler notation for such spaces: we drop the $\tau = 1$ subscript and write $\mathcal{F}^\sigma = \mathcal{F}_1^\sigma$ and $\mathcal{F} = \mathcal{F}_1$.

Remark 4.3 More precise definitions of the moduli space $\mathcal{F}_\mathbb{R}$ in the case that the domain is a strip or a cylinder are as follows: For a given Hamiltonian H the moduli space $\mathcal{F} = \mathcal{F}_1$ is the space of solutions of the Floer equation $(du - X_H \otimes dt)^{0,1} = 0$ modulo the $\mathbb{R}$–action by translations in the source. This interpretation of $\mathcal{F}$ is compatible with breaking.

When we interpolate between two Hamiltonians, we define the moduli space $\mathcal{F}_\mathbb{R}$ as the space of solutions $\mathcal{F}_{\tau_0}$, $\tau_0 \in (0,1)$, for a fixed value $\tau = \tau_0 \in (0,1)$. This interpretation of $\mathcal{F}_\mathbb{R}$ is compatible with breaking as follows. Consider a one-step Hamiltonian H_1 and a two-step Hamiltonian $H_0 \geq H_1$, and a fixed $\tau_0 \in (0,1)$ such that $\mathcal{F}_\mathbb{R} = \mathcal{F}_{\tau_0}$ for cylinders and strips, where $\mathcal{F}_{\tau_0}$ is the space of solutions of the Floer continuation equation

$$(du - X_{(1-\phi_{T(\tau_0)}(s))H_1 + \phi_{T(\tau_0)}(s)H_0} \otimes dt)^{0,1} = 0,$$

where ϕ_T is the function used in (2-8) in order to define the 1–form

$$B_\tau = X_{(1-\phi_{T(\tau)}(s))H_1 + \phi_{T(\tau)}(s)H_0} \otimes dt.$$

If a strip splits off from such a domain and if the interpolation region (the support of the derivative of ϕ_T) lies in this strip, then since the functions ϕ_T are defined as shifts $\phi_T(s) = \phi(s-T)$ of a given function ϕ (see the discussion preceding (2-8)), there is a unique translation of the parametrization of the domain of the split-off strip or cylinder such that we get a solution in $\mathcal{F}_{\tau_0} = \mathcal{F}_\mathbb{R}$.

Theorem 4.4 *Assume $hm+k \geq 2$. For generic families of almost complex structures and Hamiltonians, the moduli spaces $\mathcal{F}_\tau(\gamma, \boldsymbol{\eta})$, $\mathcal{F}_\tau(a; \boldsymbol{b}, \boldsymbol{\eta})$, $\mathcal{F}_\mathbb{R}(\gamma; \boldsymbol{\eta})$ and $\mathcal{F}_\mathbb{R}(a; \boldsymbol{b}, \boldsymbol{\eta})$ are manifolds of dimensions*

$$\begin{aligned}
\dim \mathcal{F}_\tau(\gamma; \boldsymbol{\eta}) &= \dim \mathcal{F}_\mathbb{R}(\gamma; \boldsymbol{\eta}) - 1 \\
&= (\mathrm{CZ}(\gamma) + (n-3)) - \sum_{j=1}^{k} (\mathrm{CZ}(\eta_j) + (n-3)) - 1, \\
\dim \mathcal{F}_\tau(a; \boldsymbol{b}, \boldsymbol{\eta}) &= \dim \mathcal{F}_\mathbb{R}(a; \boldsymbol{b}, \boldsymbol{\eta}) - 1 \\
&= (\mathrm{CZ}(a) - 2) - \sum_{j=1}^{m} (\mathrm{CZ}(b_j) - 2) - \sum_{j=1}^{k} (\mathrm{CZ}(\eta_j) + (n-3)) - 1.
\end{aligned}$$

For generic fixed $\sigma \in \Delta^{hm+k-1}$, the corresponding moduli spaces $\mathcal{F}^\sigma(\gamma; \boldsymbol{\eta})$ and $\mathcal{F}^\sigma(a; \boldsymbol{b}, \boldsymbol{\eta})$ are manifolds of dimensions

$$\begin{aligned}
\dim \mathcal{F}^\sigma_\tau(\gamma; \boldsymbol{\eta}) &= \dim \mathcal{F}^\sigma_\mathbb{R}(\gamma; \boldsymbol{\eta}) - 1 \\
&= (\mathrm{CZ}(\gamma) + (n-2)) - \sum_{j=1}^{k} (\mathrm{CZ}(\eta_j) + (n-2)) - 1, \\
\dim \mathcal{F}^\sigma_\tau(a; \boldsymbol{b}, \boldsymbol{\eta}) &= \dim \mathcal{F}^\sigma_\mathbb{R}(a; \boldsymbol{b}, \boldsymbol{\eta}) - 1 \\
&= (\mathrm{CZ}(a) - 1) - \sum_{j=1}^{m} (\mathrm{CZ}(b_j) - 1) - \sum_{j=1}^{k} (\mathrm{CZ}(\eta_j) + (n-2)) - 1.
\end{aligned}$$

Furthermore, for generic data, the projection of the moduli spaces $\mathcal{F}_\mathbb{R}$ and $\mathcal{F}^\sigma_\mathbb{R}$ to the line $\mathbb{R}$ (interpolating between the Hamiltonians) is a Morse function with distinct critical values.

Remark 4.5 In the case where $hm+k = 1$, ie the domain is a strip or a cylinder, the parameter σ is irrelevant since the simplex consists of a single point, and the dimensions of the relevant moduli spaces are

$$\begin{aligned}
&\dim \mathcal{F}_1(\gamma; \eta) = \mathrm{CZ}(\gamma) - \mathrm{CZ}(\eta) - 1, \quad \dim \mathcal{F}_\mathbb{R}(\gamma; \eta) = \mathrm{CZ}(\gamma) - \mathrm{CZ}(\eta), \\
&\dim \mathcal{F}_1(a; b) = \mathrm{CZ}(a) - \mathrm{CZ}(b) - 1, \quad \dim \mathcal{F}_\mathbb{R}(a; b) = \mathrm{CZ}(a) - \mathrm{CZ}(b).
\end{aligned}$$

Proof of Theorem 4.4 To see this we first note that the operator we study is Fredholm. The expected dimension of the moduli space is then given by the sum of the index of the operator acting on a fixed surface and the dimension of auxiliary parameter spaces (ie the space of conformal structures on the domain and the space which parametrizes the choice of 1–forms).

Consider first the case when $h = 1$. We denote the index of the operator on the fixed surface by $\operatorname{index}(a; \boldsymbol{b}, \boldsymbol{\eta})$. To compute it, we glue on capping operators at all punctures. Additivity of the index under gluing at a nondegenerate chord or orbit together with the Riemann–Roch formula then gives

$$n = \operatorname{index}(a; \boldsymbol{b}, \boldsymbol{\eta}) + n - \mathrm{CZ}(a) + \sum_{j=1}^{m} \mathrm{CZ}(b_j) + \sum_{j=1}^{k} (\mathrm{CZ}(\eta_j) + n).$$

The dimension is then obtained by adding the dimension of the space of conformal structures and that of the space of 1–forms:

$$\dim \mathcal{F}(a; \boldsymbol{b}, \boldsymbol{\eta}) = \operatorname{index}(a; \boldsymbol{b}, \boldsymbol{\eta}) + (m-2) + 2k + (m+k-1).$$

When the form B^σ is fixed, we simply subtract $m + k - 1$, the dimension of the simplex. The calculation in the case $h = m = 0$ is similar and gives

$$\dim \mathcal{F}(\gamma; \boldsymbol{\eta}) = (\mathrm{CZ}(\gamma) + n) - \sum_{j=1}^{k} (\mathrm{CZ}(\eta_j) + n) + 2k - 3 + (k-1),$$

where $2k - 3$ is the dimension of the space of conformal structures on the sphere with $k + 1$ punctures where there is a varying asymptotic marker at one of the punctures. In the case where the form B^σ is fixed we subtract the dimension of the simplex, $k - 1$.

Finally, to see that these are manifolds, we need to establish surjectivity of the linearized operator for generic data. This is well known in the current setup and follows from the unique continuation property of pseudoholomorphic curves in combination with an application of the Sard–Smale theorem. The key points are that J (and H) are allowed to depend on all parameters and that (X, L) is exact so that no bubbling of pseudoholomorphic spheres or disks occurs; see eg [6, Appendix] and [32, Section 9.2].

The last statement is a straightforward consequence of the Sard–Smale theorem. □

Remark 4.6 Note that letting the markers at the negative ends be determined by that at the positive end is compatible with splitting, which is essential for the description of moduli space boundaries. Also, in the case that the domain is a cylinder our moduli spaces are the same as the usual moduli spaces of Floer cylinders defined by the fixed domain $\mathbb{R} \times S^1$ with the distinguished line $\mathbb{R} \times \{1\}$.

We next show that there are no solutions of the Floer equation with only high-energy asymptotes if the 0–order term corresponds to a constant section of $\mathcal{E}$ that lies sufficiently close the boundary of the simplex.

Lemma 4.7 *For any $E > 0$, there exists $\epsilon > 0$ such that if σ is a constant section of $\mathcal{E}$ that lies in an ϵ neighborhood of the boundary of the simplex, and if $\mathfrak{a}(a) < E$ and $\mathfrak{a}(\gamma) < E$, then for any nontrivial words $\boldsymbol{b}$ and $\boldsymbol{\eta}$ of high-energy chords and orbits, respectively, and for any $\tau \in [0, 1]$, the moduli spaces $\mathcal{F}_\tau^\sigma(a; \boldsymbol{b}, \boldsymbol{\eta})$ and $\mathcal{F}_\tau^\sigma(\gamma; \boldsymbol{\eta})$ are empty.*

Proof This is an immediate consequence of the ℓ–level condition on the nonnegative 1–form β and Lemma 3.2. $\square$

4.2 Compactness and gluing

For simpler notation, we write $\mathcal{F}_\tau$, $\mathcal{F}_\tau^\sigma$, $\mathcal{F}_\mathbb{R}$ and $\mathcal{F}_\mathbb{R}^\sigma$ with unspecified punctures as common notation for either type of moduli space (corresponding to either $h = 0$ or $h = 1$) in Theorem 4.4. We also write $\mathcal{F}_\tau^+$ and $\mathcal{F}_\mathbb{R}^+$ for components of $\mathcal{F}_\tau$ and $\mathcal{F}_\mathbb{R}$ where all asymptotic chords and orbits are of high energy. Recall that, if B^σ is a splitting compatible field of 1–forms determined by a constant splitting compatible section σ of $\mathcal{E}$ then, over a several-level curve, B^σ determines 1–forms depending on constant sections over its pieces.

Theorem 4.8 *The spaces $\mathcal{F}_\tau^\sigma$ and $\mathcal{F}_\tau^+$ admit compactifications as manifolds with boundary with corners, where the boundary corresponds to several-level curves in $\mathcal{F}_\tau^\sigma$ and $\mathcal{F}_\tau^+$ respectively, joined at Hamiltonian chords or orbits.*

Proof The fact that any sequence of curves in $\mathcal{F}_\tau^\sigma$ has a subsequence that converges to a several-level curve is a well-known form of Gromov–Floer compactness for (X, L) exact. In order to find a neighborhood of the several-level curves in the boundary of the moduli space we use Floer gluing. That the Floer equation is compatible with degeneration in the moduli space of curves is a consequence of the gluing compatibility condition for the family of 1–forms B^σ, $\sigma \in \mathcal{E}$. Both compactness and gluing are treated in [32, Chapters 4 and 10] and in [40, Chapter 9]; see also eg [24, Appendix A] for a treatment of family gluing.

For $\mathcal{F}_\tau^+$, by Lemma 4.7 note that there are no solutions near the boundary of the simplex so the only possible boundary are broken curves joined at high-energy chords or orbits. $\square$

We next consider compactifications of moduli spaces $\mathcal{F}_\mathbb{R}^+$ which consist of solutions of the Floer equation with the interpolation form B_τ as τ varies over $(0, 1)$. Similar results hold for moduli spaces $\mathcal{F}_\mathbb{R}^\sigma$, but we focus on the high-energy case since that is all we use later and since we then need not involve any low-energy chords and orbits.

Theorem 4.9 *The moduli spaces $\mathcal{F}_{\mathbb{R}}^{+}$ admit compactifications as manifolds with boundary with corners, where the boundary corresponds to several-level curves joined at Hamiltonian chords and orbits of the following form:*

- *Exactly one level S (possibly of several components) lies in $\mathcal{F}_{\mathbb{R}}^{+}$.*
- *At the positive punctures of S are attached several-level curves in $\mathcal{F}_1^{+}$ in X_1 that solve the Floer equation*

$$(du - B_1)^{0,1} = 0.$$

- *At the negative punctures of S are attached several-level curves in $\mathcal{F}_0^{+}$ in X_1 corresponding to the Floer equation*

$$(du - B_0)^{0,1} = 0.$$

In fact, the curves in $\mathcal{F}_0^{+}$ in X_1 can be canonically identified with the curves in $\mathcal{F}^{+}$ in X_0 that solve the Floer equation with B_0 constructed from the Hamiltonian that equals H_0 on X_0 that continues to grow linearly over the end of X_0. (Note for this identification that all chords and orbits have positive action; compare the definition of the map in Section 5.2.)

Proof The proof is a repetition of the proof of Theorem 4.8, except for the last statement. The last statement follows from Lemma 3.1 which shows that a curve with positive puncture at a chord or orbit in C_{X_0} (notation as in Lemma 2.7) lies inside $\{r \le 1\}$, where $r = e^t$ is the coordinate in the symplectization end of X_0. □

5 Definition of the Hamiltonian simplex DGA

In this section we define the Hamiltonian simplex DGA. In order to simplify grading and dimension questions we assume that $\pi_1(X) = 0$, $c_1(X) = 0$ and that the Maslov class μ_L of the Lagrangian submanifold L vanishes; see Section 7 for a discussion of the general case.

5.1 DGA for fixed Hamiltonian

Let H be a one-step time-independent Hamiltonian, let H_t an associated time-dependent perturbation of it, let B be an $\mathcal{E}$–family of 1–forms associated to H and fix a family of almost complex structures; see Section 2.6.

Define the algebra $\mathcal{SC}^{+}(X, L; H)$ to be the algebra generated by high-energy Hamiltonian chords c of H, graded by

$$|c| = \mathrm{CZ}(c) - 2,$$

and by high-energy 1–periodic orbits γ of H, graded by

$$|\gamma| = \mathrm{CZ}(\gamma) + (n-3).$$

We impose the condition that orbits sign-commute with chords and that orbits sign-commute with orbits. See also Remark A.2.

Define the map

$$\delta\colon \mathcal{SC}^+(X,L;H) \to \mathcal{SC}^+(X,L;H), \quad \delta = \delta_1 + \delta_2 + \cdots + \delta_m + \cdots,$$

to satisfy the graded Leibniz rule and as follows on generators. For a Hamiltonian chord a,

$$\delta_r(a) = \sum_{|a|-|\boldsymbol{b}|-|\boldsymbol{\eta}|=1} \frac{1}{k!}\,|\mathcal{F}^+(a;\boldsymbol{b},\boldsymbol{\eta})|\,\boldsymbol{\eta}\boldsymbol{b},$$

where the sum ranges over all words $\boldsymbol{b} = b_1\cdots b_m$ and $\boldsymbol{\eta} = \eta_1\cdots\eta_k$ which satisfy the grading condition and are such that $m+k=r$. Here $|\mathcal{F}^+|$ denotes the algebraic number of elements in the oriented 0–dimensional manifold $\mathcal{F}^+$. Similarly, for a Hamiltonian orbit γ,

$$\delta_r(\gamma) = \sum_{|\gamma|-|\boldsymbol{\eta}|=1} \frac{1}{r!}\,|\mathcal{F}^+(\gamma;\boldsymbol{\eta})|\,\boldsymbol{\eta},$$

where the sum ranges over all words $\boldsymbol{\eta} = \eta_1\cdots\eta_r$ which satisfy the grading condition.

Remark 5.1 For $r=1$ above, the map δ_1 counts elements in the moduli spaces of strips or cylinders which means that it counts solutions modulo $\mathbb{R}$–translations in the source; see Remark 4.3. In particular δ_1 is exactly the usual differential on the high-energy part of symplectic homology $\mathrm{SH}^+(X,L)$.

Lemma 5.2 *The map $\delta\colon \mathcal{SC}^+(X,L;H) \to \mathcal{SC}^+(X,L;H)$ has degree -1 and is a differential, ie $\delta\circ\delta = 0$.*

Proof This is a consequence of Theorem 4.8: the terms in $\delta\circ\delta$ are in bijective sign-preserving correspondence with the boundary components of the (oriented) 1–dimensional compactified moduli spaces $\mathcal{F}^+$. □

Remark 5.3 Repeating the above constructions replacing the moduli spaces $\mathcal{F}^+$ with $\mathcal{F}^\sigma$ for some generic constant splitting compatible section of $\mathcal{E}$, we get a differential on the DGA $\mathcal{SC}^+(X,L;H)$ with grading shifted up by 1, denoted by $\mathcal{SC}^+(X,L;H)[-1]$. This differential can be homotoped to a differential without higher-degree terms by taking σ sufficiently close to the boundary of the simplex.

5.2 Cobordism maps for fixed Hamiltonians

Consider a symplectic cobordism (X_{10}, L_{10}), and fix a two-step Hamiltonian H_0 and a one-step Hamiltonian H_1. As in Theorem 4.9, we think of H_0 also as a Hamiltonian on X_0 only (basically removing the second step making it a one-step Hamiltonian).

Define the map

$$\Phi\colon \mathcal{SC}^+(X_1, L_1; H_1) \to \mathcal{SC}^+(X_0, L_0; H_0), \quad \Phi = \Phi_1 + \Phi_2 + \cdots, \tag{5-1}$$

as the algebra map given by the following count on generators. (In the right-hand side, we use the identification above and think of (X_1, L_1) with H_0 as being (X_0, L_0) with the corresponding one-step H_0.)

- For chords a,

$$\Phi_r(a) = \sum_{|a|-|\boldsymbol{b}|-|\boldsymbol{\eta}|=0} \frac{1}{k!} \, |\mathcal{F}^+_{\mathbb{R}}(a; \boldsymbol{b}, \boldsymbol{\eta})| \, \boldsymbol{\eta}\boldsymbol{b},$$

  where the sum ranges over all $\boldsymbol{b} = b_1 \cdots b_m$ and $\boldsymbol{\eta} = \eta_1 \cdots \eta_k$ with $m + k = r$.
- For orbits γ,

$$\Phi_r(\gamma) = \sum_{|\gamma|-|\boldsymbol{\eta}|=0} \frac{1}{r!} \, |\mathcal{F}^+_{\mathbb{R}}(\gamma; \boldsymbol{\eta})| \, \boldsymbol{\eta}.$$

Remark 5.4 As a consequence of Lemma 3.1, the target $\mathcal{SC}^+(X_0, L_0; H_0)$ of the cobordism map Φ can be interpreted as the quotient of $\mathcal{SC}(X_1, L_1; H_0)$ by the ideal generated by chords and/or orbits which have negative action. Accordingly, one can factor Φ as a composition

$$\mathcal{SC}^+(X_1, L_1; H_1) \to \mathcal{SC}^+(X_1, L_1; H_0) \to \mathcal{SC}^+(X_0, L_0; H_0),$$

in which the second map is the projection and the first map is defined by the same formulas as Φ in which we replace the moduli spaces $\mathcal{F}^+_{\mathbb{R}}$ by $\mathcal{F}_{\mathbb{R}}$.

Remark 5.5 In the definition of the moduli space $\mathcal{F}^+_{\mathbb{R}}$ in case the curve is a cylinder or a strip we are interpolating between Hamiltonians in a fixed region around $s = 0$ in the cylinder or strip; see Remark 4.3. Thus Φ_1 induces the usual Viterbo transfer map on symplectic or wrapped Floer homology. Indeed, the Viterbo transfer map is just a continuation map.

Theorem 5.6 *The map $\Phi\colon \mathcal{SC}^+(X_1, L_1; H_1) \to \mathcal{SC}^+(X_0, L_0; H_0)$ is a chain map; ie $\delta\Phi = \Phi\delta$.*

Proof By Theorems 4.4 and 4.9, contributions to $\Phi\delta - \delta\Phi$ correspond to the boundary of an oriented 1–dimensional moduli space. □

5.3 The Hamiltonian simplex DGA

In order for the DGA introduced above to capture all aspects of the Reeb dynamics at the boundary of the Liouville pair (X, L) we need to successively increase the slope of the Hamiltonian. Consider a family of one-level Hamiltonians H_a where $H_{a_1} > H_{a_0}$ if $a_1 > a_0$ and $H_a(r, y) = ar + b$ in the cylindrical end $[0, \infty) \times Y$. Inserting a trivial cobordism, we change H_{a_1} to a two-step Hamiltonian H'_{a_1} with the slope a_1 at the end of the trivial cobordism as well. Then $H'_{a_1} > H_{a_0}$. We define the *Hamiltonian simplex DGA*

$$\mathcal{SC}^+(X, L) = \varinjlim_{a\to\infty} \mathcal{SC}^+(X, L; H_a),$$

where the direct limit is taken with respect to the directed system given by the cobordism maps

$$\Phi\colon \mathcal{SC}^+(X, L; H_{a_0}) \to \mathcal{SC}^+(X, L; H'_{a_1}) = \mathcal{SC}^+(X, L; H_{a_1}).$$

See Sections 4.2 and 5.4 for the last equality. Its homology is

$$\mathcal{SH}^+(X, L) = \varinjlim_{a\to\infty} H(\mathcal{SC}^+(X, L; H_a)).$$

Remark 5.7 One can alternatively define the Hamiltonian simplex DGA $\mathcal{SC}^+(X, L)$ as the homotopy limit of the directed system $\{\mathcal{SC}^+(X, L; H_a)\}$, obtained by the algebraic mapping telescope construction as in [4, Section 3g]; see also [28, Chapter 3, page 312].

5.4 Homotopies of cobordism maps

In this subsection, we study invariance properties of the cobordism maps defined in Section 5.2. As a consequence we find that the homotopy type of the Hamiltonian simplex DGA is independent of Hamiltonian, 0–order perturbation term and field of almost complex structure, and depends only on the underlying Liouville pair (X, L).

Let (X_{10}, L_{10}) be a cobordism of pairs and consider a 1–parameter deformation of the data used to define the cobordism map parametrized by $s \in I$. We denote the corresponding cobordism maps by

$$\Phi_s\colon \mathcal{SC}^+(X_1, L_1) \to \mathcal{SC}^+(X_0, L_0), \quad s \in I.$$

Here we take the deformation of the data to be supported in the middle region of the cobordism. In other words the symplectic form, the field of almost complex structures, and the Hamiltonians and associated 0–order terms in the Floer equation vary in the compact region in the cobordism between $0 \times Y_0$ and $0 \times Y_1$ but are left unchanged inside $0 \times Y_0$ and outside $0 \times Y_1$; see Figure 7.

For fixed $s \in I$ we get an interpolation form B^s_τ, $\tau \in I$, and moduli spaces $\mathcal{F}^s_{\mathbb{R}}$, as in Section 4.2. Exactly as there, we suppress from the notation the punctures, and also the constant section σ on which B^s_τ depends. We write the corresponding parametrized moduli spaces as

$$\mathcal{F}^I_{\mathbb{R}} = \bigcup_{s \in I} \mathcal{F}^s_{\mathbb{R}}.$$

We will show below that the chain maps Φ_0 and Φ_1 are chain homotopic. The proof is however rather involved. To explain why we start with a general discussion pointing out the main obstruction to a simple proof. The chain maps Φ_0 and Φ_1 are defined by counting (-1)–disks in $\mathbb{R}$–families of Floer equations, or in other words rigid 0–dimensional curves in $\mathcal{F}^0_{\mathbb{R}}$ and $\mathcal{F}^1_{\mathbb{R}}$, respectively. A standard transversality argument shows that for generic 1–parameter families $s \in I$, the 0–dimensional components of the moduli spaces $\mathcal{F}^I_{\mathbb{R}}$ constitute a transversely cut out oriented 0–manifold. From the point of view of parametrized Floer equations this 0–manifold consists of isolated (-2)–disks, where one parameter is $\tau \in I$ and the other is $s \in I$.

Remark 5.8 In our notation below we always include the simplex parameters in the dimension counts but view both the interpolation parameter $\tau \in I$ and the deformation parameter $s \in I$ as extra parameters. With this convention we call a curve of formal dimension d a (d)*–curve*.

In analogy with the definition of the chain maps induced by cobordisms, counting (-2)–curves during a generic deformation of cobordism data should give a chain homotopy between the chain maps Φ_0 and Φ_1 at the ends of the deformation interval I. However, counting (-2)–curves is not entirely straightforward in the present setup because of the following transversality problem: since the curves considered may have several negative punctures mapping to the same Hamiltonian chord or orbit, an isolated (-2)–curve can be glued to the negative ends of a (d)–curve (asymptotic to Reeb chords or orbits of the Hamiltonian H_1 in X_1), $d > 0$ a number of $(d+1)$ times, resulting in a several-level curve of formal dimension

$$d + (d+1)(-2+1) = -1,$$

on the boundary of the space of (-1)–curves but not accounted for in the chain homotopy equation. In order for the boundary of the space of (-1)–curves to be compatible with the chain homotopy equation, the (-2)–curve should appear only once in combination with the (0)–curve that gives the differentials.

To resolve this problem, we restrict attention to a small time interval around the critical (-2)–curve moment and "time-order" the negative ends of the curves in the moduli space of Floer holomorphic curves in the positive end, ie Floer holomorphic curves in X_1 with respect to the Hamiltonian H_1. Similar arguments are used in eg [18; 4]. In these constructions there are differences between interior and boundary punctures. In case the positive puncture of the (-2)–curve is a chord (boundary puncture) the time ordering argument is simpler since there is a natural order of the boundary punctures in the disks where the (-2)–curve can be attached, and that ordering can be used in building the perturbation scheme. In the orbit case (interior puncture) there is no natural ordering and we are forced to add a homotopy of homotopies argument on top of the ordering perturbation. We sketch these constructions below but point out that actual details do depend on the existence of a suitable perturbation scheme that will not be discussed here; see Remark 1.3.

We now turn to the proof that Φ_0 and Φ_1 are chain homotopic. Consider first the case in which there are no (-2)–curve instances in the interval $[0,1]$. Then the 1–dimensional component of $\mathcal{F}^I_{\mathbb{R}}$ gives an oriented cobordism between the 0–dimensional moduli spaces used to define the cobordism maps and hence $\Phi_0 = \Phi_1$. A general deformation can be perturbed slightly into general position and then it contains only a finite number of transverse (-2)–curve instances. By subdividing the family it is then sufficient to show that Φ_0 and Φ_1 are homotopic for deformation intervals that contain exactly one such transverse (-2)–curve. The following result expresses the effect of a (-2)–curve algebraically. The proof is rather involved and occupies the rest of this section.

Lemma 5.9 *Assume that the deformation interval contains exactly one (-2)–curve. Then the DGA maps Φ_0 and Φ_1 are chain homotopic; ie there exists a degree-$(+1)$ map $K\colon \mathcal{SC}^+(X_1, L_1) \to \mathcal{SC}^+(X_0, L_0)$ such that*

$$\Phi_1 = \Phi_0 e^{(K\circ d_1 - d_0\circ K)}, \tag{5-2}$$

where d_1 and d_0 are the differentials on $\mathcal{SC}^+(X_1, L_1)$ and $\mathcal{SC}^+(X_0, L_0)$, respectively.

Remark 5.10 The exponential in (5-2) is the usual power series of operators.

Remark 5.11 For the chord algebra $\mathcal{SC}^+(L)$, Lemma 5.9 follows from an extended version of [18, Lemma B.15] (that takes orientations of the moduli spaces into account), which is stated in somewhat different terminology. In the proof below, we will adapt the terminology used there to the current setup so as to include (parametrized) orbits as well. Here, it should be mentioned that [18, Lemma B.15], and consequently also the current result, depend on a perturbation scheme for so-called M-polyfolds (the most basic level of polyfolds), the details of which are not yet worked out, and hence it should be viewed as a proof strategy rather than a proof in the strict sense.

We prove Lemma 5.9 in two steps. In the first step we relate Φ_0 and Φ_1 using an abstract perturbation that time orders the negative punctures in all moduli spaces of curves with punctures at chords and orbits in C_{X_1}. In the case that there are only chords there is a natural order of the negative punctures given by the boundary orientation of the disk and in that case the relation between Φ_0 and Φ_1 derived using the natural ordering perturbation can be turned into an algebraic relation. In the case that there are also orbits there is no natural ordering and to derive an algebraic formula we use all possible orderings and study homotopies relating different ordering perturbations.

Consider the first step. We construct a perturbation that orders the negative punctures of any curve in $\mathcal{F}_1^I$ (which is just a product with $\mathcal{F}_1^s \times I$ for any fixed $s \in I$) with negative punctures at chords or orbits in C_{X_1}. We choose this ordering so that when restricted to the boundary punctures of any disk it respects the ordering of the negative punctures induced by the orientation of the boundary of the disk and the positive puncture. We need to carry out this perturbation energy level by energy level. Consider first the lowest action generator γ of H_1 with action bigger than the chord or orbit at the positive puncture of the (-2)–curve. We perturb curves with positive puncture at γ and with negative punctures at generators in C_{X_1} by abstractly perturbing the Floer equation

$$(du - B_1)^{0,1} = 0$$

near the negative punctures. Near chords and orbits in C_{X_1} the data of the Floer equation is independent of both the $\mathbb{R}$–parameter and of $s \in I$. (Recall that the deformations are supported in the compact cobordism.) Thus, if the abstract time ordering perturbation is chosen sufficiently small then there are no (d)–curves for $d < 0$ after perturbation and the moduli space of (d)–curves for $d \geq 0$ after abstract perturbation is canonically isomorphic to the corresponding moduli space before abstract perturbation. Assume that such a perturbation is fixed.

Let $\mathcal{G}(X_1, L_1)$ denote the set of generators of $\mathcal{SC}^+(X_1, L_1)$. For $\gamma \in \mathcal{G}(X_1, L_1)$ we write $d_1^\varepsilon \gamma$ for the sum of monomials that contribute to the differential of γ, ie sum over I–components of the moduli spaces in $\mathcal{F}_1^I$, equipped with the additional structure of ordering of the generators as dictated by ε.

Lemma 5.12 *There is a map $K_\varepsilon \colon \mathcal{G}^+(X_1, L_1) \to \mathcal{SC}^+(X_0, L_0)$ such that for any generator γ (chord or orbit),*

$$\Phi_1(\gamma) - \Phi_0(\gamma) = \Omega_{K_\varepsilon}^\varepsilon(d_1^\varepsilon \gamma) + d_0 \Omega_{K_\varepsilon}^\varepsilon(\gamma). \tag{5-3}$$

Here, $\Omega^{\varepsilon}_{K_\varepsilon}$ acts on monomials with an extra ordering of generators. For a monomial of chords and orbits $\boldsymbol{\beta} = \beta_1 \cdots \beta_k$ we have

$$\Omega^{\varepsilon}_{K_\varepsilon}(\boldsymbol{\beta}) = \sum_{j=1}^{k} (-1)^{\tau_j} \Phi_{\sigma(1,j)}(\beta_1) \cdots \Phi_{\sigma(j-1,j)}(\beta_{j-1}) K_\varepsilon(\beta_j) \Phi_{\sigma(j+1,j)}(\beta_{j+1}) \cdots \Phi_{\sigma(k,j)}(\beta_k),$$

where $\tau_j = |\beta_1| + \cdots + |\beta_{j-1}|$, $\sigma(i,j) = 1$ if β_i is before β_j in the order perturbation ε and $\sigma(i,j) = 0$ if β_i is after β_j.

Proof Consider the parametrized moduli space

$$\mathcal{F}^{I}_{\mathbb{R}}(\gamma; \boldsymbol{\beta})$$

as above. Recall that $\mathcal{SC}^+(X_j, L_j)$ is defined as a direct limit using the action filtration corresponding to increasing slopes of Hamiltonians. We work below a fixed energy level with a fixed slope of our Hamiltonians and assume that the unique (-2)–curve forms a transversely cut-out 0–manifold.

We use the (-2)–curve to construct a chain homotopy. To this end we next extend the ordering perturbation ε to all curves in $\mathcal{F}^{I}_{\mathbb{R}}(\gamma; \boldsymbol{\beta})$. Before we start the actual construction, we point out that our perturbation starts from the very degenerate situation where all negative punctures lie at the same time. Thus one cannot avoid that new (-2)–curves arise when the perturbation is turned on. Gluing these to the perturbed moduli space of curves with negative asymptotes in C_{X_1} then gives new (-1)–curves with positive puncture at γ. We next show how to take these (-1)–curves into account.

We now turn to the description of the perturbation scheme. It is organized energy level by energy level in such a way that the size of the time separation of negative punctures of curves with positive and negative punctures in C_{X_1} is determined by the action of the Reeb chord at the positive puncture. In particular the time distances between positive punctures of the newly created (-2)–curves at a given energy level are of the size of the time separation at this energy level. As we move to the next energy level, the time separation is a magnitude larger, so that the following holds. Consider a curve on the new energy level with a negative puncture q followed in the order by a negative puncture q'. Then q passes all the positive punctures of the (-2)–curves created on lower energy levels before q' enters the region where (-2)–curves exist. Consequently, only one lower energy (-2)–curve at a time can be attached to a curve on the next energy level. As the energy level is given by the action at the positive puncture and the action at any negative puncture is smaller than that at the positive puncture, the energy level E' of any (-2)–curve attached at a negative puncture to a curve at energy level E satisfies $E' < E$. Consequently, there is only one (-2)–curve attached to any curve.

Consider the parametrized 1–dimensional moduli space $\mathcal{F}^{I}_{\mathbb{R}}(\gamma;\boldsymbol{\beta})$ of (-1)–curves defined using the perturbation scheme just described. The boundary of $\mathcal{F}^{I}_{\mathbb{R}}(\gamma;\boldsymbol{\beta})$ then consists of the 0–manifolds $\mathcal{F}^{0}_{\mathbb{R}}(\gamma;\boldsymbol{\beta})$ and $\mathcal{F}^{1}_{\mathbb{R}}(\gamma;\boldsymbol{\beta})$ as well as broken curves that consist of one (-2)–curve at some $s \in I$ and several (-1)–curves with a (0)–curve in the upper or lower end attached. For a generator γ, let $K_{\epsilon}(\gamma)$ denote the count of (-2)–curves *after* the ordering perturbation scheme described above is turned on:

$$K_{\epsilon}(\gamma) = \sum_{|\gamma|-|\boldsymbol{\beta}|=-1} \frac{1}{m(\boldsymbol{\beta})!}\, |\mathcal{F}^{I}_{\mathbb{R}}(\gamma;\boldsymbol{\beta})|\, \boldsymbol{\beta},$$

where $m(\boldsymbol{\beta})$ is the number of orbit generators in the monomial $\boldsymbol{\beta}$. To finish the proof we check that the (-2)–curves in the ordering perturbation scheme accurately accounts for the broken curves at the ends of the 1–dimensional moduli space. By construction, the separation of negative ends increases by a magnitude when we increase the energy level, and only one negative puncture of a curve in $\mathcal{F}^{I}_{1}$ can pass a (-2)–curve moment at a time. At the punctures which are ahead of this puncture with respect to ε, curves in Φ_1 are attached, and at punctures which are behind it, curves in Φ_0 are attached. Thus, counting the boundary points of oriented 1–manifolds we conclude that (5-3) holds. □

Lemma 5.12 expresses Φ_1 in terms of Φ_0 in a way that depends on an ordering of the negative asymptotics, Reeb chords and orbits. On the chord algebra $\mathcal{SC}^{+}(L)$ we use the ordering naturally induced by the orientation of the boundary of the disk, which is also part of the noncommutative structure of the underlying algebra, and the formula in Lemma 5.12 is a chain homotopy of noncommutative DGAs. However, on the orbit part of the algebra there is no naturally induced ordering of the negative asymptotics and the chosen ordering is an additional choice that is not part of the underlying algebraic structure. In order to get an expression with the desired algebraic properties also for the orbit part of $\mathcal{SC}^{+}(X,L)$, we study how the (-2)–curves counted by K_{ε} depend on the choice of ordering perturbation ε. To this end we consider almost ordering perturbations ε_u, $u \in I$ which are time-ordering perturbations of the sort considered above of the negative ends of Floer curves in the positive end of the cobordism. Here an almost ordering is a true ordering except at isolated instances in I when two ends are allowed to cross through with nonzero time derivative. It is clear that any two orderings can be connected through a 1–parameter family of almost orderings. Fix such a 1–parameter family ϵ_u, $u \in I$, of almost orderings that connects orderings ε_0 and ε_1. Let $K_{\varepsilon_0}(\gamma)$ and $K_{\varepsilon_1}(\gamma)$ denote the count of (-2)–curves with positive puncture at γ for the ordering perturbations ε_0 and ε_1, respectively. More precisely, we think of the whole 1–parameter family of moduli spaces associated to the orderings ε_u, $u \in I$, as follows. Recall that in the construction above the counts of (-2)–curves were obtained by following the curves in the positive end with ordered negative punctures through a 1–parameter family that passes the original (-2)–curve moment. Here we consider a 1–parameter family, parametrized by $u \in I$, of such 1–parameter family of curves with negative ends (almost) ordered by ϵ_u passing the (-2)–curve moment. Geometrically we think of this path of paths as corresponding to a unit disk D that interpolates between two paths corresponding to the orderings ε_0 and ε_1. More precisely, the boundary segment in the lower half plane in the boundary of the disk D between -1 and 1 is the path with ordering ε_0, the boundary segment in the upper half plane that with ordering ε_1, and the disk is foliated by the paths interpolating between these two.

Lemma 5.13 *Generically, there is a* 1*–dimensional locus* Γ *in* D *corresponding to* (-2)*–curves with transverse self-intersections and with boundary corresponding to* (-3)*–curves splittings, and at any* (-3)*–curve moment, the path has a definite ordering* (*ie no two negative ends are at the same time coordinate*)*. Furthermore, after deformation of* D*, we may assume that there are no self-intersections of* Γ (*but that the disk still interpolates between the paths* ε_1 *and* ε_0).

Proof The first part of the lemma is a straightforward transversality result. View the ordering paths as paths in larger-dimensional spaces of problems where time coordinates are associated to the negative ends. Choosing these finite-dimensional perturbations generically there is a transversely cut out (-2)–curve hypersurface in the larger spaces. The (-2)–curves in D now correspond to intersections of the (-2)–curve hypersurfaces with D considered as paths of paths in the larger spaces. For generic D this then gives a curve Γ with a natural compactification and with normal crossings. Endpoints of Γ correspond to one (-3)–curve breaking off. Double points of Γ correspond to two (-2)–curves which can be attached at the same disk with negative punctures in C_{X_1}.

We next deform the disk D in order to remove the double points of Γ. This is straightforward: closed components of Γ bound disks in D and can hence be shrunk by isotopy. Intersections of other types can be pushed across the boundary of D. This push results in two new intersections between the (-2)–curve hypersurface and a component of ∂D. These two intersections correspond to two copies of the same (-2)–curve with opposite signs and can be taken to lie arbitrarily close to each other. There is a third (-2)–curve between these two copies. However, by our original choice of abstract ordering perturbations all these three disks have positive puncture at almost the same moment in the 1–parameter family in ∂D. For curves along ∂D with negative ends where these disks can be attached, the time-separation of these negative ends is then larger than the separation between the two (-1)–curves of opposite signs, and hence their contributions cancel. $\square$

Consider the two counts of (-2)–curves K_ε and K_τ corresponding to two ordering perturbations ε and τ. Lemma 5.13 shows that there is a disk D in which the 1–manifold of (-2)–curves is embedded. Furthermore, if there are no (-3)–curves in D the 1–manifold of (-2)–curves gives a cobordism between the (-2)–curves along the boundary arcs and in this case $K_\varepsilon = K_\tau$. Thus, in order to relate in the general case, we only need to study what happens when the ordering path crosses a (-3)–curve moment. Moreover, there is a fixed ordering of negative ends $\varepsilon' = \varepsilon$ or $\varepsilon' = \tau$ mapping to orbits in C_{X_1} at such moments. Our next result expresses this change algebraically.

Lemma 5.14 *In the above setup, there is an operator $K_{\varepsilon\tau}$ such that*

$$K_\varepsilon(\gamma) - K_\tau(\gamma) = \Omega^{\varepsilon'}_{K_{\varepsilon\tau}}(d_1^{\varepsilon'}(\gamma)) + d_0(\Omega^{\varepsilon'}_{K_{\varepsilon\tau}}(\gamma)). \tag{5-4}$$

Proof The difference between $K_\varepsilon(\gamma)$ and $K_\tau(\gamma)$ corresponds to the intersection of D and the codimension-2 variety of (-3)–curves. The corresponding split curves are accounted for by the terms in the right-hand side of (5-4). $\square$

Proof of Lemma 5.9 By Lemma 5.12 we have

$$\Phi_1(\gamma) - \Phi_0(\gamma) = \Omega^\varepsilon_{K_\varepsilon}(d_1^\varepsilon\gamma) + d_0\Omega^\varepsilon_{K_\varepsilon}(\gamma),$$

where ε corresponds to any ordering perturbation. We first show that we can replace K_ε in this formula with K_τ for any ordering perturbation τ. To this end we use Lemma 5.14 which shows that with τ as there and $\varepsilon = \varepsilon'$ (otherwise exchange the roles of τ and ε), we have

$$\begin{aligned}\Omega^\varepsilon_{K_\varepsilon - K_\tau}&(d_1^\varepsilon\gamma) + d_0\Omega^\varepsilon_{K_\varepsilon - K_\tau}(\gamma)\\ &= \Omega^\varepsilon_{K_{\varepsilon\tau}d_1^\varepsilon}(d_1^\varepsilon\gamma) + \Omega^\varepsilon_{d_0K_{\varepsilon\tau}}(d_1^\varepsilon\gamma) + d_0d_0K_{\varepsilon\tau}(\gamma) + d_0\Omega^\varepsilon_{K_{\varepsilon\tau}}(d_1^\varepsilon\gamma).\end{aligned}$$

Here the third term in the right-hand side vanishes. We study the sum of the remaining three terms in the right-hand side.

The operator $\Omega^\varepsilon_{K_{\varepsilon\tau}d_1^\varepsilon}$ acts as follows on monomials $\beta_1\cdots\beta_k$: act by d_1^ε on β_j, attach $K_{\varepsilon\tau}$ at one of the arising negative punctures, and attach Φ_0 at all punctures before this puncture in ε and Φ_1 at all punctures after. The sum

$$d_0\Omega^\varepsilon_{K_{\varepsilon\tau}}(d_1^\varepsilon\gamma) + \Omega^\varepsilon_{d_0K_{\varepsilon\tau}}(d_1^\varepsilon\gamma)$$

counts configurations of the following form: act by d_1^ε on γ, attach $K_{\varepsilon\tau}$ at one of its negative punctures, attach Φ_0 or Φ_1 at all remaining punctures, according to the ordering ε, then act by d_0 at the resulting negative punctures that do not come from $K_{\varepsilon\tau}$. (The terms in which d_0 acts on negative punctures of $K_{\varepsilon\tau}$ are counted twice with opposite signs in the above sum and hence cancel.) Using the chain map property of Φ_j we rewrite this instead as first acting with d_1 on the positive puncture where Φ_j was attached and then attaching Φ_j (and also remove the d_0 at the corresponding negative ends). We thus conclude that we can write the sum of the remaining terms in the right-hand side as follows:

$$\Omega^\varepsilon_{K_{\varepsilon\tau}d_1^\varepsilon}(d_1^\varepsilon\gamma)+\Omega^\varepsilon_{d_0K_{\varepsilon\tau}}(d_1^\varepsilon\gamma)+d_0\Omega^\varepsilon_{K_{\varepsilon\tau}}(d_1^\varepsilon\gamma)=\Omega^\varepsilon_{K_{\varepsilon\tau}}(d_1^\varepsilon d_1^\varepsilon\gamma)=0,$$

where the first term in the left-hand side counts the terms where $K_{\varepsilon\tau}$ is attached at a negative end in the lower-level curve in $d_1^\varepsilon d_1^\varepsilon$ and the sum of the last two counts the terms where it is attached at a negative end in the upper level. To see that $d_1^\varepsilon d_1^\varepsilon\gamma=0$ note that it counts the end points of an oriented compact 1–manifold.

We thus find that

$$\Omega^\varepsilon_{K_\varepsilon}(d_1^\varepsilon\gamma)+d_0\Omega^\varepsilon_{K_\varepsilon}(\gamma)=\Omega^\varepsilon_{K_\tau}(d_1^\varepsilon\gamma)+d_0\Omega^\varepsilon_{K_\tau}(\gamma).$$

Using this formula successively and noticing that if there are no (-3)–curves K does not change over D, we find that

$$\Omega^\varepsilon_{K_\tau}(d_1^\varepsilon\gamma)+d_0\Omega^\varepsilon_{K_\tau}(\gamma)=\Omega^\tau_{K_\tau}(d_1^\tau\gamma)+d_0\Omega^\tau_{K_\tau}(\gamma).$$

Thus for a specific ordering perturbation ε we can move all the Φ_0–factors across and using the splitting repeatedly we express the right-hand side of (5-2) as the sum over all r–level trees, $r\ge 0$. Here r–level trees are defined inductively as follows. A 0–level tree is a Φ_0–curve. A 1–level tree is a curve contributing to d_1^ε with a (-2)–curve attached at one of its negative punctures and Φ_0–curves at all others. An r–level tree is a curve contributing to d_1^ε with a (-2)–curve attached at one of its negative punctures. At punctures after that, trees with $<r$ levels are attached.

By the above we may take the (-2)–curves $K_\varepsilon=K$ to be independent of the ordering perturbation chosen and averaging over all ordering perturbations then gives

$$\Phi_1(\gamma)=\Phi_0\,e^{(Kd_1-d_0K)}(\gamma)$$

by definition of the exponential. □

Corollary 5.15 *The chain maps induced by deformation equivalent cobordisms are chain homotopic.* □

5.5 Composition of cobordism maps

We next study compositions of cobordism maps. Let (X_0, L_0), (X_1, L_1) and (X_2, L_2) be Liouville pairs, and let (X_{01}, L_{01}) and (X_{12}, L_{12}) be two cobordisms between (X_0, L_0) and (X_1, L_1) and between (X_1, L_1) and (X_2, L_2), respectively. We can then glue the cobordisms to form a cobordism (X_{02}, L_{02}) from (X_0, L_0) to (X_2, L_2). This gives three cobordisms maps Φ_{01}, Φ_{12} and Φ_{02}, and we have the following result relating them:

Theorem 5.16 *The chain maps $\Phi_{01} \circ \Phi_{12}$ and Φ_{02} are homotopic.*

Proof The maps Φ_{01}, Φ_{12} and Φ_{02} are induced by interpolations of Hamiltonians H_0 and H_1, H_1 and H_2, and H_0 and H_2, respectively. For the proof we consider these interpolations simultaneously. More precisely consider the Floer moduli space with two interpolation regions, and three Hamiltonians as shown in Figure 8. Recall that our Floer moduli spaces used to define the cobordism map interpolate between two Hamiltonians in a region determined by a level set of the superharmonic function in the domains moving along $\mathbb{R}$; see Section 4. Here we use similar moduli spaces but with two moving interpolation regions, parametrized by $\mathbb{R} \times (0, \infty)$. Here the first coordinate determines the location of the first interpolation region where we interpolate between H_0 and H_1, the second coordinate determines the separation between the levels where we interpolate and near the second level we interpolate between H_1 and H_2. We then note that when the second coordinate is sufficiently large then all Floer solutions are close to broken curves and conversely broken curves can be glued to solutions. Consequently the chain map induced by two interpolation regions that are sufficiently far separated equals the composition $\Phi_{01} \circ \Phi_{12}$. At the other end, where the second coordinate equals 0 we interpolate directly from H_0 to H_2 and we get the chain map Φ_{02}. The results in Section 5.4 imply that the maps are homotopic. $\square$

Corollary 5.17 *The DGA $\mathcal{SC}^+(X, L)$ is invariant under deformations of (X, L) as well as choice of Hamiltonian and almost complex structure.*

Proof Apply the homotopy of chain maps to the obvious deformation that takes the composition of the cobordism induced by a 1–parameter family of deformations of the data and the inverse 1–parameter family to the trivial cobordism. $\square$

6 Isomorphism with contact homology

In this section we prove that the Hamiltonian simplex DGA $\mathcal{SC}^+(X, L)$ is quasi-isomorphic to the (nonequivariant) contact homology DGA $\mathcal{A}(Y, \Lambda)$ of its ideal boundary (Y, Λ). The quasi-isomorphism is obtained using the cobordism map Φ

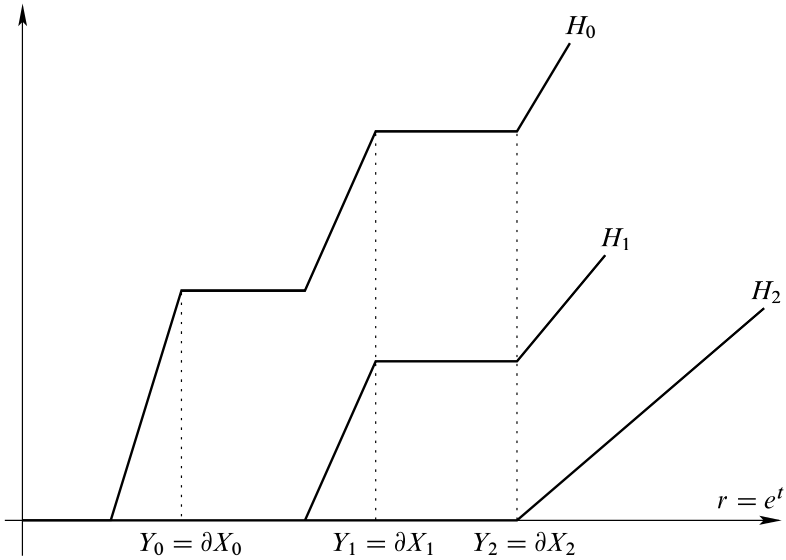


Figure 8: Hamiltonians for composition of cobordism maps

in Section 5.2 for vanishing Hamiltonian $H_1 = 0$. For the versions of contact homology where orbits and chords are not mixed this result implies that $\mathcal{SC}^+(L)$ is quasi-isomorphic to the Legendrian contact homology DGA of Λ, and that $\mathcal{SC}^+(X)$ is isomorphic to the (nonequivariant) contact homology DGA of Y. These results extend the corresponding isomorphisms between the high-energy symplectic homology of X and the nonequivariant linearized contact homology of Y [10], or between the high-energy wrapped Floer homology of L and the linearized Legendrian homology of Λ; see eg [19] and [23, Theorem 7.2].

The nonequivariant orbit contact homology DGA is a natural generalization of the nonequivariant linearized contact homology, but is not described in the literature. We include a short description of the construction in Section 6.1. In Section 6.2, we discuss the better known equivariant case that in our setup corresponds to the Hamiltonian simplex DGA associated to a time-independent Hamiltonian and time-independent almost complex structure J near the punctures. It should be mentioned that the transversality problems for the Floer equation in this setting are similar to the transversality problems for punctured holomorphic spheres in the symplectization end.

6.1 Nonequivariant contact homology orbit DGAs

We give a brief description of nonequivariant contact homology. In essence this is simply a Morse–Bott theory for holomorphic disks and spheres with several negative interior punctures, where each Reeb orbit is viewed as a Morse–Bott manifold. (The chords are treated as usual, so our result for the Legendrian DGA is unaffected by the discussion here.)

Consider the contact manifold Y which is the ideal contact boundary of X. To each Reeb orbit in Y we will associate two decorated Reeb orbits $\hat{\gamma}$ and $\check{\gamma}$; see [10; 7]. The gradings of these decorated orbits are

$$|\check{\gamma}| = \mathrm{CZ}(\gamma) + (n-3) \quad \text{and} \quad |\hat{\gamma}| = \mathrm{CZ}(\gamma) + (n-2).$$

The differential in nonequivariant contact homology counts rigid Morse–Bott curves. These are several-level holomorphic buildings where the asymptotic markers satisfy evaluation conditions with respect to a marked point on each Reeb orbit. Unlike in previous sections we here study curves in the symplectization. However, we still would like to use input from the filling. More precisely, as in [7; 10] we will consider anchored curves. This means that all our curves have additional interior and boundary punctures where rigid holomorphic spheres and rigid holomorphic disks, respectively are attached. We will not mention the anchoring below but keep it implicit.

Recall first that in $D_{1;m,k}$ the positive boundary puncture determines an asymptotic marker at any interior negative puncture and that in $D_{0;0,k}$ any asymptotic marker at the positive puncture determines markers at all negative punctures. If q is a puncture, we write ev_q for the point on the Reeb orbit which is determined by the asymptotic marker. We next define Morse–Bott curves.

Fix a point x on each geometric Reeb orbit. A several-level holomorphic curve with components $S_0, \dots, S_m$ with domain in $\mathcal{D}_{h;hm,k}$ is a *Morse–Bott building* if the following hold:

- If the top-level curve has a positive interior puncture p, then the following hold:
  - If the asymptotic orbit is $\check{\gamma}$, then $\mathrm{ev}_p = x$.
  - If the asymptotic orbit is $\hat{\gamma}$, then ev_p is arbitrary.
- For each component S_j and for each negative interior puncture q of S_j, the following hold:
  - If there is a curve S_m with positive interior puncture at p attached to S_j at q, then the oriented asymptotic Reeb orbit induces the cyclic order $(x, \mathrm{ev}_q, \mathrm{ev}_p)$ on the marked point and the two asymptotic markers.
  - If there is no curve attached at q and the asymptotic orbit is $\hat{\gamma}$, then $\mathrm{ev}_q = x$.
  - If there is no curve attached at q and the asymptotic orbit is $\check{\gamma}$, then ev_q is arbitrary.

Let $\mathcal{A}(Y,\Lambda)$ denote the graded unital algebra generated by the Reeb chords of Λ and decorated Reeb orbits, where as above chords and decorated orbits sign-commute and where decorated orbits sign-commute with each other. The differential on $\mathcal{A}(Y,\Lambda)$ is given by a holomorphic curve count. Using notation analogous to the above we write $\mathcal{M}(a;\boldsymbol{b},\widetilde{\boldsymbol{\eta}})$, $\boldsymbol{b}=b_1\cdots b_m$ and $\widetilde{\boldsymbol{\eta}}=\widetilde{\eta}_1\cdots\widetilde{\eta}_k$, where $\widetilde{\eta}_j$ is a decorated orbit, for the moduli space of anchored Morse–Bott curves $u\colon D_{1;m,k}\to(\mathbb{R}\times Y,\mathbb{R}\times\Lambda)$ with positive boundary puncture where the map is asymptotic at ∞ to the holomorphic Reeb chord strip $\mathbb{R}\times a$, and m negative boundary punctures and k negative interior punctures where the map is asymptotic to the Reeb chord strips $\mathbb{R}\times b_j$ and the Reeb orbit cylinders $\mathbb{R}\times\eta_j$ at $-\infty$. Similarly, we write $\mathcal{M}(\widetilde{\gamma};\widetilde{\boldsymbol{\eta}})$ for the moduli space of anchored Morse–Bott curves $u\colon D_{0;0,k}\to\mathbb{R}\times Y$ with positive interior puncture where the map is asymptotic at ∞ to the holomorphic Reeb orbit cylinder $\mathbb{R}\times\gamma$, and k negative interior punctures where the map is asymptotic to the Reeb orbit cylinders $\mathbb{R}\times\eta_j$ at $-\infty$. Note that in the definition of the moduli spaces of Morse–Bott curves the $\mathbb{R}$–action in the target is divided out at each level of the corresponding buildings. In particular, if a building consists of a single level we divide by the $\mathbb{R}$–action in the target as usual in SFT.

Define the differential d to satisfy the Leibniz rule and to be given as follows on generators: for chords,

$$da=\sum_{|a|-|\boldsymbol{b}|-|\widetilde{\boldsymbol{\eta}}|=1}\frac{1}{k!}\,|\mathcal{M}(a;\boldsymbol{b},\widetilde{\boldsymbol{\eta}})|\,\widetilde{\boldsymbol{\eta}}\boldsymbol{b},$$

and for orbits,

$$d\widetilde{\gamma}=\sum_{|\widetilde{\gamma}|-|\widetilde{\boldsymbol{\eta}}|=1}\frac{1}{k!}\,|\mathcal{M}(\widetilde{\gamma};\widetilde{\boldsymbol{\eta}})|\,\widetilde{\boldsymbol{\eta}}.$$

Here $|\mathcal{M}|$ denotes a sign count of elements of a rigid moduli space with respect to a system of coherent orientations and k is the number of orbits in the monomial $\widetilde{\boldsymbol{\eta}}$. See [11, Section 4.4] for a discussion of orientations for fibered products that is relevant in the case at hand. Then, much like in Lemma 5.2, we have $d^2=0$.

Remark 6.1 Instead of using the Morse–Bott framework above, one can give an alternative definition of the nonequivariant DGA $\mathcal{A}(Y,\Lambda)$ by considering gluing compatible almost complex structures which are time-dependent and periodic in cylindrical end coordinates near interior punctures, ie $J=J_t$, $t\in S^1$. The relevant moduli spaces would then have to be defined in terms of asymptotic incidence conditions determined by a choice of reference point on each periodic Reeb orbit.

We next describe the moduli spaces used to establish the isomorphism between contact homology and Hamiltonian simplex DGAs. The constructions correspond to a version of the construction presented in Section 5.2 where $H_1 = 0$ and $B_1 = 0$, and where the cobordism is replaced by the trivial cobordism, ie the symplectization of Y. As in Section 5.2, we consider a 2–step Hamiltonian H_0 with an associated family B_0 of nonpositive 1–forms with values in Hamiltonian vector fields parametrized by splitting compatible constant sections in $\mathcal{E}$.

The moduli spaces from Section 4.1 which we used in order to define the cobordism map need to be reinterpreted as follows in this context.

First, we define *Morse–Bott buildings with free negative ends* exactly as Morse–Bott buildings, defined above, except that we do not impose any condition on the evaluation maps at the negative interior punctures where no curve is attached.

Second, let a be a Reeb chord, γ a Reeb orbit, $\boldsymbol{b}$ a word of Hamiltonian chords and $\boldsymbol{\eta}$ a word of Hamiltonian orbits. We define the moduli spaces $\underline{\mathcal{F}_{\mathbb{R}}}(a; \boldsymbol{b}, \boldsymbol{\eta})$ and $\underline{\mathcal{F}_{\mathbb{R}}}(\gamma; \boldsymbol{\eta})$ as the moduli spaces $\mathcal{F}_{\mathbb{R}}(a'; \boldsymbol{b}, \boldsymbol{\eta})$ and $\mathcal{F}_{\mathbb{R}}(\gamma'; \boldsymbol{\eta})$ for a' a Hamiltonian chord and γ' a Hamiltonian orbit, in Section 4.1, with the following modifications: any element in $\underline{\mathcal{F}_{\mathbb{R}}}(a; \boldsymbol{b}, \boldsymbol{\eta})$ is asymptotic at the positive puncture at ∞ to the holomorphic Reeb chord strip $\mathbb{R} \times a$, and any element in $\underline{\mathcal{F}_{\mathbb{R}}}(\gamma; \boldsymbol{\eta})$ is asymptotic at the positive puncture at ∞ to the holomorphic Reeb orbit cylinder $\mathbb{R} \times \gamma$. Note that these conditions make sense since the 1–forms B_τ^σ are equal to 0 near the positive puncture. Here $\tau \in (0, 1)$ and $\sigma \in \mathcal{E}$ is a splitting compatible constant section over $\mathcal{D}$. Note also that we do not impose any constraint on the asymptotic marker in the case of an interior positive puncture, this marker is allowed to vary and induces the location of the markers at all negative punctures.

Third, let a be a Reeb chord, $\tilde{\gamma}$ a decorated Reeb orbit, $\boldsymbol{b} = b_1 \cdots b_m$ a word of Hamiltonian chords and $\boldsymbol{\eta} = \eta_1 \cdots \eta_k$ a word of Hamiltonian orbits.

We define the moduli space $\mathcal{F}_{\mathbb{R}}(a; \boldsymbol{b}, \boldsymbol{\eta})$ to consist of a Morse–Bott building with free negative ends whose top-level curve is asymptotic at its positive puncture at ∞ to the holomorphic Reeb chord strip $\mathbb{R} \times a$, together with curves in the moduli spaces defined in the second step above, attached at all its negative punctures. Whenever such a curve with positive puncture p is attached at an interior negative puncture q of the Morse–Bott building with free ends, we require that the common oriented asymptotic Reeb orbit induces the cyclic order $(x, \mathrm{ev}_q, \mathrm{ev}_p)$ on the marked point x and the images of the two asymptotic markers ev_q and ev_p. Finally, we require that the word obtained by reading the boundary negative punctures of the resulting multilevel curve is equal to $\boldsymbol{b}$, and the word determined by the interior negative punctures is equal to $\boldsymbol{\eta}$. We point out that the Morse–Bott building with free negative ends is allowed to be a trivial Reeb chord strip $\mathbb{R} \times a$.

We define, in a similar way, the moduli space $\mathcal{F}_{\mathbb{R}}(\tilde{\gamma};\boldsymbol{\eta})$ to consist of a Morse–Bott building with free negative ends and with positive asymptote at the decorated Reeb orbit $\tilde{\gamma}$, together with curves in the moduli spaces defined in the second step above attached at all its negative punctures. For each such curve with positive puncture p which is attached at a negative puncture q of the Morse–Bott building with free ends, we require again that the common oriented asymptotic Reeb orbit induces the cyclic order $(x, \mathrm{ev}_q, \mathrm{ev}_p)$ on the marked point x and the images of the two asymptotic markers ev_q and ev_p. Finally, we require that the word determined by the negative punctures of the resulting multilevel curve is equal to $\boldsymbol{\eta}$. We point out that the Morse–Bott building with free negative ends is allowed to be a trivial cylinder over the Reeb orbit underlying $\tilde{\gamma}$.

Define the algebra map $\Phi\colon \mathcal{A}(Y,\Lambda) \to \mathcal{SC}^+(X,L;H_0,J_0)$ as follows on generators: for chords,

$$\Phi(a) = \sum_{|a|-|\boldsymbol{b}|-|\boldsymbol{\eta}|=0} \frac{1}{k!}\,|\underline{\mathcal{F}_{\mathbb{R}}}(a;\boldsymbol{b},\boldsymbol{\eta})|\,\boldsymbol{\eta}\boldsymbol{b},$$

and for orbits,

$$\Phi(\tilde{\gamma}) = \sum_{|\tilde{\gamma}|-|\boldsymbol{\eta}|=0} \frac{1}{k!}\,|\underline{\mathcal{F}_{\mathbb{R}}}(\tilde{\gamma};\boldsymbol{\eta})|\,\boldsymbol{\eta}.$$

Passing to the direct limit as the slope of H_0 goes to infinity we obtain an induced map

$$\Phi\colon \mathcal{A}(Y,\Lambda) \to \mathcal{SC}^+(X,L).$$

Remark 6.2 Let the word $\boldsymbol{\eta}$ consist of a single letter η. The moduli spaces $\underline{\mathcal{F}_{\mathbb{R}}}(\tilde{\gamma};\eta)$ then coincide with the moduli spaces giving the isomorphism map between nonequivariant linearized contact homology and symplectic homology in [10, Section 6]. (Note that the latter isomorphism used an intermediate neck-stretching procedure which is unnecessary in our setup since anchored curves appear naturally in the compactification of the relevant 1–dimensional moduli spaces.) Similarly, in case the word $\boldsymbol{b}$ consists of a single letter b and the word $\boldsymbol{\eta}$ is empty, the moduli spaces $\underline{\mathcal{F}_{\mathbb{R}}}(a;b)$ coincide with the moduli spaces giving the isomorphism map between wrapped Floer homology and linearized Legendrian contact homology in [23, Theorem 7.2].

Theorem 6.3 *The induced map $\Phi\colon \mathcal{A}(Y,\Lambda) \to \mathcal{SC}^+(X,L)$ is a chain isomorphism.*

Proof The fact that Φ is a chain map follows as usual by identifying contributing terms of $d\Phi - \Phi d$ with the endpoints of a 1–dimensional moduli space. The isomorphism statement is a consequence of the fact that interpolating strips of Reeb chords and interpolating cylinders of Reeb orbits contribute 1, together with a standard action-filtration argument. Here the interpolating strips and cylinders are simply reparametrizations of trivial strips over Reeb chords and cylinders over Reeb orbits cut off in the r–slice where the corresponding Hamiltonian chord or orbit lies. See [10, Proof of Proposition 4, pages 660–662] for orbits and [23, Theorem 7.2] for chords. □

Remark 6.4 In the proof of Theorem 6.3 the curves which split-off at positive infinity do not have weights since the 1–forms B^{σ}_{τ} are zero near the positive puncture for all values of the parameters σ and τ. Note that Reeb chords and (decorated) orbits have identical gradings in contact homology and in the Hamiltonian simplex DGA (for a small time-independent perturbation of the Hamiltonian as in [10]). The grading shift corresponding to the simplex in the Hamiltonian DGA corresponds on the contact homology side to the Morse–Bott degeneracy in the symplectization direction at a negative puncture.

To see how this works consider (X, L) as above and recall that the (nonequivariant) linearized contact homology of the boundary (Y, Λ) is isomorphic to the high-energy symplectic homology of (X, L) and that the isomorphism is given by a count of rigid holomorphic cylinders and strips along which we interpolate from zero Hamiltonian at the positive end, where the curves are asymptotic to Reeb chords and orbits, to nonzero Hamiltonian at the negative end where the curves are asymptotic to Hamiltonian chords and orbits. Consider now the higher-degree parts (quadratic, cubic, etc.) of the differential in the (nonequivariant) contact homology DGA. We would like to interpret also this part of the differential in terms of symplectic homology, by composing it with the above isomorphism map. Consider thus a curve in the symplectization with one positive and several negative punctures that contributes to the contact homology differential; ie the curve is rigid up to translations. Composing this curve with the isomorphism map corresponds geometrically to gluing an interpolating cylinder or strip at each negative end.

This is a standard gluing problem in SFT, and provided there is one gluing (or translation) parameter at each negative puncture the Floer–Picard lemma applies and the gluing results in curves moving in a unique 1–dimensional moduli space. Note however that the independent gluing parameters at the negative ends give rise to different 1–form-perturbations of the Cauchy–Riemann equation on domains in the same conformal class. The actual 1–form is determined by the values of all gluing parameters but the shift in the symplectization direction identifies forms that differ by an overall translation in the whole domain. The domains with such families of 1–forms are related to the curve with varying weights in the Hamiltonian simplex DGA, sliding the interpolation region at a puncture towards minus infinity corresponds to lowering the weight at that puncture. In this sense the translation degree of freedom at the ends of SFT curves corresponds to the weight degree of freedom in the Hamiltonian simplex DGA.

6.2 Equivariance and autonomous Hamiltonians

In order to relate the usual (equivariant) contact homology $\widetilde{\mathcal{A}}(Y,\Lambda)$ of the ideal contact boundary to a Hamiltonian simplex DGA we can use more or less the same argument as in the nonequivariant case. The starting point here is to set up an equivariant version of the Hamiltonian simplex DGA. To this end we use a time-independent one-step Hamiltonian and define a version $\widetilde{\mathcal{SC}}^+(X,L)$ of the Hamiltonian simplex DGA generated by unparametrized Hamiltonian orbits. To establish transversality for this theory one needs to use abstract perturbations. Assuming that such a perturbation scheme — that also extends to curves in the symplectization with no Hamiltonian — has been fixed, we can repeat the constructions of Section 6.1 word by word to prove:

Theorem 6.5 *The map $\widetilde{\Phi}\colon \widetilde{\mathcal{A}}(Y,\Lambda)\to\widetilde{\mathcal{SC}}^+(X,L)$ is a chain isomorphism.*

Proof Analogous to Theorem 6.3. □

7 Examples and further developments

In this section we first discuss examples where the Hamiltonian simplex DGA is known via the isomorphism to contact homology. Then we discuss how the theory can be generalized to connect Hamiltonian Floer theory to other parts of SFT.

7.1 Knot contact homology

Our first class of examples comes from Legendrian contact homology. By Theorem 6.3, this corresponds to the chord case of our secondary DGA.

Given a knot $K\subset S^3$, one considers its conormal bundle $\nu K\subset X=T^*S^3$. This is an exact Lagrangian that is conical at infinity, that has vanishing Maslov class, and whose wrapped Floer homology $\mathrm{WH}(\nu K)$ was shown in [1] to be equal to the homology of the space $\mathcal{P}_K S^3$ of paths in S^3 with endpoints on K, ie

$$\mathrm{WH}(\nu K)\simeq H(\mathcal{P}_K S^3).$$

One can prove that the homotopy type of the space $\mathcal{P}_K S^3$ does not change as the knot is deformed in a 1–parameter family possibly containing immersions. Since any two knots in S^3 can be connected by a path that consists of embeddings except at a finite number of values of the deformation parameter, where it consists of immersions with a single double point, we infer that $\mathcal{P}_K S^3$ has the same homotopy type as $\mathcal{P}_U S^3$, where $U\subset S^3$ is the unknot. As a matter of fact, the homotopy equivalence can be chosen to be compatible with the evaluation maps at the endpoints, showing that $H(\mathcal{P}_K S^3) \simeq H(\mathcal{P}_U S^3)$ as algebras with respect to the Pontryagin–Chas–Sullivan product. Moreover, we also have isomorphisms $H(\mathcal{P}_K S^3, K) \simeq H(\mathcal{P}_U S^3, U)$ induced by homotopy equivalences. We infer that $\mathrm{WH}(\nu K)$ and its high-energy version $\mathcal{WH}^+(\nu K) \simeq H(\mathcal{P}_K S^3, K)$ are too weak as invariants in order to distinguish knots.

In contrast, for the superficially different case $K \subset \mathbb{R}^3$, the Legendrian contact homology of νK, also called *knot contact homology*, was proved in [20] to coincide with the combinatorial version of [33] and, as such, to detect the unknot. Theorem 6.3 can be extended in a straightforward way to cover the case of $T^*\mathbb{R}^3$ in order to show that Legendrian contact homology of νK is isomorphic to the homology of the Hamiltonian simplex DGA $\mathcal{SC}^+(\nu K)$. In particular, the higher coproducts constituting the differential on $\mathcal{SC}^+(\nu K)$ are rich and interesting operations. This contrasts to the naive higher coproducts defined without varying the weights which are rather trivial. In terms of $\mathcal{P}_K$ the operations of the Hamiltonian simplex DGA correspond to fixing points on the paths with endpoints on K, constraining these points to map to the knot K, and then averaging over the locations of these points. This gives a string topological interpretation of knot contact homology, where chains of strings split as the strings cross the knot as studied in [14].

As a final remark, the coefficient ring of knot contact homology involves a relative second homology group that in the unit cotangent bundle of $\mathbb{R}^3$ contains also the class of the fiber, which is killed in the full cotangent bundle. This extra variable is key to the relation between knot contact homology and the topological string (see [5]) and indicates that it would be important to study the extension of the theory described in the current paper to a situation where the contact data at infinity does not have any symplectic fillings.

7.2 A_∞, L_∞ and the diagonal

As already mentioned in the introduction, the Hamiltonian simplex DGA $\mathcal{SC}^+(X)$ in the orbit case can be viewed as the cobar construction on the vector space generated by the high-energy orbits, viewed as an ∞–Lie coalgebra with the sequence of operations $(d_1, d_2, \dots)$. Note that ∞–Lie coalgebras are dual to L_∞, or ∞–Lie algebras.

In a similar vein, given a Lagrangian $L \subset X$ the Hamiltonian simplex DGA $\mathcal{SC}^+(L)$ in the chord case can be viewed as the cobar construction on the vector space generated by the high-energy chords, viewed as an ∞–coalgebra, a type of structure that is dual to A_∞–algebras.

It turns out that one can produce an ∞–algebra structure in the orbit case by implementing exactly the same construction subject to the additional condition that all punctures lie on a circle on the sphere. This condition is invariant under conformal transformations and yields well-defined moduli spaces, which effectively appear as submanifolds inside the moduli space that define the operations on $\mathcal{SC}^+(X)$. The resulting DGA is not an ∞–Lie coalgebra, but simply an ∞–coalgebra.

Doubling chords to orbits and holomorphic disks to holomorphic spheres with punctures on a circle using Schwarz reflection, it is straightforward to show that the resulting DGA coincides with $\mathcal{SC}^+(\Delta_X)$, the Hamiltonian simplex DGA of the Lagrangian diagonal $\Delta_X \subset X \times X$. This fact parallels the well-known isomorphism between periodic Hamiltonian Floer homology and Lagrangian Floer homology of the diagonal.

This example shows in particular that the relationship between the Hamiltonian simplex DGAs in the closed and in the open case is subtler than its linear counterpart.

7.3 Chern class, Maslov class and exactness

We discuss in this section some of the standing assumptions in the paper.

A first set of assumptions imposed in Section 5 is that $\pi_1(X) = 0$, $c_1(X) = 0$ and $\mu_L = 0$. These are the simplest technical assumptions under which the theory has a unique $\mathbb{Z}$–grading. If $\pi_1(X) = 0$ but $c_1(X)$ or μ_L are nonzero, the closed theory would be uniquely graded modulo the positive generator of $c_1(X) \cdot H_2(X)$, and the open theory would be uniquely graded modulo the positive generator of $\mu_L \cdot H_2(X, L)$. There are also ways to dispose of the condition $\pi_1(X) = 0$ at the expense of possibly further weakening the grading; see the discussion in [25]. In any case, the grading is not unique if $\pi_1(X)$ is nontrivial.

A standing assumption of a quite different and much more fundamental kind is that the manifold X and the Lagrangian L be exact. This is a simple way to rule out, a priori, the bubbling-off of pseudoholomorphic spheres in X, respectively of pseudoholomorphic discs with boundary on L. The advantage of this simple setup is that it allows us to focus on the new algebraic structure. The theory would need to be significantly adapted should one like to consider nonexact situations.

7.4 Dependence on the filling

The Hamiltonian simplex DGA depends, a priori, on the filling, and this is reflected in the definition of the nonequivariant contact homology DGA $\mathcal{A}(Y, \Lambda)$ by the fact that its differential involves curves which are anchored in (X, L).

The Hamiltonian simplex DGA can be defined also in the absence of a filling under the *index-positivity* assumptions explained for example in [17], namely if Y admits a contact form such that every closed Reeb orbit is nondegenerate and has Conley–Zehnder index $\mathrm{CZ}(\gamma)+n-3>1$ and every Reeb chord is nondegenerate and has Conley–Zehnder index $\mathrm{CZ}(c)>1$. The nonequivariant contact homology DGA $\mathcal{A}(Y,\Lambda)$ can be defined in the absence of a filling under the same assumptions.

This is to be contrasted with the definition of the contact homology DGA from SFT [25], which does not require the existence of a filling, though it is subject to the same caveats regarding the existence of an abstract perturbation scheme as explained in Remark 1.3. The reason is that, within the setup of Floer theory, bubbling-off at the negative end of the symplectization always produces curves which satisfy a Cauchy–Riemann equation without zero order perturbation. These objects are external to the framework of our Hamiltonian simplex DGA, whereas they are incorporated in the definition of the differential for the contact homology DGA.

To resolve this discrepancy one needs to clarify the relationship between the nonequivariant Hamiltonian simplex DGA and the contact homology DGA. One direction of study would be to build a *mixed* theory combining the two. Another direction is discussed further below.

7.5 Further developments

At a linear level, S^1–equivariant symplectic homology is obtained from its nonequivariant counterpart using (an ∞–version of) the BV-operator [12]. The BV-operator is an operation governed by the fundamental class of the moduli space of spheres with two punctures and varying asymptotic markers at the punctures. Note that this moduli space is homeomorphic to a circle and the BV-operator has degree $+1$ as a homological operation, which corresponds to the fact that the fundamental class of the moduli space lives in degree 1. It was proved in [12] that the high-energy, or positive part of S^1–equivariant symplectic homology recovers linearized cylindrical contact homology of the contact boundary Y. One advantage of the S^1–equivariant point of view over the symplectic field theory (SFT) point of view is that it does not require any abstract perturbation theory.

Question What is the additional structure on the nonequivariant Hamiltonian simplex DGA $\mathcal{SC}^+(X)$ that allows to recover the equivariant Hamiltonian simplex DGA described in Section 6.2?

Though one can construct an ∞–version of the BV-operator in the DGA setting that we consider in this paper by methods similar to those of [12], it is not clear whether this is enough in order to recover the equivariant DGA from the nonequivariant one. It may be that one needs more information coming from the structure of an algebra over the operad of framed little 2–disks that exists on any Hamiltonian Floer theory.

From the point of view of SFT, the natural next steps are to understand the algebraic structure that is determined on Hamiltonian Floer theory by moduli spaces of genus-0 curves with an arbitrary number of positive punctures, respectively by moduli spaces of curves with an arbitrary number of positive punctures and arbitrary genus. These would provide in particular nonequivariant analogues of the rational SFT and full SFT.

Appendix: Determinant bundles and signs

In this appendix, we give a more detailed discussion of how the sign rules of the Hamiltonian simplex DGA derive from orientations of determinant bundles. The material here has been discussed at many places in this context; see for example [40, Section 11; 32, Appendix A.2; 9; 43; 26; 21].

If V is a finite-dimensional vector space, then $\bigwedge^{\max} V = \bigwedge^{\dim V} V$ is its highest exterior power. For the 0–dimensional vector space, $\bigwedge^{\max}(0) = \mathbb{R}$. If

$$0 \to V_1 \xrightarrow{f_1} V_2 \xrightarrow{f_2} \cdots \xrightarrow{f_n} V_{n+1} \to 0$$

is an exact sequence of finite-dimensional vector spaces, then there is a canonical isomorphism

$$\bigotimes_{k \text{ odd}} \textstyle\bigwedge^{\max} V_k \;\cong\; \bigotimes_{k \text{ even}} \bigwedge^{\max} V_k$$

that *does* depend on the maps $f_1, \dots, f_n$. For example, if $\dim V_1$ is odd and the map f_1 is changed to $-f_1$, then the isomorphism changes sign.

If X and Y are Banach spaces and $D\colon X \to Y$ is a Fredholm operator, then the *determinant line* $\det(D)$ *of* D is the 1–dimensional vector space

$$\det(D) = \textstyle\bigwedge^{\max}(\operatorname{coker} D)^* \otimes \bigwedge^{\max} \ker D.$$

We think of $\det(D)$ as a graded vector space supported in degree $\operatorname{index}(D)$.

We next discuss stabilization. We first stabilize in the source. Let $D\colon X \to Y$ be a Fredholm operator, V a finite-dimensional real vector space and $\Phi\colon V \to Y$ a linear map. The *stabilization of* D *by* Φ is the Fredholm operator $D^V = D \oplus \Phi\colon X \times V \to Y$, $(x, v) \mapsto Dx + \Phi v$. The exact sequence

$$0 \to \ker D \to \ker D^V \to V \xrightarrow{\Phi} \operatorname{coker} D \to \operatorname{coker} D^V \to 0$$

gives a canonical isomorphism that depends on the map Φ:

$$\det(D^V) \cong \det D \otimes \textstyle\bigwedge^{\max} V.$$

For example, if the map Φ changes sign and if $\dim \operatorname{coker} D - \dim \operatorname{coker} D^V$ is odd, then the isomorphism changes sign.

Similarly we can stabilize in the target. If W is a finite-dimensional vector space and $\Psi\colon X \to W$ is a continuous linear map, then with $D_W = (D, \Psi)\colon X \to Y \times W$, $x \mapsto (Dx, \Psi x)$, we get

$$0 \to \ker D_W \to \ker D \xrightarrow{\Psi} W \to \operatorname{coker} D_W \to \operatorname{coker} D \to 0,$$

which gives a canonical isomorphism that depends on Ψ:

$$\det(D_W) \cong \big(\textstyle\bigwedge^{\max} W\big)^* \otimes \det D.$$

For example, if the map Ψ changes sign and $\dim \ker D - \dim \ker D_W$ is odd, then the isomorphism changes sign.

Finally, combining the two, if $\alpha\colon V \to W$ is a linear map, then the map

$$D^V_W\colon X \times V \to Y \times W, \quad D^V_W(x, v) = (Dx + \Phi v, \Psi x + \alpha v)$$

gives a canonical isomorphism that depends on Φ, Ψ and α:

$$\det(D^V_W) \cong \big(\textstyle\bigwedge^{\max} W\big)^* \otimes \det D \otimes \textstyle\bigwedge^{\max} V. \tag{A-1}$$

Remark A.1 For the isomorphism above, one also needs to specify conventions for orientations of direct sums corresponding to stabilizations. The details of these conventions, however, do not affect our discussion here.

If $D \in \mathcal{F}(X, Y)$, then by stabilizing in the domain, one may make all operators in a neighborhood of D surjective and that together with the above isomorphism allows for the definition of a locally trivial line bundle $\underline{\det} \to \mathcal{F}(X, Y)$ over the space of Fredholm operators acting from X to Y with fiber over D equal to $\det(D)$.

Assume $\mathcal{D}\colon \mathcal{O} \to \mathcal{F}(X, Y)$ is a continuous map defined on some topological space $\mathcal{O}$. Consider the pull-back bundle $\mathcal{D}^*\underline{\det} \to \mathcal{O}$ and note that it admits a trivialization provided the first Stiefel–Whitney class vanishes, $w_1(\mathcal{D}^*\underline{\det}) = 0$.

If V and W are finite-dimensional vector spaces, we consider in line with the discussion above the bundle $\mathcal{O}^V_W = \mathcal{O} \times \operatorname{Hom}(V, Y) \times \operatorname{Hom}(X, W) \times \operatorname{Hom}(V, W)$ and the map $\mathcal{D}^V_W\colon \mathcal{O}^V_W \to \mathcal{F}(X \times V, Y \times W)$ defined as follows: $\mathcal{D}^V_W(p, \Phi, \Psi, \alpha)$ is the linear map which takes $(x, v) \in X \times V$ to

$$(\mathcal{D}(p)x + \Phi v,\ \Psi x + \alpha v) \in X \times W.$$

Using the natural base point in $\mathrm{Hom}(V,Y)\times\mathrm{Hom}(X,W)\times\mathrm{Hom}(V,W)$ given by $(\Phi,\Psi,\alpha)=0$ and the natural isomorphism (A-1), we transport orientations of $\bigwedge^{\max}\underline{W}^*\otimes\mathcal{D}^*\underline{\det}\otimes\bigwedge^{\max}\underline{V}\to\mathcal{O}$ to orientations of $(\mathcal{D}^V_W)^*\underline{\det}\to\mathcal{O}^V_W$, and back. Here $\underline{V}\to\mathcal{O}$ and $\underline{W}\to\mathcal{O}$ denote the trivial bundles $\mathcal{O}\times V\to\mathcal{O}$ and $\mathcal{O}\times W\to\mathcal{O}$ respectively.

We now apply this setup to spaces of (stabilized) Cauchy–Riemann operators used in the definition of the Hamiltonian simplex DGA. Indeed, the linearized operator for our family of Cauchy–Riemann equations parametrized by the simplex Δ^{m-1} is of the type $D^{T_w\Delta^{m-1}}$. Since $T_w\Delta^{m-1}=\ker\ell_\nu$ with $\ell_\nu\colon\mathbb{R}^m\to\mathbb{R}$, $\ell_\nu(\zeta)=\langle\zeta,\nu\rangle$, where ν is the vector $\nu=(1,1,\dots,1)$, we have a canonical isomorphism $\det(D^{T_w\Delta^{m-1}})\simeq\det(D^{\mathbb{R}^m}_{\mathbb{R}})$, with $D^{\mathbb{R}^m}_{\mathbb{R}}(x,\zeta)=\big(D^{T_w\Delta^{m-1}}(x,\pi\zeta),\ell_\nu(\zeta)\big)$ and $\pi\colon\mathbb{R}^m\to T_w\Delta^{m-1}$ the orthogonal projection parallel to ν. We can thus view the linearization of our parametrized Cauchy–Riemann problem as an element of a suitable space $\mathcal{O}^{\mathbb{R}^m}_{\mathbb{R}}$ of Fredholm operators of Cauchy–Riemann type.

The negative orbit and chord capping operators $o_-(\gamma)$ and $o_-(c)$ belong to natural spaces $\mathcal{O}_-(\gamma)$ and $\mathcal{O}_-(c)$ of Cauchy–Riemann operators with fixed asymptotic behavior determined by the linearized Hamiltonian flow along γ and respectively c, acting between appropriate Sobolev spaces of sections $W^{1,p}\to L^p$, $p>2$ (see [11, Section 4.4] for the orbit case and [40, Section 11; 4, Section 9] for the chord case). These spaces of Cauchy–Riemann operators with fixed asymptotes are contractible, and consequently the determinant line bundle can be trivialized over each of them. We similarly define natural spaces of Cauchy–Riemann operators $\mathcal{O}_+(\gamma)$ and $\mathcal{O}_-(c)$ containing the positive orbit and chord capping operators $o_+(\gamma)$ and $o_+(c)$.

Our procedure for the construction of coherent orientations for the parametrized Cauchy–Riemann equation is then the following:

(i) Given the canonical orientation on $\mathbb{C}$, we orient the determinant bundles over the spaces $\mathcal{O}(\mathbb{C}P^1)$ of Cauchy–Riemann operators over $\mathbb{C}P^1$ by the canonical orientation of complex linear operators. Since all the Euclidean spaces $\mathbb{R}^n$ are canonically oriented, this induces orientations of the determinant bundles over all spaces $\mathcal{O}^{\mathbb{R}^k}_{\mathbb{R}^\ell}(\mathbb{C}P^1)$ for arbitrary $k,\ell\in\mathbb{Z}_{\geq 0}$.

Similarly, following [26], the choice of a relative spin structure on the Lagrangian L determines an orientation of the determinant bundle over all spaces of Cauchy–Riemann operators $\mathcal{O}(D^2)$ defined on the pull-back of TX over the disk D^2 by arbitrary smooth maps $u\colon(D^2,\partial D^2)\to(X,L)$, with totally real boundary conditions given by $u|^*_{\partial D^2}TL$. This then induces orientations of the determinant bundles over all spaces $\mathcal{O}^{\mathbb{R}^k}_{\mathbb{R}^\ell}(D^2)$ for arbitrary $k,\ell\in\mathbb{Z}_{\geq 0}$.

(ii) We choose orientations of the determinant lines over the spaces $\mathcal{O}_-(\gamma)$ and $\mathcal{O}_-(c)$, which determine in turn orientations of the determinant lines over all spaces $\mathcal{O}_-{}^{\mathbb{R}}_0(\gamma)$ and $\mathcal{O}_-{}^{\mathbb{R}}_0(c)$.

(iii) By gluing, we obtain orientations of the determinant lines over the spaces $\mathcal{O}_+(\gamma)$ and $\mathcal{O}_+(c)$, which determine in turn orientations of the determinant lines over all spaces $\mathcal{O}_+{}^{\mathbb{R}}_0(\gamma)$ and $\mathcal{O}_+{}^{\mathbb{R}}_0(c)$.

(iv) If $\boldsymbol{b} = b_1 \cdots b_m$ is a word of Hamiltonian chords and $\boldsymbol{\eta} = \eta_1 \cdots \eta_k$ a word of Hamiltonian orbits, then we write

$$\mathcal{O}_+{}^{\mathbb{R}}_0(\boldsymbol{b}, \boldsymbol{\eta}) = \mathcal{O}_+{}^{\mathbb{R}}_0(b_1) \times \cdots \times \mathcal{O}_+{}^{\mathbb{R}}_0(b_m) \times \mathcal{O}_+{}^{\mathbb{R}}_0(\eta_1) \times \cdots \times \mathcal{O}_+{}^{\mathbb{R}}_0(\eta_k)$$

and

$$\mathcal{O}_+{}^{\mathbb{R}}_0(\boldsymbol{\eta}) = \mathcal{O}_+{}^{\mathbb{R}}_0(\eta_1) \times \cdots \times \mathcal{O}_+{}^{\mathbb{R}}_0(\eta_k).$$

Given a Hamiltonian chord a, we write $\mathcal{O}(a; \boldsymbol{b}, \boldsymbol{\eta})$ for the space of Cauchy–Riemann operators defined on a punctured disc with one positive boundary puncture, m negative boundary punctures, and k negative interior punctures, with Lagrangian boundary conditions given by the pull-back of TL via a map on the disk into X with boundary in L, and with asymptotic behavior at the punctures according to the Hamiltonian chords and orbits a, $\boldsymbol{b}$ and $\boldsymbol{\eta}$. Similarly, given a Hamiltonian orbit γ we write $\mathcal{O}(\gamma; \boldsymbol{\eta})$ for the space of Cauchy–Riemann operators defined on a sphere with one positive puncture and k negative punctures, and with asymptotic behavior at the punctures determined by the linearized flow along the Hamiltonian orbits γ, $\boldsymbol{\eta}$. We then have spaces $\mathcal{O}_0^{\mathbb{R}^{m+k+1}}(a; \boldsymbol{b}, \boldsymbol{\eta})$ and $\mathcal{O}_0^{\mathbb{R}^{k+1}}(\gamma; \boldsymbol{\eta})$, and $\mathcal{O}_{\mathbb{R}}^{\mathbb{R}^{m+k}}(a; \boldsymbol{b}, \boldsymbol{\eta})$ and $\mathcal{O}_{\mathbb{R}}^{\mathbb{R}^{k}}(\gamma; \boldsymbol{\eta})$.

(v) Cauchy–Riemann operators which are stabilized by finite-dimensional spaces at the source can be glued much like usual, ie nonstabilized, Cauchy–Riemann operators; see eg [24, Section 4.3]. The gluing operations

$$\mathcal{O}_-{}^{\mathbb{R}}_0(a) \times \mathcal{O}_0^{\mathbb{R}^{m+k+1}}(a; \boldsymbol{b}, \boldsymbol{\eta}) \times \mathcal{O}_+{}^{\mathbb{R}}_0(\boldsymbol{b}, \boldsymbol{\eta}) \to \mathcal{O}_0^{\mathbb{R}^{2(m+k+1)}}(D^2)$$

and

$$\mathcal{O}_-{}^{\mathbb{R}}_0(\gamma) \times \mathcal{O}_0^{\mathbb{R}^{k+1}}(\gamma; \boldsymbol{\eta}) \times \mathcal{O}_+{}^{\mathbb{R}}_0(\boldsymbol{\eta}) \to \mathcal{O}_0^{\mathbb{R}^{2(k+1)}}(\mathbb{C}P^1)$$

induce isomorphisms of determinant bundles which are canonical up to homotopy. From our previous choices we obtain orientations of all the spaces $\mathcal{O}_0^{\mathbb{R}^{m+k+1}}(a; \boldsymbol{b}, \boldsymbol{\eta})$ and $\mathcal{O}_0^{\mathbb{R}^{k+1}}(\gamma; \boldsymbol{\eta})$. After restricting to the slice given by the zero stabilization map, we obtain as explained above orientations of all the spaces $\mathcal{O}_{\mathbb{R}}^{\mathbb{R}^{m+k}}(a; \boldsymbol{b}, \boldsymbol{\eta})$ and $\mathcal{O}_{\mathbb{R}}^{\mathbb{R}^{k}}(\gamma; \boldsymbol{\eta})$. These orientations are used in order to count rigid holomorphic curves with signs in the relevant moduli spaces.

Our choice of coherent orientations gives the following graded commutativity property. As in Section 4.1, let $|c| = \mathrm{CZ}(c) - 2$ and $|\gamma| = \mathrm{CZ}(\gamma) + n - 3$. Let $\boldsymbol{b} = b_1 \cdots b_m$

and $\boldsymbol{\eta} = \eta_1 \cdots \eta_k$ be words in Hamiltonian chords and orbits respectively as above. Consider spaces of stabilized Cauchy–Riemann operators $\mathcal{O}_0^{\mathbb{R}^{m+k+1}}(a; \boldsymbol{b}, \boldsymbol{\eta})$ and $\mathcal{O}_0^{\mathbb{R}^{k+1}}(\gamma; \boldsymbol{\eta})$ for Hamiltonian chords and orbits a and γ. Given $1 \le i \le k-1$, let $\boldsymbol{\eta}^i = \eta_1 \cdots \eta_{i-1}\eta_{i+1}\eta_i\eta_{i+2}\cdots\eta_k$. There are canonical identifications

$$\mathcal{O}_0^{\mathbb{R}^{m+k+1}}(a; \boldsymbol{b}, \boldsymbol{\eta}) \cong \mathcal{O}_0^{\mathbb{R}^{m+k+1}}(a; \boldsymbol{b}, \boldsymbol{\eta}^i) \quad \text{and} \quad \mathcal{O}_0^{\mathbb{R}^{k+1}}(\gamma; \boldsymbol{\eta}) \cong \mathcal{O}_0^{\mathbb{R}^{k+1}}(\gamma; \boldsymbol{\eta}^i)$$

obtained by relabeling the i^{th} and $(i+1)^{\text{st}}$ interior punctures of the domain. Accordingly, the determinant line bundles over these spaces of operators are canonically identified. Each of them comes with an induced orientation as above, and these orientations differ by the sign

$$(-1)^{|\eta_i||\eta_{i+1}|}.$$

Indeed, these orientations differ by the same sign as the orientations of the determinant lines over $\mathcal{O}+_0^{\mathbb{R}}(\gamma_i) \times \mathcal{O}_0^{\mathbb{R}}(\gamma_{i+1})$ and $\mathcal{O}+_0^{\mathbb{R}}(\gamma_{i+1}) \times \mathcal{O}_0^{\mathbb{R}}(\gamma_i)$, identified via the obvious exchange of factors. By [40, page 150], the latter sign is equal to

$$(-1)^{\operatorname{index}(D^{\mathbb{R}}_{0\ i}) \times \operatorname{index}(D^{\mathbb{R}}_{0\ i+1})} = (-1)^{|\eta_i||\eta_{i+1}|},$$

where $D^{\mathbb{R}}_{0\ i} \in \mathcal{O}+_0^{\mathbb{R}}(\gamma_i)$ and $D^{\mathbb{R}}_{0\ i+1} \in \mathcal{O}+_0^{\mathbb{R}}(\gamma_{i+1})$. This holds because

$$\operatorname{index}(D^{\mathbb{R}}_{0\ i}) = \mathrm{CZ}(\eta_i) + n + 1 \equiv |\eta_i| \pmod 2$$

and

$$\operatorname{index}(D^{\mathbb{R}}_{0\ i+1}) = \mathrm{CZ}(\eta_{i+1}) + n + 1 \equiv |\eta_{i+1}| \pmod 2;$$

see Section 4.1.

This shows that orbits sign-commute in the Hamiltonian simplex DGA of Section 5.

Remark A.2 In the Hamiltonian simplex DGA of Section 5 orbits sign-commute with chords. That is *not* a consequence of coherent orientations. It is just an algebraic choice that reflects the interpretation of orbits as coefficients for the algebra generated by chord generators. Indeed, we can always order the negative punctures of a holomorphic curve by first considering boundary punctures and then considering interior punctures (analogous to normal ordering of operators).

References

[1] **A Abbondandolo**, **A Portaluri**, **M Schwarz**, *The homology of path spaces and Floer homology with conormal boundary conditions*, J. Fixed Point Theory Appl. 4 (2008) 263–293 MR

[2] **A Abbondandolo**, **M Schwarz**, *On the Floer homology of cotangent bundles*, Comm. Pure Appl. Math. 59 (2006) 254–316 MR

[3] **M Abouzaid**, *Symplectic cohomology and Viterbo's theorem*, from "Free loop spaces in geometry and topology" (J Latschev, A Oancea, editors), IRMA Lect. Math. Theor. Phys. 24, Eur. Math. Soc., Zürich (2015) 271–485 MR

[4] **M Abouzaid**, **P Seidel**, *An open string analogue of Viterbo functoriality*, Geom. Topol. 14 (2010) 627–718 MR

[5] **M Aganagic**, **T Ekholm**, **L Ng**, **C Vafa**, *Topological strings, D–model, and knot contact homology*, Adv. Theor. Math. Phys. 18 (2014) 827–956 MR

[6] **F Bourgeois**, *Contact homology and homotopy groups of the space of contact structures*, Math. Res. Lett. 13 (2006) 71–85 MR

[7] **F Bourgeois**, **T Ekholm**, **Y Eliashberg**, *Effect of Legendrian surgery*, Geom. Topol. 16 (2012) 301–389 MR

[8] **F Bourgeois**, **Y Eliashberg**, **H Hofer**, **K Wysocki**, **E Zehnder**, *Compactness results in symplectic field theory*, Geom. Topol. 7 (2003) 799–888 MR

[9] **F Bourgeois**, **K Mohnke**, *Coherent orientations in symplectic field theory*, Math. Z. 248 (2004) 123–146 MR

[10] **F Bourgeois**, **A Oancea**, *An exact sequence for contact- and symplectic homology*, Invent. Math. 175 (2009) 611–680 MR

[11] **F Bourgeois**, **A Oancea**, *Symplectic homology, autonomous Hamiltonians, and Morse–Bott moduli spaces*, Duke Math. J. 146 (2009) 71–174 MR

[12] **F Bourgeois**, **A Oancea**, *S^1–equivariant symplectic homology and linearized contact homology*, Int. Math. Res. Not. (online publication June 2016) 89 pages

[13] **K Cieliebak**, **T Ekholm**, **J Latschev**, *Compactness for holomorphic curves with switching Lagrangian boundary conditions*, J. Symplectic Geom. 8 (2010) 267–298 MR

[14] **K Cieliebak**, **T Ekholm**, **J Latschev**, **L Ng**, *Knot contact homology, string topology, and the cord algebra*, preprint (2016) arXiv

[15] **K Cieliebak**, **Y Eliashberg**, *From Stein to Weinstein and back: Symplectic geometry of affine complex manifolds*, Amer. Math. Soc. Colloq. Pub. 59, Amer. Math. Soc., Providence, RI (2012) MR

[16] **K Cieliebak**, **A Floer**, **H Hofer**, **K Wysocki**, *Applications of symplectic homology, II: Stability of the action spectrum*, Math. Z. 223 (1996) 27–45 MR

[17] **K Cieliebak**, **A Oancea**, *Symplectic homology and the Eilenberg–Steenrod axioms*, preprint (2015) arXiv

[18] **T Ekholm**, *Rational symplectic field theory over $\mathbb{Z}_2$ for exact Lagrangian cobordisms*, J. Eur. Math. Soc. 10 (2008) 641–704 MR

[19] **T Ekholm**, *Rational SFT, linearized Legendrian contact homology, and Lagrangian Floer cohomology*, from "Perspectives in analysis, geometry, and topology" (I Itenberg, B Jöricke, M Passare, editors), Progr. Math. 296, Springer (2012) 109–145 MR

[20] **T Ekholm**, **J B Etnyre**, **L Ng**, **M G Sullivan**, *Knot contact homology*, Geom. Topol. 17 (2013) 975–1112 MR

[21] **T Ekholm**, **J Etnyre**, **M Sullivan**, *Orientations in Legendrian contact homology and exact Lagrangian immersions*, Internat. J. Math. 16 (2005) 453–532 MR

[22] **T Ekholm**, **J Etnyre**, **M Sullivan**, *Legendrian contact homology in $P \times \mathbb{R}$*, Trans. Amer. Math. Soc. 359 (2007) 3301–3335 MR

[23] **T Ekholm**, **K Honda**, **T Kálmán**, *Legendrian knots and exact Lagrangian cobordisms*, J. Eur. Math. Soc. 18 (2016) 2627–2689 MR

[24] **T Ekholm**, **I Smith**, *Exact Lagrangian immersions with one double point revisited*, Math. Ann. 358 (2014) 195–240 MR

[25] **Y Eliashberg**, **A Givental**, **H Hofer**, *Introduction to symplectic field theory*, from "Visions in mathematics: GAFA 2000" (N Alon, J Bourgain, A Connes, M Gromov, V Milman, editors), Geom. Funct. Anal. Special Volume, II (2000) 560–673 MR

[26] **K Fukaya**, **Y-G Oh**, **H Ohta**, **K Ono**, *Lagrangian intersection Floer theory: anomaly and obstruction*, AMS/IP Studies in Advanced Mathematics 46, Amer. Math. Soc., Providence, RI (2009) MR

[27] **M Goresky**, **N Hingston**, *Loop products and closed geodesics*, Duke Math. J. 150 (2009) 117–209 MR

[28] **A Hatcher**, *Algebraic topology*, Cambridge Univ. Press (2002) MR

[29] **H Hofer**, **K Wysocki**, **E Zehnder**, *Sc-smoothness, retractions and new models for smooth spaces*, Discrete Contin. Dyn. Syst. 28 (2010) 665–788 MR

[30] **H Hofer**, **K Wysocki**, **E Zehnder**, *Polyfold and Fredholm theory, I: Basic theory in M–polyfolds*, preprint (2014) arXiv

[31] **T Kimura**, **J Stasheff**, **A A Voronov**, *On operad structures of moduli spaces and string theory*, Comm. Math. Phys. 171 (1995) 1–25 MR

[32] **D McDuff**, **D Salamon**, *J–holomorphic curves and symplectic topology*, Amer. Math. Soc. Colloq. Pub. 52, Amer. Math. Soc., Providence, RI (2004) MR

[33] **L Ng**, *Framed knot contact homology*, Duke Math. J. 141 (2008) 365–406 MR

[34] **J Pardon**, *Contact homology and virtual fundamental cycles*, preprint (2015) arXiv

[35] **J Pardon**, *An algebraic approach to virtual fundamental cycles on moduli spaces of pseudo-holomorphic curves*, Geom. Topol. 20 (2016) 779–1034 MR

[36] **A F Ritter**, *Topological quantum field theory structure on symplectic cohomology*, J. Topol. 6 (2013) 391–489 MR

[37] **J Robbin**, **D Salamon**, *The Maslov index for paths*, Topology 32 (1993) 827–844 MR

[38] **D A Salamon**, **J Weber**, *Floer homology and the heat flow*, Geom. Funct. Anal. 16 (2006) 1050–1138 MR

[39] **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 Press, Somerville, MA (2008) 211–253 MR

[40] **P Seidel**, *Fukaya categories and Picard–Lefschetz theory*, Eur. Math. Soc., Zürich (2008) MR

[41] **C Viterbo**, *Functors and computations in Floer homology with applications, II*, preprint (1996) Updated version available at http://tinyurl.com/viterbo-prepubs

[42] **C Viterbo**, *Functors and computations in Floer homology with applications, I*, Geom. Funct. Anal. 9 (1999) 985–1033 MR

[43] **A Zinger**, *The determinant line bundle for Fredholm operators: construction, properties, and classification*, Math. Scand. 118 (2016) 203–268 MR

Department of Mathematics, University of Uppsala
Box 480, SE-751 06 Uppsala, Sweden

Sorbonne Universités, UPMC Univ. Paris 06
UMR 7586, Institut de Mathématiques de Jussieu-Paris Rive Gauche
Case 247, 4 place Jussieu, 75005 Paris, France

tobias.ekholm@math.uu.se, alexandru.oancea@imj-prg.fr

Proposed: Yasha Eliashberg Received: 20 August 2015
Seconded: András I. Stipsicz, Ciprian Manolescu Revised: 16 June 2016

msp

Geometry & Topology 21 (2017) 2231–2241

The stable cohomology of the Satake compactification of $\mathcal{A}_g$

JIAMING CHEN
EDUARD LOOIJENGA

Charney and Lee have shown that the rational cohomology of the Satake–Baily–Borel compactification $\mathcal{A}_g^{\mathrm{bb}}$ of $\mathcal{A}_g$ stabilizes as $g \to \infty$ and they computed this stable cohomology as a Hopf algebra. We give a relatively simple algebrogeometric proof of their theorem and show that this stable cohomology comes with a mixed Hodge structure of which we determine the Hodge numbers. We find that the mixed Hodge structure on the primitive cohomology in degrees $4r+2$ with $r \geq 1$ is an extension of $\mathbb{Q}(-2r-1)$ by $\mathbb{Q}(0)$; in particular, it is not pure.

14G35, 32S35

1 The theorem

Let $\mathcal{A}_g = \mathcal{A}_g(\mathbb{C})$ denote the coarse moduli space of principally polarized complex abelian varieties of genus g endowed with the analytic (Hausdorff) topology. Recall that the Satake–Baily–Borel compactification $j_g\colon \mathcal{A}_g \subset \mathcal{A}_g^{\mathrm{bb}}$ realizes $\mathcal{A}_g$ as a Zariski open dense subset in a normal projective variety $\mathcal{A}_g^{\mathrm{bb}}$. Forming the product of two principally polarized abelian varieties defines a morphism of moduli spaces $\mathcal{A}_g \times \mathcal{A}_{g'} \to \mathcal{A}_{g+g'}$ which extends to these compactifications: we have a commutative diagram

$$\begin{array}{ccc} \mathcal{A}_g \times \mathcal{A}_{g'} & \longrightarrow & \mathcal{A}_{g+g'} \\ {\scriptstyle j_g \times j_{g'}}\big\downarrow & & \big\downarrow{\scriptstyle j_{g+g'}} \\ \mathcal{A}_g^{\mathrm{bb}} \times \mathcal{A}_{g'}^{\mathrm{bb}} & \longrightarrow & \mathcal{A}_{g+g'}^{\mathrm{bb}} \end{array} \tag{1}$$

By taking $g' = 1$ and choosing a point of $\mathcal{A}_1$, we get the "stabilization maps"

$$\begin{array}{ccc} \mathcal{A}_g & \longrightarrow & \mathcal{A}_{g+1} \\ {\scriptstyle j_g}\big\downarrow & & \big\downarrow{\scriptstyle j_{g+1}} \\ \mathcal{A}_g^{\mathrm{bb}} & \longrightarrow & \mathcal{A}_{g+1}^{\mathrm{bb}} \end{array} \tag{2}$$

whose homotopy type does not depend on the point we choose, for $\mathcal{A}_1$ is isomorphic to the affine line and hence connected. Since we are only concerned with homotopy classes and commutativity up to homotopy, we can for the definition of the map $\mathcal{A}_g^{\mathrm{bb}} \to \mathcal{A}_{g+1}^{\mathrm{bb}}$

Published: 19 May 2017 DOI: 10.2140/gt.2017.21.2231

even choose this point to be represented by the singleton $\mathcal{A}_0$. Then this map is a homeomorphism onto the Satake boundary (since $\mathcal{A}_1^{\mathrm{bb}} \cong \mathbb{P}^1$ the maps are not just homotopic, but even induce the same map on Chow groups). We shall see that this gives rise to two Hopf algebras with a mixed Hodge structure.

Before we proceed, let us recall that $\mathcal{A}_g$ is a locally symmetric variety associated to the $\mathbb{Q}$–algebraic group $\mathcal{S}p_g$ and that the $\mathbb{Q}$–rank of $\mathcal{S}p_g$ is g. According to Borel and Serre [4, Corollary 11.4.3] the virtual cohomological dimension of $\mathrm{Sp}(2g,\mathbb{Z})$ equals $\dim_{\mathbb{R}} \mathcal{A}_g - g$. This implies that the rational cohomology of $\mathcal{A}_g$, and more generally, the cohomology of a sheaf $\mathcal{F}$ on $\mathcal{A}_g$ defined by a representation of $\mathrm{Sp}(2g,\mathbb{Z})$ on a $\mathbb{Q}$–vector space, vanishes in degrees $> \dim_{\mathbb{R}} \mathcal{A}_g - g$. Since $\mathcal{A}_g$ is an orbifold, this is via Poincaré–Lefschetz duality equivalent to $H^k_c(\mathcal{A}_g;\mathcal{F})$ being zero for $k < g$. We shall use this basic fact in the proofs of Lemmas 1.1 and 2.1.

Lemma 1.1 *The stabilization maps $\mathcal{A}_g \hookrightarrow \mathcal{A}_{g+1}$ (multiplication by a fixed elliptic curve) and $\mathcal{A}_g^{\mathrm{bb}} \to \mathcal{A}_{g+1}^{\mathrm{bb}}$ (mapping onto the boundary) defined above induce on rational cohomology an isomorphism in degrees $< g$ and are injective in degree g.*

Proof Recall that $\mathcal{A}_g$ is a locally symmetric variety associated to the $\mathbb{Q}$–algebraic group $\mathcal{S}p_g$ and that the $\mathbb{Q}$–rank of $\mathcal{S}p_g$ is g. The first assertion then follows from a theorem of Borel [2, Theorems 7.5 and 11.1]. The second stability assertion is equivalent to the vanishing of the relative cohomology $H^k(\mathcal{A}_{g+1}^{\mathrm{bb}},\mathcal{A}_g^{\mathrm{bb}};\mathbb{Q})$ for $k \le g$. As this is just $H^k_c(\mathcal{A}_{g+1};\mathbb{Q})$, this follows from the Borel–Serre result quoted above. □

We then form the stable rational cohomology spaces

$$H^k(\mathcal{A}_\infty;\mathbb{Q}) := \varprojlim_g H^k(\mathcal{A}_g;\mathbb{Q}), \quad H^k(\mathcal{A}_\infty^{\mathrm{bb}};\mathbb{Q}) := \varprojlim_g H^k(\mathcal{A}_g^{\mathrm{bb}};\mathbb{Q}),$$

where the notation is only suggestive, for there is here no pretense of introducing spaces $\mathcal{A}_\infty$ and $\mathcal{A}_\infty^{\mathrm{bb}}$. If we take the direct sum over k we get a $\mathbb{Q}$–algebra in either case. It follows from the homotopy commutativity of the diagram (2) above that the inclusions j_g define a graded $\mathbb{Q}$–algebra homomorphism

$$j_\infty^*\colon H^\bullet(\mathcal{A}_\infty^{\mathrm{bb}};\mathbb{Q}) \to H^\bullet(\mathcal{A}_\infty;\mathbb{Q}).$$

The multiplication maps exhibited in diagram (1) are (almost by definition) compatible with the stabilization maps and hence induce a graded coproduct on either algebra so that j_∞^* becomes a homomorphism of (graded bicommutative) Hopf algebras. Since the multiplication maps and the stability maps are morphisms in the category of complex algebraic varieties, these Hopf algebras come with a natural mixed Hodge structure such that j_∞^* is also a morphism in the mixed Hodge category. The Hopf algebra $H^\bullet(\mathcal{A}_\infty;\mathbb{Q})$ is well-known and due to Borel [2, 11.4]: it has as its primitive elements classes $\mathrm{ch}_{2r+1}\in H^{4r+2}(\mathcal{A}_\infty;\mathbb{Q})$, $r\geq 0$, where ch_{2r+1} restricts to $\mathcal{A}_g$ as the rational $(2r+1)^{\mathrm{st}}$ Chern character of the Hodge bundle on $\mathcal{A}_g$, and so $H^\bullet(\mathcal{A}_\infty;\mathbb{Q}) = \mathbb{Q}[\mathrm{ch}_1,\mathrm{ch}_3,\mathrm{ch}_5,\dots]$ with ch_{2r+1} of type $(2r+1,2r+1)$ (if we are happy with multiplicative generators, we can just as well replace ch_{2r+1} by the corresponding Chern class c_{2r+1}, for c_{2r+1} is expressed universally in $\mathrm{ch}_1,\mathrm{ch}_3,\mathrm{ch}_5,\dots,\mathrm{ch}_{2r+1}$ and vice versa). The principal and essentially only result of this paper is Theorem 1.2. Its first assertion is due Charney and Lee [5, Theorem 4.2], who derive this from a determination of a limit of homotopy types. We shall obtain this in a relatively elementary manner by means of algebraic geometry and the classical vanishing results of Borel and of Borel and Serre. Our approach has the advantage that it helps us to understand the new classes that appear here geometrically, to the extent that this enables us to determine their Hodge type. We address the homotopy discussion of Charney and Lee and a generalization thereof in another paper [6] that will not be used here.

Theorem 1.2 *The graded Hopf algebra $H^\bullet(\mathcal{A}_\infty^{\mathrm{bb}};\mathbb{Q})$ has for every integer $r\geq 1$ a primitive generator y_r of degree $4r+2$ and for every integer $r\geq 0$ a primitive generator $\widetilde{\mathrm{ch}}_{2r+1}$ of degree $4r+2$ such that the map $j_\infty^*\colon H^\bullet(\mathcal{A}_\infty^{\mathrm{bb}};\mathbb{Q})\to H^\bullet(\mathcal{A}_\infty;\mathbb{Q})$ sends $\widetilde{\mathrm{ch}}_{2r+1}$ to ch_{2r+1} and is zero on y_r when $r\geq 1$. In particular, if $\tilde{c}_{2r+1}\in H^{4r+2}(\mathcal{A}_\infty^{\mathrm{bb}};\mathbb{Q})$ denotes the lift of $c_{2r+1}\in H^{4r+2}(\mathcal{A}_\infty;\mathbb{Q})$ that is obtained from our choice of the $\widetilde{\mathrm{ch}}_1,\dots,\widetilde{\mathrm{ch}}_{2r+1}$ (as a universal polynomial in these classes), then $H^\bullet(\mathcal{A}_\infty^{\mathrm{bb}};\mathbb{Q})=\mathbb{Q}[y_1,y_2,y_3\dots,\tilde{c}_1,\tilde{c}_3,\tilde{c}_5,\dots]$ as a commutative $\mathbb{Q}$–algebra.*

The mixed Hodge structure on $H^\bullet(\mathcal{A}_\infty^{\mathrm{bb}};\mathbb{Q})$ is such that y_r is of bidegree $(0,0)$ and $\widetilde{\mathrm{ch}}_{2r+1}$ (or equivalently, $\tilde{c}_{2r+1}$) is of bidegree $(2r+1,2r+1)$.

Remark 1.3 So for $r\geq 1$, the primitive part $H^{4r+2}_{\mathrm{pr}}(\mathcal{A}_\infty^{\mathrm{bb}};\mathbb{Q})$ of the Hopf algebra $H^\bullet(\mathcal{A}_\infty^{\mathrm{bb}};\mathbb{Q})$ is two-dimensional in degree $4r+2$ and defines a Tate extension

$$0\to\mathbb{Q}\to H^{4r+2}_{\mathrm{pr}}(\mathcal{A}_\infty^{\mathrm{bb}};\mathbb{Q})\to\mathbb{Q}(-2r-1)\to 0,$$

with $\mathbb{Q}$ spanned by y_r and $\mathbb{Q}(-2r-1)$ spanned by ch_{2r+1}. We discuss the nature of this extension briefly in Remark 3.1.

Acknowledgement We thank the referee for helpful comments on an earlier version. These led to an improved exposition.

2 Determination of the stable cohomology as a Hopf algebra

According to [8, Chapter V, Theorem 2.3(3)], $\mathcal{A}_g^{\mathrm{bb}}\smallsetminus\mathcal{A}_g$ is as a variety isomorphic to $\mathcal{A}_{g-1}^{\mathrm{bb}}$. In particular, we have a partition into locally closed subvarieties: $\mathcal{A}_g^{\mathrm{bb}} = \mathcal{A}_g\sqcup\mathcal{A}_{g-1}\sqcup\cdots\sqcup\mathcal{A}_0$.

We will use the fact that the higher direct images $R^\bullet j_{g*}\mathbb{Q}_{\mathcal{A}_g}$ are locally constant on each stratum $\mathcal{A}_r$. Each point of $\mathcal{A}_r$ has a neighborhood basis whose members meet $\mathcal{A}_g$ in a virtual classifying space of an arithmetic group $P_g(r)$ defined below (for a more detailed discussion we refer to [12, Example 3.5]; see also [6, Section 4]), so that $R^\bullet j_{g*}\mathbb{Q}_{\mathcal{A}_g}$ can be identified with the rational cohomology of $P_g(r)$.

Let H stand for $\mathbb{Z}^2$ (with basis denoted (e,e')) and endowed with the symplectic form characterized by $\langle e,e'\rangle = 1$. We also put $I := \mathbb{Z}e$. We regard H^g as a direct sum of symplectic lattices with g summands. In terms of the decomposition $H^g = H^r \oplus H^{g-r}$, $P_g(r)$ is the group of symplectic transformations in H^g that are the identity on $H^r \oplus 0$ and preserve $H^r \oplus I^{g-r}$. The orbifold fundamental group of $\mathcal{A}_r$ is isomorphic to $\mathrm{Sp}(H^r)$ (the isomorphism is of course given up to conjugacy) and its representation on a stalk of $R^\bullet j_{g*}\mathbb{Q}_{\mathcal{A}_g}|\mathcal{A}_r$ corresponds to its obvious action (given by conjugation) on $P_g(r)$. Note that this action is algebraic in the sense that it extends to a representation of the underlying affine algebraic group (which assigns to a commutative ring R the group $\mathrm{Sp}(H^r\otimes R)$). If $p\in\mathcal{A}_r$ and U_p is a regular neighborhood of p in $\mathcal{A}_g^{\mathrm{bb}}$ such that the natural map $H^\bullet(U_p\cap\mathcal{A}_g;\mathbb{Q})\to(R^\bullet j_{g*}\mathbb{Q}_{\mathcal{A}_g})_p$ is an isomorphism, then for every $r\le s\le g$ and $q\in U_p\cap\mathcal{A}_s$ the restriction map yields a map of $\mathbb{Q}$–algebras $(R^\bullet j_{g*}\mathbb{Q}_{\mathcal{A}_g})_p\to(R^\bullet j_{g*}\mathbb{Q}_{\mathcal{A}_g})_q$. Under the above identification this is represented by the $\mathrm{Sp}(H^r)$–orbit of the obvious inclusion $P_g(s)\hookrightarrow P_g(r)$. Similarly, the restriction to $\mathcal{A}_r\times\mathcal{A}_{r'}\subset\mathcal{A}_g\times\mathcal{A}_{g'}$ of the natural sheaf homomorphism

$$R^\bullet j_{g+g'*}\mathbb{Q}_{\mathcal{A}_{g+g'}}|\mathcal{A}_g^{\mathrm{bb}}\times\mathcal{A}_{g'}^{\mathrm{bb}}\to R^\bullet(j_g\times j_{g'})_*\mathbb{Q}_{\mathcal{A}_g\times\mathcal{A}_{g'}}\cong R^\bullet j_{g*}\mathbb{Q}_{\mathcal{A}_g}\boxtimes R^\bullet j_{g'*}\mathbb{Q}_{\mathcal{A}_{g'}}$$

(we invoked the Künneth isomorphism) is induced by the obvious embedding

$$P_g(r)\times P_{g'}(r')\hookrightarrow P_{g+g'}(r+r'),$$

or rather its $\mathrm{Sp}(H^{r+r'})$–orbit.

The proof of the first assertion of our main theorem rests on careful study of the Leray spectral sequence for the inclusion $j_g\colon\mathcal{A}_g\subset\mathcal{A}_g^{\mathrm{bb}}$,

$$E_2^{p,q}=H^p(\mathcal{A}_g^{\mathrm{bb}},R^qj_{g*}\mathbb{Q})\ \Rightarrow\ H^{p+q}(\mathcal{A}_g;\mathbb{Q}). \tag{3}$$

Such a spectral sequence can be set up in the category of mixed Hodge modules (see [13]), so that this is in fact a spectral sequence of mixed Hodge structures.

Lemma 2.1 *Let $r\le g$. Then the natural map*

$$H^p(\mathcal{A}_g^{\mathrm{bb}},R^\bullet j_{g*}\mathbb{Q})\to H^p(\mathcal{A}_r^{\mathrm{bb}},R^\bullet j_{g*}\mathbb{Q})$$

is an isomorphism for $p<r$ and is injective for $p=r$.

Proof It suffices to show that when $r < g$, the natural map $H^p(\mathcal{A}_{r+1}^{\mathrm{bb}}, R^\bullet j_{g*}\mathbb{Q}) \to H^p(\mathcal{A}_r^{\mathrm{bb}}, R^\bullet j_{g*}\mathbb{Q})$ has this property. For this we consider the exact sequence

$$\cdots \to H_c^p(\mathcal{A}_{r+1}, R^\bullet j_{g*}\mathbb{Q}) \to H^p(\mathcal{A}_{r+1}^{\mathrm{bb}}, R^\bullet j_{g*}\mathbb{Q}) \to H^p(\mathcal{A}_r^{\mathrm{bb}}, R^\bullet j_{g*}\mathbb{Q}) \\ \to H_c^{p+1}(\mathcal{A}_{r+1}, R^\bullet j_{g*}\mathbb{Q}) \to \cdots .$$

The restriction $R^q j_{g*}\mathbb{Q}|\mathcal{A}_{r+1}$ is a local system whose monodromy comes from an action of the algebraic group $\mathcal{S}p(H^r)$. Following the Borel–Serre result mentioned above, $H_c^i(\mathcal{A}_{r+1}, R^\bullet j_{g*}\mathbb{Q})$ vanishes for $i \le r$ and so the lemma follows. □

By viewing I^{g-r} as the subquotient $(H^r \oplus I^{g-r})/(H^r \oplus 0)$ of H^g, we see that there is a natural homomorphism of arithmetic groups $P_g(r) \to \mathrm{GL}(I^{g-r}) = \mathrm{GL}(g-r, \mathbb{Z})$.

Lemma 2.2 *The homomorphism $P_g(r) \to \mathrm{GL}(g-r, \mathbb{Z})$ induces an isomorphism on rational cohomology in degrees $< \frac{1}{2}(g-r-1)$. In that range the rational cohomology of $\mathrm{GL}(g-r, \mathbb{Z})$ is stable and is canonically isomorphic to the cohomology of $\mathrm{GL}(\mathbb{Z}) := \bigcup_r \mathrm{GL}(r, \mathbb{Z})$. The inclusion $P_g(r) \times P_{g'}(r') \subset P_{g+g'}(r + r')$ induces on rational cohomology in the stable range (relative to both factors) the coproduct in the Hopf algebra $H^\bullet(\mathrm{GL}(\mathbb{Z}); \mathbb{Q})$.*

Proof According to Borel [3, Theorem 4.4], the cohomology of the arithmetic group $\mathrm{GL}(r, \mathbb{Z})$ with values in an irreducible representation of the underlying algebraic group $\mathcal{SL}_r^{\pm}$ (the group of invertible matrices of determinant ± 1) is zero in degrees $< \frac{1}{2}(r-1)$, unless the representation is trivial. Let $N_g(r)$ be the kernel of $P_g(r) \to \mathrm{GL}(g-r, \mathbb{Z})$. This is a nilpotent subgroup whose center, when written additively, may be identified with the symmetric quotient $\mathrm{Sym}_2(I^{g-r})$ of $I^{g-r} \otimes I^{g-r}$. The quotient of $N_g(r)$ by this center is abelian, and when written additively, naturally identified with the lattice $H^r \otimes I^{g-r}$. So in view of the Leray spectral sequence

$$H^p(\mathrm{GL}(g-r, \mathbb{Z}), H^q(N_g(r), \mathbb{R})) \;\Rightarrow\; H^{p+q}(P_g(r), \mathbb{R}),$$

it suffices to show that $H^q(N_g(r); \mathbb{R})$ does not contain the trivial representation of $\mathrm{SL}^{\pm 1}(g-r, \mathbb{R})$ in positive degrees $q < \frac{1}{2}(g-r-1)$. This follows from another Leray spectral sequence

$$H^s(I^{g-r} \otimes H^r, H^t(\mathrm{Sym}_2 I^{g-r}, \mathbb{R})) \;\Rightarrow\; H^{s+t}(N_g(r), \mathbb{R}).$$

The left-hand side is isomorphic to

$$\textstyle\bigwedge^s \mathrm{Hom}(I^{g-r} \otimes H^r, \mathbb{R}) \otimes \bigwedge^t \mathrm{Hom}(\mathrm{Sym}_2 I^{g-r}, \mathbb{R})$$

as a representation of $\mathrm{SL}^{\pm 1}(g-r, \mathbb{R})$. The invariant theory of $\mathrm{SL}(g-r; \mathbb{R})$ tells us that the trivial representations in the tensor algebra generated by $\mathrm{Hom}(I^{g-r}, \mathbb{R})$ come from the formation of powers of the determinant $\bigwedge^{g-r} \mathrm{Hom}(I^{g-r}, \mathbb{R}) \cong \mathbb{R}$ (see for example [9, Proposition F.10]). Since the displayed representation of $\mathrm{SL}^{\pm 1}(g-r,\mathbb{R})$ is a quotient of this tensor algebra, it will not contain the trivial representation when $0<s+2t<g-r$. Hence the first part of the lemma follows. The second assertion merely quotes a theorem of Borel [2, Theorems 7.5 and 11.1], and the last assertion is easy. □

Corollary 2.3 *For $q<\frac{1}{2}(g-r-1)$, $R^q j_{g*}\mathbb{Q}|\mathcal{A}_r^{\mathrm{bb}}$ is a constant local system whose stalk is canonically isomorphic to $H^q(\mathrm{GL}(\mathbb{Z}),\mathbb{Q})$. This identification is compatible with the multiplicative structure. It is also compatible with the coproduct in the sense that when $0\le r'\le g'$, then in degrees $<\frac{1}{2}\min\{g-r-1,g'-r'-1\}$, the natural map*

$$R^\bullet j_{g+g'*}\mathbb{Q}_{\mathcal{A}_{g+g'}}|\mathcal{A}_r^{\mathrm{bb}}\times\mathcal{A}_{r'}^{\mathrm{bb}}\to(R^\bullet j_{g*}\mathbb{Q}_{\mathcal{A}_g}|\mathcal{A}_r^{\mathrm{bb}})\boxtimes(R^\bullet j_{g'*}\mathbb{Q}_{\mathcal{A}_{g'}}|\mathcal{A}_{r'}^{\mathrm{bb}})$$

is stalkwise identified with the coproduct on $H^\bullet(\mathrm{GL}(\mathbb{Z});\mathbb{Q})$.

Proof of the first assertion of Theorem 1.2 We have shown (Lemma 2.1 and Corollary 2.3) that when $p<r$ and $q<\frac{1}{2}(g-r-1)$ we have

$$E_2^{p,q}=H^p(\mathcal{A}_g^{\mathrm{bb}},R^q j_{g*}\mathbb{Q})=H^p(\mathcal{A}_r^{\mathrm{bb}},\mathbb{Q})\otimes H^q(\mathrm{GL}(\mathbb{Z});\mathbb{Q})$$

The Leray spectral sequences (3) for j_{g*} and j_{g+1*} are compatible and so we may form a limit: we fix p and q, but we let r and $g-r$ tend to infinity. This then yields a spectral sequence

$$E_2^{p,q}=H^p(\mathcal{A}_\infty^{\mathrm{bb}};\mathbb{Q})\otimes H^q(\mathrm{GL}(\mathbb{Z});\mathbb{Q})\ \Rightarrow\ H^{p+q}(\mathcal{A}_\infty;\mathbb{Q}). \tag{4}$$

This spectral sequence is not just multiplicative, but also compatible with the coproduct. So the differentials take primitive elements to primitive elements (or zero) and the spectral sequence restricts to one of graded vector spaces by restricting to the primitive parts. The primitive part of $E_2^{p,q}$ is zero unless $p=0$ or $q=0$. A theorem of Borel [2, 11.4] tells us that $H^\bullet(\mathrm{GL}(\mathbb{Z});\mathbb{Q})_{\mathrm{pr}}$ has for every positive integer r a generator a_r in degree $4r+1$ (and is zero in all other positive degrees) and that $H^\bullet(\mathcal{A}_\infty;\mathbb{Q})_{\mathrm{pr}}$ has for every odd integer s a primitive generator ch_s in degree $2s$ (and is zero in all other positive degrees). This implies that $d^k(1\otimes a_r)=0$ for $k=2,3,\dots 4r+1$, but that $y_r:=d^{4r+2}(1\otimes a_r)$ will be nonzero and primitive. We also see that for $s>0$ odd, $H^{2s}(\mathcal{A}_g^{\mathrm{bb}};\mathbb{Q})$ must contain a lift $\widetilde{\mathrm{ch}}_s$ of ch_s. Since $H^\bullet(\mathcal{A}_\infty^{\mathrm{bb}};\mathbb{Q})$ is a Hopf algebra, it then follows that the Hopf algebra $H^\bullet(\mathcal{A}_\infty^{\mathrm{bb}};\mathbb{Q})$ is primitively generated by $y_1,y_2,\dots,\widetilde{\mathrm{ch}}_1,\widetilde{\mathrm{ch}}_3,\widetilde{\mathrm{ch}}_5,\dots$. So as a commutative $\mathbb{Q}$–algebra it is freely generated by $y_1,y_2,\dots,\tilde{c}_1,\tilde{c}_3,\tilde{c}_5,\dots$. □

The spectral sequence (4) suggests that the space $\mathcal{A}_\infty$ (which we did not define) has the homotopy type of a $B\mathrm{GL}(\mathbb{Z})$–bundle over $\mathcal{A}_\infty^{\mathrm{bb}}$ (which we did not define either). Indeed, Charney and Lee provide in [5, Theorem 3.2] an appropriate homotopy substitute for such a fibration (which they attribute to Giffen), namely, a homotopy fibration whose fiber is a model of $B\mathrm{GL}(\mathbb{Z})^+$ (where "$+$" is the Quillen plus construction) and whose total space is $\mathbb{Q}$–homotopy equivalent to $B\mathrm{Sp}(\mathbb{Z})^+$, so that the base (which admits an explicit description as the classifying space of a category) may be regarded as a $\mathbb{Q}$–homotopy type representing $\mathcal{A}_\infty^{\mathrm{bb}}$.

Remark 2.4 The long exact sequence for the pair $(\mathcal{A}_g^{\mathrm{bb}}, \mathcal{A}_g)$ shows that the cohomology $H^\bullet(\mathcal{A}_g^{\mathrm{bb}}, \mathcal{A}_g; \mathbb{Q})$ stabilizes as well with g and is equal to the ideal in $\mathbb{Q}[y_1, y_2, \dots, \widetilde{c}_1, \widetilde{c}_3, \widetilde{c}_5, \dots]$ generated by the y_r. We shall therefore denote this ideal by $H^\bullet(\mathcal{A}_\infty^{\mathrm{bb}}, \mathcal{A}_\infty; \mathbb{Q})$. We use the occasion to point out that the y–classes are canonically defined, but that this is not at all clear for the $\widetilde{c}$–classes (for more on this, see Remark 3.1).

Remark 2.5 We can account geometrically for the classes y_r as follows. Denote the single point of $\mathcal{A}_0 \subset \mathcal{A}_g^{\mathrm{bb}}$ by ∞ (the worst cusp of $\mathcal{A}_g^{\mathrm{bb}}$), and take g so large that the natural maps

$$H^{4r+1}(\mathrm{GL}(\mathbb{Z}); \mathbb{Q}) \to H^{4r+1}(\mathrm{GL}(g, \mathbb{Z}); \mathbb{Q}) \to (R^{4r+1} j_{g*}\mathbb{Q})_\infty,$$
$$H^{4r+2}(\mathcal{A}_\infty^{\mathrm{bb}}, \mathcal{A}_\infty; \mathbb{Q}) \to H^{4r+2}(\mathcal{A}_g^{\mathrm{bb}}, \mathcal{A}_g; \mathbb{Q})$$

are isomorphisms. Choose a regular neighborhood U_∞ of ∞ in $\mathcal{A}_g^{\mathrm{bb}}$ so that if we put $\mathring{U}_\infty := U_\infty \cap \mathcal{A}_g$, the natural maps

$$(R^{4r+1} j_{g*}\mathbb{Q})_\infty \leftarrow H^{4r+1}(\mathring{U}_\infty; \mathbb{Q}) \xrightarrow{\delta} H^{4r+2}(U_\infty, \mathring{U}_\infty; \mathbb{Q})$$

are also isomorphisms. If we identify $a_r \in H^{4r+1}(\mathrm{GL}(\mathbb{Z}); \mathbb{Q})$ with its image in $H^{4r+1}(\mathring{U}_\infty; \mathbb{Q})$, then $\delta(a_r) \in H^{4r+2}(U_\infty, \mathring{U}_\infty; \mathbb{Q})$ is precisely the image of y_r under the restriction map

$$H^{4r+2}(\mathcal{A}_\infty^{\mathrm{bb}}, \mathcal{A}_\infty; \mathbb{Q}) \cong H^{4r+2}(\mathcal{A}_g^{\mathrm{bb}}, \mathcal{A}_g; \mathbb{Q}) \to H^{4r+2}(U_\infty, \mathring{U}_\infty; \mathbb{Q}).$$

We may also get a homology class this way: the Hopf algebra $H_\bullet(\mathrm{GL}(\mathbb{Z}); \mathbb{Q})$ has a primitive generator in $H_{4r+1}(\mathrm{GL}(\mathbb{Z}); \mathbb{Q})$ that is dual to a_r, and if we represent this generator as $(4r+1)$–cycle B_r in $\mathring{U}_\infty$, then B_r bounds both in U_∞ (almost canonically) and in $\mathcal{A}_g$ (not canonically). The two bounding $(4r+2)$–chains make up a $(4r+2)$–cycle in $\mathcal{A}_g^{\mathrm{bb}}$ whose class $z_r \in H_{4r+2}(\mathcal{A}_g^{\mathrm{bb}}; \mathbb{Q})$ pairs nontrivially with the image of y_r in $H^{4r+2}(\mathcal{A}_g^{\mathrm{bb}}; \mathbb{Q})$.

3 The mixed Hodge structure on the primitive stable cohomology

Proof that the y–classes are of weight zero In view of Remark 2.5 it is enough to show that the image of $H^\bullet(\mathrm{GL}(\mathbb{Z}); \mathbb{Q})$ in the stalk $(R^\bullet j_{g*}\mathbb{Q})_\infty$ has weight zero. For this we will need a toroidal resolution of U_∞ as described in [1, Chapter III, Section 1], but we will try to get by with the minimal input necessary (for a somewhat more detailed review of this construction one may consult [6]).

Consider the symmetric quotient $\mathrm{Sym}_2\,\mathbb{Z}^g$ of $\mathbb{Z}^g\otimes\mathbb{Z}^g$ and regard it as a lattice in the space $\mathrm{Sym}_2\,\mathbb{R}^g$ of quadratic forms on $\mathrm{Hom}(\mathbb{Z}^g,\mathbb{R})$. The positive definite quadratic forms make up a cone $C_g\subset\mathrm{Sym}_2\,\mathbb{R}^g$ that is open and convex and is as such spanned by its intersection with $\mathrm{Sym}_2\,\mathbb{Z}^g$. Let $C_g^+\supset C_g$ be the convex cone spanned by $\overline{C}_g\cap\mathrm{Sym}_2\,\mathbb{Z}^g$; this is just the set of semipositive quadratic forms on $\mathrm{Hom}(\mathbb{Z}^g,\mathbb{R})$, whose kernel is spanned by its intersection with $\mathrm{Hom}(\mathbb{Z}^g,\mathbb{Z})$. The obvious action of $\mathrm{GL}(g,\mathbb{Z})$ on $\mathrm{Sym}_2\,\mathbb{Z}^g$ preserves both cones and is proper on C_g.

Consider the algebraic torus $T_g:=\mathbb{C}^\times\otimes_\mathbb{Z}\mathrm{Sym}_2\,\mathbb{Z}^g$. If we apply the "log norm" lgnm: $z\in\mathbb{C}^\times\mapsto\log|z|\in\mathbb{R}$ to the first tensor factor, we get a $\mathrm{GL}(g,\mathbb{Z})$–equivariant homomorphism $\mathrm{lgnm}_{T_g}\colon T_g\to\mathrm{Sym}_2\,\mathbb{R}^g$ with kernel the compact torus $U(1)\otimes_\mathbb{Z}\mathrm{Sym}_2\,\mathbb{Z}^g$. We denote by $\mathcal{T}_g\subset T_g$ the preimage of C_g so that we have defined a proper $\mathrm{GL}(g,\mathbb{Z})$–equivariant homomorphism of semigroups $\mathrm{lgnm}_{\mathcal{T}_g}\colon\mathcal{T}_g\to C_g$. Since $\mathrm{GL}(g,\mathbb{Z})$ acts properly on C_g it does so on $\mathcal{T}_g$ and hence the orbit space $\mathring{V}:=\mathrm{GL}(g,\mathbb{Z})\backslash\mathcal{T}_g$ has the structure of a complex-analytic orbifold. There is a natural extension of $V\supset\mathring{V}$ in the complex analytic category (it is in fact the Stein hull of $\mathring{V}$ in case $g>1$) that comes with a distinguished point that we will (for good reasons) also denote by ∞ and which is such that $\mathring{V}$ is open-dense in V and $(V,V\smallsetminus\mathring{V})$ is topologically the open cone over a pair of spaces with vertex ∞. It has the property that there exists an open embedding of U_∞ in V that takes ∞ to ∞ and identifies U_∞ with a regular neighborhood of ∞ in V in such a way that $\mathring{U}_\infty=U_\infty\cap\mathring{V}$. This justifies our focus on the triple $(V,\mathring{V};\infty)$. All else we need to know about V is that the toroidal extension of $\mathring{V}$ that we are about to consider provides a resolution of V as an orbifold.

The universal cover of $\mathcal{T}_g$ is contractible (with covering group $\mathrm{Sym}_2\,\mathbb{Z}^g$) and hence the universal cover of $\mathring{V}$ as an orbifold is also contractible and has covering group $\mathrm{GL}(g,\mathbb{Z})\ltimes\mathrm{Sym}_2\,\mathbb{Z}^g$ (it is in fact a virtual classifying space for this group). Similarly, the orbit space $\mathcal{I}_g:=\mathrm{GL}(g,\mathbb{Z})\backslash C_g$ exists as a real-analytic orbifold and is a virtual classifying space for $\mathrm{GL}(g,\mathbb{Z})$. The map lgnm_{T_g} induces a projection $\nu\colon\mathring{V}\to\mathcal{I}_g$ and the classes that concern us lie in the image of

$$H^\bullet(\mathrm{GL}(\mathbb{Z});\mathbb{Q})\to H^\bullet(\mathrm{GL}(g,\mathbb{Z});\mathbb{Q})\to H^\bullet(\mathcal{I}_g;\mathbb{Q})\xrightarrow{\nu^*}H^\bullet(\mathring{V};\mathbb{Q}). \tag{5}$$

A *nonsingular admissible decomposition* of C_g^+ is a collection $\{\sigma\}_{\sigma\in\Sigma}$ of closed cones in C_g^+, each of which is spanned by a partial basis of $\mathrm{Sym}_2\,\mathbb{Z}^g$, such that the collection is closed under "taking faces" and "taking intersections" and whose relative interiors are pairwise disjoint with union C_g^+. Let Σ be such a decomposition that is also $\mathrm{GL}(g,\mathbb{Z})$–invariant and is fine enough in the sense that every $\mathrm{GL}(g,\mathbb{Z})$–orbit in C_+^g meets every member of Σ in at most one point. Such decompositions exist [1, Chapter II, Corollary 5.23]. (One usually also requires that $\mathrm{GL}(g,\mathbb{Z})$ has only finitely many orbits in Σ, but this is in fact implied by the other conditions; see [11, Theorem 3.8].) The associated torus embedding $T_g\subset T_g^\Sigma$ is then nonsingular and comes with an action of $\mathrm{GL}(g,\mathbb{Z})$. We denote by $\mathcal{T}_g^\Sigma$ the interior of the closure of $\mathcal{T}_g$ in T_g^Σ. This is an open $\mathrm{GL}(g,\mathbb{Z})$–invariant subset of T_g^Σ on which $\mathrm{GL}(g,\mathbb{Z})$ acts properly, so that $V^\Sigma:=\mathrm{GL}(g,\mathbb{Z})\backslash\mathcal{T}_g^\Sigma$ exists as an analytic orbifold. It is of the type alluded to above: we have a natural proper morphism $f\colon V^\Sigma\to V$ that is complex-algebraic over V and is an isomorphism over $\mathring{V}$. Moreover, the exceptional set is a simple normal crossing divisor in the orbifold sense.

As for every torus embedding, there is also a real counterpart in the sense that $\mathrm{lgnm}_{\mathcal{T}_g}$ extends in a $\mathrm{GL}(g,\mathbb{Z})$–equivariant manner to a proper and surjective map $\mathrm{lgnm}_{\mathcal{T}_g^\Sigma}\colon\mathcal{T}_g^\Sigma\to C_g^\Sigma$, where C_g^Σ is a certain stratified locally compact Hausdorff space which contains C_g as an open dense subset. In the present case C_g^Σ is simply a manifold with corners, because Σ is nonsingular. The strata of C_g^Σ are indexed by Σ, with the stratum defined by σ being the image of C_g under the projection along the real subspace of $\mathrm{Sym}_2\mathbb{R}^g$ spanned by σ. So each stratum of C_g^Σ appears as a convex open subset of some vector space and it is all of this vector space precisely when the relative interior of σ is contained in C_g. This is also equivalent to the stratum having compact closure in C_g^Σ.

Let us define a *wall* of C_g^Σ to be the closure of a stratum defined by a ray (ie a one-dimensional member) of Σ. So a wall is compact if and only if the associated ray lies in $C_g\cup\{0\}$. We denote by $\partial_{\mathrm{pr}}C_g^\Sigma$ the union of these compact walls. This is a closed subset of C_g^Σ and its covering by such compact walls is a *Leray covering*: the covering is locally finite and each nonempty intersection is contractible (and is in fact the closure of a stratum). Its nerve is easily expressed in terms of Σ. Let us say that a member of Σ is *proper* if it is contained in $C_g\cup\{0\}$. The proper members of Σ make up a subset $\Sigma_{\mathrm{pr}}\subset\Sigma$ that is also closed under "taking faces" and "taking intersections" and their union makes up a $\mathrm{GL}(g,\mathbb{Z})$–invariant cone contained in $C_g\cup\{0\}$. If we projectivize that cone we get a simplicial complex in the real projective space of $\mathrm{Sym}_2\mathbb{R}^g$ that we denote by $P(\Sigma_{\mathrm{pr}})$. A vertex of $P(\Sigma_{\mathrm{pr}})$ corresponds of course to a ray of Σ_{pr}, and this in turn defines a compact wall of C_g^Σ. In this way $P(\Sigma_{\mathrm{pr}})$ can be identified in a $\mathrm{GL}(g,\mathbb{Z})$–equivariant manner with the nerve complex of the covering of $\partial_{\mathrm{pr}}C_g^\Sigma$ by the compact walls of C_g^Σ. A standard argument shows that we have a $\mathrm{GL}(g,\mathbb{Z})$–equivariant homotopy equivalence between $\partial_{\mathrm{pr}}C_g^\Sigma$ and the nerve $P(\Sigma_{\mathrm{pr}})$ of this covering.

Each stratum closure in C_g^Σ can be retracted in a canonical manner onto its intersection with $\partial_{\mathrm{pr}} C_g^\Sigma$ and we thus find a $\mathrm{GL}(g,\mathbb{Z})$–equivariant deformation retraction $C_g^\Sigma \to \partial_{\mathrm{pr}} C_g^\Sigma$. This shows at the same time that the inclusion $C_g \subset C_g^\Sigma$ is a $\mathrm{GL}(g,\mathbb{Z})$–equivariant homotopy equivalence. So if we put

$$\mathcal{I}_g^\Sigma := \mathrm{GL}(g,\mathbb{Z})\backslash C_g^\Sigma \quad \text{and} \quad \partial_{\mathrm{pr}}\mathcal{I}_g^\Sigma := \mathrm{GL}(g,\mathbb{Z})\backslash \partial_{\mathrm{pr}} C_g^\Sigma,$$

then we end up with homotopy equivalences $\mathcal{I}_g \subset \mathcal{I}_g^\Sigma \supset \partial_{\mathrm{pr}}\mathcal{I}_g^\Sigma$. We also have a homotopy equivalence $\partial_{\mathrm{pr}}\mathcal{I}_g^\Sigma \sim \mathrm{GL}(g,\mathbb{Z})\backslash P(\Sigma_{\mathrm{pr}})$.

Taking the preimage under lgnm makes walls of C_g^Σ correspond to irreducible components of the toric boundary $\mathcal{T}_g^\Sigma \smallsetminus \mathcal{T}_g$ and a wall of C_g^Σ is compact if and only if the associated irreducible component is. So the preimage $\partial_{\mathrm{pr}}\mathcal{T}_g^\Sigma$ of $\partial_{\mathrm{pr}} C_g^\Sigma$ is the union of the compact irreducible components of the toric boundary. It is clear that $P(\Sigma_{\mathrm{pr}})$ is also the nerve of the covering of $\partial_{\mathrm{pr}}\mathcal{T}_g^\Sigma$ by its irreducible components. The image of $\partial_{\mathrm{pr}}\mathcal{T}_g^\Sigma$ in V (in other words, its $\mathrm{GL}(g,\mathbb{Z})$–orbit space) is the normal crossing divisor $f^{-1}(\infty)$. The inclusion $f^{-1}(\infty) \subset V^\Sigma$ is also a deformation retract. So in the commutative diagram

$$\begin{tikzcd}
\mathring{V} \arrow[r, hook] \arrow[d, two heads] & V^\Sigma \arrow[d, two heads] & f^{-1}(\infty) \arrow[l, hook'] \arrow[d, two heads] \\
\mathcal{I}_g \arrow[r, hook] & \mathcal{I}_g^\Sigma & \partial_{\mathrm{pr}}\mathcal{I}^\Sigma \arrow[l, hook']
\end{tikzcd}$$

the inclusion on the top right and those at the bottom are homotopy equivalences. It follows that the composite map in diagram (5) factors through the rational cohomology of $\mathcal{I}_g^\Sigma$ and hence also through the rational cohomology of V^Σ and that the nonzero classes in $H^\bullet(V^\Sigma;\mathbb{Q}) \cong H^\bullet(f^{-1}(\infty);\mathbb{Q})$ that we thus obtain come from the nerve of the covering of $f^{-1}(\infty)$ by its irreducible components. Such classes are known to be of weight zero [7, Proposition 8.1.20]. $\square$

Remark 3.1 Goresky and Pardon [10, Corollary 11.9] have constructed a lift c_r^{bb} of the *real* Chern class $c_r \in H^{2r}(\mathcal{A}_g;\mathbb{R})$ to $H^{2r}(\mathcal{A}_g^{\mathrm{bb}};\mathbb{R})$. The second author [12, Theorem 2.8] recently proved that c_r^{bb} (and hence also the corresponding Chern character $\mathrm{ch}_r^{\mathrm{bb}}$) lies in $F^r H^{2r}(\mathcal{A}_g^{\mathrm{bb}};\mathbb{R})$. So the class of the Tate extension in Remark 1.3 is up to a rational number given by the value of c_{2r+1}^{bb} on the class $z_r \in H_{4r+2}(\mathcal{A}_g^{\mathrm{bb}};\mathbb{Q})$ found in Remark 2.5 (two choices of z_r differ by a class of the form $j_{g*}(w)$ with $w \in H_{4r+2}(\mathcal{A}_g;\mathbb{Q})$ and c_{2r+1}^{bb} takes on such a class the rational value $c_{2r+1}(w)$). Arvind Nair, after learning of our theorem, informed us that his techniques enable him to show that this extension class is nonzero. Subsequently a different proof (based on the Beilinson regulator) was given in [12, Theorem 5.1].

References

[1] **A Ash**, **D Mumford**, **M Rapoport**, **Y-S Tai**, *Smooth compactifications of locally symmetric varieties*, 2nd edition, Cambridge Univ. Press (2010) MR

[2] **A Borel**, *Stable real cohomology of arithmetic groups*, Ann. Sci. École Norm. Sup. 7 (1974) 235–272 MR

[3] **A Borel**, *Stable real cohomology of arithmetic groups, II*, from "Manifolds and Lie groups" (S Murakami, J Hano, K Okamoto, A Morimoto, H Ozeki, editors), Progr. Math. 14, Birkhäuser, Boston (1981) 21–55 MR

[4] **A Borel**, **J-P Serre**, *Corners and arithmetic groups*, Comment. Math. Helv. 48 (1973) 436–491 MR

[5] **R Charney**, **R Lee**, *Cohomology of the Satake compactification*, Topology 22 (1983) 389–423 MR

[6] **J Chen**, **E Looijenga**, *The homotopy type of the Baily–Borel and allied compactifications*, preprint (2015) arXiv

[7] **P Deligne**, *Théorie de Hodge, III*, Inst. Hautes Études Sci. Publ. Math. 44 (1974) 5–77 MR

[8] **G Faltings**, **C-L Chai**, *Degeneration of abelian varieties*, Ergeb. Math. Grenzgeb. 22, Springer (1990) MR

[9] **W Fulton**, **J Harris**, *Representation theory*, Graduate Texts in Mathematics 129, Springer (1991) MR

[10] **M Goresky**, **W Pardon**, *Chern classes of automorphic vector bundles*, Invent. Math. 147 (2002) 561–612 MR

[11] **E Looijenga**, *Discrete automorphism groups of convex cones of finite type*, Compos. Math. 150 (2014) 1939–1962 MR

[12] **E Looijenga**, *Goresky–Pardon lifts of Chern classes and associated Tate extensions*, Compos. Math. 153 (2017) 1349–1371

[13] **M Saito**, *Modules de Hodge polarisables*, Publ. Res. Inst. Math. Sci. 24 (1988) 849–995 MR

Institut de mathématiques de Jussieu-Paris Rive Gauche, Université Paris 7 Diderot
Campus de Grands Moulins, Bureau 745, Bâtiment Sophie Germain, 75205 Paris Cedex 13, France

Yau Mathematical Sciences Center, Tsinghua University
Jin Chun Yuan West Building, Haidian District, Beijing, 100084, China

`jiaming.chen@imj-prg.fr`, `eduard@math.tsinghua.edu.cn`

Proposed: Benson Farb   Received: 3 November 2015
Seconded: Jim Bryan, Dan Abramovich   Accepted: 10 August 2016

Geometry & Topology 21 (2017) 2243–2280

Collar lemma for Hitchin representations

GYE-SEON LEE
TENGREN ZHANG

There is a classical result known as the collar lemma for hyperbolic surfaces. A consequence of the collar lemma is that if two closed curves A and B on a closed orientable hyperbolizable surface intersect each other, then there is an explicit lower bound for the length of A in terms of the length of B, which holds for every hyperbolic structure on the surface. In this article, we prove an analog of the classical collar lemma in the setting of Hitchin representations.

57M50; 30F60, 32G15

1 Introduction

Let S be a closed connected oriented topological surface of genus $g \geq 2$, and let Γ be its fundamental group. The Teichmüller space of S, which we denote by $\mathcal{T}(S)$, is the space of hyperbolic structures on S, ie the space of isotopy classes of hyperbolic metrics on S. Via the holonomy representation, $\mathcal{T}(S)$ can be identified with a component of the space of conjugacy classes of representations from Γ to $\mathrm{PSL}(2,\mathbb{R})$. One advantage of doing so is that it allows us to generalize $\mathcal{T}(S)$ in the following way. It is a standard fact in representation theory that for any $n \geq 2$, there is a unique (up to conjugation) irreducible representation $\iota_n\colon \mathrm{PSL}(2,\mathbb{R}) \to \mathrm{PSL}(n,\mathbb{R})$. This gives, via postcomposition, an embedding

$$\mathcal{T}(S) \hookrightarrow \mathcal{X}_n(S) := \mathrm{Hom}(\Gamma, \mathrm{PSL}(n,\mathbb{R}))/\mathrm{PSL}(n,\mathbb{R}).$$

The image of this embedding is known as the *Fuchsian locus* and the component of $\mathcal{X}_n(S)$ containing the Fuchsian locus is the n^{th} *Hitchin component*, which we denote by $\mathrm{Hit}_n(S)$. By definition, $\mathrm{Hit}_2(S) = \mathcal{T}(S)$, so Hitchin representations can be thought of as generalizations of Fuchsian representations.

For the hyperbolic structures in $\mathcal{T}(S)$, there is a classical result due to Keen [14] known as the collar lemma. It gives an effective lower bound on the width of the maximal collar neighborhood of a simple closed curve in a hyperbolic surface, which grows to ∞ as the length of the simple closed curve is shrunk to 0. A consequence of the collar lemma is that if two closed curves η and γ in a hyperbolic surface have

Published: 19 May 2017 DOI: 10.2140/gt.2017.21.2243

nonvanishing geometric intersection number and γ is simple, then there is an explicit lower bound on the length of η in terms of the length of γ. This is a powerful tool that has been used to understand surfaces. For example, it was used to study the length spectrum of Riemann surfaces; see Buser [5].

The goal of this paper is to generalize a version of the classical collar lemma to Hitchin representations. By Labourie [15], for any Hitchin representation ρ and any nonidentity element X in Γ, we know that $\rho(X)$ is diagonalizable over $\mathbb{R}$ with eigenvalues that have pairwise distinct moduli. *For the rest of this paper, we will denote by $x^+, x^- \in \partial_\infty\Gamma$ the attracting and repelling fixed points, respectively, of $X \in \Gamma\setminus\{\mathrm{id}\}$.* With this notation, we now state the main theorem of this paper.

Theorem 1.1 *Let A, B be elements in Γ such that a^+, b^+, a^-, b^- lie in $\partial_\infty\Gamma$ in that cyclic order. Also, let $\rho \in \mathrm{Hit}_n(S)$ and let $\alpha_n < \cdots < \alpha_1$ and $\beta_n < \cdots < \beta_1$ be the moduli of the eigenvalues of $\rho(A)$ and $\rho(B)$, respectively. For every $k = 0, \dots, n-2$, the following hold:*

(1) $$\frac{\alpha_1}{\alpha_n} > \frac{\beta_{k+1}}{\beta_{k+1}-\beta_{k+2}}.$$

(2) *Let η and γ be closed curves in S corresponding to A and B, respectively, and let $i(\eta,\gamma)$ be the geometric intersection number between η and γ. If γ is simple, then*
$$\frac{\alpha_1}{\alpha_n} > \left(\frac{\beta_{k+1}}{\beta_{k+1}-\beta_{k+2}}\right)^u \cdot \left(\frac{\beta_{n-k-1}}{\beta_{n-k-1}-\beta_{n-k}}\right)^{i(\eta,\gamma)-u}$$
for some nonnegative integer $u \le i(\eta,\gamma)$ that is independent of k.

Observe that Theorem 1.1(2) does not depend on the choice of orientation on η or γ. We can also say what the constant u in Theorem 1.1(2) is. Choose orientations on η and γ and let $\hat{\imath}(\eta,\gamma)$ be the algebraic intersection number between η and γ. Then
$$u = \tfrac{1}{2}\big(i(\eta,\gamma) + |\hat{\imath}(\eta,\gamma)|\big).$$

In the setting of Hitchin representations, the width of a collar neighborhood is not well defined since Hitchin representations in general do not give a metric on S. However, for every Hitchin representation ρ, we do still have a natural notion of length for free homotopy classes of closed curves in S. Given any representation ρ in $\mathrm{Hit}_n(S)$ and any closed curve γ in S, we can define the *ρ–length* of γ to be
$$l_\rho(\gamma) = \log\Big(\frac{\lambda_1}{\lambda_n}\Big),$$

where λ_1 and λ_n are the largest and smallest moduli of the eigenvalues of $\rho(X)$, respectively, and $X \in \Gamma$ corresponds to the closed curve γ equipped with a choice of orientation. Observe that the ρ–length does not depend on the choice of orientation on γ or the choice of X, and is constant on each free homotopy class of closed curves in S.

If $\rho \in \mathrm{Hit}_2(S)$, then $l_\rho(\gamma)$ is exactly the hyperbolic length of the geodesic homotopic to γ, measured in the hyperbolic metric corresponding to ρ. Also, Choi and Goldman [7] proved that representations in $\mathrm{Hit}_3(S)$ are exactly holonomies of convex $\mathbb{RP}^2$ structures on S. Moreover, each such convex $\mathbb{RP}^2$ structure also induces a natural Finsler metric, known as the Hilbert metric, on S. One can then verify, in the case when $\rho \in \mathrm{Hit}_3(S)$, that $l_\rho(\gamma)$ is the length of the geodesic homotopic to γ, measured in the Hilbert metric induced by the convex $\mathbb{RP}^2$ structure corresponding to ρ.

With this notion of ρ–length, we have the following corollary of Theorem 1.1, which one can think of as a generalization of the collar lemma.

Corollary 1.2 *Let S be a surface of genus $g \geq 2$, and let η and γ be two essential closed curves in S. Then, for any $n \geq 2$ and any $\rho \in \mathrm{Hit}_n(S)$, the following hold:*

(1) *If $i(\eta, \gamma) \neq 0$, then*
$$\frac{1}{\exp(l_\rho(\eta))} < 1 - \frac{1}{\exp(l_\rho(\gamma)/(n-1))}.$$

(2) *If $i(\eta, \gamma) \neq 0$ and γ is simple, then there are nonnegative integers u and v with $u \geq v$ and $u + v = i(\eta, \gamma)$ such that*
$$\frac{1}{\exp(l_\rho(\eta))} < \left(1 - \frac{1}{\exp(l_\rho(\gamma)/(n-1))}\right)^u \left(1 - \frac{1}{\exp(l_\rho(\gamma))}\right)^v.$$

(3) *Let $\delta_n > 0$ be the unique real solution to the equation $e^{-x} + e^{-x/(n-1)} = 1$. If η is a nonsimple closed curve, then*
$$l_\rho(\eta) > \delta_n.$$

The quantity u in the above corollary is the same u as in Theorem 1.1. Observe that δ_n is an increasing unbounded sequence, and $\delta_2 = \log(2)$. Also, the expressions on the right hand side of the inequalities in parts (1) and (2) of Corollary 1.2 are maximized when $n = 2$. Hence, we can replace n by 2 in the right hand side of all three inequalities in Corollary 1.2, and they will still hold.

In the case of $\mathcal{T}(S)$, the first inequality in Corollary 1.2 can be rewritten as
$$\big(\exp(l_\rho(\eta)) - 1\big)\big(\exp(l_\rho(\gamma)) - 1\big) > 1.$$

This is weaker than a version of the classical collar lemma, which is the inequality

$$\sinh\big(\tfrac{1}{2}l_\rho(\eta)\big)\sinh\big(\tfrac{1}{2}l_\rho(\gamma)\big) > 1, \tag{1-1}$$

although in both inequalities, $l_\rho(\eta)$ grows logarithmically with $1/l_\rho(\gamma)$. In general, it is not known if the inequality (1-1) holds for all Hitchin components. However, it is a consequence of recent work of Tholozan [18] that it in fact holds for $\mathrm{Hit}_3(S)$. See Section 3.3 for more details.

Choi [6] proved an analog of the Margulis lemma for convex $\mathbb{RP}^2$ surfaces. As a consequence, he showed the existence of a collar neighborhood in the convex $\mathbb{RP}^2$ surface about a simple closed curve of sufficiently short length, and found (nonexplicit) lower bounds for the width of this collar neighborhood in terms of the length of the simple closed curve. This analog of the Margulis lemma was later extended by Cooper, Long and Tillman [8] to all convex real projective manifolds. Burger and Pozzetti [4] also recently proved a statement analogous to Theorem 1.1 for maximal representations into $\mathrm{PSp}(2k,\mathbb{R})$.

The image of the irreducible representation $\iota_n\colon \mathrm{PSL}(2,\mathbb{R}) \to \mathrm{PSL}(n,\mathbb{R})$ lies in a conjugate of the subgroup $\mathrm{PSO}(k,k+1) \subset \mathrm{PSL}(2k+1,\mathbb{R})$ when $n = 2k+1$, and a conjugate of $\mathrm{PSp}(2k,\mathbb{R}) \subset \mathrm{PSL}(2k,\mathbb{R})$ when $n = 2k$. Hence, we can define Hitchin components in

$$\mathrm{Hom}(\Gamma,\mathrm{PSO}(k,k+1))/\,\mathrm{PSO}(k,k+1), \quad \mathrm{Hom}(\Gamma,\mathrm{PSp}(2k,\mathbb{R}))/\,\mathrm{PSp}(2k,\mathbb{R})$$

in the same way as we did for $\mathrm{PSL}(n,\mathbb{R})$. Denote these Hitchin components by $\mathrm{Hit}_n(S)'$. Since the image of ι_7 in particular lies in the exceptional Lie group $G_2 \subset \mathrm{PSO}(3,4)$, we can also define a Hitchin component $\mathrm{Hit}(S,G_2)$ in $\mathrm{Hom}(\Gamma,G_2)/G_2$. Note that $\mathrm{Hit}_n(S)'$ and $\mathrm{Hit}(S,G_2)$ can be naturally identified with a subset of $\mathrm{Hit}_n(S)$ and $\mathrm{Hit}_7(S)'$, respectively. In the case when $\rho \in \mathrm{Hit}_n(S)$ happens to be an element of $\mathrm{Hit}_n(S)'$, we can strengthen Theorem 1.1(2), which we state as the following corollary.

Corollary 1.3 *Let A and B be elements in Γ such that a^+, b^+, a^-, b^- lie in $\partial_\infty\Gamma$ in that cyclic order. Let $\rho \in \mathrm{Hit}_n(S)'$ and let $\alpha_n < \cdots < \alpha_1$ and $\beta_n < \cdots < \beta_1$ be the moduli of the eigenvalues of $\rho(A)$ and $\rho(B)$, respectively. Finally, let η and γ be closed curves on S corresponding to A and B, respectively. If γ is simple, then for every $k = 0,\dots,n-2$,*

$$\alpha_1^2 > \left(\frac{\beta_{k+1}}{\beta_{k+1}-\beta_{k+2}}\right)^{i(\eta,\gamma)}.$$

Hitchin representations into $\mathrm{PSp}(2k,\mathbb{R})$ are special examples of maximal representations. In this case, the inequality in the above corollary is stronger than the one given by Burger and Pozzetti [4].

The proof of our results relies heavily on the seminal work of Labourie [15], who showed that every Hitchin representation into $\mathrm{PSL}(n,\mathbb{R})$ (and hence into $\mathrm{PSp}(2k,\mathbb{R})$, $\mathrm{PSO}(k,k+1)$ and G_2) naturally comes with an equivariant Frenet curve; see Theorem 2.5. While Hitchin representations can be defined for any split real group, properties of the limit curve of these Hitchin representations are still poorly understood in general. As such, we are unable to generalize our techniques to prove an analog of Theorem 1.1 for Hitchin representations into split real groups other than $\mathrm{PSL}(n,\mathbb{R})$, $\mathrm{PSp}(2k,\mathbb{R})$, $\mathrm{PSO}(k,k+1)$ and G_2.

Unfortunately, for $\rho\in \mathrm{Hit}_n(S)$ when $n\geq 4$, it is not known whether there exists a metric on S that induces l_ρ as its length function. However, we can still interpret Corollary 1.2 geometrically by considering the $\mathrm{PSL}(n,\mathbb{R})$ symmetric space $\widetilde{M}$. Normalize the Riemannian metric on $\widetilde{M}$ so that for any $Z\in\mathrm{PSL}(n,\mathbb{R})$ with real eigenvalues,

$$\inf\{d_{\widetilde{M}}(o,Z\cdot o):o\in\widetilde{M}\}=\sqrt{2\sum_{i=1}^{n}(\log\lambda_i)^2},$$

where $\lambda_1,\dots,\lambda_n$ are the moduli of the eigenvalues of Z and $d_{\widetilde{M}}$ is the distance function on $\widetilde{M}$ induced by the normalized Riemannian metric. Let $M:=\rho(\Gamma)\backslash\widetilde{M}$, and for any closed curve ω in M, let $l_M(\omega)$ be the length of ω measured in the Riemannian metric on M induced by the normalized Riemannian metric on $\widetilde{M}$. Then the following corollary holds.

Corollary 1.4 *Let η and γ be two essential closed curves in S and let X and Y be elements in Γ corresponding to η and γ, respectively. For any $\rho\in\mathrm{Hit}_n(S)$, let η' and γ' be two closed curves in M that correspond to $X,Y\in\Gamma$, respectively. Then the statements in Corollary 1.2 hold, with $l_\rho(\eta)$ and $l_\rho(\gamma)$ replaced by $l_M(\eta')$ and $l_M(\gamma')$, respectively.*

It is an important remark that this corollary (and hence Corollary 1.2) is not simply a quantitative version of the Margulis lemma on $\mathrm{PSL}(n,\mathbb{R})$ because the closed curves η' and γ' do not need to intersect, even when $i(\eta,\gamma)\neq 0$.

Theorem 1.1 is a property that is special to Hitchin representations. In fact, for any pair of simple closed curves in S, one can find a sequence of quasi-Fuchsian representations

$$\rho_i\colon \Gamma\to\mathrm{PSO}(3,1)^+\subset\mathrm{PSL}(4,\mathbb{R})$$

such that the lengths of the geodesics in $\rho_i(\Gamma)\backslash\widetilde{M}$ corresponding to both of these two simple closed curves converge to 0 along this sequence. In particular, Corollary 1.2 does not hold on the space of quasi-Fuchsian representations. This is explained in greater detail in Section 3.2.

Theorem 1.1 can also be generalized to the setting where we allow S to be compact but not necessarily closed; see Corollary 3.4.

As a final consequence of Theorem 1.1, we have the following properness result.

Corollary 1.5 *Let $\mathcal{C} := \{\gamma_1, \dots, \gamma_k\}$ be a collection of closed curves in S that contains a pants decomposition, such that the complement of $\mathcal{C}$ in S is a union of discs. Then the map*

$$\mathrm{Hit}_n(S) \to \mathbb{R}^k, \quad \rho \mapsto (l_\rho(\gamma_1), \dots, l_\rho(\gamma_k)),$$

is proper.

In other words, in order for a sequence $\{\rho_i\}_{i=1}^\infty$ in $\mathrm{Hit}_n(S)$ to escape, the ρ_i–length of some curve in $\mathcal{C}$ must grow to ∞. We will give the proof of this corollary in the Appendix because it uses some technical results from Zhang [19]. Refer to Section 3.1 for more corollaries of Theorem 1.1.

Acknowledgements This work started as an attempt to answer a question that Scott Wolpert asked the second author: does a collar lemma exist for convex $\mathbb{RP}^2$ surfaces? We are grateful to him for asking the question. We also thank Daniele Alessandrini, Ara Basmajian, Richard Canary, François Labourie and Anna Wienhard for many useful conversations with the authors. Finally, we would like to thank the referees for carefully reading this paper and suggesting several improvements.

Lee was supported by the DFG research grant "Higher Teichmüller Theory". Zhang was partially supported by US National Science Foundation grants DMS 1006298, DMS 1306992 and DMS 1307164. The authors acknowledge support from US National Science Foundation grants DMS 1107452, 1107263, 1107367 "RNMS: GEometric structures And Representation varieties" (the GEAR Network).

2 Proof of Theorem 1.1

We start this section by discussing some useful topological properties of Γ and its boundary in Section 2.1. Then for the sake of demonstrating the proof without too many technical details, we prove Theorem 1.1(1) for the special case $\mathrm{Hit}_3(S)$ in Section 2.2. Next, we develop the technical tools that we need in Section 2.3, and apply them in Section 2.4 to prove Theorem 1.1 in its full generality.

2.1 Properties of the boundary of the group

It is well known that Γ is Gromov hyperbolic, so the Cayley graph of Γ has a natural boundary, which we denote by $\partial_\infty\Gamma$, and the action of Γ on its Cayley graph extends to an action on $\partial_\infty\Gamma$. Moreover, if we choose $\rho \in \mathcal{T}(S)$, ie a hyperbolic structure on S, we get a ρ–equivariant identification of $\partial_\infty\Gamma$ with the boundary $\partial\mathbb{H}^2$ of the hyperbolic plane $\mathbb{H}^2$.

For any hyperbolic element $A \in \mathrm{PSL}(2,\mathbb{R})$, the *axis* of A, which we denote by L_A, is the unique geodesic in $\mathbb{H}^2$ whose endpoints are the repelling and attracting fixed points of A in $\partial\mathbb{H}^2$. The proof of the main theorem relies crucially on an important property of the action of Γ on $\partial_\infty\Gamma$, which we state as Lemma 2.2. These are well-known facts about surface groups, but for lack of a good reference, we will give the proof here.

Lemma 2.1 *Let B and B' be noncommuting elements in $\mathrm{PSL}(2,\mathbb{R})$ that generate a subgroup consisting only of hyperbolic isometries. If the translation lengths of B and B' are the same and $L_{B'} \cap L_B = \varnothing$, then $(B \cdot L_{B'}) \cap L_{B'} = \varnothing$.*

Proof Since B and B' do not commute, $L_B \neq L_{B'}$. Since the commutator $[B, B']$ is not parabolic, B and B' cannot share a fixed point. Hence, by changing coordinates and replacing B and B' with their inverses if necessary, we can assume that L_B and $L_{B'}$ are as in Figure 1, and that B and B' translate along their axes in the directions drawn.

Let L be the geodesic in $\mathbb{H}^2$ that is perpendicular to both $L_{B'}$ and L_B, and let R be the reflection about L. There is a unique geodesic K that is perpendicular to L_B and whose distances to L and $B \cdot L$ are equal. Let S be the reflection about K, and note that $B = SR$. Also, observe that the distance between K and L is realized only by the points $K \cap L_B$ and $L \cap L_B$, and is half the translation length of B, which we denote by T. Furthermore, $(B \cdot L_{B'}) \cap L_{B'} = (SR \cdot L_{B'}) \cap L_{B'} = (S \cdot L_{B'}) \cap L_{B'}$ is empty if and only if $K \cap L_{B'}$ is empty.

Thus, it is sufficient to show that $K \cap L_{B'}$ is empty. Suppose for contradiction that it is not. As before, there is a unique geodesic K' such that $B' = S'R$, where S' is the reflection about K'. Since the translation lengths of B and B' are the same, the symmetry between B and B' ensures that $K' \cap L_B$ is also nonempty.

Now, note that $K' \cap L_{B'}$ lies between $K \cap L_{B'}$ and $L \cap L_{B'}$ because

$$d(K \cap L_{B'}, L \cap L_{B'}) > d(K \cap L_B, L \cap L_B) = \tfrac{1}{2}T = d(K' \cap L_{B'}, L \cap L_{B'}).$$

Similarly, $K \cap L_B$ lies between $K' \cap L_B$ and $L \cap L_B$. This implies that K and K' have a common point of intersection, p; see Figure 1. Observe that $B'B^{-1} = S'RR^{-1}S^{-1} = S'S$ fixes p, but that is impossible because $B'B^{-1}$ is not elliptic. □

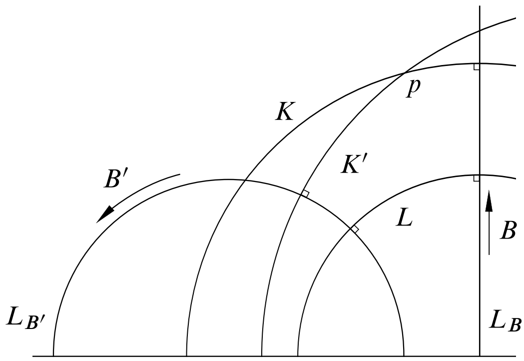


Figure 1: An impossible configuration of K and K' in Lemma 2.1

Lemma 2.2 *Let A, B and B' be pairwise noncommuting elements in Γ such that B and B' are conjugate. If*

$$a^+,\ b'^+,\ b^+,\ a^-,\ b^-,\ b'^-$$

lie in $\partial_\infty\Gamma$ in that cyclic order, then

$$a^+,\ b'^+,\ B\cdot a^+,\ b^+,\ a^-,\ b^-,\ B^{-1}\cdot a^+,\ b'^-$$

lie in $\partial_\infty\Gamma$ in that cyclic order; see Figure 2.

Proof Let s_0 be the open subsegment of $\partial_\infty\Gamma$ with endpoints b'^- and b^+ that does not contain b^-, and let s_1 be the open subsegment of $\partial_\infty\Gamma$ with endpoints b'^+ and b^+ that does not contain b^-. Observe that $B\cdot b'^-$ lies in s_0 and $B\cdot b'^+$ lies in s_1.

Choose a hyperbolic metric on S. This identifies $\partial_\infty\Gamma$ with $\partial\mathbb{H}^2$ and Γ with a discrete, torsion-free subgroup of $\mathrm{PSL}(2,\mathbb{R})$. Since

$$a^+,\ b'^+,\ b^+,\ a^-,\ b^-,\ b'^-$$

lie in $\partial_\infty\Gamma$ in that cyclic order, L_B and $L_{B'}$ have to be disjoint. Moreover, B and B' have the same translation lengths and do not commute. Hence, we can apply Lemma 2.1 to conclude that $B\cdot L_{B'}$ and $L_{B'}$ are disjoint. This implies that both $B\cdot b'^-$ and $B\cdot b'^+$ have to lie in s_1. Since a^+ lies in s_0 between b'^- and b'^+, we have that $B\cdot a^+$ must lie in s_1 between $B\cdot b'^-$ and $B\cdot b'^+$. In particular,

$$a^+,\ b'^+,\ B\cdot a^+,\ b^+,\ a^-$$

lie in $\partial_\infty\Gamma$ in that cyclic order; see Figure 2.

A similar argument, using B^{-1} instead of B, shows that

$$a^-,\ b^-,\ B^{-1}\cdot a^+,\ b'^-,\ a^+$$

lie in $\partial_\infty\Gamma$ in that cyclic order. This proves the lemma. □

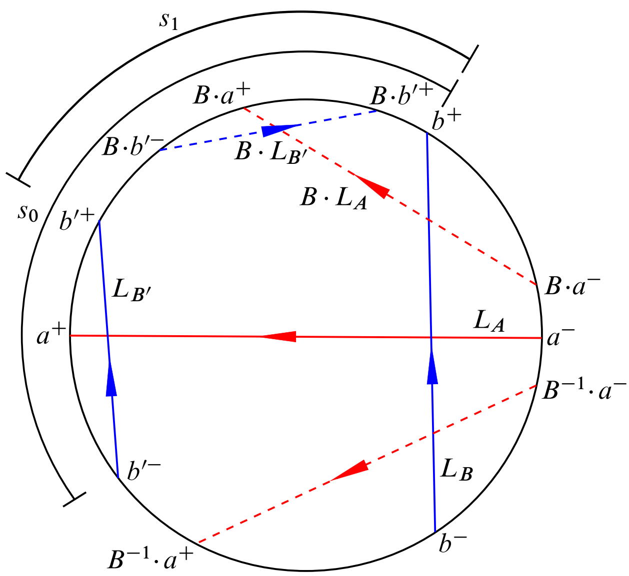


Figure 2: The cyclic order of the attracting and repelling fixed points of A, B, B' and BAB^{-1} along $\partial_\infty \Gamma$ in Lemma 2.2

2.2 Proof in the $\mathrm{PSL}(3,\mathbb{R})$ case

In order to demonstrate the main ideas of the proof without involving too many technicalities, we will first prove Theorem 1.1(1) in the special case when $n = 3$, ie $\rho\colon \Gamma \to \mathrm{PSL}(3,\mathbb{R}) = \mathrm{SL}(3,\mathbb{R})$ is a Hitchin representation.

By Choi and Goldman [7], we know that in this case, ρ is the holonomy of a convex $\mathbb{RP}^2$ structure on S. In other words, there is a strictly convex domain Ω_ρ in $\mathbb{RP}^2$ which is preserved by the Γ–action on $\mathbb{RP}^2$ induced by ρ, and on which the Γ–action is properly discontinuous and cocompact. Moreover, $\rho(X)$ is diagonalizable with positive pairwise distinct eigenvalues for any nonidentity element $X \in \Gamma$ (see Goldman [11, Theorem 3.2]), so $\rho(X)$ has an attracting and repelling fixed point in $\partial\Omega_\rho$. Since the Hilbert metric in Ω_ρ is invariant under projective transformations and the geodesics of the Hilbert metric are lines, one can use the Švarc–Milnor lemma [3, Proposition 8.19] to construct a continuous map

$$\xi^{(1)}\colon \partial_\infty \Gamma \to \partial\Omega_\rho$$

which identifies the attracting fixed point of any $X \in \Gamma \backslash \{\mathrm{id}\}$ to the attracting fixed point of $\rho(X)$.

Pick any four projective lines in $\mathbb{RP}^2$ that intersect at a common point, such that no three of the four agree. There is a classical projective invariant of these four projective

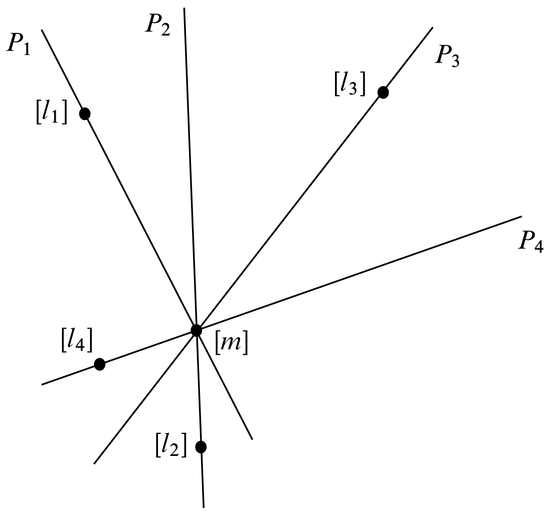


Figure 3: A choice of vectors l_i to compute the cross ratio (P_1, P_2, P_3, P_4)

lines, called the *cross ratio*, which can be defined as follows. Let the four projective lines be P_1, P_2, P_3, P_4 and let m be a vector in $\mathbb{R}^3$ such that $[m]$, the projective point corresponding to the $\mathbb{R}$–span of m, is the common point of intersection of the P_i. For each i, choose a vector $l_i \in \mathbb{R}^3$ so that $[l_i] \neq [m]$ and $[l_i]$ lies in P_i; see Figure 3. By choosing a linear identification

$$f\colon \textstyle\bigwedge^3 \mathbb{R}^3 \to \mathbb{R},$$

we can evaluate the expression

$$(P_1, P_2, P_3, P_4) := \frac{m \wedge l_1 \wedge l_3}{m \wedge l_1 \wedge l_2} \cdot \frac{m \wedge l_4 \wedge l_2}{m \wedge l_4 \wedge l_3}$$

as an extended real number. One can then verify that the cross ratio (P_1, P_2, P_3, P_4) does not depend on the choice of m, l_1, l_2, l_3, l_4 or the choice of identification f.

This definition of the cross ratio agrees with the classical notion of the cross ratio of four points on a line in the following way. By taking the dual, the four lines $P_1, \dots, P_4$ become four points $p_1, \dots, p_4 \in (\mathbb{RP}^2)^*$, and they lie in the projective line in $(\mathbb{RP}^2)^*$ that is dual to the point $[m]$ in $\mathbb{RP}^2$. One can then check that (P_1, P_2, P_3, P_4) is exactly the cross ratio of the four collinear points $p_1, \dots, p_4$.

Proof of Theorem 1.1(1) when $n = 3$ Observe that

$$a^+,\ A \cdot b^+,\ b^+,\ a^-,\ b^-,\ A \cdot b^-$$

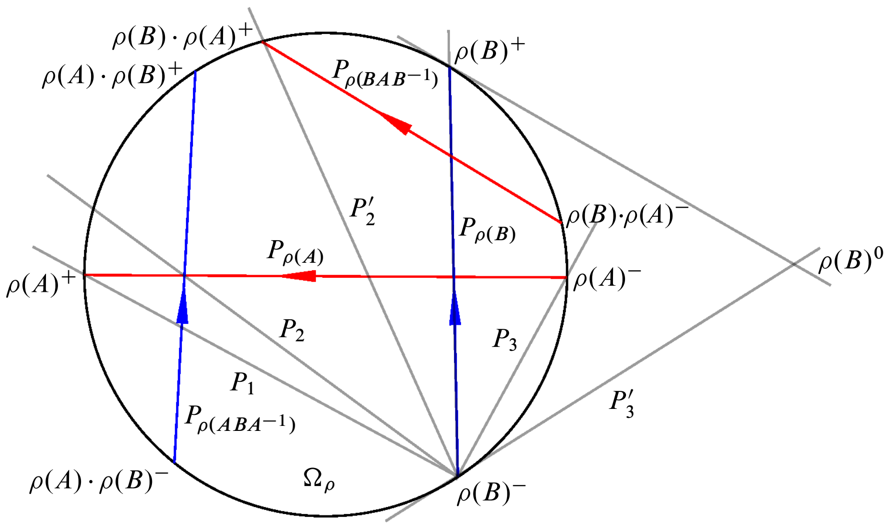


Figure 4: A schematic for the comparison between the cross ratios $(P_1, P_2, P_{\rho(B)}, P_3)$ and $(P_1, P_2', P_{\rho(B)}, P_3')$

lie in $\partial_\infty\Gamma$ in that cyclic order. By Lemma 2.2, we see that

$$a^+,\ A\cdot b^+,\ B\cdot a^+,\ b^+,\ a^-,\ b^-,\ B^{-1}\cdot a^+,\ A\cdot b^-$$

lie in $\partial_\infty\Gamma$ in that cyclic order, because $A\cdot b^+$ and $A\cdot b^-$ are the attracting and repelling fixed points of ABA^{-1}, respectively.

Choose any $\rho\in\mathrm{Hit}_3(S)$. For any nonidentity element $X\in\Gamma$, let $\rho(X)^+$, $\rho(X)^0$ and $\rho(X)^-$ be the three fixed points for $\rho(X)$, where $\rho(X)^+$ is attracting and $\rho(X)^-$ is repelling. Denote by $P_{\rho(X)}$ the line segment in Ω_ρ with endpoints $\rho(X)^+$ and $\rho(X)^-$.

Now, let

- P_1 be the line through $\rho(B)^-$ and $\rho(A)^+$,
- P_2 be the line through $\rho(B)^-$ and $P_{\rho(A)}\cap P_{\rho(ABA^{-1})}$,
- P_3 be the line through $\rho(B)^-$ and $\rho(A)^-$,
- P_2' be the line through $\rho(B)^-$ and $\rho(B)\cdot\rho(A)^+$,
- P_3' be the line through $\rho(B)^-$ and $\rho(B)^0$.

By using $\xi^{(1)}$ to identify $\partial_\infty\Gamma$ with $\partial\Omega_\rho$, we have that

$$\rho(A)^+,\ \rho(A)\cdot\rho(B)^+,\ \rho(B)\cdot\rho(A)^+,\ \rho(B)^+,\ \rho(A)^-,\ \rho(B)^-$$

lie in $\partial\Omega_\rho$ in that cyclic order; see Figure 4. It is a classically known property of the cross ratio (see Proposition 2.10) that

$$(P_1, P_2, P_{\rho(B)}, P_3) > (P_1, P_2', P_{\rho(B)}, P_3').$$

It is an easy cross ratio computation (see Lemmas 2.8 and 2.9) that

$$(P_1, P_2', P_{\rho(B)}, P_3') = \frac{\beta_1}{\beta_1 - \beta_2} \quad \text{and} \quad (P_1, P_2, P_{\rho(B)}, P_3) = \frac{\alpha_1}{\alpha_3}.$$

Hence, we have

$$\frac{\alpha_1}{\alpha_3} > \frac{\beta_1}{\beta_1 - \beta_2}.$$

Similarly, by reversing the roles of $\rho(B)^-$ and $\rho(B)^+$, and using $\rho(B)^{-1}$ in place of $\rho(B)$, we can also show that

$$\frac{\alpha_1}{\alpha_3} > \frac{\beta_2}{\beta_2 - \beta_3}.$$

This proves Theorem 1.1(1) in the case when $n = 3$. $\square$

2.3 Properties of Frenet curves of Hitchin representations

Next, we want to generalize the proof given in Section 2.2 to any Hitchin representation. We will devote this section to developing the tools needed to do so. In the rest of the paper, we use the same notation for points in $\mathbb{RP}^{n-1}$ and for lines in $\mathbb{R}^n$. It should be clear to which we are referring from the context.

Denote by $\mathcal{F}(\mathbb{R}^n)$ the space of complete flags in $\mathbb{R}^n$. Labourie [15] and Guichard [12] gave a beautiful characterization of representations in $\mathrm{Hit}_n(S)$ as representations that admit an equivariant Frenet curve $\partial_\infty\Gamma \to \mathcal{F}(\mathbb{R}^n)$. When $n = 3$, the Frenet curve, postcomposed with the projection from $\mathcal{F}(\mathbb{R}^3)$ to $\mathbb{RP}^2$, is exactly the map $\xi^{(1)}\colon \partial_\infty\Gamma \to \partial\Omega_\rho$ described in Section 2.2. This characterization will be the main tool we use to extend our proof in Section 2.2 to the general case.

We will start by first defining the Frenet property.

Notation 2.3 Let $\xi\colon S^1 \to \mathcal{F}(\mathbb{R}^n)$ be a continuous closed curve and denote the Grassmannian of k–dimensional subspaces of $\mathbb{R}^n$ by $\mathrm{Gr}(k,n)$. For any $k = 1, \dots, n-1$ and any point $x \in S^1$, let $\xi(x)^{(k)} := \pi_k(\xi(x))$, where $\pi_k\colon \mathcal{F}(\mathbb{R}^n) \to \mathrm{Gr}(k,n)$ is the obvious projection.

Definition 2.4 A closed curve $\xi\colon S^1 \to \mathcal{F}(\mathbb{R}^n)$ is *Frenet* if for every set of distinct points $x_1, \dots, x_k$ in S^1, for every $x \in S^1$, and for all positive integers $n_1, \dots, n_k$ such that $m := \sum_{i=1}^{k} n_i \le n$,

$$\dim \sum_{i=1}^{k} \xi(x_i)^{(n_i)} = m \quad \text{and} \quad \lim_{\substack{x_i \to x, \forall i \\ x_i \neq x_j, \forall i \neq j}} \sum_{i=1}^{k} \xi(x_i)^{(n_i)} = \xi(x)^{(m)}.$$

The Frenet property ensures ξ has good continuity properties and is "maximally transverse". Combining the work of Labourie [15, Theorem 1.4] and Guichard [12, théorème 1], one can characterize the representations in the $\mathrm{Hit}_n(S)$ as those that preserve an equivariant Frenet curve.

Theorem 2.5 (Guichard, Labourie) *A representation ρ in*

$$\mathrm{Hom}(\Gamma, \mathrm{PSL}(n, \mathbb{R})) / \mathrm{PSL}(n, \mathbb{R})$$

lies in $\mathrm{Hit}_n(S)$ if and only if there exists a ρ–equivariant Frenet curve $\xi\colon \partial_\infty \Gamma \to \mathcal{F}(\mathbb{R}^n)$. If ξ exists, then it is unique.

We will now prove several properties of these Frenet curves that will be needed. These are special cases of more general properties that appear in Section 2 of Zhang [19]. However, for the sake of completeness, we will reproduce the proofs.

Lemma 2.6 *Let a, m_0, b, m_1, m_2 and m_3 be distinct points on $\partial_\infty \Gamma$ in that cyclic order, and let $\rho \in \mathrm{Hit}_n(S)$ with corresponding Frenet curve ξ. Also, let $P := \mathbb{P}\big(\xi(a)^{(1)} + \xi(b)^{(1)}\big)$. Then the following hold:*

(1) *Let k_0, k_1, k_2 and k_3 be nonnegative integers that sum to $n-2$, and let $M := \sum_{i=0}^{3} \xi(m_i)^{(k_i)}$. The map*

$$f_M\colon \partial_\infty \Gamma \to P$$

given by

$$f_M\colon x \mapsto \begin{cases} \mathbb{P}\big(\xi(x)^{(1)} + \sum_{i=0}^{3} \xi(m_i)^{(k_i)}\big) \cap P & \text{if } x \neq m_j, \\ \mathbb{P}\big(\xi(m_j)^{(k_j+1)} + \sum_{i \neq j} \xi(m_i)^{(k_i)}\big) \cap P & \text{if } x = m_j, \end{cases}$$

is a homeomorphism with $f_M(a) = \xi(a)^{(1)}$ and $f_M(b) = \xi(b)^{(1)}$.

(2) *Let k_0, k_1 and k_2 be nonnegative integers that sum to $n-1$, and let s be the closed subsegment of $\partial_\infty \Gamma$ with endpoints a and b that does not contain m_0.*

Also, let $M := \xi(m_0)^{(k_0)}$. Then there is some closed subsegment ω of P with endpoints $\xi(a)^{(1)}$ and $\xi(b)^{(1)}$ such that the map

$$g_M\colon s \to \omega$$

given by

$$g_M\colon x \mapsto \begin{cases} \mathbb{P}\big(\xi(x)^{(k_2)} + \xi(m_1)^{(k_1)} + \xi(m_0)^{(k_0)}\big) \cap P & \text{if } x \neq m_1, \\ \mathbb{P}\big(\xi(m_1)^{(k_1+k_2)} + \xi(m_0)^{(k_0)}\big) \cap P & \text{if } x = m_1, \end{cases}$$

is a homeomorphism with $g_M(a) = \xi(a)^{(1)}$ and $g_M(b) = \xi(b)^{(1)}$.

Proof Before we start the proof, observe that for any nonnegative integers $t_0, \dots, t_4$ such that $\sum_{i=0}^4 t_i = n-1$, the intersection $\mathbb{P}\big(\xi(x)^{(t_4)} + \sum_{i=0}^3 \xi(m_i)^{(t_i)}\big) \cap P$ is a single point; otherwise, $\mathbb{P}\big(\xi(a)^{(1)} + \xi(b)^{(1)}\big) \subset \mathbb{P}\big(\xi(x)^{(t_4)} + \sum_{i=0}^3 \xi(m_i)^{(t_i)}\big)$, which contradicts the Frenet property of ξ.

(1) Since ξ is Frenet, f_M is continuous. Moreover, because the domain and target of f_M are both topologically a circle, it is sufficient to show that f_M is injective. Suppose for contradiction that there exist $x \neq x'$ such that $f_M(x) = f_M(x')$. We will assume that $x, x' \neq m_i$ for all $i = 0, 1, 2, 3$ as the other cases are similar. Then

$$\begin{aligned} \sum_{i=0}^3 \xi(m_i)^{(k_i)} + \xi(x)^{(1)} &= \sum_{i=0}^3 \xi(m_i)^{(k_i)} + f_M(x) \\ &= \sum_{i=0}^3 \xi(m_i)^{(k_i)} + f_M(x') \\ &= \sum_{i=0}^3 \xi(m_i)^{(k_i)} + \xi(x')^{(1)}, \end{aligned}$$

which is impossible because ξ is Frenet. The fact that $f_M(a) = \xi(a)^{(1)}$ and $f_M(b) = \xi(b)^{(1)}$ is easily verified.

(2) First, observe that g_M viewed as a map from s to P is continuous. Also, for any x in s, we have that $g_M(x) = \xi(a)^{(1)}$ if and only if $x = a$ and $g_M(x) = \xi(b)^{(1)}$ if and only if $x = b$. This proves that the image of g_M is a subsegment ω of P with endpoints $\xi(a)^{(1)}$ and $\xi(b)^{(1)}$.

To finish the proof, we only need to show that g_M is injective. Choose x and x' in the interior of s with $x \neq x'$, and assume without loss of generality that a, x', x and b lie along s in that order. Again, we assume that $x, x' \neq m_1$ as the other cases are similar. For any positive integer $i \leq k_2$, let

$$M_i := \xi(x)^{(i-1)} + \xi(x')^{(k_2-i)} + \xi(m_1)^{(k_1)} + \xi(m_0)^{(k_0)}.$$

By (1), we know that $f_{M_i}(x)$ lies on ω strictly between $f_{M_i}(x')$ and $f_{M_i}(b) = \xi(b)^{(1)}$. Also, observe that $f_{M_i}(x) = f_{M_{i+1}}(x')$. This implies that $f_{M_{k_2}}(x)$ lies on ω strictly between $f_{M_1}(x')$ and $\xi(b)^{(1)}$. In particular, $g_M(x) = f_{M_{k_2}}(x) \neq f_{M_1}(x') = g_M(x')$, so g_M is injective. □

In the proof of the $n = 3$ case given in Section 2.2, the classical cross ratio in $\mathbb{RP}^2$ was the main computational tool used to obtain our estimates. We will now define a generalization of the cross ratio for $\mathbb{RP}^{n-1}$.

Definition 2.7 Let $P_1, \dots, P_4$ be four hyperplanes in $\mathbb{R}^n$ that intersect along a $(n-2)$–dimensional subspace $M = \mathrm{Span}\{m_1, \dots, m_{n-2}\} \subset \mathbb{R}^n$, such that no three of the four P_i agree. For $i = 1, \dots, 4$, let $L_i = [l_i]$ be a line through the origin in P_i that does not lie in M. Define the *cross ratio* by

$$(P_1, P_2, P_3, P_4) := \frac{m_1 \wedge \cdots \wedge m_{n-2} \wedge l_1 \wedge l_3}{m_1 \wedge \cdots \wedge m_{n-2} \wedge l_1 \wedge l_2} \cdot \frac{m_1 \wedge \cdots \wedge m_{n-2} \wedge l_4 \wedge l_2}{m_1 \wedge \cdots \wedge m_{n-2} \wedge l_4 \wedge l_3}.$$

In the above definition, choose an identification between $\bigwedge^n(\mathbb{R}^n)$ and $\mathbb{R}$ to evaluate the fraction on the right as a real number. One can check that this number does not depend on the identification chosen, the choice of basis $\{m_1, \dots, m_{n-2}\}$ for M, the choice of L_i in P_i, or the choice of representatives l_i for L_i. When convenient, we sometimes use the notation

$$(L_1, L_2, L_3, L_4)_M := (P_1, P_2, P_3, P_4).$$

Also, at times, in our notation for the cross ratio, we replace the subspaces L_i, P_i and M of $\mathbb{R}^n$ with their projectivizations. As with the $n = 3$ case, this definition of the cross ratio agrees with the classical cross ratio of four points along a projective line in $(\mathbb{RP}^{n-1})^*$.

The following two lemmas summarize some basic properties of this cross ratio.

Lemma 2.8 *Let $L_1, \dots, L_5$ be pairwise distinct lines in $\mathbb{R}^n$ through 0 and let M and M' be $(n-2)$–dimensional subspaces of $\mathbb{R}^n$ not containing L_i for any $i = 1, \dots, 5$, such that no three of the five $M + L_i$ agree and no three of the five $M' + L_i$ agree.*

(1) $(X \cdot L_1, \dots, X \cdot L_4)_{X \cdot M} = (L_1, \dots, L_4)_M$ *for any* $X \in \mathrm{PSL}(n, \mathbb{R})$.

(2) *If* L_1, L_2, L_3, L_4 *lie in a plane, then* $(L_1, L_2, L_3, L_4)_M = (L_1, L_2, L_3, L_4)_{M'}$.

(3) $(L_1, L_2, L_3, L_4)_M = (L_4, L_3, L_2, L_1)_M$.

(4) $(L_1, L_2, L_3, L_5)_M \cdot (L_1, L_3, L_4, L_5)_M = (L_1, L_2, L_4, L_5)_M$.

(5) $(L_1, L_2, L_3, L_4)_M \cdot (L_1, L_3, L_2, L_4)_M = 1$.

(6) $(L_1, L_2, L_3, L_4)_M = 1 - (L_1, L_2, L_4, L_3)_M$.

Proof (1), (3), (4) and (5) follow immediately from the definition of the cross ratio. To prove (2), observe that there is a projective transformation X that fixes L_1, L_2 and L_3, and maps M to M'. Since L_4 lies in the plane containing L_1, L_2 and L_3, X must also fix L_4. This allows us to use (1) to get (2).

To prove (6), assume that $M + L_1, \dots, M + L_4$ are distinct; the other cases are similar. Choose a basis $e_1, \dots, e_n$ for $\mathbb{R}^n$ so that

$$M = \operatorname{Span}\{e_1, \dots, e_{n-2}\}, \quad L_1 = [e_{n-1}], \quad L_4 = [e_n], \quad L_2 = \left[\sum_{i=1}^{n} e_i\right], \quad L_3 = \left[\sum_{i=1}^{n} \alpha_i e_i\right]$$

for some real numbers $\alpha_1, \dots, \alpha_n$. The assumption that $M + L_1, \dots, M + L_4$ are pairwise distinct implies that α_{n-1} and α_n are nonzero real numbers. One can then easily compute that

$$(L_1, L_2, L_3, L_4)_M = \frac{\alpha_n}{\alpha_{n-1}} \quad \text{and} \quad (L_1, L_2, L_4, L_3)_M = \frac{\alpha_{n-1} - \alpha_n}{\alpha_{n-1}}. \qquad \square$$

In view of Lemma 2.8(2), we will denote $(L_1, L_2, L_3, L_4)_M$ by (L_1, L_2, L_3, L_4) in the case when L_1, L_2, L_3 and L_4 lie in the same plane.

Lemma 2.9 *Let $X \in \mathrm{PSL}(n, \mathbb{R})$ be diagonalizable with n real eigenvalues $\lambda_1, \dots, \lambda_n$ (these are only well defined up to sign) of pairwise distinct moduli, such that $|\lambda_n| < \cdots < |\lambda_1|$. Let L_i and L_j be fixed lines through the origin in $\mathbb{R}^n$ corresponding to the eigenvalues λ_i and λ_j, respectively, with $i < j$, and let L be a line through the origin in the plane $L_i + L_j$ such that $L_i \neq L \neq L_j$. Then*

$$(L_j, L, X \cdot L, L_i) = \frac{\lambda_i}{\lambda_j}.$$

Proof Choose a basis $e_1, \dots, e_n$ for $\mathbb{R}^n$ so that $[e_k]$ is a fixed line through the origin of $\rho(X)$ corresponding to the eigenvalue λ_k. In this basis, $\rho(X)$ is the diagonal matrix $[x_{u,v}]$, where

$$x_{u,v} = \begin{cases} 0 & \text{if } u \neq v, \\ \lambda_u & \text{if } u = v. \end{cases}$$

Let M be the $(n-2)$–dimensional subspace $\operatorname{Span}\{e_1, \dots, \widehat{e_i}, \dots, \widehat{e_j}, \dots, e_n\}$ of $\mathbb{R}^n$. Via a projective transformation that fixes $e_1, \dots, e_n$, we can assume $L = [e_i + e_j]$. The lemma follows from an easy computation using the cross ratio definition. $\square$

The next task is to understand how the cross ratio interacts with Frenet curves.

Proposition 2.10 *Let $\rho \in \mathrm{Hit}_n(S)$, and let ξ be the corresponding Frenet curve. Also, let a, b, c, m_0, d and m_1 be distinct points along $\partial_\infty \Gamma$ in that cyclic order, and let k_0 and k_1 be nonnegative integers that sum to $n-2$. For any $x \in \partial_\infty \Gamma$, define*

$$P_x = \begin{cases} \xi(x)^{(1)} + \xi(m_0)^{(k_0)} + \xi(m_1)^{(k_1)} & \text{if } x \neq m_0, m_1, \\ \xi(m_i)^{(k_i+1)} + \xi(m_{1-i})^{(k_{1-i})} & \text{if } x = m_i. \end{cases}$$

Then the following hold:

(1) $(P_a, P_b, P_{m_0}, P_d) > (P_a, P_b, P_{m_0}, P_{m_1})$.

(2) $(P_a, P_b, P_{m_0}, P_d) > (P_a, P_c, P_{m_0}, P_d)$.

Proof We will only show the proof of (1); the same proof together with Lemma 2.8 gives (2). Let

$$\begin{aligned} L_{m_0} &= P_{m_0} \cap \big(\xi(a)^{(1)} + \xi(b)^{(1)}\big), \\ L_{m_1} &= P_{m_1} \cap \big(\xi(a)^{(1)} + \xi(b)^{(1)}\big), \\ L_d &= P_d \cap \big(\xi(a)^{(1)} + \xi(b)^{(1)}\big). \end{aligned}$$

Choose vectors $l_{m_0}, l_{m_1}, l_a, l_b, l_d \in \mathbb{R}^n$ such that

$$[l_{m_0}] = L_{m_0}, \quad [l_{m_1}] = L_{m_1}, \quad [l_a] = \xi(a)^{(1)}, \quad [l_b] = \xi(b)^{(1)}, \quad [l_d] = L_d.$$

By Lemma 2.6(1), we can ensure, by replacing each l_i with $-l_i$ if necessary, that

$$\begin{aligned} l_{m_0} &= \alpha l_a + (1-\alpha) l_b, \\ l_d &= \beta l_a + (1-\beta) l_b, \\ l_{m_1} &= \gamma l_a + (1-\gamma) l_b \end{aligned}$$

for $0 < \alpha < \beta < \gamma < 1$. Then we can compute

$$\begin{aligned} (P_a, P_b, P_{m_0}, P_d) &= \frac{1-\alpha}{1-\alpha/\beta} \\ &> \frac{1-\alpha}{1-\alpha/\gamma} \\ &= (P_a, P_b, P_{m_0}, P_{m_1}). \end{aligned}$$ □

2.4 Proof in the PSL($n, \mathbb{R}$) case

We will now use the technical facts established in Section 2.3 to prove Theorem 1.1. For the rest of this section, fix $\rho \in \mathrm{Hit}_n(S)$ and let ξ be its corresponding Frenet curve. The next lemma is the main computation in the proof of Theorem 1.1.

Lemma 2.11 *Let B be a nonidentity element in Γ. Pick $k = 0, \dots, n-2$, and for any $x \in \partial_\infty \Gamma$, define*

$$P_x = P_x^{(k)} := \begin{cases} \xi(x)^{(1)} + \xi(b^+)^{(k)} + \xi(b^-)^{(n-k-2)} & \text{if } x \neq b^+, b^-, \\ \xi(b^+)^{(k+1)} + \xi(b^-)^{(n-k-2)} & \text{if } x = b^+, \\ \xi(b^+)^{(k)} + \xi(b^-)^{(n-k-1)} & \text{if } x = b^-. \end{cases}$$

Suppose that x_1, x_2 and x_3 are points in $\partial_\infty \Gamma$ such that

$$x_1,\ x_2,\ B \cdot x_1,\ b^+,\ x_3,\ b^-$$

lie on $\partial_\infty \Gamma$, in that cyclic order. Then

$$(P_{x_1}, P_{x_2}, P_{b^+}, P_{x_3}) > \frac{\beta_{k+1}}{\beta_{k+1} - \beta_{k+2}},$$

where $0 < \beta_n < \cdots < \beta_1$ are the eigenvalues of $\rho(B)$.

Proof By Proposition 2.10 and parts (5) and (6) of Lemma 2.8, we have

$$\begin{aligned} (P_{x_1}, P_{x_2}, P_{b^+}, P_{x_3}) &> (P_{x_1}, P_{B\cdot x_1}, P_{b^+}, P_{b^-}) \\ &= \frac{1}{(P_{x_1}, P_{b^+}, P_{B\cdot x_1}, P_{b^-})} \\ &= \frac{1}{1 - (P_{b^+}, P_{x_1}, P_{B\cdot x_1}, P_{b^-})}. \end{aligned} \tag{2-1}$$

Note that for all $j = 1, \dots, n$,

$$L_j := \xi(b^+)^{(j)} \cap \xi(b^-)^{(n-j+1)}$$

is the fixed line through the origin in $\mathbb{R}^n$ of $\rho(B)$ corresponding to the eigenvalue β_j. Also, observe that P_{b^+} and P_{b^-} intersect the plane $\xi(b^+)^{(k+2)} \cap \xi(b^-)^{(n-k)}$ at L_{k+1} and L_{k+2}, respectively. Let

$$L := P_{x_1} \cap \big(\xi(b^+)^{(k+2)} \cap \xi(b^-)^{(n-k)}\big),$$

and it is clear that $P_{B\cdot x_1} \cap \big(\xi(b^+)^{(k+2)} \cap \xi(b^-)^{(n-k)}\big) = \rho(B) \cdot L$. Thus, we can use Lemma 2.9, to conclude that

$$(P_{b^+}, P_{x_1}, P_{B\cdot x_1}, P_{b^-}) = (L_{k+1}, L, \rho(B) \cdot L, L_{k+2}) = \frac{\beta_{k+2}}{\beta_{k+1}}.$$

Combining this with inequality (2-1) proves the lemma. □

Applying Lemma 2.11 to our setting, we can now prove Theorem 1.1.

Proof of Theorem 1.1 (1) Let ω be the subsegment of $\mathbb{P}\big(\xi(a^+)^{(1)}+\xi(a^-)^{(1)}\big)$ with endpoints $\xi(a^+)^{(1)}$ and $\xi(a^-)^{(1)}$ that has nonempty intersection with $\mathbb{P}\big(\xi(b^+)^{(k+1)}+\xi(b^-)^{(n-k-2)}\big)$. Define

$$\begin{aligned} p &:= \mathbb{P}\big(\xi(b^+)^{(k+1)}+\xi(b^-)^{(n-k-2)}\big)\cap\omega,\\ q &:= \mathbb{P}\big(\xi(A\cdot b^+)^{(1)}+\xi(b^+)^{(k)}+\xi(b^-)^{(n-k-2)}\big)\cap\omega, \end{aligned}$$

and note that q exists by Lemma 2.6(1). Also, observe that

$$\rho(A)\cdot p=\mathbb{P}\big(\xi(A\cdot b^+)^{(k+1)}+\xi(A\cdot b^-)^{(n-k-2)}\big)\cap\omega,$$

so Lemma 2.6(2) implies that $\rho(A)\cdot p$ lies between $\xi(a^+)^{(1)}$ and q in ω. Lemma 2.8, Lemma 2.9 and Proposition 2.10 together then allow us to conclude that

$$\begin{aligned} \frac{\alpha_1}{\alpha_n} &= \big(\xi(a^+)^{(1)},\rho(A)\cdot p,p,\xi(a^-)^{(1)}\big)\\ &> \big(\xi(a^+)^{(1)},q,p,\xi(a^-)^{(1)}\big)\\ &= (P_{a^+},P_{A\cdot b^+},P_{b^+},P_{a^-}), \end{aligned}$$

where

$$P_x=P_x^{(k)}:=\begin{cases}\xi(x)^{(1)}+\xi(b^+)^{(k)}+\xi(b^-)^{(n-k-2)} & \text{if } x\neq b^+,b^-,\\ \xi(b^+)^{(k+1)}+\xi(b^-)^{(n-k-2)} & \text{if } x=b^+,\\ \xi(b^+)^{(k)}+\xi(b^-)^{(n-k-1)} & \text{if } x=b^-.\end{cases}$$

By Lemma 2.2, we know that a^+, $A\cdot b^+$, $B\cdot a^+$, b^+, a^-, b^- lie along $\partial_\infty\Gamma$ in that cyclic order. This allows us to apply Lemma 2.11 with x_1, x_2 and x_3 as a^+, $A\cdot b^+$ and a^-, respectively, to obtain the desired inequality.

(2) Let r^- and r^+ be the closed subsegments of $\partial_\infty\Gamma$ with endpoints a^- and a^+ such that b^- lies in r^-, while b^+ lies in r^+. Orient both r^- and r^+ from a^- to a^+. Define $\mathcal{B}$ to be the set of unordered pairs $\{b'^+,b'^-\}$ in the Γ–orbit of $\{b^+,b^-\}$ such that b'^+ lies in r^+ between b^+ and $A\cdot b^+$, while b'^- lies in r^- between b^- and $A\cdot b^-$.

Every pair in $\mathcal{B}$ is the set of attracting and repelling fixed points for some B' in Γ that is conjugate to B. Since γ is simple, for every $\{b'^+,b'^-\}$ and $\{b''^+,b''^-\}$ in $\mathcal{B}$, we know that b'^+ precedes b''^+ (in the orientation on r^+) if and only if b'^- precedes b''^- (in the orientation of r^-). The orientations on r^- and r^+ thus induce an ordering on $\mathcal{B}$. Also, observe that $|\mathcal{B}|=i(\eta,\gamma)+1$, so we can label the pairs in $\mathcal{B}$ according to the order; ie

$$\mathcal{B}=\big\{\{b_1^+,b_1^-\},\dots,\{b_{m+1}^+,b_{m+1}^-\}\big\},$$

where $b_1^+=b^+$, $b_1^-=b^-$, $b_{m+1}^+=A\cdot b^+$, $b_{m+1}^-=A\cdot b^-$ and $m=i(\eta,\gamma)$.

For each i, let B_i be the element in Γ that is conjugate to either B or B^{-1} such that its attracting and repelling fixed points are b_i^+ and b_i^-, respectively. By Lemma 2.2, a^+, b_{i+1}^+, $B_i\cdot a^+$, b_i^+, a^-, b_i^- lie along $\partial_\infty\Gamma$ in that cyclic order, so we can apply Lemma 2.11 with x_1, x_2 and x_3 as a^+, b_{i+1}^+ and a^-, respectively, to conclude that

$$(P_{a^+,i}, P_{b_{i+1}^+,i}, P_{b_i^+,i}, P_{a^-,i}) > \frac{\beta_{k+1}}{\beta_{k+1}-\beta_{k+2}} \tag{2-2}$$

if B_i is conjugate to B, and

$$(P_{a^+,i}, P_{b_{i+1}^+,i}, P_{b_i^+,i}, P_{a^-,i}) > \frac{\beta_{n-k-1}}{\beta_{n-k-1}-\beta_{n-k}} \tag{2-3}$$

if B_i is conjugate to B^{-1}, where

$$P_{x,i} = P_{x,i}^{(k)} := \begin{cases} \xi(x)^{(1)} + \xi(b_i^+)^{(k)} + \xi(b_i^-)^{(n-k-2)} & \text{if } x \neq b_i^+, b_i^-, \\ \xi(b_i^+)^{(k+1)} + \xi(b_i^-)^{(n-k-2)} & \text{if } x = b_i^+, \\ \xi(b_i^+)^{(k)} + \xi(b_i^-)^{(n-k-1)} & \text{if } x = b_i^-. \end{cases}$$

Fix any $k=0,\dots,n-2$, and let ω be the subsegment of $\mathbb{P}\big(\xi(a^+)^{(1)}+\xi(a^-)^{(1)}\big)$ with endpoints $\xi(a^+)^{(1)}$, $\xi(a^-)^{(1)}$ whose intersection with $\mathbb{P}\big(\xi(b_i^+)^{(k+1)}+\xi(b_i^-)^{(n-k-2)}\big)$ is nonempty. For $i=1,\dots,m+1$, define

$$p_i := \mathbb{P}\big(\xi(b_i^+)^{(k+1)} + \xi(b_i^-)^{(n-k-2)}\big) \cap \omega,$$

and for $i=1,\dots,m$, define

$$q_i := \mathbb{P}\big(\xi(b_{i+1}^+)^{(1)} + \xi(b_i^+)^{(k)} + \xi(b_i^-)^{(n-k-2)}\big) \cap \omega.$$

Observe that Lemma 2.6(2) implies that p_i and q_i are well defined, and that $\xi(a^-)^{(1)}$, p_1, q_1, p_2, q_2, $\dots$, p_m, q_m, p_{m+1}, $\xi(a^+)^{(1)}$ lie in ω in that order. Hence, by similar arguments as those used in the proof of (1), we have

$$\begin{aligned}\big(\xi(a^+)^{(1)}, p_{i+1}, p_i, \xi(a^-)^{(1)}\big) &> \big(\xi(a^+)^{(1)}, q_i, p_i, \xi(a^-)^{(1)}\big) \\ &= (P_{a^+,i}, P_{b_{i+1}^+,i}, P_{b_i^+,i}, P_{a^-,i}).\end{aligned}$$

We can then use Lemmas 2.9 and 2.8 to obtain

$$\begin{aligned}\frac{\alpha_1}{\alpha_n} &= \big(\xi(a^+)^{(1)}, p_{m+1}, p_1, \xi(a^-)^{(1)}\big) \\ &= \prod_{i=1}^m \big(\xi(a^+)^{(1)}, p_{i+1}, p_i, \xi(a^-)^{(1)}\big) \\ &> \prod_{i=1}^m (P_{a^+,i}, P_{b_{i+1}^+,i}, P_{b_i^+,i}, P_{a^-,i}).\end{aligned} \tag{2-4}$$

Let $\mathcal{B}_+ := \{i : B_i \text{ is conjugate to } B\}$ and $\mathcal{B}_- := \{i : B_i \text{ is conjugate to } B^{-1}\}$, and let $u := |\mathcal{B}_+|$. Then combining the inequalities (2-2), (2-3) and (2-4) yields

$$\begin{aligned}\frac{\alpha_1}{\alpha_n} &> \prod_{i\in\mathcal{B}_+}(P_{a^+,i}, P_{b^+_{i+1},i}, P_{b^+_i,i}, P_{a^-,i}) \cdot \prod_{i\in\mathcal{B}_-}(P_{a^+,i}, P_{b^+_{i+1},i}, P_{b^+_i,i}, P_{a^-,i}) \\ &> \left(\frac{\beta_{k+1}}{\beta_{k+1}-\beta_{k+2}}\right)^u \cdot \left(\frac{\beta_{n-k-1}}{\beta_{n-k-1}-\beta_{n-k}}\right)^{i(\eta,\gamma)-u}. \end{aligned}$$ □

3 Further remarks

In this section, we prove some corollaries of Theorem 1.1, show that it does not hold for quasi-Fuchsian representations, and perform a comparison with the classical collar lemma.

3.1 Corollaries

Theorem 1.1 has some interesting consequences. The first is an analog of the classical collar lemma for Hitchin representations.

Corollary 1.2 *Let S be a surface of genus $g \geq 2$, and let η and γ be two essential closed curves in S. Denote the geometric intersection number between η and γ by $i(\eta,\gamma)$. Then for any $n \geq 2$ and any $\rho \in \mathrm{Hit}_n(S)$, the following hold:*

(1) *If $i(\eta,\gamma) \neq 0$, then*

$$\frac{1}{\exp(l_\rho(\eta))} < 1 - \frac{1}{\exp(l_\rho(\gamma)/(n-1))}.$$

(2) *If $i(\eta,\gamma) \neq 0$ and γ is simple, then there are nonnegative integers u and v with $u \geq v$ and $u + v = i(\eta,\gamma)$ such that*

$$\frac{1}{\exp(l_\rho(\eta))} < \left(1 - \frac{1}{\exp(l_\rho(\gamma)/(n-1))}\right)^u \left(1 - \frac{1}{\exp(l_\rho(\gamma))}\right)^v.$$

(3) *Let $\delta_n > 0$ be the unique real solution to the equation $e^{-x} + e^{-x/(n-1)} = 1$. If η is a nonsimple closed curve, then*

$$l_\rho(\eta) > \delta_n.$$

Proof In this proof, we will use the same notation as we used in Theorem 1.1.

(1) Choose orientations on η and γ. The hypothesis on η and γ imply that there are group elements A and B in Γ corresponding to η and γ, respectively, such that

$$a^+,\ b^+,\ a^-,\ b^-$$

lie along $\partial_\infty \Gamma$ in that cyclic order. Let $0 < \alpha_n < \cdots < \alpha_1$ and $0 < \beta_n < \cdots < \beta_1$ be the eigenvalues of $\rho(A)$ and $\rho(B)$, respectively. By Theorem 1.1(1), we know that for all $k = 0, \dots, n-2$,

$$\frac{\alpha_1}{\alpha_n} > \frac{\beta_{k+1}}{\beta_{k+1} - \beta_{k+2}},$$

which implies that

$$\frac{\beta_{k+2}}{\beta_{k+1}} < 1 - \frac{\alpha_n}{\alpha_1}.$$

Taking the product over all $k = 0, \dots, n-2$, we get

$$\frac{\alpha_n}{\alpha_1} + \left(\frac{\beta_n}{\beta_1}\right)^{1/(n-1)} < 1.$$

Since $l_\rho(\eta) = \log(\alpha_1/\alpha_n)$ and $l_\rho(\gamma) = \log(\beta_1/\beta_n)$, the above inequality gives us (1).

(2) By Theorem 1.1(2), we know that there is some nonnegative integer $u \le i(\eta, \gamma)$ such that for any $k = 0, \dots, n-2$, we have

$$\frac{\alpha_n}{\alpha_1} < \left(1 - \frac{\beta_{k+2}}{\beta_{k+1}}\right)^u \left(1 - \frac{\beta_{n-k}}{\beta_{n-k-1}}\right)^{i(\eta,\gamma)-u}.$$

In particular, we also have that for any $k = 0, \dots, n-2$,

$$\frac{\alpha_n}{\alpha_1} < \left(1 - \frac{\beta_{k+2}}{\beta_{k+1}}\right)^{i(\eta,\gamma)-u} \left(1 - \frac{\beta_{n-k}}{\beta_{n-k-1}}\right)^u,$$

so we can assume that

$$\frac{\alpha_n}{\alpha_1} < \left(1 - \frac{\beta_{k+2}}{\beta_{k+1}}\right)^u \left(1 - \frac{\beta_{n-k}}{\beta_{n-k-1}}\right)^v$$

for some nonnegative integers u and v such that $u \ge v$ and $u + v = i(\eta, \gamma)$. This implies that

$$\frac{\alpha_n}{\alpha_1} < \left(1 - \frac{\beta_{k+2}}{\beta_{k+1}}\right)^u \left(1 - \frac{\beta_n}{\beta_1}\right)^v,$$

which we can rewrite as

$$\frac{\beta_{k+2}}{\beta_{k+1}} < 1 - \frac{(\alpha_n/\alpha_1)^{1/u}}{(1 - \beta_n/\beta_1)^{v/u}}.$$

By taking the product of the above inequality over $k = 0, \dots, n-2$, we have

$$\left(\frac{\beta_n}{\beta_1}\right)^{1/(n-1)} < 1 - \frac{(\alpha_n/\alpha_1)^{1/u}}{(1 - \beta_n/\beta_1)^{v/u}},$$

which can be rewritten as

$$\frac{\alpha_n}{\alpha_1} < \left(1 - \left(\frac{\beta_n}{\beta_1}\right)^{1/(n-1)}\right)^u \left(1 - \frac{\beta_n}{\beta_1}\right)^v,$$

from which (2) follows.

(3) Choose an orientation on η. Since η is nonsimple, there are group elements A and B corresponding to η such that

$$a^+,\ b^+,\ a^-,\ b^-$$

lie along $\partial_\infty\Gamma$ in that cyclic order. Let $0 < \alpha_n < \cdots < \alpha_1$ and $0 < \beta_n < \cdots < \beta_1$ be the eigenvalues of $\rho(A)$ and $\rho(B)$, respectively. Note that $\rho(B)$ is either conjugate to $\rho(A)$ or $\rho(A)^{-1}$, so $\beta_n/\beta_1 = \alpha_n/\alpha_1$. Hence, the same computation as the proof of (1) then yields the inequality

$$\frac{\alpha_n}{\alpha_1} + \left(\frac{\alpha_n}{\alpha_1}\right)^{1/(n-1)} < 1,$$

which is equivalent to

$$\left(1 - \frac{\alpha_n}{\alpha_1}\right)^{n-1} - \frac{\alpha_n}{\alpha_1} > 0. \tag{3-1}$$

Consider the polynomial $P_n(x) = x^{n-1} + x - 1$. Note that for $n \geq 2$, we have that $P_n(x)$ is strictly increasing on the interval $[0, 1]$, $P_n(0) = -1$ and $P_n(1) = 1$. Hence, P_n has a unique zero in the interval $(0, 1)$, which we denote by x_n. It then follows that

$$\{x \in [0, 1] : P_n(x) > 0\} = (x_n, 1].$$

Also, observe that

$$P_n\left(1 - \frac{\alpha_n}{\alpha_1}\right) = \left(1 - \frac{\alpha_n}{\alpha_1}\right)^{n-1} - \frac{\alpha_n}{\alpha_1}$$

and $0 < 1 - \alpha_n/\alpha_1 < 1$. Since α_n/α_1 satisfies the inequality (3-1), we have

$$x_n < 1 - \frac{\alpha_n}{\alpha_1} < 1.$$

This implies that

$$l_\rho(\eta) = \log\left(\frac{\alpha_1}{\alpha_n}\right) > \delta_n := -\log(1 - x_n). \qquad \square$$

An easy consequence of Corollary 1.2 is the following.

Corollary 3.1 *For any $n \geq 2$ and any $\rho \in \mathrm{Hit}_n(S)$, there are at most $3g-3$ closed curves in S of ρ–length at most δ_n.*

In the case of $\mathcal{T}(S)$, one can replace the number $\delta_2 = \log(2)$ with $4 \cdot \sinh^{-1}(1)$; see Buser [5, Theorem 4.2.2].

Proof By Corollary 1.2(1), if η and γ are closed curves in S such that $i(\eta,\gamma) \neq 0$, then $l_\rho(\eta)$ and $l_\rho(\gamma)$ cannot both be smaller than δ_n. Moreover, Corollary 1.2(3) tells us that any closed curve of ρ–length less than δ_n has to be simple. Thus, the set of closed curves of ρ–length less than δ_n has to be a pairwise disjoint collection of simple closed curves, so the size of this collection is at most $3g-3$. □

Let $\widetilde{M}$ be the $\mathrm{PSL}(n,\mathbb{R})$ symmetric space, and let $d_{\widetilde{M}}$ be the distance function given by the Riemannian metric on $\widetilde{M}$. It is well known that for any $Z \in \Gamma$, the translation length of $\rho(Z)$, namely $\inf\{d_{\widetilde{M}}(o,\rho(Z)\cdot o) : o \in \widetilde{M}\}$, is

$$c_n \sqrt{\sum_{i=1}^{n} (\log \lambda_i)^2}$$

for some positive constant c_n depending only on n. Here, $0 < \lambda_n < \cdots < \lambda_1$ are the eigenvalues of $\rho(Z)$. (See Chapter II.10 of Bridson and Haefliger [3].) For our purposes, we normalize the metric on $\widetilde{M}$ so that $c_n = \sqrt{2}$, ie so that the image of the totally geodesic embedding of $\mathbb{H}^2$ in $\widetilde{M}$ induced by ι_n: $\mathrm{PSL}(2,\mathbb{R}) \to \mathrm{PSL}(n,\mathbb{R})$ has sectional curvature $-6/(n(n-1)(n+1))$. Then for any discrete, faithful representation ρ: $\Gamma \to \mathrm{PSL}(n,\mathbb{R})$, and for any rectifiable closed curve ω in $M := \rho(\Gamma)\backslash\widetilde{M}$, let $l_M(\omega)$ be the length of ω in the Riemannian metric on M.

In the case when $\rho \in \mathrm{Hit}_n(S)$, we can use Corollary 1.2, to obtain a relationship between the lengths of curves in the quotient locally symmetric space M.

Corollary 1.4 *Let η and γ be two essential closed curves in S and let X and Y be elements in Γ corresponding to η and γ, respectively. For any $\rho \in \mathrm{Hit}_n(S)$, let η' and γ' be two closed curves in $\rho(\Gamma)\backslash\widetilde{M} =: M$ that correspond to $X, Y \in \Gamma = \pi_1(M)$, respectively. Then the statements in Corollary 1.2 hold, with $l_\rho(\eta)$ and $l_\rho(\gamma)$ replaced by $l_M(\eta')$ and $l_M(\gamma')$, respectively.*

Proof Pick any $Z \in \Gamma \setminus \{\mathrm{id}\}$, and let ω in S and ω' in M be closed curves corresponding to Z. Observe then that the translation length of $\rho(Z)$ in $\widetilde{M}$ is a lower bound for $l_M(\omega')$.

Also, since

$$2\sum_{i=1}^{n} x_i^2 - (x_1 - x_n)^2 = (x_1 + x_n)^2 + 2(x_2^2 + \cdots + x_{n-1}^2) \geq 0,$$

we have that

$$(x_1 - x_n)^2 \le 2\sum_{i=1}^{n} x_i^2.$$

This allows us to compute

$$l_\rho(\omega) = \log\Big(\frac{\lambda_1}{\lambda_n}\Big) \le \sqrt{2\sum_{i=1}^{n}(\log\lambda_i)^2} \le l_M(\omega'),$$

where $0 < \lambda_n < \cdots < \lambda_1$ are the eigenvalues of $\rho(Z)$. $\square$

Let $f\colon S \to M := \rho(\Gamma)\backslash\widetilde{M}$ be a π_1–injective map such that $f(\eta)$ and $f(\gamma)$ are rectifiable curves in the Riemannian metric on M. It then follows from Corollary 1.4 that the statements in Corollary 1.2 hold, with $l_\rho(\eta)$ and $l_\rho(\gamma)$ replaced with $l_M(f(\eta))$ and $l_M(f(\gamma))$, respectively. In particular, we have a collar lemma for the image of the harmonic maps corresponding to Hitchin representations that were given by Corlette [9].

Corollary 1.2 also allow us to deduce consequences that are similar to Corollary 1.4, but with the Hilbert metric on the symmetric space instead of the Riemannian one. The symmetric space $\widetilde{M}$ can be given a Hilbert metric in the following way. Let $S(n,\mathbb{R})$ be the space of symmetric $n \times n$ matrices with real entries and let $P(n,\mathbb{R})$ be the set of positive-definite matrices in $S(n,\mathbb{R})$. Let $\mathbb{P}(P)$ and $\mathbb{P}(S)$ be the projectivizations of $P(n,\mathbb{R})$ and $S(n,\mathbb{R})$, respectively, and observe that $\mathbb{P}(P)$ is a properly convex domain in $\mathbb{P}(S) \simeq \mathbb{RP}^{N-1}$, where $N = \frac{1}{2}(n(n+1))$. This allows us to equip $\mathbb{P}(P)$ with a Hilbert metric.

Moreover, we can define a $\mathrm{PSL}(n,\mathbb{R})$–action on $\mathbb{P}(S)$ by $g \cdot A := gAg^T$ for any $g \in \mathrm{PSL}(n,\mathbb{R})$ and any $A \in \mathbb{P}(S)$. Note that this action preserves the projective structure on $\mathbb{P}(S)$, and also preserves $\mathbb{P}(P)$. In fact, $\mathrm{PSL}(n,\mathbb{R})$ acts transitively on $\mathbb{P}(P)$, and the stabilizer of the projective class of the identity matrix in $\mathbb{P}(P)$ is $\mathrm{PSO}(n)$, so the symmetric space $\widetilde{M}$ can be identified with $\mathbb{P}(P)$. This equips $\widetilde{M}$ with a Hilbert metric. Denote $\widetilde{M}$ equipped with the Hilbert metric by $\widetilde{M}'$, and for any discrete, faithful representation $\rho\colon \Gamma \to \mathrm{PSL}(n,\mathbb{R})$, let $l_{M'}$ be the length function on $M' := \rho(\Gamma)\backslash\widetilde{M}'$ induced by the Hilbert metric. Corollary 1.2 then also implies the following corollary.

Corollary 3.2 *Let η and γ be two essential closed curves in S and let X and Y be elements in Γ corresponding to η and γ, respectively. For any $\rho \in \mathrm{Hit}_n(S)$, let η' and γ' be two closed curves in M' that correspond to $X, Y \in \Gamma = \pi_1(M')$, respectively. Then the statements in Corollary 1.2 hold, with $l_\rho(\eta)$ and $l_\rho(\gamma)$ replaced with $\frac{1}{2}l_{M'}(\eta')$ and $\frac{1}{2}l_{M'}(\gamma')$, respectively.*

Proof For any $Z \in \Gamma \setminus \{\mathrm{id}\}$, let $0 < \lambda_n < \cdots < \lambda_1$ be the eigenvalues of $\rho(Z)$. We can assume without loss of generality that $\rho(Z)$ is a diagonal. Let E_{ij} be the $n \times n$ matrix with 1 at position (i, j) and zero everywhere else, and let $B_{ij} = E_{ij} + E_{ji}$. Obviously, $\{B_{ij}\}_{i \leq j}$ is a basis of $S(n, \mathbb{R}) = \mathbb{R}^N$, and it is easy to verify that $\rho(Z) \cdot B_{ij} = \lambda_i \lambda_j B_{ij}$. That means B_{ij} is an eigenvector of $\rho(Z)$ with eigenvalue $\lambda_i \lambda_j$. Consequently, the translation length of $\rho(Z)$ is

$$\log\left(\frac{\lambda_1^2}{\lambda_n^2}\right) = 2\log\left(\frac{\lambda_1}{\lambda_n}\right);$$

see Cooper, Long and Tillmann [8, Proposition 2.1]. The corollary follows easily. □

As mentioned in the introduction, if we restrict to Hitchin representations that lie in $\mathrm{Hit}_n(S)' \subset \mathrm{Hit}_n(S)$, then we can strengthen Theorem 1.1(2).

Corollary 1.3 *Let A and B be elements in Γ such that a^+, b^+, a^-, b^- lie in $\partial_\infty \Gamma$ in that cyclic order. Let $\rho \in \mathrm{Hit}_n(S)'$ and let $\alpha_n < \cdots < \alpha_1$ and $\beta_n < \cdots < \beta_1$ be the moduli of the eigenvalues of $\rho(A)$ and $\rho(B)$, respectively. Finally, let η and γ be closed curves on S corresponding to A and B, respectively. If γ is a simple closed curve in S, then for every $k = 0, \ldots, n-2$,*

$$\alpha_1^2 > \left(\frac{\beta_{k+1}}{\beta_{k+1} - \beta_{k+2}}\right)^{i(\eta,\gamma)}.$$

Proof Since $\rho(B)$ is a diagonalizable element in $\mathrm{PSO}(k, k+1)$ or $\mathrm{PSp}(2k, \mathbb{R})$, we see that $\beta_{k+1} = 1/\beta_{n-k}$ for $k = 0, \ldots, n-1$, and $\alpha_n = 1/\alpha_1$. Apply this to Theorem 1.1(2). □

From this corollary, the same proof as Corollary 1.2(2) allows us to obtain the following stronger inequality in the case when $\rho \in \mathrm{Hit}_n(S)'$.

Corollary 3.3 *Let η and γ be two essential closed curves in S such that γ is simple and $i(\eta, \gamma) \neq 0$. Then for any $\rho \in \mathrm{Hit}_n(S)'$,*

$$\frac{1}{\exp(l_\rho(\eta))} < \left(1 - \frac{1}{\exp(l_\rho(\gamma)/(n-1))}\right)^{i(\eta,\gamma)}.$$

Our results can be generalized to surfaces with boundaries in the following way. Let S' be a connected, oriented, topological surface with boundary, such that the double of S' is S, a closed connected, oriented topological surface of genus $g \geq 2$. Let Γ' be the fundamental group of S', and note that by choosing appropriate basepoints in the universal covers of S and S', the inclusion $S' \subset S$ induces an inclusion $\Gamma' \subset \Gamma$, which in turn induces an inclusion $\partial\Gamma' \subset \partial\Gamma$. In particular, $\partial\Gamma'$ inherits a natural cyclic order from $\partial\Gamma$.

The inclusion $\Gamma' \subset \Gamma$ also allows us to define the restriction map

$$\mathrm{res}\colon \mathrm{Hit}_n(S) \to \mathcal{X}_n(S') := \mathrm{Hom}(\Gamma', \mathrm{PSL}(n,\mathbb{R}))/\,\mathrm{PSL}(n,\mathbb{R})$$

by $\mathrm{res}\colon [\rho] \mapsto [\rho|_{\Gamma'}]$. Using this, define the *n^{th} Hitchin component of S'* to be

$$\mathrm{Hit}_n(S') := \mathrm{res}(\mathrm{Hit}_n(S)).$$

(See the introduction of Labourie and McShane [16] for an alternative definition.) While $\mathrm{Hit}_n(S')$ is still topologically a cell, it is no longer a connected component of $\mathcal{X}_n(S')$, so it is properly contained in its closure $\overline{\mathrm{Hit}_n(S')}$ in $\mathcal{X}_n(S')$.

Corollary 3.4 *Theorem 1.1 holds with Γ replaced with Γ', $\mathrm{Hit}_n(S)$ replaced with $\overline{\mathrm{Hit}_n(S')}$, and the strict inequalities $>$ replaced with weak inequalities $\geq$.*

Proof For any closed curve γ in S, let $X \in \Gamma$ be a corresponding group element. First, note that the moduli of the eigenvalues of $\rho(X)$ and $\mathrm{res}(\rho)(X)$ agree, so Theorem 1.1 clearly holds for $\rho \in \mathrm{Hit}_n(S')$.

Since $\alpha_n < \cdots < \alpha_1$ and $\beta_n < \cdots < \beta_1$ on $\mathrm{Hit}_n(S')$, these moduli of eigenvalues are still well defined on $\overline{\mathrm{Hit}_n(S')}$, and satisfy the weak inequalities $\alpha_n \leq \cdots \leq \alpha_1$ and $\beta_n \leq \cdots \leq \beta_1$. Furthermore, as functions on $\overline{\mathrm{Hit}_n(S')}$, they are continuous. As such, the inequalities in Theorem 1.1 hold on $\overline{\mathrm{Hit}_n(S')}$, with $>$ replaced with $\geq$. □

3.2 Counterexample for non-Hitchin representations

Note that in our proof, we used very strongly that the representations we consider are in $\mathrm{Hit}_n(S)$ because we used properties of the Frenet curve to obtain our estimates. In fact, the collar lemma is special to Hitchin representations, and does not hold even on the space of discrete and faithful representations from Γ to $\mathrm{PSL}(n,\mathbb{R})$.

To see this, consider the space of conjugacy classes of quasi-Fuchsian representations from Γ to $\mathrm{PSL}(2,\mathbb{C}) = \mathrm{PSO}(3,1)^+ \subset \mathrm{PSL}(4,\mathbb{R})$, which is the group of orientation-preserving isometries of $\mathbb{H}^3$. These are discrete and faithful representations whose limit set in the Riemann sphere $\partial\mathbb{H}^3$ is a Jordan curve. It is well known that each quasi-Fuchsian representation ρ induces a convex cocompact hyperbolic structure on the three-manifold $S \times I$. Also, for any nonidentity element X in Γ, the closed geodesic γ in $S\times I$ (equipped with the hyperbolic metric induced by ρ) corresponding to X has ρ–length

$$l_\rho(\gamma)=\log\Bigl(\frac{\lambda_1}{\lambda_4}\Bigr),$$

where λ_1 and λ_4 are the moduli of eigenvalues of $\rho(X)$ with largest and smallest modulus, respectively.

It is a theorem of Bers [1, Theorem 1] that the space of quasi-Fuchsian representations can be naturally identified with $\mathcal{T}(S)\times\mathcal{T}(\overline{S})$, where $\overline{S}$ is S with the opposite orientation. For any quasi-Fuchsian representation ρ let (ρ^+,ρ^-) denote the pair of Fuchsian representations that corresponds to ρ, such that $\rho^+\in\mathcal{T}(S)$ and $\rho^-\in\mathcal{T}(\overline{S})$. Then for any essential closed curve γ in S, let γ_ρ be the geodesic representative of γ in the hyperbolic metric on $S\times I$ corresponding to ρ, and let γ_{ρ^+} and γ_{ρ^-} be the geodesic representatives of γ in the hyperbolic metrics on S and $\overline{S}$ corresponding to ρ^+ and ρ^-, respectively. By Epstein, Marden and Markovic [10, Theorem 3.1], we know that

$$l_\rho(\gamma_\rho)\le\min\{2\cdot l_{\rho^+}(\gamma_{\rho^+}),\ 2\cdot l_{\rho^-}(\gamma_{\rho^-})\}.$$

For any pair of simple closed curves η and γ, and for any $\epsilon>0$, let ρ be a quasi-Fuchsian representation such that

$$l_{\rho^+}(\eta_{\rho^+})<\tfrac{1}{2}\epsilon\quad\text{and}\quad l_{\rho^-}(\gamma_{\rho^-})<\tfrac{1}{2}\epsilon.$$

Hence, $l_\rho(\eta_\rho)$ and $l_\rho(\gamma_\rho)$ are both smaller than ϵ. This implies that the analog of Corollary 1.2 does not hold on the space of discrete and faithful, or even Anosov, representations from Γ to $\mathrm{PSL}(4,\mathbb{R})$. (See Guichard and Wienhard [13] for more background on Anosov representations.)

3.3 Comparison with the classical collar lemma

Let ρ be a representation in the Fuchsian locus of $\mathrm{Hit}_n(S)$ and let h be the corresponding Fuchsian representation in $\mathcal{T}(S)$. Also, let X be a nonidentity element in Γ and let γ be a curve in S corresponding to X. If λ^{-1} and λ are the two eigenvalues of $h(X)$, then $\lambda^{-n+1},\lambda^{-n+3},\dots,\lambda^{n-3},\lambda^{n-1}$ are the n eigenvalues of $\rho(X)$. Hence we can get the lengths

$$l_h(\gamma)=2\log(\lambda)\quad\text{and}\quad l_\rho(\gamma)=2(n-1)\log(\lambda).$$

Since $h\in\mathcal{T}(S)$, the classical collar lemma holds. In other words, for any pair of curves η and γ in S such that γ is simple and $i(\eta,\gamma)>0$, we have

$$I_{\eta,\gamma}(h):=\sinh\bigl(\tfrac{1}{2}l_h(\eta)\bigr)\sinh\bigl(\tfrac{1}{2}l_h(\gamma)\bigr)>1; \tag{3-2}$$

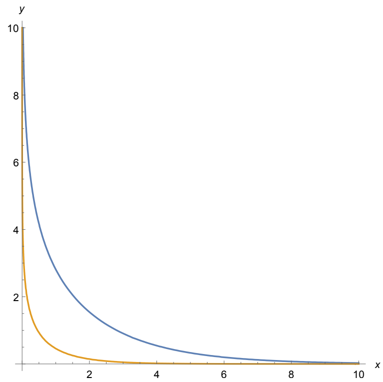


Figure 5: The upper curve $\sinh\left(\frac{1}{2}x\right)\sinh\left(\frac{1}{2}y\right) = 1$ and the lower curve $(e^x - 1)(e^y - 1) = 1$

see Buser [5, Corollary 4.1.2]. This inequality is sharp, in the sense that for any S, there are simple curves η and γ in S and a sequence of Fuchsian representations $\{h_i\}$ such that $I_{\eta,\gamma}(h_i)$ converges to 1. For more details, refer to Section 6 of Matelski [17].

On the other hand, Corollary 1.2(1), specialized to the $n = 2$ case, is the inequality

$$(e^{l_h(\eta)} - 1)(e^{l_h(\gamma)} - 1) > 1.$$

This is weaker than the inequality (3-2) because

$$e^x - 1 > \tfrac{1}{2}e^{-x/2}(e^x - 1) = \sinh\left(\tfrac{1}{2}x\right)$$

for every $x > 0$; see Figure 5. Moreover, we are unable to show that the inequality (3-2) fails in $\mathrm{Hit}_n(S)$ for any $n > 2$. This led us to conjecture in an earlier version of this paper, that for any ρ in $\mathrm{Hit}_n(S)$, there is some representation ρ' in the Fuchsian locus of $\mathrm{Hit}_n(S)$ such that

$$l_\rho(\gamma) \geq l_{\rho'}(\gamma) \quad \text{for any } \gamma \in \Gamma.$$

This conjecture implies that

$$\sinh\left(\frac{l_\rho(\eta)}{2(n-1)}\right)\sinh\left(\frac{l_\rho(\gamma)}{2(n-1)}\right) > 1$$

for any $\rho \in \mathrm{Hit}_n(S)$, which is sharp on every $\mathrm{Hit}_n(S)$ because it is sharp when restricted to the Fuchsian locus.

Recently, Tholozan proved (Section 0.4 of [18]) that the conjecture holds in the case when $n = 3$. Furthermore, Labourie disproved our conjecture in the case when $n \geq 4$. We will give his argument here.

Proposition 3.5 (Labourie) *When $n \geq 4$, there is some $\rho \in \mathrm{Hit}_n(S)$ such that for any ρ' in the Fuchsian locus of $\mathrm{Hit}_n(S)$, there is some closed curve γ in S such that $l_\rho(\gamma) < l_{\rho'}(\gamma)$.*

Proof For any closed curve γ in S, let $L_\gamma\colon \mathrm{Hit}_n(S) \to \mathbb{R}$ denote the map given by $L_\gamma(\rho) = l_\rho(\gamma)$. As before, let $\mathrm{Hit}_n(S)'$ be the $\mathrm{PSp}(2k,\mathbb{R})$ or $\mathrm{PSO}(k,k+1)$ Hitchin components when $n = 2k$ or $n = 2k+1$, respectively, and recall that $l_\rho(\gamma) = 2\log\lambda_1(\rho(X))$ for all $\rho \in \mathrm{Hit}_n(S)'$, where $X \in \Gamma$ is a group element corresponding to γ and $\lambda_1(\rho(X))$ is the modulus of eigenvalue of $\rho(X)$ with largest modulus. Proposition 10.3 of Bridgeman, Canary, Labourie and Sambarino [2] then implies that for any $\rho \in \mathrm{Hit}_n(S)'$, the set of differentials $\{dL_\gamma : \gamma \text{ a closed curve in } S\}$ generates the entire cotangent space of $\mathrm{Hit}_n(S)'$ at ρ.

Observe that if $n \geq 4$, then $\mathrm{Hit}_n(S)' \subset \mathrm{Hit}_n(S)$ properly contains the Fuchsian locus. Thus, it is sufficient to prove the proposition on $\mathrm{Hit}_n(S)'$. Suppose for contradiction that the proposition is false on $\mathrm{Hit}_n(S)'$. Choose a point ρ_0 in the Fuchsian locus, and take a smooth path ρ_t for $t \in (-\epsilon,\epsilon)$ with $\epsilon > 0$, whose nonzero tangent vector $U \in T_{\rho_0}\mathrm{Hit}_n(S)'$ is not tangential to the Fuchsian locus. Along the path, choose a sequence of representations $\{\rho_{t_i}\}_{i=1}^\infty$ which converges to ρ_0 as $i \to \infty$ so that $t_i > 0$ for all i.

Since the proposition is false on $\mathrm{Hit}_n(S)'$, there exists the corresponding sequence of Fuchsian representations ρ'_{t_i} such that $L_\gamma(\rho_{t_i}) \geq L_\gamma(\rho'_{t_i})$ for any closed curve γ in S. Also, since ρ_{t_i} converges to ρ_0, we see that $L_\gamma(\rho'_{t_i})$ is bounded for all γ, so the sequence $\{\rho'_{t_i}\}_{i=1}^\infty$ lie in a compact subset of the Fuchsian locus. By picking subsequence, we can assume without loss of generality that $\{\rho'_{t_i}\}_{i=1}^\infty$ converges to some ρ'_0 in the Fuchsian locus. The continuity of L_γ then implies that $L_\gamma(\rho_0) \geq L_\gamma(\rho'_0)$ for all γ, so $\rho_0 = \rho'_0$ because both ρ_0 and ρ'_0 lie in the Fuchsian locus.

Thus, the sequence $\{\rho'_{t_i}\}_{i=1}^\infty$ converges to ρ_0 as well. Choose a Riemannian metric on a neighborhood of ρ_0 in $\mathrm{Hit}_n(S)'$. By taking a further subsequence of $\{\rho_{t_i}\}_{i=1}^\infty$, we can also assume that either one of the following cases hold:

(i) $\rho'_{t_i} = \rho_0$ for all i;

(ii) the unit vectors at ρ_0 that are tangential to the geodesic between ρ'_{t_i} and ρ_0 converge to some unit vector $V \neq 0$ in $T_{\rho_0}\mathrm{Hit}_n(S)'$ that is tangential to the Fuchsian locus.

If (i) holds, then we have that $dL_\gamma(U) \geq 0$ for all γ. On the other hand, if (ii) holds, then for all closed curves γ in S,

$$
\begin{aligned}
dL_\gamma(U) &= \frac{d}{dt} L_\gamma(\rho_t)|_{t=0} \\
&= \lim_{i\to\infty} \frac{L_\gamma(\rho_{t_i}) - L_\gamma(\rho_0)}{t_i} \\
&\geq \lim_{i\to\infty} \frac{L_\gamma(\rho'_{t_i}) - L_\gamma(\rho'_0)}{t_i} \\
&= \Big(\lim_{i\to\infty} s_i\Big) \cdot dL_\gamma(V)
\end{aligned}
$$

for some sequence of positive numbers $\{s_i\}_{i=1}^\infty$. More precisely, if V_i denotes the tangent vector whose exponential is ρ'_{t_i} and $\|V_i\|$ is the norm of V_i with respect to the chosen Riemannian metric, then $s_i = \|V_i\|/t_i$.

Note that if $\lim_{i\to\infty} s_i = \infty$, then $dL_\gamma(V) \leq 0$ for all γ, which is impossible since $V \neq 0$ is tangential to the Fuchsian locus. Hence, $dL_\gamma(U + sV) \geq 0$ for some $s \leq 0$.

In either case, there is some vector $W \in T_{\rho_0} \operatorname{Hit}_n(S)'$ (possibly the zero vector) that is tangential to the Fuchsian locus such that $dL_\gamma(U + W) \geq 0$ for all γ. Furthermore, since $U + W \neq 0$, the fact that the differentials dL_γ generate the cotangent space of $\operatorname{Hit}_n(S)'$ at ρ_0 implies that $dL_\gamma(U + W) > 0$ for some γ. By a similar argument, we can also show that there is some vector $W' \in T_{\rho_0} \operatorname{Hit}_n(S)'$ that is tangential to the Fuchsian locus such that $dL_\gamma(-U + W') \geq 0$ for all γ, and this inequality holds strictly for some γ.

Adding these two inequalities together gives $dL_\gamma(W + W') \geq 0$ for all γ, and $dL_\gamma(W + W') > 0$ for some γ. However, this is impossible since $W + W'$ is tangential to the Fuchsian locus. □

Note that Labourie's argument to disprove our conjecture relied very heavily on the fact that $\operatorname{Hit}_n(S)' \subset \operatorname{Hit}_n(S)$ properly contains the Fuchsian locus. There is thus still hope that the following modified conjecture might be true.

Conjecture 3.6 *Let ρ be a representation in $\operatorname{Hit}_n(S)$. Then there is some representation ρ' in $\operatorname{Hit}_n(S)'$ such that*

$$l_\rho(\gamma) \geq l_{\rho'}(\gamma)$$

for any closed curve γ in S.

Appendix: Proof of Corollary 1.5

In this appendix, we will prove the properness result stated as Corollary 1.5. We begin by recalling some results from Zhang [19] that we will need.

Let $\mathcal{P} := \{\gamma_1, \dots, \gamma_{3g-3}\}$ be an oriented pants decomposition of S, ie a maximal collection of pairwise nonintersecting, pairwise nonhomotopic, homotopically nontrivial, oriented simple closed curves on S. These curves cut S into $2g-2$ pairs of pants, which we label by $P_1, \dots, P_{2g-2}$, and also gives us a real analytic diffeomorphism

$$\mathrm{Hit}_n(S) \to (\mathbb{R}^+)^{(3g-3)(n-1)} \times \mathbb{R}^{(3g-3)(n-1)} \times \mathbb{R}^{(2g-2)(n-1)(n-2)},$$

which one should think of as a generalization of the Fenchel–Nielsen coordinates on the Teichmüller space $\mathcal{T}(S)$; see [19, Proposition 3.5].

The first $(3g-3)(n-1)$ positive numbers are called the *boundary invariants*. For any $\rho \in \mathrm{Hit}_n(S)$, these are the numbers

$$\beta_{\gamma_j,k} := \log\left(\frac{\lambda_k(\rho(X_j))}{\lambda_{k+1}(\rho(X_j))}\right),$$

where $k = 1, \dots, n-1$ and $j = 1, \dots, 3g-3$. Here, $X_j \in \Gamma$ is a group element that corresponds to γ_j, and $\lambda_1(\rho(X_j)), \dots, \lambda_n(\rho(X_j))$ are the moduli of eigenvalues of $\rho(X_j)$ arranged in decreasing order. Note that each of the $3g-3$ curves in $\mathcal{P}$ are associated $n-1$ of these numbers. They capture the eigenvalue data of the holonomy about each of the curves in $\mathcal{P}$, and are a generalization of the Fenchel–Nielsen length coordinates.

The next $(3g-3)(n-1)$ real numbers are called the *gluing parameters*, and these are also associated to the curves in $\mathcal{P}$. Informally, the $n-1$ gluing parameters associated to each curve in $\mathcal{P}$ is the data specifying how one should "glue" the representations on adjacent pair of pants together along a common boundary component. Hence, these generalize the Fenchel–Nielsen twist coordinates. Just like the Fenchel–Nielsen twist coordinates, to specify these gluing parameters formally, we need to make additional topological choices to define what is "zero gluing". In this case, this additional topological choice we make is a pair of distinct points $a_j, b_j \in \partial_\infty\Gamma$ so that x_j^-, a_j, x_j^+, b_j lie in $\partial_\infty\Gamma$ in that cyclic order.

For simplicity, we will fix such a choice once and for all in the following way. Let P_1 and P_2 be the two pairs of pants that have γ_j as a common boundary component (it is possible for $P_1 = P_2$). For $i = 1, 2$, choose A_i, B_i and C_i to be elements in Γ corresponding the boundary components of P_i so that $C_i \cdot B_i \cdot A_i = \mathrm{id}$ and $A_1 = A_2^{-1} = X_j$. Let a_j be the repelling fixed point of B_1 and b_j be the repelling fixed point of C_2. The gluing parameters are then

$$g_{\gamma_j,k} := \log\bigl(-(P_{k,1}, P_{k,2}, P_{k,4}, P_{k,3})\bigr)$$

for $k = 1, \dots, n-1$, where $\xi\colon \partial_\infty\Gamma \to \mathcal{F}(\mathbb{R}^n)$ is the Frenet curve corresponding to ρ, and

$$\begin{aligned}
P_{k,1} &:= \xi(x_j^+)^{(k)} + \xi(x_j^-)^{(n-k-1)},\\
P_{k,2} &:= \xi(x_j^+)^{(k-1)} + \xi(x_j^-)^{(n-k-1)} + \xi(a_j)^{(1)},\\
P_{k,3} &:= \xi(x_j^+)^{(k-1)} + \xi(x_j^-)^{(n-k)},\\
P_{k,4} &:= \xi(x_j^+)^{(k-1)} + \xi(x_j^-)^{(n-k-1)} + \xi(b_j)^{(1)}
\end{aligned}$$

are four hyperplanes in $\mathbb{R}^n$ that intersect at $M_k := \xi(x_j^+)^{(k-1)} + \xi(x_j^-)^{(n-k-1)}$.

Finally, the remaining $(2g-2)(n-1)(n-2)$ real numbers are called the *internal parameters*, and are associated to the pairs of pants $P_1, \dots, P_{2g-2}$. To each P_j, we associate $(n-1)(n-2)$ internal parameters, and they parametrize the Hitchin representations on a pair of pants after the boundary invariants are fixed. These are defined in great detail in Section 3 of [19]. For our purposes though, we do not need to know what these parameters are, but only the following proposition.

Proposition A.1 *Fix a pair of pants P_{j_0} given by $\mathcal{P}$. Let $\{\rho_i\}$ be a sequence in $\mathrm{Hit}_n(S)$ such that*

- *the boundary invariants corresponding to ∂P_{j_0} remain bounded away from 0 and ∞ along $\{\rho_i\}$, and*
- *some internal parameter corresponding to P_{j_0} grows to ∞ or $-\infty$ along $\{\rho_i\}$.*

Let γ be a closed curve in S with the property that any closed curve homotopic to γ has nonempty intersection with P_{j_0}. Then $\lim_{i\to\infty} l_{\rho_i}(\gamma) = \infty$.

Proof The proof of this proposition is a slight modification of the proof of the main theorem given in Section 5.1 of [19]. In Section 3.2 of [19], there is a description of a particular way to cut each P_j into two ideal triangles that share all three edges. Doing this over all P_j gives us $6g-6$ edges. Here, we view each of these edges $e = [a,b]$ as a Γ–orbit of a pair of distinct points $a, b \in \partial_\infty\Gamma$.

Let $\rho \in \mathrm{Hit}_n(S)$ and ξ the corresponding Frenet curve. As was done in Section 4.4 of [19], one can associate a particular positive number $K[a,b]$ to each of these $6g-6$ edges $[a,b]$. Using this, define

$$K(\rho, j_0) := \min\{K[a,b] : [a,b] \subset P_{j_0}\}.$$

The same argument as given in Section 5.1 of [19] proves that

$$\lim_{i\to\infty} K(\rho_i, j_0) = \infty.$$

Let $X \in \Gamma$ be a group element corresponding to γ. Let $e = [a,b]$ be an edge in P_{j_0} such that there is a lift $\widetilde{e} = \{a,b\}$ with the property that x^-, a, x^+, b lie in $\partial_\infty \Gamma$ in that cyclic order. Such an edge exists by the hypothesis we imposed on γ. For any $p = 0, \dots, n-1$, one can define subsegments $c_p(\widetilde{e})$ of the projective line $\mathbb{P}\big(\xi(x^-)^{(1)} + \xi(x^+)^{(1)}\big) \subset \mathbb{RP}^{n-1}$ associated to each lift $\widetilde{e} = \{a,b\}$ of $e = [a,b]$. These are called the *crossing* (p)*–subsegments*; see Definition 4.7 of [19]. Using the cross ratio, we can define a notion of length for these subsegments, which we denote by $l(c_p(\widetilde{e}))$; see Definition 4.8 of [19].

By the proof of [19, Proposition 4.16], we see that

$$\frac{1}{n} \sum_{p=0}^{n-1} l(c_p(\widetilde{e})) \geq K(\rho, j_0).$$

Furthermore, by Lemmas 4.9 and 4.10 of [19], we have

$$l_\rho(\gamma) \geq l(c_p(\widetilde{e}))$$

for all $p = 0, \dots, n-1$, which allows us to conclude that

$$l_\rho(\gamma) \geq K(\rho, j_0).$$

Combining this with the fact that $\lim_{i\to\infty} K(\rho_i, j_0) = \infty$ gives the proposition. □

With the above proposition, we are ready to prove Corollary 1.5. Let $\{\rho_i\}$ be a sequence in $\mathrm{Hit}_n(S)$, let $\mathcal{C} := \{\gamma_1, \dots, \gamma_k\}$ satisfy the hypothesis of Corollary 1.5 and let $\mathcal{P} := \{\gamma_1, \dots, \gamma_{3g-3}\} \subset \mathcal{C}$ be a pants decomposition. Observe that the hypothesis on $\mathcal{C}$ ensures the following:

- For any $\gamma \in \mathcal{P}$, there is some $\gamma' \in \mathcal{C}$ that intersects γ transversely.
- For each pair of pants P given by $\mathcal{P}$, there is some $\gamma \in \mathcal{C}$ such that any closed curve homotopic to γ has nonempty intersection with P.

The pants decomposition $\mathcal{P}$ then gives us a parametrization of $\mathrm{Hit}_n(S)$ as described above. We will prove Corollary 1.5 in the following steps.

(1) If there is some boundary invariant $\beta_{\gamma_j,k}$ such that $\lim_{i\to\infty} \beta_{\gamma_j,k}(\rho_i) = \infty$, then $\lim_{i\to\infty} l_{\rho_i}(\gamma_j) = \infty$.

(2) If there is some boundary invariant $\beta_{\gamma_j,k}$ such that $\lim_{i\to\infty} \beta_{\gamma_j,k}(\rho_i) = 0$, then $\lim_{i\to\infty} l_{\rho_i}(\gamma) = \infty$ for any closed curve γ that intersects γ_j transversely.

(3) If all the boundary invariants remain bounded away from 0 and ∞ and some internal parameter associated to a pair of pants P grows to $\pm\infty$, then $\lim_{i\to\infty} l_{\rho_i}(\gamma) = \infty$ for any closed curve γ with the property that any closed curve homotopic to γ has nonempty intersection with P.

(4) If all the boundary invariants remain bounded away from 0 and ∞ and there is some gluing parameter $g_{\gamma_j,k}$ such that $\lim_{i\to\infty} g_{\gamma_j,k}(\rho_i) = \pm\infty$, then $\lim_{i\to\infty} l_{\rho_i}(\gamma) = \infty$ for any γ that intersects γ_j transversely.

Note that together, the four statements above prove Corollary 1.5. Statement (1) is obvious because

$$l_\rho(\gamma_j) = \sum_{k=1}^{n-1} \beta_{\gamma_j,k}(\rho),$$

and all the boundary invariants are positive. Also, statement (3) is a restatement of Proposition A.1, and statement (2) is an immediate consequence of Theorem 1.1(1), which is a main result in this paper. The rest of this appendix will be the proof of statement (4).

Let $X_j, X \in \Gamma$ correspond to γ_j and γ, respectively, such that x_j^-, x^-, x_j^+, x^+ lie in $\partial_\infty\Gamma$ in that cyclic order. We previously chose a pair of points $a_j, b_j \in \partial_\infty\Gamma$ so that x_j^-, a_j, x_j^+, b_j lie in $\partial_\infty\Gamma$ in that cyclic order in order to define the gluing parameters $g_{\gamma_j,k}$ associated to γ_j. If we choose $X_j^l \cdot a_j$ and $X_j^m \cdot b_j$ in place of a_j and b_j, we get another collection of gluing parameters, which we denote by $g^{l,m}_{\gamma_j,k}$. The next lemma explains the relationship between $g_{\gamma_j,k} = g^{0,0}_{\gamma_j,k}$ and $g^{l,m}_{\gamma_j,k}$. Its proof is an easy computation which we omit.

Lemma A.2 *Let $\rho \in \mathrm{Hit}_n(S)$, and let $\lambda_1, \dots, \lambda_n$ be the moduli of eigenvalues of $\rho(X_j)$ arranged in decreasing order. For any integers l and m, we have*

$$g^{l,m}_{\gamma_j,k}(\rho) = (l-m)\log\Big(\frac{\lambda_k}{\lambda_{k+1}}\Big) + g_{\gamma_j,k}(\rho).$$

In particular, when the boundary invariants corresponding to γ_j are bounded away from 0 and ∞ along a sequence of representations $\{\rho_i\}$ in $\mathrm{Hit}_n(S)$, then we have $\lim_{i\to\infty} g_{\gamma_j,k}(\rho_i) = \pm\infty$ if and only if $\lim_{i\to\infty} g^{l,m}_{\gamma_j,k}(\rho_i) = \pm\infty$. Statement (4) then follows immediately from this observation and the following proposition.

Proposition A.3 *Let $\rho \in \mathrm{Hit}_n(S)$ and let γ_j, γ, x_j^-, x_j^+, x^-, x^+, a_j, b_j be as above. Let l and m be integers such that x_j^-, $X_j^{l-1}\cdot a_j$, x^-, $X_j^l \cdot a_j$, x_j^+, $X_j^{m+1}\cdot b_j$, x^+, $X_j^m \cdot b_j$ lie in $\partial_\infty\Gamma$ in that cyclic order. Then*

$$3l_\rho(\gamma) \geq g^{l-1,m+1}_{\gamma_j,k}(\rho) \quad \textit{and} \quad 3l_\rho(\gamma) \geq -g^{l,m}_{\gamma_j,k}(\rho)$$

for all $k = 1, \dots, n-1$.

Proof The technique used in this proof is the same as that used in the proof of [19, Lemma 4.18]. For any $k=1,\dots,n-1$, let

$$
\begin{aligned}
P_{k,0} &:= \xi(x_j^+)^{(k)} + \xi(x_j^-)^{(n-k-1)},\\
P_{k,1} &:= \xi(x_j^+)^{(k-1)} + \xi(x_j^-)^{(n-k-1)} + \xi(x^-)^{(1)},\\
P_{k,2} &:= \xi(x_j^+)^{(k-1)} + \xi(x_j^-)^{(n-k-1)} + \xi(X_j^{l-1}\cdot a_j)^{(1)},\\
P'_{k,2} &:= \xi(x_j^+)^{(k-1)} + \xi(x_j^-)^{(n-k-1)} + \xi(X_j^{l}\cdot a_j)^{(1)},\\
P_{k,3} &:= \xi(x_j^+)^{(k-1)} + \xi(x_j^-)^{(n-k)},\\
P_{k,4} &:= \xi(x_j^+)^{(k-1)} + \xi(x_j^-)^{(n-k-1)} + \xi(X_j^{m}\cdot b_j)^{(1)},\\
P'_{k,4} &:= \xi(x_j^+)^{(k-1)} + \xi(x_j^-)^{(n-k-1)} + \xi(X_j^{m+1}\cdot b_j)^{(1)},\\
P_{k,5} &:= \xi(x_j^+)^{(k-1)} + \xi(x_j^-)^{(n-k-1)} + \xi(x^+)^{(1)}.
\end{aligned}
$$

Also, for all i, let

$$
\begin{aligned}
L'_{k,i} &:= P'_{k,i} \cap \big(\xi(x^-)^{(1)} + \xi(x^+)^{(1)}\big),\\
L_{k,i} &:= P_{k,i} \cap \big(\xi(x^-)^{(1)} + \xi(x^+)^{(1)}\big),
\end{aligned}
$$

and let

$$
\begin{aligned}
L_{k,a_j} &:= \big(\xi(X_j^{l}\cdot a_j)^{(k-1)} + \xi(X_j^{l-1}\cdot a_j)^{(n-k)}\big) \cap \big(\xi(x^-)^{(1)} + \xi(x^+)^{(1)}\big),\\
L_{k,b_j} &:= \big(\xi(X_j^{m+1}\cdot b_j)^{(k-1)} + \xi(X_j^{m}\cdot b_j)^{(n-k)}\big) \cap \big(\xi(x^-)^{(1)} + \xi(x^+)^{(1)}\big).
\end{aligned}
$$

It follows from [19, Lemma 2.5] that

$$\xi(x^-)^{(1)},\ L_{k,a_j},\ L_{k,2},\ L_{k,3},\ L_{k,4},\ L_{k,b_j},\ \xi(x^+)^{(1)}$$

lie in the projective line $\xi(x^-)^{(1)} + \xi(x^+)^{(1)}$ in that cyclic order. Also, by [19, Lemma 4.11], we know

$$3l_\rho(\gamma) \ge \log\big(\xi(x^-)^{(1)}, L_{k,a_j}, L_{k,b_j}, \xi(x^+)^{(1)}\big),$$

which implies that

$$
\begin{aligned}
3l_\rho(\gamma) &\ge \log\big(\xi(x^-)^{(1)}, L_{k,2}, L_{k,3}, \xi(x^+)^{(1)}\big),\\
3l_\rho(\gamma) &\ge \log\big(\xi(x^-)^{(1)}, L_{k,3}, L_{k,4}, \xi(x^+)^{(1)}\big)
\end{aligned}
$$

by Lemma 2.9. Using Lemmas 2.8 and 2.6, we can also deduce that

$$\begin{aligned}
\big(\xi(x^-)^{(1)}, L_{k,2}, L_{k,3}, \xi(x^+)^{(1)}\big) &= \big(\xi(x^-)^{(1)}, L_{k,2}, L_{k,3}, \xi(x^+)^{(1)}\big)_{M_k} \\
&= (P_{k,1}, P_{k,2}, P_{k,3}, P_{k,5}) \\
&\geq (P_{k,0}, P_{k,2}, P_{k,3}, P'_{k,4}) \\
&= 1 - (P_{k,0}, P_{k,2}, P'_{k,4}, P_{k,3}) \\
&= 1 + e^{g^{l-1,m+1}_{\gamma_j,k}} \\
&\geq e^{g^{l-1,m+1}_{\gamma_j,k}}
\end{aligned}$$

and

$$\begin{aligned}
\big(\xi(x^-)^{(1)}, L_{k,3}, L_{k,4}, \xi(x^+)^{(1)}\big) &= \big(\xi(x^-)^{(1)}, L_{k,3}, L_{k,4}, \xi(x^+)^{(1)}\big)_{M_k} \\
&= (P_{k,1}, P_{k,3}, P_{k,4}, P_{k,5}) \\
&\geq (P'_{k,2}, P_{k,3}, P_{k,4}, P_{k,0}) \\
&= 1 - \frac{1}{(P_{k,0}, P'_{k,2}, P_{k,4}, P_{k,3})} \\
&= 1 + e^{-g^{l,m}_{\gamma_j,k}} \\
&\geq e^{-g^{l,m}_{\gamma_j,k}},
\end{aligned}$$

where $M_k := \xi(x_j^+)^{(k-1)} + \xi(x_j^-)^{(n-k-1)}$. □

References

[1] **L Bers**, *Simultaneous uniformization*, Bull. Amer. Math. Soc. 66 (1960) 94–97 MR

[2] **M Bridgeman**, **R Canary**, **F Labourie**, **A Sambarino**, *The pressure metric for Anosov representations*, Geom. Funct. Anal. 25 (2015) 1089–1179 MR

[3] **M R Bridson**, **A Haefliger**, *Metric spaces of non-positive curvature*, Grundl. Math. Wissen. 319, Springer, New York (1999) MR

[4] **M Burger**, **M B Pozzetti**, *Maximal representations, non Archimedean Siegel spaces, and buildings*, preprint (2015) arXiv

[5] **P Buser**, *Geometry and spectra of compact Riemann surfaces*, Progress in Mathematics 106, Birkhäuser, Boston (1992) MR

[6] **S Choi**, *The Margulis lemma and the thick and thin decomposition for convex real projective surfaces*, Adv. Math. 122 (1996) 150–191 MR

[7] **S Choi**, **W M Goldman**, *Convex real projective structures on closed surfaces are closed*, Proc. Amer. Math. Soc. 118 (1993) 657–661 MR

[8] **D Cooper**, **D D Long**, **S Tillmann**, *On convex projective manifolds and cusps*, Adv. Math. 277 (2015) 181–251 MR

[9] **K Corlette**, *Flat G–bundles with canonical metrics*, J. Differential Geom. 28 (1988) 361–382 MR

[10] **D B A Epstein**, **A Marden**, **V Markovic**, *Quasiconformal homeomorphisms and the convex hull boundary*, Ann. of Math. 159 (2004) 305–336 MR

[11] **W M Goldman**, *Convex real projective structures on compact surfaces*, J. Differential Geom. 31 (1990) 791–845 MR

[12] **O Guichard**, *Composantes de Hitchin et représentations hyperconvexes de groupes de surface*, J. Differential Geom. 80 (2008) 391–431 MR

[13] **O Guichard**, **A Wienhard**, *Anosov representations: domains of discontinuity and applications*, Invent. Math. 190 (2012) 357–438 MR

[14] **L Keen**, *Collars on Riemann surfaces*, from "Discontinuous groups and Riemann surfaces" (L Greenberg, editor), Ann. of Math. Studies 79, Princeton Univ. Press (1974) 263–268 MR

[15] **F Labourie**, *Anosov flows, surface groups and curves in projective space*, Invent. Math. 165 (2006) 51–114 MR

[16] **F Labourie**, **G McShane**, *Cross ratios and identities for higher Teichmüller–Thurston theory*, Duke Math. J. 149 (2009) 279–345 MR

[17] **J P Matelski**, *A compactness theorem for Fuchsian groups of the second kind*, Duke Math. J. 43 (1976) 829–840 MR

[18] **N Tholozan**, *Entropy of Hilbert metrics and length spectrum of Hitchin representations in* $\mathrm{PSL}(3,\mathbb{R})$, preprint (2015) arXiv

[19] **T Zhang**, *Degeneration of Hitchin representations along internal sequences*, Geom. Funct. Anal. 25 (2015) 1588–1645 MR

Mathematisches Institut, Ruprecht-Karls-Universität Heidelberg
Im Neuenheimer Feld 205, D-69120 Heidelberg, Germany

Mathematics Department, California Institute of Technology
Mail Code 253-37, 1200 East California Boulevard, Pasadena, CA 91125, United States

gyeseonlee@gmail.com, tengren@caltech.edu

http://www.mathi.uni-heidelberg.de/~lee/,
http://sites.google.com/site/tengren85/

Proposed: Ian Agol Received: 22 December 2015
Seconded: Simon Donaldson, András I. Stipsicz Revised: 10 July 2016

Geometry & Topology 21 (2017) 2281–2352

Top-dimensional quasiflats in CAT(0) cube complexes

JINGYIN HUANG

We show that every n–quasiflat in an n–dimensional CAT(0) cube complex is at finite Hausdorff distance from a finite union of n–dimensional orthants. Then we introduce a class of cube complexes, called *weakly special* cube complexes, and show that quasi-isometries between their universal covers preserve top-dimensional flats. This is the foundational result towards the quasi-isometric classification of right-angled Artin groups with finite outer automorphism group.

Some of our arguments also extend to CAT(0) spaces of finite geometric dimension. In particular, we give a short proof of the fact that a top-dimensional quasiflat in a Euclidean building is Hausdorff close to a finite union of Weyl cones, which was previously established by Kleiner and Leeb (1997), Eskin and Farb (1997) and Wortman (2006) by different methods.

20F67; 20F65, 20F69

1 Introduction

1.1 Summary of results

A quasiflat of dimension d in a metric space X is a quasi-isometric embedding $\phi\colon \mathbb{E}^d \to X$, ie there exist positive constants L, A such that for all $x, y \in \mathbb{E}^d$,

$$L^{-1}d(x,y) - A \le d(\phi(x),\phi(y)) \le Ld(x,y) + A.$$

Top-dimensional (or maximal) flats and quasiflats in spaces of higher rank are analogues of geodesics and quasigeodesics in Gromov hyperbolic spaces, which play a key role in understanding the large scale geometry of these spaces. In particular, several quasi-isometric rigidity results were established on the study of such flats or quasiflats. Here is a list of examples:

- Euclidean buildings and symmetric spaces of noncompact type; see Mostow [34], Kleiner and Leeb [30], Eskin and Farb [17], Kramer and Weiss [32].
- Universal covers of certain Haken manifolds (see Kapovich and Leeb [27]); higher-dimensional graph manifolds (see Frigerio, Lafont and Sisto [18]); two-dimensional tree groups and their higher dimensional analogues (see Behrstock, Januszkiewicz and Neumann [7; 5].

Published: 19 May 2017 DOI: 10.2140/gt.2017.21.2281

- CAT(0) 2–complexes; see Bestvina, Kleiner and Sageev [9], with applications to the quasi-isometric rigidity of atomic right-angled Artin groups in their paper [8].
- Flats generated by Dehn twists in mapping class groups; see Behrstock, Kleiner, Minsky and Mosher [6].

In this paper, we will mainly focus on top-dimensional quasiflats and flats in CAT(0) cube complexes. All cube complexes in this paper will be finite-dimensional. Our first main result shows how the cubical structure interacts with quasiflats.

Theorem 1-1 *If X is a* CAT(0) *cube complex of dimension n, then for every n–quasiflat Q in X, there is a finite collection $O_1, \dots, O_k$ of n–dimensional orthant subcomplexes in X such that*

$$d_H\left(Q, \bigcup_{i=1}^{k} O_k\right) < \infty,$$

where d_H denotes the Hausdorff distance.

An *orthant* O of X is a convex subset which is isometric to the Cartesian product of finitely many half-lines $\mathbb{R}_{\geq 0}$. If O is both a subcomplex and an orthant, then O is called an *orthant subcomplex*. We caution the reader that the definition of orthant subcomplex here is slightly different from other places, ie we require an orthant subcomplex to be convex with respect to the CAT(0) metric.

The 2–dimensional case of Theorem 1-1 was proved in [9]. We will use this theorem as one of the main ingredients to study the coarse geometry of right-angled Artin groups (see Corollary 1-4 below and the remarks after). Also note that recently Behrstock, Hagen, and Sisto have obtained a quasiflat theorem of quite a different flavor in [4]. Their result does not imply our result and vice versa.

Based on Theorem 1-1, we study how the top-dimensional flats behave under quasi-isometries. In general, quasi-isometries between CAT(0) complexes of the same dimension do not necessarily preserve top dimension flats up to finite Hausdorff distance, even if the underlying spaces are cocompact. However, motivated by Haglund and Wise [20], we can define a large class of cube complexes such that top-dimensional flats behave nicely with respect to quasi-isometries between universal covers of these complexes. Our class contains all compact nonpositively curved special cube complexes up to finite cover [20, Proposition 3.10].

Definition 1-2 A cube complex W is *weakly special* if and only if it has the following properties:

(1) W is nonpositively curved.

(2) No hyperplane *self-osculates* or *self-intersects*.

The notions of self-osculation and self-intersection were introduced in [20, Definition 3.1].

Theorem 1-3 *Let W_1' and W_2' be two compact weakly special cube complexes with $\dim(W_1') = \dim(W_2') = n$, and let W_1 and W_2 be the universal covers of W_1' and W_2', respectively. If $f\colon W_1 \to W_2$ is an (L, A)–quasi-isometry, then there exists a constant $C = C(L, A)$ such that for any top-dimensional flat $F \subset W_1$, there exists a top-dimensional flat $F' \subset W_2$ with $d_H(f(F), F') < C$.*

We now apply this result to *right-angled Artin groups* (RAAGs). Recall that for every finite simplicial graph Γ with its vertex set denoted by $\{v_i\}_{i\in I}$, one can define a group using the following presentation:

$$\langle \{v_i\}_{i\in I} \mid [v_i, v_j] = 1 \text{ if } v_i \text{ and } v_j \text{ are adjacent}\rangle.$$

This is called the *right-angled Artin group with defining graph* Γ, and we denote it by $G(\Gamma)$. Each $G(\Gamma)$ can be realized as the fundamental group of a nonpositively curved cube complex $\bar{X}(\Gamma)$, which is called the Salvetti complex (see Charney [13] for a precise definition). The 2–skeleton of the Salvetti complex is the usual presentation complex for $G(\Gamma)$. The universal cover of $\bar{X}(\Gamma)$ is a CAT(0) cube complex, which we denote by $X(\Gamma)$.

Corollary 1-4 *Let Γ_1, Γ_2 be finite simplicial graphs, and let $\phi\colon X(\Gamma_1) \to X(\Gamma_2)$ be an (L, A)–quasi-isometry. Then:*

(1) $\dim(X(\Gamma_1)) = \dim(X(\Gamma_2))$.

(2) *There is a constant $D = D(L, A)$ such that for any top-dimensional flat F_1 in $X(\Gamma_1)$, we can find a flat F_2 in $X(\Gamma_2)$ such that $d_H(\phi(F_1), F_2) < D$.*

This is the foundation for a series of work on quasi-isometric classification and rigidity of RAAGs by the author and Kleiner [21; 22; 24; 23].

Remark We could also use Theorem 1-3 to obtain an analogous statement for quasi-isometries between the Davis complexes of certain right-angled Coxeter groups, but in general the dimensions of maximal flats in a Davis complex are strictly smaller than the dimension of complex itself, so we need extra condition on the right-angled Coxeter groups; see Corollary 5-18 for a precise statement.

Corollary 1-4 implies that ϕ maps chains of top-dimensional flats to chains of top-dimensional flats, and this gives rise to several quasi-isometry invariants for RAAGs. More precisely, we consider a graph $\mathcal{G}_d(\Gamma)$ where the vertices are in 1–1 correspondence to top-dimensional flats in $X(\Gamma)$ and two vertices are connected by an edge if and only if the coarse intersection of the corresponding flats has dimension $\geq d$. The connectedness of $\mathcal{G}_d(\Gamma)$ can be read off from Γ, which gives us the desired invariants.

Definition 1-5 Let $d \geq 1$ be an integer. Let Γ be a finite simplicial graph and let $F(\Gamma)$ be the flag complex that has Γ as its 1–skeleton. Γ has *property* (P_d) if and only if:

(1) Any two top-dimensional simplices Δ_1 and Δ_2 in $F(\Gamma)$ are connected by a $(d-1)$*–gallery*.

(2) For any vertex $v \in F(\Gamma)$, there is a top-dimensional simplex $\Delta \subset F(\Gamma)$ such that Δ contains at least d vertices that are adjacent to v.

A sequence of n–dimensional simplices $\{\Delta_i\}_{i=1}^{p}$ in $F(\Gamma)$ is a k*–gallery* if $\Delta_i \cap \Delta_{i+1}$ contains a k–dimensional simplex for $1 \leq i \leq p-1$.

Theorem 1-6 *$\mathcal{G}_d(\Gamma)$ is connected if and only if Γ has property (P_d). In particular, for any $d \geq 1$, property (P_d) is a quasi-isometry invariant for RAAGs.*

Remark Another interesting fact in the case $d = 1$ is that one can tell whether Γ admits a nontrivial join decomposition by looking at the diameter of $\mathcal{G}_1(\Gamma)$. This basically follows from the argument in Dani and Thomas [14]. See Theorem 5-30 for a precise statement. Thus in the case of $X(\Gamma)$, one can determine whether the space splits as a product by looking at the intersection pattern of top-dimensional flats. We ask whether this is true in general: if Z is a cocompact geodesically complete CAT(0) space that has n–flats but not $(n+1)$–flats, can one determine whether Z splits as a product of two unbounded CAT(0) spaces by looking at the intersection pattern of n–flats in Z?

Actually, a large portion of our discussion generalizes to n–dimensional quasiflats in CAT(0) spaces of geometric dimension $= n$ (the notion of geometric dimension and its relation to other notions of dimension are discussed in Kleiner [28]). This will be discussed in the appendix and see Theorem A-18 and Theorem A-19 for a summary.

In particular, this leads to a short proof of the following result, which was previously established in Kleiner and Leeb [30], Eskin and Farb [17], and Wortman [37] by different methods, and it is one of the main ingredients in proving quasi-isometric rigidity for Euclidean buildings.

Theorem 1-7 *Let Y be a Euclidean building of rank n, and let $Q \subset Y$ be an n–quasiflat. Then there exist finitely many Weyl cones $\{W_i\}_{i=1}^h$ such that*

$$d_H\left(Q, \bigcup_{i=1}^{h} W_i\right) < \infty.$$

On the way to Theorem 1-7, we also give a more accessible proof of the following weaker version of one of the main results in Kleiner and Lang [29].

Theorem 1-8 *Let $q\colon Y \to Y'$ be a quasi-isometric embedding, where Y and Y' are* CAT(0) *spaces of geometric dimension $\le n$. Then q induces a monomorphism $q_*\colon H_{n-1}(\partial_T Y) \to H_{n-1}(\partial_T Y')$. If q is a quasi-isometry, then q_* is an isomorphism.*

Here $\partial_T Y$ and $\partial_T Y'$ denote the Tits boundary of Y and Y' respectively.

1.2 Sketch of proofs

1.2.1 Proof of Theorems 1-1 and 1-7 The proof of Theorem 1-1 has five steps, as below. The first one follows Bestvina, Kleiner and Sageev [9] closely, but the others are different, since part of the argument in [9] depends heavily on special features of dimension 2, and does not generalize to the n–dimensional case.

Let X be a CAT(0) piecewise Euclidean polyhedral complex with $\dim(X) = n$, and let $Q\colon \mathbb{E}^n \to X$ be a top-dimensional quasiflat in X.

Step 1 Following [9], one can replace the top-dimensional quasiflat, which usually contains local wiggles, by a minimizing object which is more rigid.

More precisely, let us assume without of loss of generality that Q is a continuous quasi-isometric embedding. Let $[\mathbb{E}_n]$ be the fundamental class in the n^{th} locally finite homology group of $\mathbb{E}^n$ and let $[\sigma] = Q_*([\mathbb{E}_n])$. Let S be the support set (Definition 3-1) of $[\sigma]$. It turns out that S has nice local property (it is a subcomplex with geodesic extension property) and asymptotic property (it looks like a cone from far away). Moreover, $d_H(S, Q) < \infty$.

In the next few steps, we study the structure of S by looking at its "boundary".

Recall that X has a Tits boundary $\partial_T X$, whose points are asymptotic classes of geodesic rays in X, and the asymptotic angle between two geodesic rays induces a metric on $\partial_T X$. See Section 2.2 for a precise definition. We define the boundary of S, denoted $\partial_T S$, to be the subset of $\partial_T X$ corresponding to geodesic rays inside S.

Step 2 We produce a collection of orthants in X from S. More precisely, we find an embedded simplicial complex $K \subset \partial_T X$ such that $\partial_T S \subset K$. Moreover, K is made of right-angled spherical simplices, each of which is the boundary of an isometrically embedded orthant in X. This step depends on the cubical structure of X, and is discussed in Section 4.1.

Step 3 We show $\partial_T S$ is actually a cycle. Namely, it is the "boundary cycle at infinity" of the homology class $[\sigma]$. This step does not depend on the cubical structure of X and is actually true in greater generality by the much earlier, but still unpublished work of Kleiner and Lang [29]. However, their paper was based on metric current theory. Under the assumption of Theorem 1-1, we are able to give a self-contained account which only requires homology theory; see Section 4.2.

Step 4 We deduce from the previous two steps that $\partial_T S$ is a cycle made of $(n-1)$–dimensional all-right spherical simplices. Moreover, each simplex is the boundary of an orthant in X.

Step 5 We finish the proof by showing S is Hausdorff close to the union of these orthants. See Section 4.3 for the last two steps.

If X is a Euclidean building, then it is already clear that the cycle at infinity can be represented by a cellular cycle, since the Tits boundary is a polyhedral complex (a spherical building). The problem is that X itself may not be a polyhedral complex. There are several ways to get around this point. Here we deal with it by generalizing several results of [9] to CAT(0) spaces of finite geometric dimension, which is of independent interest.

1.2.2 Proof of Theorem 1-3 Let W_1 and W_2 be the universal covers of two weakly special cube complexes. We also assume $\dim(W_1) = \dim(W_2) = n$. Our starting point is similar to the treatment in Kleiner and Leeb [30] and Bestvina, Kleiner and Sageev [8]. Let $\mathcal{KQ}(W_i)$ be the lattice generated by finite unions, coarse intersections and coarse subtractions of top-dimensional quasiflats in W_i, modulo finite Hausdorff distance. Any quasi-isometry $q\colon W_1 \to W_2$ will induce a bijection $q_\sharp\colon \mathcal{KQ}(W_1) \to \mathcal{KQ}(W_2)$.

It suffices to study the combinatorial structure of $\mathcal{KQ}(W_i)$. By Theorem 1-1, each element $[A] \in \mathcal{KQ}(W_i)$ is made of a union of top-dimensional orthants, together with several lower dimensional objects. We denote the number of top-dimensional orthants in $[A]$ by $\|[A]\|$. $[A]$ is *essential* if $\|[A]\| > 0$, and $[A]$ is *minimal essential* if for any $[B] \in \mathcal{KQ}(W_i)$ with $[B] \subset [A]$ (ie B is coarsely contained in A) and $[B] \neq [A]$, we have $\|[B]\| = 0$.

It suffices to study the minimal essential elements of $\mathcal{KQ}(W_i)$, since every element in $\mathcal{KQ}(W_i)$ can be decomposed into minimal essential elements together with several lower dimensional objects. In the case of universal covers of special cube complexes, these elements have nice characterizations and behave nicely with respect to quasi-isometries:

Theorem 1-9 *If* $[A] \in \mathcal{KQ}(W_i)$ *is minimal essential, then there exists a convex subcomplex* $K \subset W_i$ *which is isometric to* $(\mathbb{R}_{\geq 0})^k \times \mathbb{R}^{n-k}$ *such that* $[K] = [A]$.

Theorem 1-10 $|q_\sharp([A])| = |[A]|$ *for any minimal essential element* $[A] \in \mathcal{KQ}(W_1)$.

Theorem 1-3 essentially follows from the above two results.

1.3 Organization of the paper

In Section 2 we will recall several basic facts about $\mathrm{CAT}(\kappa)$ spaces and $\mathrm{CAT}(0)$ cube complexes. We also collect several technical lemmas in this section, which will be used later.

In Section 3 we will review the discussion in [9] which will enable us to replace the top-dimensional quasiflat by the support set of the corresponding homology class. In Section 4 we will study the geometry of this support set and prove Theorem 1-1. In Sections 5.1 and 5.2, we look at the behavior of top-dimensional flats in the universal covers of weakly special cube complexes and prove Theorem 1-3 and Corollary 1-4. In Section 5.3, we use Corollary 1-4 to establish several quasi-isometric invariants for RAAGs.

In the appendix, we generalize some results of Sections 3 and 4 to $\mathrm{CAT}(0)$ spaces of finite geometric dimension and prove Theorem 1-8 and Theorem 1-7.

Acknowledgements This paper is part of the author's PhD thesis and it was finished under the supervision of B Kleiner. The author would like to thank B Kleiner for all the helpful suggestions and stimulating discussions. The author is grateful to B Kleiner and U Lang for sharing the preprint [29], which influenced several ideas in this paper. The author also thanks the referee for extremely helpful comments and clarifications.

2 Preliminaries

We start with some basic notation. The open balls and closed balls of radius r in a metric space will be denoted by $B(p,r)$ and $\overline{B}(p,r)$ respectively. The sphere of radius r centered at p is denoted by $S(p,r)$. The open r–neighborhood of a set A in a metric space is denoted by $N_r(A)$. The diameter of A is denoted by $\mathrm{diam}(A)$.

For a metric space K, $C_\kappa K$ denotes the κ–cone over K; see [11, Definition I.5.6]. When $\kappa = 0$, we call it the Euclidean cone over K and denote it by CK for simplicity. All products of metric spaces in this paper will be l^2–products.

The closed and open stars of a vertex v in a polyhedral complex are denoted by $\overline{st}(v)$ and $st(v)$ respectively. We use "$*$" for the join of two polyhedral complexes and "$\circ$" for the join of two graphs.

2.1 M_k–polyhedral complexes with finite shapes

In this section, we summarize some results about M_k–polyhedral complexes with finitely many isometry types of cells from [11, Chapter I.7], see also [10].

An M_k–polyhedral complex is obtained by taking a disjoint union of a collection of convex polyhedra from the complete simply connected n–dimensional Riemannian manifolds with constant curvature equal to k (n is not fixed) and gluing them along isometric faces. The complex is endowed with the quotient metric (see [11, Definition I.7.37]). Note that the topology induced by the quotient metric may be different from the topology as a cell complex.

An M_1–polyhedral complex is also called a piecewise spherical complex. If the complex is made of right-angled spherical simplices, then it is also called an all-right spherical complex. A M_0–polyhedral complex is also called a piecewise Euclidean complex.

We are mainly interested in the case where the collection of convex polyhedra we use to build the complex has only finitely many isometry types. Following [11], we denote the isometry classes of cells in K by $\mathrm{Shape}(K)$. Note that we can barycentrically subdivide any M_k–polyhedral complex twice to get an M_k–simplicial complex.

For an M_k–polyhedral complex K and a point $x \in K$, we denote the unique open cell of K which contains x by $\mathrm{supp}(x)$ and the closure of $\mathrm{supp}(x)$ by $\mathrm{Supp}(x)$. We also denote the geometric link of x in K by $\mathrm{Lk}(x, K)$; see [11, Section I.7.38]. In this paper, we always truncate the usual length metric on $\mathrm{Lk}(x, K)$ by π. If an ϵ–ball $B(x, \epsilon)$ around x has the properties that

- $B(x, \epsilon)$ is contained in the open star of x in K,
- $B(x, \epsilon)$ is isometric to the ϵ–ball centered at the cone point in $C_k(\mathrm{Lk}(x, K))$,

then we call $B(x, \epsilon)$ a *cone neighborhood of* x.

Theorem 2-1 [11, Theorem I.7.39] *Suppose K is an M_k–polyhedral complex with* $\mathrm{Shape}(K)$ *finite. Then for every $x \in K$, there exists a positive number ϵ (depending on x) such that $B(x, \epsilon)$ is a cone neighborhood of x.*

Theorem 2-2 [11, Theorem I.7.50] *Suppose K is an M_k–polyhedral complex with* Shape(K) *finite. Then K is a complete geodesic metric space.*

Lemma 2-3 *If K is an M_k–polyhedral complex with* Shape(K) *finite, then there exist positive constants c_1 and c_2, which depend on* Shape(K)*, such that every geodesic segment in K of length L is contained in a subcomplex which is a union of at most $c_1 L + c_2$ closed cells.*

This lemma follows from [11, Corollaries I.7.29 and I.7.30].

2.2 CAT(κ) spaces

Please see [11] for an introduction to CAT(κ) spaces.

Let X be a CAT(0) space and pick $x, y \in X$. We denote by $\overline{xy}$ the unique geodesic segment joining x and y. For any $y, z \in X \setminus \{x\}$, we denote the comparison angle between $\overline{xy}$ and $\overline{xz}$ at x by $\overline{\angle}_x(y,z)$ and the Alexandrov angle by $\angle_x(y,z)$.

The Alexandrov angle induces a distance on the space of germs of geodesics emanating from x. The completion of this metric space is called the *space of directions* at x and is denoted by $\Sigma_x X$. The tangent cone at x, denoted $T_x X$, is the Euclidean cone over $\Sigma_x X$. Following [9], we define the logarithmic map $\log_p\colon X \setminus \{x\} \to \Sigma_x X$ by sending $y \in X \setminus \{x\}$ to the point in $\Sigma_x X$ represented by $\overline{xy}$. Similarly, one can define $\log_x\colon X \to T_x X$. For a constant speed geodesic $l\colon [a,b] \to X$, we denote by $l^-(t)$ and $l^+(t)$ respectively the incoming and outgoing directions in $\Sigma_{l(t)} X$ for $t \in [a,b]$. Note that if X is a CAT(0) M_k–polyhedral complex with finitely many isometry types of cells, then $\Sigma_x X$ is naturally isometric to $\mathrm{Lk}(x, X)$, so we will identify these two objects.

Let us denote the Tits boundary of X by $\partial_T X$. We also have a well-defined map $\log_x\colon \partial_T X \to \Sigma_x X$. For $\xi_1, \xi_2 \in \partial_T X$, recall that the Tits angle $\angle_T(\xi_1, \xi_2)$ between them is defined as

$$\angle_T(\xi_1,\xi_2) = \sup_{x \in X} \angle_x(\xi_1,\xi_2).$$

This induces a metric on $\partial_T X$, which is called the *angular metric*. There are several different ways to define $\angle_T(\xi_1,\xi_2)$ (see [30, Section 2.3] or [11, Chapter II.9]):

Lemma 2-4 *Let X be a complete* CAT(0) *space and let ξ_1, ξ_2 be as above. Pick a base point $p \in X$, and let l_1 and l_2 be two unit speed geodesic rays emanating from p such that $l_i(\infty) = \xi_i$ for $i = 1, 2$. Then:*

(1) $\angle_T(\xi_1,\xi_2)=\lim_{t,t'\to\infty}\overline{\angle}_p(l_1(t),l_2(t'))$

(2) $\angle_T(\xi_1,\xi_2)=\lim_{t\to\infty}\angle_{l_1(t)}(\xi_1,\xi_2)$

(3) $2\sin(\angle_T(\xi_1,\xi_2)/2)=\lim_{t\to\infty}d(l_1(t),l_2(t))/t$

The space $(\partial_T X,\angle_T)$ is $\mathrm{CAT}(1)$; see [11, Chapter II.9]. We denote the Tits cone, which is the Euclidean cone over $\partial_T X$, by $C_T X$. Note that $C_T X$ is $\mathrm{CAT}(0)$. Denote the cone point of $C_T X$ by o. Then for each $p\in X$, there is a well-defined 1–Lipschitz logarithmic map $\log_p\colon C_T X\to X$ sending a geodesic ray $\overline{o\xi}\subset C_T X$ ($\xi\in\partial_T X$) to the geodesic ray $\overline{p\xi}\subset X$. This also gives rise to two other 1–Lipschitz logarithmic maps,

$$\log_p\colon C_T X\to T_p X,\quad \log_p\colon \partial_T X\to \Sigma_p X.$$

We always have $\angle_p(\xi_1,\xi_2)\le\angle_T(\xi_1,\xi_2)$, and the following flat sector lemma (see [30, Section 2.3] or [11, Chapter II.9]) describes when the equality holds.

Lemma 2-5 *Let X, ξ_1, ξ_2, l_1, l_2 and p be as above. If $\angle_T(\xi_1,\xi_2)=\angle_p(\xi_1,\xi_2)<\pi$, then the convex hull of l_1 and l_2 in X is isometric to a sector of angle $\angle_p(\xi_1,\xi_2)$ in the Euclidean plane.*

Any convex subset $C\subset X$ is also a $\mathrm{CAT}(0)$ space (with the induced metric) and there is an isometric embedding $i\colon\partial_T C\to\partial_T X$. There is a well-defined nearest point projection $\pi_C\colon X\to C$, which has the following properties.

Lemma 2-6 *Let X, C and π_C be as above. Then:*

(1) *π_C is 1–Lipschitz.*

(2) *For $x\notin C$ and $y\in C$ such that $y\neq\pi_C(x)$, we have $\angle_{\pi_C(x)}(x,y)\ge\frac{\pi}{2}$.*

See [11, Chapter II.2] for a proof of the above lemma.

Two convex subset C_1 and C_2 are *parallel* if $d(\,\cdot\,,C_1)|_{C_2}$ and $d(\,\cdot\,,C_2)|_{C_1}$ are constant. In this case, the convex hull of C_1 and C_2 is isometric to $C_1\times[0,d(C_1,C_2)]$. Moreover, $\pi_{C_1}|_{C_2}$ and $\pi_{C_2}|_{C_1}$ are isometric inverse to each other; see [30, Section 2.3.3] or [11, Chapter II.2].

Let $Y\subset X$ be a closed convex subset. We define the *parallel set* of Y, denoted by P_Y, to be the union of all convex subsets which are parallel to Y. P_Y is not a convex set in general, but when Y has the geodesic extension property, P_Y is closed and convex.

Now we turn to $\mathrm{CAT}(1)$ spaces. In this paper, $\mathrm{CAT}(1)$ spaces are assumed to have diameter $\le\pi$ (we truncate the length metric on the space by π). We say a subset of a $\mathrm{CAT}(1)$ space is *convex* if it is π–convex.

For a CAT(1) space Y and $p \in Y$, $K \subset Y$, we define the *antipodal set of z in K* to be $\mathrm{Ant}(p, K) := \{v \in K \mid d(v, p) = \pi\}$.

Let Y and Z be two metric spaces. Their *spherical join*, denoted by $Y * Z$, is the quotient space of $Y \times Z \times \left[0, \frac{\pi}{2}\right]$ under the identifications $(y, z_1, 0) \sim (y, z_2, 0)$ and $\left(y_1, z, \frac{\pi}{2}\right) \sim \left(y_2, z, \frac{\pi}{2}\right)$. One can write the elements in $Y * Z$ as formal sums $(\cos\alpha)y + (\sin\alpha)z$, where $\alpha \in \left[0, \frac{\pi}{2}\right]$, $y \in Y$ and $z \in Z$. Let

$$w_1 = (\cos\alpha_1)y_1 + (\sin\alpha_1)z_1, \quad w_2 = (\cos\alpha_2)y_2 + (\sin\alpha_2)z_2.$$

Their distance in $Y * Z$ is defined to be

$$d_{Y*Z}(w_1, w_2) = \cos\alpha_1 \cos\alpha_2 \cos\bigl(d^{\pi}_Y(y_1, y_2)\bigr) + \sin\alpha_1 \sin\alpha_2 \sin\bigl(d^{\pi}_Z(z_1, z_2)\bigr),$$

where d^{π}_Y is the metric on Y truncated by π, similarly for d^{π}_Z.

When Y is only one point, $Y * Z$ is the spherical cone over Z. When Y consists of two points a distance π from each other, $Y * Z$ is the spherical suspension of Z. The spherical join of two CAT(1) spaces is still CAT(1).

Definition 2-7 (cell structure on the link) Let X be an M_κ–polyhedral complex and pick a point $x \in X$. Suppose σ_x is the unique closed cell which contains x as its interior point. Then $\mathrm{Lk}(x, X)$ is isometric to $\mathrm{Lk}(x, \sigma_x) * \mathrm{Lk}(\sigma_x, X) = \mathbb{S}^{k-1} * \mathrm{Lk}(\sigma_x, X)$, where k is the dimension of σ_x. Note that $\mathrm{Lk}(\sigma_x, X)$ has a natural M_1–polyhedral complex structure which is induced from the ambient space X.

When X is made of Euclidean rectangles, $\mathrm{Lk}(\sigma_x, X)$ is an all-right spherical complex. Moreover, there is a canonical way to triangulate $\mathrm{Lk}(x, \sigma_x)$ into an all-right spherical complex which is isomorphic to an octahedron as simplicial complexes. The vertices of $\mathrm{Lk}(x, \sigma_x)$ come from segments passing through x which are parallel to edges of σ_x. Thus $\mathrm{Lk}(\sigma_x, X)$ has a natural all-right spherical complex structure. In general, there is no canonical way to triangulate $\mathrm{Lk}(x, \sigma_x)$. However, there are still cases when we want to treat $\mathrm{Lk}(x, X)$ as a piecewise spherical complex. In such cases, one can pick an arbitrary all-right spherical complex structure on $\mathrm{Lk}(x, \sigma_x)$.

If X is CAT(0), then we can identify $\Sigma_x X$ with $\mathrm{Lk}(x, X)$. In this case, $\Sigma_x X$ is understood to be equipped with the above polyhedral complex structure.

Any two points of distance less than π from each other in a CAT(1) space are joined by a unique geodesic. A generalization of this fact would be the following.

Lemma 2-8 *Let Y be* CAT(1) *and let $\Delta \subset Y$ be an isometrically embedded spherical k–simplex with its vertices denoted by $\{v_i\}_{i=0}^k$. Pick $v \in \Delta$ and $v' \in Y$. If $d(v', v_i) \le d(v, v_i)$ for all i, then $v = v'$.*

By spherical simplices, we always means those which are not too large, ie those contained in an open hemisphere.

Proof We proceed by induction. When $k = 1$, it follows from the uniqueness of geodesics. In general, since $\triangle = \triangle_1 * \triangle_2$, where $\triangle_1$ is spanned by vertices $\{v_i\}_{i=0}^{k-2}$ and $\triangle_2$ is spanned by v_{k-1} and v_k, there exists $w \in \triangle_2$ such that $v \in \triangle_1 * \{w\}$. Triangle comparison implies $d(v', w) \leq d(v, w)$, so we can apply the induction assumption to the $(k-1)$–simplex $\triangle_1 * \{w\}$, which implies $v = v'$. □

Lemma 2-9 *Let Y be a $\mathrm{CAT}(1)$ piecewise spherical complex with finitely many isometry types of cells, and let $K \subset Y$ be a subcomplex which is a spherical suspension (in the induced metric). Pick a suspension point $v \in K$. Then all points in $\mathrm{Supp}(v)$ are suspension points of K and we have a splitting $K = \mathbb{S}^k * K'$, where $k = \dim(\mathrm{Supp}(v))$ and $\mathbb{S}^k$ is the standard sphere of dimension k.*

Proof By Theorem 2-1, v has a small neighborhood isometric to the ϵ–ball centered at the cone point in the spherical cone over $\Sigma_v K$. Since v is a suspension point, $K = \mathbb{S}^0 * \mathrm{Lk}(v, K) = \mathbb{S}^0 * \Sigma_v K$. However, $\Sigma_v K = \Sigma_v \mathrm{Supp}(v) * K' = \mathbb{S}^{k-1} * K'$ for some K', thus $K = \mathbb{S}^k * K'$. Also every point in $\mathrm{Supp}(v)$ belongs to the $\mathbb{S}^k$–factor, hence is a suspension point. □

2.3 CAT(0) cube complexes

All cube complexes in this paper are assumed to be finite-dimensional.

Every cube complex X (a polyhedral complex whose building blocks are cubes) has a canonical cubical metric: endow each n–cube with the standard metric of the unit cube in Euclidean n–space $\mathbb{E}^n$, then glue these cubes together to obtain a piecewise Euclidean metric on X. This metric is complete and geodesic if X is of finite dimension, and is $\mathrm{CAT}(0)$ if the link of each vertex is a flag complex [11; 19].

Now we come to the notion of hyperplane, which is the cubical analogue of "track" introduced in [16]. A *hyperplane* h in a cube complex X is a subset such that:

(1) h is connected.

(2) For each cube $C \subset X$, $h \cap C$ is either empty or a union of mid-cubes of C.

(3) h is minimal, ie if there exists $h' \subset h$ satisfying (1) and (2), then $h = h'$.

Recall that a *mid-cube* of $C = [0, 1]^n$ is a subset of form $f_i^{-1}\left(\frac{1}{2}\right)$, where f_i is one of the coordinate functions.

For each edge $e \in X$, there exists a unique hyperplane which intersects e in one point. This is called the hyperplane *dual* to the edge e. Following [20], we say a hyperplane h *self-intersects* if there exists a cube C such that $C \cap h$ contains at least two different mid-cubes. A hyperplane h *self-osculates* if there exist two different edges e_1 and e_2 such that (1) $e_1 \cap e_2 \neq \varnothing$; (2) e_1 and e_2 are not consecutive edges in a 2–cube; (3) $e_i \cap h \neq \varnothing$ for $i = 1, 2$.

Let X be a CAT(0) cube complex, and let $e \subset X$ be an edge. Denote the hyperplane dual to e by h_e. Suppose $\pi_e \colon X \to e \cong [0, 1]$ is the CAT(0) projection. It is known that:

(1) h_e is embedded, ie the intersection of h_e with every cube in X is either a mid-cube, or an empty set.

(2) h_e is a convex subset of X, and h_e with the induced cell structure from X is also a CAT(0) cube complex.

(3) $h_e = \pi_e^{-1}\left(\frac{1}{2}\right)$.

(4) $X \setminus h_e$ has exactly two connected components; they are called *halfspaces*.

(5) If N_h is a union of closed cells in X which has nontrivial intersection with h_e, then N_h is a convex subcomplex of X and N_h is isometric to $h_e \times [0, 1]$. We call N_h the *carrier* of h_e. Note that $N_h = P_e$, where P_e is the parallel set of e.

We refer to [36] for more information about hyperplanes.

Now we investigate the coarse intersection of convex subcomplexes. The following lemma adjusts [8, Lemma 2.3] to our cubical setting.

Lemma 2-10 *Let X be a* CAT(0) *cube complex of dimension n, and let C_1, C_2 be convex subcomplexes. Suppose $\triangle = d(C_1, C_2)$, $Y_1 = \{y \in C_1 \mid d(y, C_2) = \triangle\}$ and $Y_2 = \{y \in C_2 \mid d(y, C_1) = \triangle\}$. Then:*

(1) *Y_1 and Y_2 are not empty.*

(2) *Y_1 and Y_2 are convex; π_{C_1} maps Y_2 isometrically onto Y_1 and π_{C_2} maps Y_1 isometrically onto Y_2; the convex hull of $Y_1 \cup Y_2$ is isometric to $Y_1 \times [0, \triangle]$.*

(3) *Y_1 and Y_2 are subcomplexes.*

(4) *There exists $A = A(\Delta, n, \epsilon)$ such that if $p_1 \in C_1$, $p_2 \in C_2$ and $d(p_1, Y_1) \geq \epsilon > 0$, $d(p_2, Y_2) \geq \epsilon > 0$, then*

$$d(p_1, C_2) \geq \triangle + A d(p_1, Y_1), \quad d(p_2, C_1) \geq \triangle + A d(p_2, Y_2). \tag{2-11}$$

Proof For assertion (1), since X has finite dimension, X has only finitely many isometry types of cells; we use the "quasicompact" argument of Bridson [10]. Suppose we have sequences of points $\{x_n\}_{n=1}^{\infty}$ in C_1 and $\{y_n\}_{n=1}^{\infty}$ in C_2 such that

$$d(x_n, y_n) < \Delta + \tfrac{1}{n}. \tag{2-12}$$

Then by Lemma 2-3, there exists an integer N such that for every n, the geodesic joining x_n and y_n is contained in a subcomplex K_n which is a union of at most N closed cells. Write $C_{1n} = C_1 \cap K_n$ and $C_{2n} = C_2 \cap K_n$, which are also subcomplexes. Since there are only finitely many isomorphisms types among $\{K_n\}_{n=1}^{\infty}$, we can assume, up to a subsequence, that there exist a finite complex K_∞ and subcomplexes $C_{1\infty}$, $C_{2\infty}$ of K_∞ such that for any n, there is a simplicial isomorphism $\varphi_n\colon K_n \to K$ with $\varphi_n(C_{1n}) = C_{1\infty}$ and $\varphi_n(C_{2n}) = C_{2\infty}$. By (2-12), $d_{K_\infty}(C_{1\infty}, C_{2\infty}) \le \Delta$ in the intrinsic metric of K_∞, so there exist $x_\infty \in C_{1\infty}$ and $y_\infty \in C_{2\infty}$ such that $d_{K_\infty}(x_\infty, y_\infty) \le \Delta$ by compactness of K_∞. It follows that $d_X(\varphi_n^{-1}(x_\infty), \varphi_n^{-1}(y_\infty)) \le d_{K_n}(\varphi_n^{-1}(x_\infty), \varphi_n^{-1}(y_\infty)) \le \Delta$.

We prove (4) with $\epsilon = 1$; the other cases are similar. A similar argument as above implies that there is a constant $A > 0$, such that if $x \in C_1$ and $d(x, Y_1) = 1$, then $d(x, C_2) > A + \Delta$. Note that the combinatorial complexity depends on Δ and n, so A also depends on Δ and n. Now for any $p_1 \in C_1$ and $d(p_1, Y_1) \ge 1$, let $p_0 = \pi_{Y_1}(p_1)$ and let $l\colon [0, d(p_0, p_1)] \to X$ be the unit speed geodesic from p_0 to p_1. We have $l(1) \in \{x \in C_1 \mid d(x, Y_1) = 1\}$, so $d(l(1), C_2)) > A + \Delta$ while $d(l(0), C_2)) = \Delta$. Then (2-11) follows from the convexity of the function $d(\,\cdot\,, C_2)$.

The assertion (2) is a standard fact in [11, Chapter II.2].

To prove (3), it suffices to prove that for every $y_1 \in Y_1$, we have $\mathrm{Supp}(y_1) \in Y_1$. Denote $y_2 = \pi_{C_2}(y_1) \in Y_2$ (hence $y_1 = \pi_{C_1}(y_2)$ by (2)) and $l\colon [0, \Delta] \to X$ the unit speed geodesic from y_1 to y_2. Recall that we use $l^-(t)$ and $l^+(t)$ to denote the incoming and outgoing directions of l in $\Sigma_{l(t)}X$ for $t \in [0, \Delta]$. Our goal is to construct a "parallel transport" of $\mathrm{Supp}(y_1)$ (which is a k–cube) along l.

Since X has only finitely many isometry types of cells, l is contained in a finite union of closed cells, and we can find a sequence of closed cubes $\{B_i\}_{i=1}^{N}$ and $0 = t_0 < t_1 < \cdots < t_{N-1} < t_N = \Delta$ such that each B_i contains $\{l(t) \mid t_{i-1} < t < t_i\}$ as interior points. We denote $\mathrm{Supp}(y_1)$ by $\Box_{t_0}$ from now on.

Starting At $l(0) = y_1$, we have a splitting $\Sigma_{y_1} X = \Sigma_{y_1} \Box_{t_0} * K_1$ for some convex subset $K_1 \subset \Sigma_{y_1} X$. Since $y_1 = \pi_{C_1}(y_2)$ and $\Box_{t_0} \subset C_1$, by Lemma 2-6 we know $d_{\Sigma_{y_1} X}(l^+(t_0), \Sigma_{y_1} \Box_{t_0}) \ge \frac{\pi}{2}$. Thus $l^+(t_0) \in K_1$ and $d_{\Sigma_{y_1} X}(v, l^+(t_0)) = \frac{\pi}{2}$ for any $v \in \Sigma_{y_1} \Box_{t_0}$. It follows that the segment $B_1 \cap l$ is orthogonal to $\Box_{t_0}$ in B_1. Since

$\Box_{t_0}$ is a subcube of B_1, by geometry of cubes, there is an isometric embedding $e_1\colon \Box_{t_0} \times [0, t_1] \to B_1$ with $e_1(y_1, t) = l(t)$. Denote $\Box_{t_1} = e_1(\Box_{t_0} \times \{t_1\})$; then $l(t_1) \in \Box_{t_1} \subseteq \mathrm{Supp}(l(t_1)) \subseteq B_1 \cap B_2$. Note that $\Box_{t_1}$ is not necessarily a subcomplex of B_1 (or B_2), but it is always parallel to some subcube of B_1 (or B_2).

Continuing By construction we know $d_{\Sigma_{l(t_1)}}(l^-(t_1), v) = \frac{\pi}{2}$ for $v \in \Sigma_{l(t_1)}\Box_{t_1}$, so $d_{\Sigma_{l(t_1)}}(l^+(t_1), \Sigma_{l(t_1)}\Box_{t_1}) \geq \frac{\pi}{2}$, since $d_{\Sigma_{l(t_1)}}(l^-(t_1), l^+(t_1)) = \pi$. However, there is a splitting $\Sigma_{l(t_1)}X = \Sigma_{l(t_1)}\Box_{t_1} * K_2$ for some convex subset $K_2 \subset \Sigma_{l(t_1)}X$. Thus $l^+(t_1) \in K_2$ and $d_{\Sigma_{l(t_1)}X}(v, l^+(t_1)) = \frac{\pi}{2}$ for any $v \in \Sigma_{l(t_1)}\Box_{t_1}$. It follows that inside B_2, the segment $B_2 \cap l$ is orthogonal to $\Box_{t_1}$. Recall that $\Box_{t_1}$ is parallel to a subcube of B_2, hence by geometry of cubes, we have an isometric embedding

$$e_2\colon \Box_{t_1} \times [t_1, t_2] \to B_2$$

with $e_2(y, t) = l(t)$ for some $y \in \Box_{t_1}$. Write $\Box_{t_2} = e_2(\Box_{t_1} \times \{t_2\})$; we know $\Box_{t_2}$ is parallel to some subcube of B_3, so one can proceed to construct an isometric embedding e_3 as before. More generally, we can build $e_i\colon \Box_{t_{i-1}} \times [t_{i-1}, t_i] \to B_i$ with $e_i(y, t) = l(t)$ for some $y \in \Box_{t_{i-1}}$ and $\Box_{t_i} = e_i(\Box_{t_{i-1}} \times \{t_i\})$ inductively. Note that $l(t_i) \in \Box_{t_i} \subseteq \mathrm{Supp}(l(t_i)) \subseteq B_i \cap B_{i+1}$ by construction.

Arriving Since $y_2 = l(t_N) \in B_N \cap C_2$, where B_N and C_2 are subcomplexes, we have $l(t_N) \in \Box_{t_N} \subseteq \mathrm{Supp}(l(t_N)) \subseteq B_N \cap C_2$ by construction. Moreover, we can concatenate the embeddings $\{e_i\}_{i=1}^N$ constructed in the previous step to obtain a map $e\colon \Box_{t_0} \times [0, \triangle] \to X$ such that

- $e(y, t) = l(t)$ for some $y \in \Box_{t_0}$;
- $e(\Box_{t_0} \times \{0\}) \subseteq C_1$;
- $e(\Box_{t_0} \times \{\triangle\}) \subseteq C_2$;
- e is 1–Lipschitz (e is actually an isometric embedding, since e is a local isometric embedding by construction).

Therefore $d(y, C_2) \leq \triangle$ for any $y \in \Box_{t_0}$ (recall that $\mathrm{Supp}(y_1) = \Box_{t_0}$), which implies assertion (3). □

Remark 2-13 (1) By the same proof, items (1), (2) and (4) in the above lemma are true for piecewise Euclidean CAT(0) complexes with finitely many isometry types of cells. However, (3) might not be true in such generality.

(2) If C_1 and C_2 are orthant subcomplexes, then by items (2) and (3), Y_1 (or Y_2) is isometric to $O \times \prod_{i=1}^k I_i$, where O is an orthant and each I_i is a finite interval. In other words, there exists an orthant subcomplex $O \subset X$ such that $d_H(Y_1, O) < \infty$.

(3) Equation (2-11) implies that for any $R_1, R_2 > 0$, we have

$$N_{R_1}(C_1) \cap N_{R_2}(C_2) \subset N_{R_1'}(Y_1), \quad N_{R_1}(C_1) \cap N_{R_2}(C_2) \subset N_{R_2'}(Y_2),$$

where

$$R_1' = \min\Big(1, \frac{R_1+R_2-\Delta}{A} + R_2\Big), \quad R_2' = \min\Big(1, \frac{R_1+R_2-\Delta}{A} + R_1\Big),$$

with $A = A(\Delta, n, 1)$. Moreover, $\partial_T C_1 \cap \partial_T C_2 = \partial_T Y_1 = \partial_T Y_2$.

The last remark implies that Y_1 and Y_2 capture the information about how C_1 and C_2 intersect coarsely. We use the notation $\mathcal{I}(C_1, C_2) = (Y_1, Y_2)$ to describe this situation, where $\mathcal{I}$ stands for the word "intersect". The next lemma gives a combinatorial description of Y_1 and Y_2.

Lemma 2-14 *Let X, C_1, C_2, Y_1 and Y_2 be as in Lemma 2-10. Pick an edge e in Y_1 (or Y_2), and let h be the hyperplane dual to e. Then $h \cap C_i \neq \varnothing$ for $i = 1, 2$. Conversely, if a hyperplane h' satisfies $h' \cap C_i \neq \varnothing$ for $i = 1, 2$, then*

$$\mathcal{I}(h' \cap C_1, h' \cap C_2) = (h' \cap Y_1, h' \cap Y_2)$$

and h' comes from the dual hyperplane of some edge e' in Y_1 (or Y_2).

Proof The first part of the lemma follows from the proof of Lemma 2-10. Let $\mathcal{I}(h' \cap C_1, h' \cap C_2) = (Y_1', Y_2')$. Pick $x \in Y_1'$ and set $x' = \pi_{h' \cap C_2}(x) \in Y_2'$. Then $\pi_{h' \cap C_1}(x') = x$. We identify the carrier of h' with $h' \times [0, 1]$. Since C_i is a subcomplex, $(h' \cap C_i) \times \big(\frac{1}{2} - \epsilon, \frac{1}{2} + \epsilon\big) = C_i \cap \big(h' \times \big(\frac{1}{2} - \epsilon, \frac{1}{2} + \epsilon\big)\big)$ for $i = 1, 2$ and $\epsilon < \frac{1}{2}$. Thus for any $y \in C_2$, one has $\angle_{x'}(x, y) \geq \frac{\pi}{2}$, which implies that $x' = \pi_{C_2}(x)$. Similarly, $x = \pi_{C_1}(x') = \pi_{C_1} \circ \pi_{C_2}(x)$, hence $x \in Y_1$ and $Y_1' \subset Y_1$. By the same argument, $Y_2' \subset Y_2$, thus $Y_i' = Y_i \cap h'$ for $i = 1, 2$ and the lemma follows. □

Definition 2-15 We call an isometrically embedded orthant O *straight* if for any $x \in O$, the space $\Sigma_x O$ is a subcomplex of $\Sigma_x X$ (see Definition 2-7 for the cell structure on $\Sigma_x X$). In particular, if the orthant is 1–dimensional, we will call it a *straight geodesic ray*. Note that O itself may not be a subcomplex.

Remark 2-16 Any k–dimensional straight orthant $O \subset X$ is Hausdorff close to an orthant subcomplex of X.

To see this, let $k' = \max_{x \in O}\{\dim(\mathrm{Supp}(x))\}$; we proceed by induction on $k' - k$. The case $k' - k = 0$ is clear. Assume $k' - k = m \geq 1$ and pick $x_0 \in O$ such that $\dim(\mathrm{Supp}(x_0)) = k'$. Then there exists $B \subset \mathrm{Supp}(x_0)$ such that $B \cong [0, 1]^k$, B is parallel to a k–dimensional subcube of $\mathrm{Supp}(x_0)$ and $O \cap \mathrm{Supp}(x_0) \subset B$. Choose a line segment $e \cong [0,1]$ in $\mathrm{Supp}(x_0)$ such that $x_0 \in e$, e is orthogonal to B and e is parallel to some edge e' of $\mathrm{Supp}(x_0)$.

Suppose h is the hyperplane dual to e and suppose $N_h \cong e \times h$ is the carrier of h. For any other point $x \in O$, the segment $\overline{x_0 x}$ is orthogonal to e by our construction, thus there exists a point $y \in e$ such that $O \subset \{y\} \times h \subset N_h$. Now we can endow $\{y\} \times h$ with the induced cubical structure and use our induction hypothesis to find an orthant complex O_1 in the k–skeleton of $\{y\} \times h$ such that $d_H(O, O_1) < \infty$. Since $N_h \cong e \times h$, we can slide O_1 along e in N_h to get an orthant subcomplex in the k–skeleton of X.

Lemma 2-17 *Let X be a* CAT(0) *cube complex. If l_1 and l_2 are two straight geodesic rays in X, then either $\angle_T(l_1, l_2) = 0$, or $\angle_T(l_1, l_2) \geq \frac{\pi}{2}$.*

Proof We can assume without loss of generality that l_1 and l_2 are in the 1–skeleton and $l_1(0)$ is a vertex of X. We parametrize these two geodesic rays by unit speed. Let $\{b_m\}_{m=1}^{\infty}$ be the collection of hyperplanes in X such that $b_m \cap l_1 = l_1\left(\frac{1}{2} + m\right)$, and let h_m be the halfspace bounded by b_m which contains l_1 up to a finite segment. Suppose N_m is the carrier of b_m.

Suppose $l_2 \cap b_m \neq \varnothing$ for infinitely many m. Since each b_m separates X, there exists an m_0 such that $l_2 \cap b_m \neq \varnothing$ for all $m \geq m_0$. Recall that l_2 is in the 1–skeleton, so for each $m \geq m_0$, there exists an edge e_m such that $e_m \subset l_2$, $e_m \subset N_m$ and $e_m \cap b_m$ is a point. Consider the function $f(t) = d(l_2(t), l_1)$ for $t \geq 0$. Then f is convex and there exist infinitely many intervals of unit length (they come from e_m for $m \geq m_0$) such that f restricted to each interval is constant, so there exists a t_0 such that $f|_{[t_0,\infty)}$ is constant, which implies $\angle_T(l_1, l_2) = 0$.

If $l_2 \cap b_m \neq \varnothing$ for finitely many m, then there exists an m_0 such that $h_{m_0} \cap l_2 = \varnothing$, which implies the CAT(0) projection of l_2 to l_1 is a finite segment. If $\angle_T(l_1, l_2) < \frac{\pi}{2}$, then $\pi_{l_1}(l_2)$ is an infinite segment by Lemma 2-4, which is a contradiction, so $\angle_T(l_1, l_2) \geq \frac{\pi}{2}$. □

We will see later on that the subset of $\partial_T X$ which is responsible for the behavior of top-dimensional quasiflats is spanned by those points represented by straight geodesic rays. The following lemma makes the word "span" precise.

Lemma 2-18 *Let X be a* CAT(0) *cube complex, and let $\{l_i\}_{i=1}^{k}$ be a collection of straight geodesic rays in X, emanating from the same base point p, such that $\angle_T(l_i, l_j) = \angle_p(l_i, l_j) = \frac{\pi}{2}$ for $i \neq j$. Then the convex hull of $\{l_i\}_{i=1}^{k}$ is a k–dimensional straight orthant.*

One may compare this lemma with [3, Propositions 2.10 and 2.11].

Proof By Lemma 2-5, l_i and l_j together bound an isometrically embedded quarter plane for $i \neq j$. We prove the lemma by induction and assume the claim is true for $\{l_i\}_{i=1}^{k-1}$. We parametrize l_k by arc length and denote by O_0 the straight orthant spanned by $\{l_i\}_{i=1}^{k-1}$. Note that $O_0 \cap l_k = p$.

For $s > 0$ and $1 \le i \le k-1$, let c_i be the geodesic ray such that (1) c_i is in the quarter plane spanned by l_k and l_i; (2) c_i starts at $l_k(s)$; (3) c_i is parallel to l_i. Thus $\angle_T(c_i, c_j) = \frac{\pi}{2}$ and $\angle_{l_k(s)}(c_i, c_j) \le \frac{\pi}{2}$ for $i \neq j$. Note that $\{c_i\}_{i=1}^{k-1}$ are also straight geodesic rays, and $\{\log_{l_k(s)} c_i\}_{i=1}^{k-1}$ are distinct points in the 0–skeleton of $\Sigma_{l_k(s)} X$. It follows that actually $\angle_{l_k(s)}(c_i, c_j) = \frac{\pi}{2}$ for $i \neq j$. Hence by the induction assumption, there is a straight orthant O_s spanned by $\{c_i\}_{i=1}^{k-1}$.

By Lemma 2-8, $\partial_T O_0 = \partial_T O_s$. Let $l \subset O_s$ be a unit-speed geodesic ray emanating from $l_k(s)$. Then $d(l(t), O_0)$ is a bounded convex function. Since $\Sigma_{l_k(s)} O_s$ is an all-right spherical simplex in $\Sigma_{l_k(s)} X$ spanned by $\{\log_{l_k(s)} c_i\}_{i=1}^{k-1}$, we have $\angle_{l_k(s)}(l(t), l_k(0)) = \frac{\pi}{2}$ for any $t > 0$. Similarly, we have $\angle_{l_k(0)}(y, l_k(s)) = \frac{\pi}{2}$ for any $y \in O_0 \setminus \{l_s(0)\}$. Hence by triangle comparison, $d(l(t), O_0)$ attains its minimum at $t = 0$. Thus $d(l(t), O_0)$ has to be a constant function. Thus $d(x, O_0) \equiv s$ for any $x \in O_s$, and similarly $d(x, O_s) \equiv s$ for any $x \in O_0$, which implies the convex hull of O_0 and O_s is isometric to $O_0 \times [0, s]$; see eg [11, Chapter II.2]. Moreover, the convex hull of O_0 and O_s is contained in the convex hull of O_0 and $O_{s'}$ for $s \le s'$. So the convex hull of $\{l_i\}_{i=1}^{k}$ is a straight orthant O. □

3 Proper homology classes of bounded growth

In this section we summarize some results from [9] about locally finite homology classes of certain polynomial growth and make some generalizations to adjust the results to our situation.

3.1 Proper homology and supports of homology classes

For an arbitrary metric space Z, we define the proper (singular) homology of Z with coefficients in an abelian group G, denoted $H_*^{\mathrm{p}}(Z; G)$, as follows. Elements in the proper n–chain group $C_n^{\mathrm{p}}(Z; G)$ are of the form $\Sigma_{\lambda \in \Lambda} g_\lambda \sigma_\lambda$ (here Λ may be infinite, $g_\lambda \in G$ and the σ_λ are singular n–simplices) such that for every bounded set $K \subset Z$, the set $\{\lambda \in \Lambda \mid g_\lambda \neq \mathrm{Id} \text{ and } \sigma_\lambda(\triangle^n) \cap K \neq \varnothing\}$ is finite. The usual boundary map gives rise to a group homomorphism $\partial\colon C_n^{\mathrm{p}}(Z; G) \to C_{n-1}^{\mathrm{p}}(Z; G)$, yielding a chain complex $C_*^{\mathrm{p}}(Z; G)$, and $H_*^{\mathrm{p}}(Z; G)$ is defined to be the homology of this chain complex.

We will use Greek letters $\alpha, \beta, \dots$ to denote (proper) singular chains. We denote the union of images of singular simplices in a (proper) singular chain α by $\mathrm{Im}\,\alpha$. If α is a (proper) cycle, we denote the corresponding (proper) homology class by $[\alpha]$.

We also define the relative version of proper homology $H^{\mathrm{p}}_*(Z,Y)$ for $Y \subset Z$ in a similar way (Y is endowed with the induced metric). Then there is a long exact sequence

$$\cdots \to H^{\mathrm{p}}_n(Y) \to H^{\mathrm{p}}_n(Z) \to H^{\mathrm{p}}_n(Z,Y) \to H^{\mathrm{p}}_{n-1}(Y) \to H^{\mathrm{p}}_{n-1}(Z) \to \cdots.$$

Moreover, by the usual procedure of subdividing the chains, we know excision holds. Namely, for a subspace W such that the closure of W is in the interior of Y, the map $H^{\mathrm{p}}_*(Z-W, Y-W) \to H^{\mathrm{p}}_*(Z,Y)$ induced by inclusion is an isomorphism. As a corollary, if $B \subset Z$ is bounded, then there is a natural isomorphism $H^{\mathrm{p}}_*(Z, Z-A) \cong H_*(Z, Z-A)$, since we can replace the pair $(Z, Z-A)$ by $(O, O-B)$ by excision, where O is a bounded open neighborhood of B. Pick a point $z \in Z \setminus Y$; then there is a homomorphism $i\colon H^{\mathrm{p}}_k(Z,Y) \to H^{\mathrm{p}}_k(Z, Z\setminus\{z\}) \cong H_k(Z, Z\setminus\{z\})$ induced by the inclusion of pairs $(Z,Y) \to (Z, Z-\{z\})$. The map i is called the *inclusion homomorphism.*

If Z is also a simplicial complex or polyhedral complex, we can similarly define the proper simplicial (or cellular) homology by considering the former sum of simplices or cells such that for every bounded subset $K \subset Z$, we have only finitely many terms which intersect K nontrivially. The simplicial version (or the cellular version) of the homology theory is isomorphic to the singular version in a simplicial complex (or polyhedral complex) by the usual proof in algebraic topology.

The proper homology depends on the metric of the space, so it is not a topological invariant. By definition, every proper chain is locally finite and we have a group homomorphism $H^{\mathrm{p}}_*(Z,Y) \to H^{\mathrm{lf}}_*(Z,Y)$, where $H^{\mathrm{lf}}_*(Z,Y)$ is the locally finite homology defined in [9]. If Z is a proper metric space, then these two homology theories are the same.

A continuous map $f\colon Z_1 \to Z_2$ is *(metrically) proper* if the inverse image of every bounded subset is bounded. In this case, we have an induced map on proper homology $f_*\colon H^{\mathrm{p}}_k(Z_1, G) \to H^{\mathrm{p}}_k(Z_2, G)$.

In the rest of this paper, we will always take $G = \mathbb{Z}/2\mathbb{Z}$ and omit G when we write the homology.

Definition 3-1 For $z \in Z \setminus Y$, let $i\colon H^{\mathrm{p}}_k(Z,Y) \to H_k(Z, Z\setminus\{z\})$ be the inclusion homomorphism defined as above. For $[\sigma] \in H^{\mathrm{p}}_k(Z,Y)$, we define the *support set* of $[\sigma]$, denoted $S_{[\sigma],Z,Y}$, to be $\{z \in Z \setminus Y \mid i_*[\sigma] \neq \mathrm{Id}\}$. We will write $S_{[\sigma],Z}$ if Y is empty, and use $S_{[\sigma]}$ to denote the support set if the underlying spaces Z and Y are clear.

It follows that $S_{[\sigma]} = \big(\bigcap_{[\sigma']=[\sigma]\in H^{\mathrm{p}}_k(Z,Y)} \operatorname{Im}\sigma'\big) \setminus Y$.

If $Z \subset Z_1$, then $S_{[\sigma],Z,Y} \supseteq S_{[\sigma],Z_1,Y}$. These two sets are equal if Z is open in Z_1. If Z is a polyhedral complex and $Y = \varnothing$, then the support set is always a subcomplex. In particular, if $[\sigma] \in H^{\mathrm{p}}_n(Z)$ is a nontrivial top-dimensional class, then $[\sigma]$ can be represented by a top-dimensional polyhedral cycle, which implies the support set $S_{[\sigma]} \neq \varnothing$. But the support of a nontrivial class can be empty if it is not top-dimensional.

The support sets of (proper) homology classes behave like the support sets of currents in the following situation.

Lemma 3-2 *Let Z_1 be a metric space of homological dimension $\le n$, and let $Y_1 \subset Z_1$ be a subspace. Pick $[\sigma] \in H^{\mathrm{p}}_n(Z_1, Y_1)$. If $f\colon (Z_1, Y_1) \to (Z_2, Y_2)$ is a proper map, then $S_{f_*[\sigma]} \subset f(S_{[\sigma]})$.*

Recall that Z_1 has $(\mathbb{Z}/2\mathbb{Z})$–homological dimension $\le n$ if $H_r(U, V) = 0$ for any $r > n$ and U, V open in Z_1.

Proof Pick $y \in S_{f_*[\sigma]}$. Since $f^{-1}(y)$ is bounded, we have the following commutative diagram:

$$\begin{array}{ccc} H^{\mathrm{p}}_n(Z_1, Y_1) & \xrightarrow{f_*} & H^{\mathrm{p}}_n(Z_2, Y_2) \\ \big\downarrow{\scriptstyle i_*} & & \big\downarrow{\scriptstyle i_*} \\ H_n(Z_1, Z_1 \setminus f^{-1}(y)) & \xrightarrow{f_*} & H_n(Z_2, Z_2 \setminus \{y\}) \end{array}$$

Thus if $S_{f_*[\sigma]} \neq \varnothing$, then $[\sigma'] = i_*[\sigma] \in H_n(Z_1, Z_1 \setminus f^{-1}(y))$ is nontrivial. It suffices to show there exists $x \in f^{-1}(y)$ such that $[\sigma']$ is nontrivial when viewed as an element in $H_n(Z_1, Z_1 \setminus \{x\})$.

Fix a singular chain $\sigma' \in C_n(Z_1, Z_1 \setminus f^{-1}(y))$ which represents $[\sigma']$. We argue by contradiction and assume that $[\sigma']$ is trivial in $H_n(Z_1, Z_1 \setminus \{x\})$ for any $x \in f^{-1}(y)$. Let $K = f^{-1}(y) \cap \operatorname{Im}\sigma'$. For each $x \in K$, there exists an $\epsilon(x) > 0$ such that $\overline{B}(x, 2\epsilon(x)) \cap \operatorname{Im}\partial\sigma' = \varnothing$ and $[\sigma']$ is trivial in $H_n\big(Z_1, Z_1 \setminus \overline{B}(x, 2\epsilon(x))\big)$. Since $f^{-1}(y)$ is bounded and closed, K is compact and we can find finitely many points $\{x_i\}_{i=1}^N$ in K such that $K \subset \bigcup_{i=1}^N B(x_i, \epsilon(x_i))$. Let $U = \bigcup_{i=1}^N B(x_i, \epsilon(x_i))$ and let $U' = (Z_1 \setminus f^{-1}(y)) \cup U$. Then Im $\sigma' \subset U'$ and $[\sigma']$ is trivial in $H_n(Z_1, U')$. We put $U'' = U' \setminus \big(\bigcup_{i=1}^N \overline{B}(x_i, 2\epsilon(x_i))\big)$. Then $\operatorname{Im}\partial\sigma' \subset U''$ and $U'' \subset Z_1 \setminus f^{-1}(y)$. So if we can show that $[\sigma']$ is trivial in $H_n(Z_1, U'')$, then $[\sigma']$ must be trivial in $H_n(Z_1, Z_1 \setminus f^{-1}(y))$, which yields a contradiction.

Let us assume $N = 1$. Then there is a Mayer–Vietoris sequence

$$H_{n+1}\big(Z_1, U' \cup (Z_1 \setminus \overline{B}(x_1, 2\epsilon(x_1)))\big) \to H_n(Z_1, U'') \\ \to H_n(Z_1, U') \oplus H_n\big(Z_1, Z_1 \setminus \overline{B}(x_1, 2\epsilon(x_1))\big).$$

The first term is trivial since Z_1 has homological dimension $\le n$ and $[\sigma']$ is trivial in the last term by construction, so $[\sigma']$ has to be trivial in the second term. Using an induction argument, we can obtain the contradiction similarly for $N \ge 2$. □

Remark 3-3 (1) The assumption on Z_1 is satisfied if Z_1 is a CAT(κ) space of geometric dimension $\le n$, see [28, Theorem A].

(2) The assumption on Z_1 is satisfied if Z_1 is a locally finite n–dimensional polyhedral complex (with the topology of a cell complex) or an M_k–polyhedral complex of finite shape, since such a space supports a CAT(1) metric which induces the same topology as its original metric.

3.2 The growth condition

In Sections 3.2–3.3, Y will be a piecewise Euclidean CAT(0) complex of dimension n. The following result shows every top-dimensional quasiflat in Y is Hausdorff close to the support set of some proper homology class. Therefore to understand quasiflats, we can focus on the support sets, which have nice local and asymptotic properties.

Lemma 3-4 [9, Lemma 4.3] *If $Q \subset Y$ is an (L, A)–quasiflat of dimension n, then there exists $[\sigma] \in H_n^{\mathrm{p}}(Y)$ satisfying the following conditions:*

(1) *There exists a constant $D = D(L, A)$ such that $d_H(S, Q) \le D$, where S is the support set of $[\sigma]$.*

(2) *There exists $a = a(L, A)$ such that for every $p \in Y$,*

$$\mathcal{H}^n(B(p, r) \cap S) \le a(1 + r)^n. \tag{3-5}$$

Here $\mathcal{H}^n$ denotes the n–dimensional Hausdorff measure and d_H denotes the Hausdorff distance.

Since Y is uniformly contractible, we can approximate the (L, A)–quasi-isometric embedding $q\colon \mathbb{R}^n \to Y$ by a Lipschitz (L, A)–quasi-isometric embedding $\widetilde{q}$, which is proper. Then $[\sigma]$ is chosen to be the pushforward of the fundamental class of $\mathbb{R}^n$ under $\widetilde{q}$.

The support set of a top-dimensional homology class enjoys the following geodesic extension property.

Lemma 3-6 [9, Lemma 3.1] *Let S be the support set of some top-dimensional proper homology class in Y. Pick arbitrary $p \in Y$ and $y \in S$. Then there is a geodesic ray $\overline{y\xi} \subset S$, which fits together with the geodesic segment $\overline{py}$ to form a geodesic ray $\overline{p\xi}$.*

Note that this lemma does not imply S is convex; see [9, Remark 3.2]. However, we still can define the Tits boundary of S.

Definition 3-7 Let Z be a CAT(0) space and let $A \subset Z$ be any subset. We define the *Tits boundary* of A, denoted $\partial_T A$, to be the set of points $\xi \in \partial_T Z$ such that there exists a geodesic ray $\overline{x\xi}$ such that $\overline{x\xi} \subset A$. The Tits boundary $\partial_T A$ is endowed with the usual Tits metric. We define the *Tits cone* of A, denoted $C_T A$, to be the Euclidean cone over $\partial_T A$.

Let S be as in Lemma 3-6. Then $\partial_T S$ is nonempty if S is nonempty.

There is a similar version of the geodesic extension property for the link $\Sigma_y S \subset \Sigma_y Y$, where $y \in S$.

Lemma 3-8 *Let S be as in Lemma 3-6. Then for any point $y \in Y$, $\Sigma_y S$ is the support set of some top-dimensional homology class in $\Sigma_y Y$.*

Proof By subdividing Y in an appropriate way, we may assume y is a vertex of Y. Suppose $S = S_{[\sigma]}$. We can represent $[\sigma]$ as a cellular cycle $\sigma = \Sigma_{\lambda\in\Lambda}\eta_\lambda$, where the η_λ are closed top-dimensional cells in Y (recall that we are using $\mathbb{Z}/2\mathbb{Z}$ coefficients, so all the coefficients are either 0 or 1). Then $S = \bigcup_{\lambda\in\Lambda}\eta_\lambda$. Let $\Lambda_y = \{\lambda \in \Lambda \mid y \in \eta_\lambda\}$. Since η is a cycle, $\eta_y = \Sigma_{\lambda\in\Lambda_y}\mathrm{Lk}(y, \eta_\lambda)$ is a top-dimensional cycle in the link $\mathrm{Lk}(y, Y) \cong \Sigma_y Y$. Moreover, $S_{[\eta_y]} = \bigcup_{\lambda\in\Lambda_y}\mathrm{Lk}(y, \eta_\lambda) = \mathrm{Lk}(y, S)$. □

Lemma 3-9 *Let K be a k–dimensional* CAT(1) *piecewise spherical complex, and let $K' \subset K$ be the support set of a top-dimensional homology class. Pick arbitrary $w \in K$, $v \in K'$, and suppose $\overline{wv}$ is a local geodesic joining v and w. Then there is a (nontrivial) local geodesic segment $\overline{vv'} \subset K'$ which fits together with $\overline{wv}$ to form a local geodesic segment $\overline{wv'}$. Moreover,* $\mathrm{length}(\overline{vv'})$ *can be as large as we want.*

Now we turn to the global properties of S. Since we are in a CAT(0) space, for any $p \in Y$ and $0 < r \le R$, we have a map $\Phi_{r,R}\colon B(p, R) \to B(p, r)$ obtained by contracting points toward p by a factor of r/R. This contracting map together with Lemma 3-6 implies $B(p, r) \cap S \subset \Phi_{r,R}(B(p, R) \cap S)$ [9, Corollary 3.3, item 1].

Since $\Phi_{r,R}$ is (r/R)–Lipschitz, we have the following result.

Theorem 3-10 [9, Corollary 3.3] *Let S be the support set of some top-dimensional proper homology class in Y, and let $n = \dim(Y)$. Then:*

(1) **Monotonicity of density** *For all $0 \le r \le R$,*

$$\frac{\mathcal{H}^n(B(p,r)\cap S)}{r^n} \le \frac{\mathcal{H}^n(B(p,R)\cap S)}{R^n}. \tag{3-11}$$

(2) **Lower density bound** *For all $p \in S$ and $r > 0$,*

$$\mathcal{H}^n(B(p,r)\cap S) \ge \omega_n r^n, \tag{3-12}$$

with equality only if $B(p,r)\cap S$ is isometric to an r–ball in $\mathbb{E}^n$, where ω_n is the volume of an n–dimensional Euclidean ball of radius 1.

From (3-11) we know the quantity

$$\frac{\mathcal{H}^n(B(p,r)\cap S)}{r^n} \tag{3-13}$$

is monotone increasing with respect to r, and (3-5) tells us that if S comes from a top-dimensional quasiflat, then (3-13) is bounded above by some constant. Thus the limit exists and is finite as $r \to \infty$. More generally, we will consider those top-dimensional proper homology classes whose support sets S satisfy

$$\lim_{r\to+\infty} \frac{\mathcal{H}^n(B(p,r)\cap S)}{r^n} < \infty, \tag{3-14}$$

where $n = \dim(Y)$. We call them *proper homology classes of Cr^n growth.* These classes form a subgroup of $H^{\mathrm{p}}_n(Y)$, which will be denoted by $H^{\mathrm{p}}_{n,n}(Y)$.

The following lemma can be proved by a packing argument.

Lemma 3-15 [9, Lemma 3.12] *Pick $[\sigma] \in H^{\mathrm{p}}_{n,n}(Y)$ and let $S = S_{[\sigma]}$. Then given a base point $p \in Y$, for all $\epsilon > 0$ there is an N such that for all $r \ge 0$, the set $B(p,r)\cap S$ does not contain an ϵr–separated subset of cardinality greater than N.*

Lemma 3-16 *Let S and p be as in Lemma 3-15. Denote the cone point in $C_T S$ by o. Then*

$$\lim_{r\to+\infty} d_{\mathrm{GH}}\big(\tfrac{1}{r}(B(p,r)\cap S), B(o,1)\big) = 0. \tag{3-17}$$

Here d_{GH} denotes the Gromov–Hausdorff distance, $B(o,1)$ is the ball of radius 1 in $C_T S$ centered at o and $\frac{1}{r}(S\cap B(p,r))$ means we rescale the metric on $S \cap B(x,r)$ by a factor $\frac{1}{r}$.

Proof We follow the argument in [9]. It suffices to prove that for any $\epsilon > 0$, there exists $R > 0$ such that for any $r > R$, we can find an ϵ–isometry between $\frac{1}{r}(B(p,r) \cap S)$ and $B(o,1)$.

For $r > 0$, we denote the maximal cardinality of an ϵr–separated net in $B(p,r) \cap S$ by m_r. By Lemma 3-15, there exists N_0 such that $m_r \le N_0$ for all r. Pick R_1 such that $m_r \le m_{R_1}$ for all $r \neq R_1$ and denote the corresponding ϵR_1–net in $B(p,R_1) \cap S$ by $\{x_i\}_{i=1}^N$. By Lemma 3-6, for each i, we can extend the geodesic $\overline{px_i}$ to obtain a geodesic ray $\overline{p\xi_i}$ such that $\overline{x_i\xi_i} \subset S$. Let $l_i\colon [0,\infty) \to Y$ be a constant-speed geodesic ray joining p and ξ_i such that $l_i(R_1) = x_i$ and $l_i(0) = p$.

Since the quantity $d(l_i(t), l_j(t))/t$ is monotone increasing, $\{l_i(t)\}_{i=1}^N$ is a maximal ϵt–separated net in $B(p,R_1) \cap S$ for $t \ge R_1$. We pick $R > R_1$ such that for all $t > R$ and $1 \le i, j \le N$, we have

$$\left| \frac{d(l_i(t), l_j(t))}{t} - \lim_{t \to +\infty} \frac{d(l_i(t), l_j(t))}{t} \right| < \epsilon. \tag{3-18}$$

Now we fix $t > R$ and define a map such that for each i, $l_i(t) \in B(p,t) \cap S$ is mapped to the point $y_i \in B(o,1) \subset C_T S$ satisfying $y_i \in \overline{o\xi_i}$ and $d(y_i, o) = d(l_i(t), p)/t$. It follows from (3-18) that

$$\left| \frac{d(l_i(t), l_j(t))}{t} - d(y_i, y_j) \right| < \epsilon. \tag{3-19}$$

We claim $\{y_i\}_{i=1}^N$ is an ϵ–net in $B(o,1)$.

Pick an arbitrary $y \in B(o,1)$ and suppose $y \in \overline{o\xi}$ for $\xi \in \partial_T S$. We parametrize the geodesic ray $\overline{p\xi}$ by constant speed $= d(y,o)$ and denote this ray by l. Since there exists a geodesic $\overline{p'\xi} \subset S$ such that $d_H(\overline{p\xi}, \overline{p'\xi}) = C < \infty$, we can find $x \in \overline{p'\xi} \subset S$ with $d(x, l(t)) < C$ for every t. Thus $x \in B(p, td(y,o) + C) \cap S \subset B(p, t + C) \cap S$, which implies there exists some i such that $d(l_i(t+C), x) \le \epsilon(t+C)$. These estimates together with $d(l_i(t+C), l_i(t)) \le C$ (the ray l_i has speed ≤ 1) imply

$$d(l(t), l_i(t)) \le \epsilon t + (2+\epsilon)C. \tag{3-20}$$

Here i might depend on t, but we can choose a sequence $\{t_k\}_{k=1}^\infty$ such that $t_k \to +\infty$ and

$$d(l(t_k), l_{i_0}(t_k)) \le \epsilon t + (2+\epsilon)C \tag{3-21}$$

for every k with i_0 fixed, thus

$$d(y, y_{i_0}) = \lim_{k \to +\infty} \frac{d(l(t_k), l_{i_0}(t_k))}{t_k} \le \epsilon. \tag{3-22}$$

So $\{y_i\}_{i=1}^N$ is an ϵ–net in $B(o,1)$; this fact together with (3-19) give us the ϵ–isometry as required. □

Remark 3-23 (1) Define $\partial_{p,r} S = \{\xi \in \partial_T S \mid \overline{p\xi} \subset B(p,r) \cup S\}$. Then the above proof shows

$$\lim_{r\to+\infty} d_H(\partial_{p,r} S, \partial_T S) = 0. \tag{3-24}$$

(2) $\partial_T S$ has similar behavior to the Tits boundary of a convex subset in the following aspect. Let $l\colon [0,\infty) \to Y$ be a constant-speed geodesic ray. If there exist a constant $C < \infty$ and a sequence $t_i \to +\infty$ such that $d(l(t_i), S) < C$, then $\partial_T l$ is an accumulation point of $\partial_T S$. The proof is similar to the above argument.

3.3 ϵ–splittings

As we have seen from Lemma 3-16, the growth bound (3-14) implies that S looks more and more like a cone if one observes S from a farther and farther away point (this is called asymptotic conicality in [9]). So one would expect some regularity of S near infinity. The following key lemma from [9] will be our starting point.

Lemma 3-25 [9, Lemma 3.13] *Let S and p be as in Lemma 3-15. Then for all $\beta > 0$ there is an $r < \infty$ such that if $x \in S \setminus B(p,r)$, then the antipodal set satisfies*

$$\operatorname{diam}(\operatorname{Ant}(\log_x p, \Sigma_x S)) < \beta. \tag{3-26}$$

The proof of this lemma in [9] actually shows something more. Given a base point $p \in Y$ and $x \in S$, we define the *antipode at* ∞ of $\log_x p$ in S, denoted $\operatorname{Ant}_\infty(\log_x p, S)$, to be $\{\xi \in \partial_T S \mid \overline{x\xi} \subset S \text{ and } x \in \overline{p\xi}\}$. Then we have:

Lemma 3-27 *Let S and p be as in Lemma 3-15. Then for all $\beta > 0$ there is an $r < \infty$ such that if $x \in S \setminus B(p,r)$, then*

$$\operatorname{diam}(\operatorname{Ant}_\infty(\log_x p, S)) < \beta. \tag{3-28}$$

The diameter here is with respect to the angular metric on $\partial_T X$.

Lemma 3-25 tells us that $\Sigma_y S$ looks more and more like a suspension as $d(y,p) \to \infty$ (for $y \in S$). If we also assume $\operatorname{Shape}(Y)$ is finite, then for all $y \in S$, $\Sigma_y S$ is built from cells of finitely many isometry types. Moreover, by Theorem 3-10, there is a positive constant N such that $\Sigma_y S$ has at most N cells for any $y \in S$. Thus the family $\{\Sigma_y S\}_{y\in S}$ has only finite possible combinatorics. As $\beta \to 0$, one may expect $\Sigma_y S$ to actually be a suspension (this is called isolated suspension in [9]).

Now we restrict ourselves to the case of finite-dimensional $\mathrm{CAT}(0)$ cube complexes of finite dimension. Then the spaces of directions are finite-dimensional all-right spherical $\mathrm{CAT}(1)$ complexes (see Definition 2-7 for the definition of cell structure on the spaces of directions).

Lemma 3-29 *Suppose $\mathcal{F}$ is a family of all-right spherical* $\mathrm{CAT}(1)$ *complexes with dimension at most d. Then for every $\alpha > 0$ and $N > 0$, there is a constant $\beta = \beta(d, N, \alpha) > 0$ such that if K' satisfies the following conditions:*

1. *$K' \subset K$ is a subcomplex of some $K \in \mathcal{F}$,*
2. *the number of cells in K' is bounded above by N,*
3. *K' has the geodesic extension property in the sense of Lemma 3-9,*
4. *there exists $v \in K$ such that* $\mathrm{diam}(\mathrm{Ant}(v, K')) < \beta$,

then K' is a metric suspension (in the metric induced from K) and v lies at a distance $< \alpha$ from a suspension point of K'.

Proof We prove the lemma by contradiction. Suppose there exist $\alpha > 0$, $N > 0$ and a sequence $\{K_i'\}_{i=1}^{\infty}$ such that for each i, K_i' satisfies conditions (2) and (3), K_i' is a subcomplex of some $K_i \in \mathcal{F}$, and there exists a $v_i \in K_i$ such that

$$\mathrm{diam}(\mathrm{Ant}(v_i, K_i')) < \tfrac{1}{i}, \tag{3-30}$$

but no point in the α–neighborhood of v_i is suspension point of K_i'.

Let w_i be the point in K_i' which realizes the minimal distance to v_i in the length metric of K_i (note that the original metric on K_i is the length metric truncated by π). If l_i is the geodesic segment (in the length metric) joining v_i and w_i, then by (3-30) and the geodesic extension property of K_i', there exists a C such that $\mathrm{length}(l_i) < C$ for all i. So for any i, there exists a subcomplex $L_i \subset K_i$ such that $l_i \subset L_i$ and the number of cells in L_i are uniformly bounded by constant N_1 (by Lemma 2-3).

Let M_i be the full subcomplex spanned by $K_i' \cup L_i$, ie M_i is the union of simplices in K_i whose vertex sets are in $K_i' \cup L_i$. Then M_i is a π–convex, hence $\mathrm{CAT}(1)$, subcomplex of K_i, and the number of cells in M_i is uniformly bounded above by some constant N_2. Without loss of generality, we can replace K_i by M_i. Since M_i has only finitely many possible isometry types, after passing to a subsequence, we can assume there exist a finite $\mathrm{CAT}(1)$ complex M and a subcomplex $K' \subset M$ such that for every i, there is a simplicial isomorphism $\phi_i\colon M_i \to M$ mapping K_i' onto K' (here ϕ_i is also an isometry).

Since M is compact, there is a subsequence of $\{\phi_i(v_i)\}_{i=1}^{\infty}$ converging to a point $v \in M$. We claim $\mathrm{Ant}(v, K')$ is exactly one point. First $\mathrm{Ant}(v, K') \neq \varnothing$ by the geodesic extension property of K'. If there were two distinct points $u, u' \in \mathrm{Ant}(v, K')$, then we could extend the geodesic segment $\overline{v_i u}$, $\overline{v_i u'}$ into K', yielding a contradiction with (3-30) for large i.

Suppose $\mathrm{Ant}(v, K') = \{v'\}$. Then $\mathrm{Ant}(v', K') = v$. In fact, if this were not true, then we would have some point $w \in \mathrm{Ant}(v', K')$ such that $0 < d(v, w) < \pi$. Then we could extend the geodesic segment $\overline{vw}$ into K' to get a local geodesic $\overline{vw'}$ with $w' \in K'$ and $\mathrm{length}(\overline{vw'}) = \pi$. This would actually be a geodesic since we are in a CAT(1) space, thus $w' \in \mathrm{Ant}(v, K')$ and $w' \neq v'$, contradicting $\mathrm{Ant}(v, K') = v'$.

Now pick a point $k \in K'$, with $k \neq v$ and $k \neq v'$. Then $d(k, v') < \pi$ and $d(k, v) < \pi$. We extend the geodesic segment $\overline{v'k}$ into K' to get a geodesic of length π, then the other end must hit v since $\mathrm{Ant}(v', K') = v$. Thus $\overline{kv} \subset K'$ by the uniqueness of geodesic joining k and v. Similarly we know $\overline{kv'} \subset K'$, thus there is a geodesic segment in K' passing through k and joining v and v'. By CAT(1) geometry, K' (with the induced metric from M) splits as a metric suspension and v, v' are suspension points. However, by the assumption at the beginning of the proof, $\{\phi_i(v_i)\}_{i=1}^{\infty}$ should have distance at least α from a suspension point for every i, so v should also be α–away from a suspension point; this contradiction finishes the proof. □

Remark 3-31 (1) The above proof also shows the following result. Let K be a piecewise spherical CAT(1) complex, and let $K' \subset K$ be a subcomplex with geodesic extension property in the sense of Lemma 3-9. Pick $v \in K$. If $\mathrm{Ant}(v, K')$ is exactly one point, then $v \in K'$ and v is a suspension point of K'.

(2) By the same proof, it is not hard to see Lemma 3-29 is also true when $\mathcal{F}$ is a finite family of finite piecewise spherical CAT(0) complexes (not necessarily all-right).

From Lemmas 3-4, 3-25 and 3-29, we have the following analogue of [9, Theorem 3.11].

Theorem 3-32 *Let X be an n–dimensional* CAT(0) *cube complex, and let $S = S_{[\sigma]}$, where $\sigma \in H^p_{n,n}(X)$. Then for every $p \in X$ and every $\epsilon > 0$, there is an $r < \infty$ such that if $x \in S \setminus B(p, r)$, then $\Sigma_x S$ is a suspension and $\log_x p$ is ϵ–close to a suspension point of $\Sigma_x S$.*

Remark 3-33 By the same proof, the conclusion of Theorem 3-32 is also true if X is a proper n–dimensional CAT(0) complex with a cocompact (cellular) isometry group.

4 The structure of the top-dimensional support sets

Throughout this section, X is an n–dimensional CAT(0) cube complex. Pick a homology class $[\sigma] \in H^{\mathrm{p}}_{n,n}(X)$ and let $S = S_{[\sigma]}$. Also recall that $\Sigma_x X$ is an all-right spherical CAT(1) complex for each $x \in X$; see Definition 2-7.

Let Δ^k be the k–dimensional all-right spherical simplex, and let Δ^k_{mod} be the quotient of Δ^k by the action of its isometry group (Δ^k_{mod} is endowed with the quotient metric). Define the function $\chi\colon \Delta^k \to (0, +\infty)$ by

$$\chi(v) = \inf\{d(v, v') \mid v' \in \Delta^k \text{ and } \operatorname{Supp}(v') \cap v = \varnothing\}. \tag{4-1}$$

Recall that $\operatorname{Supp}(v')$ denotes the unique closed face of Δ^k which contains v' as an interior point. By symmetry of Δ^k, χ descends to a function $\chi\colon \Delta^k_{\mathrm{mod}} \to (0, +\infty)$.

For any $k' \geq k$, we have a canonical isometric embedding $i\colon \Delta^k_{\mathrm{mod}} \hookrightarrow \Delta^{k'}_{\mathrm{mod}}$ with $\chi = \chi \circ i$. Let $\Delta_{\mathrm{mod}} = \varinjlim \Delta^k_{\mathrm{mod}}$ be the corresponding direct limit of metric spaces.

Let Y be an all-right spherical CAT(1) complex. Then there is a well-defined 1–Lipschitz map

$$\theta\colon Y \to \Delta_{\mathrm{mod}}$$

such that θ restricted to any k–face $\Delta^k \subset Y$ is the map $\Delta^k \to \Delta^k_{\mathrm{mod}} \hookrightarrow \Delta_{\mathrm{mod}}$. Moreover, for $v \in Y$,

(1) $v \in \operatorname{Supp}(v')$ if $d(v, v') < \chi(\theta(v))$,

(2) $\chi \circ \theta$ is continuous on the interior of each face of Y.

When $Y = \Sigma_x X$ for some $x \in X$ and $v \in \Sigma_x X$, we also call $\theta(v)$ the Δ_{mod} *direction of* v.

4.1 Producing orthants

In this section, we study geodesic rays with constant Δ_{mod} direction, ie unit-speed geodesic rays $l\colon [0, \infty) \to S$ with $\theta(l^-(t)) = \theta(l^+(t)) = \theta(l^-(t')) = \theta(l^+(t'))$ for any $t \neq t'$. Here are two examples.

(1) If a geodesic ray l stays inside an orthant subcomplex of $O \subset Y$ (or more generally a straight orthant), then it has constant Δ_{mod} direction. Moreover, the Δ_{mod} direction of $\partial_T l$ in $\partial_T O$ is equal to $\theta(l^{\pm}(t))$.

(2) If Y is a product of trees, then each geodesic ray in $l \in Y$ has constant Δ_{mod} direction. Again, the Δ_{mod} direction of $\partial_T l$ in $\partial_T Y$ (in this case $\partial_T Y$ is an all-right spherical complex) is equal to $\theta(l^{\pm}(t))$.

Later, geodesic rays with constant Δ_{mod} direction will play an important role in the construction of orthants; see Lemma 4-9. First we show such geodesics exist in the support set of a top-dimensional proper cycle and there are plenty of them.

Lemma 4-2 *If Y is a k–dimensional all-right spherical* CAT(1) *complex and if $K \subset Y$ is the support set of some top-dimensional homology class, then for any $v \in K$, there exists a $v' \in K$ such that $d(v, v') = \pi$ and $\theta(v) = \theta(v')$.*

Recall that the metric on Y is the length metric on Y truncated by π.

Proof The lemma is clear when $k = 1$ by Lemma 3-9. We assume it is true for $i \le k-1$. Denote $k' = \dim(\mathrm{Supp}(v))$. We endow $\mathbb{S}^{k'}$ with the structure of an all-right spherical complex and pick $w \in \mathbb{S}^{k'}$ such that $\theta(v) = \theta(w)$. Suppose $w' = \mathrm{Ant}(w, \mathbb{S}^{k'})$ and suppose $\gamma'\colon [0,\pi] \to \mathbb{S}^{k'}$ is a unit-speed geodesic joining w and w'. It is clear that $\theta(w) = \theta(w')$. Our goal is to construct a unit-speed local geodesic $\gamma\colon [0,\pi] \to K$ such that $\gamma(0) = v$ and $\theta(\gamma(s)) = \theta(\gamma'(s))$ for all $s \in [0,\pi]$, as in the following diagram:

$$\begin{tikzcd} {[0,\pi]} \arrow[r, "\gamma'"] \arrow[d, "\gamma"'] & \mathbb{S}^{k'} \arrow[d, "\theta"] \\ K \arrow[r, "\theta"] & \Delta_{\mathrm{mod}} \end{tikzcd}$$

Then γ is actually a geodesic and we can take $v' = \gamma(\pi)$ to finish the proof.

There exists a sequence of faces $\{\Delta'_j\}_{j=1}^N$ in $\mathbb{S}^{k'}$ and $0 = t_0 < t_1 < \cdots < t_{N-1} < t_N = \pi$ such that each Δ'_j contains $\{\gamma'(t) \mid t_{j-1} < t < t_j\}$ as interior points. Let $\Delta_1 = \mathrm{Supp}(v)$. Then since $\theta(v) = \theta(w)$, we can find $v_1 \in \Delta_1$ such that there exists an isometry $\Phi\colon \Delta_1 \to \Delta'_1$ with $\Phi(v) = w$ and $\Phi(v_1) = \gamma'(t_1)$, in particular $\theta(v_1) = \theta(\gamma'(t_1))$. Define $\gamma\colon [0, t_1] \to K$ to be the geodesic segment $\overline{vv_1}$.

Recall that we have identified $\Sigma_{v_1} Y$ with $\mathrm{Lk}(v_1, Y)$; see Definition 2-7. Now let $\sigma_1 = \mathrm{Supp}(v_1)$ and $k_1 = \dim(\sigma) - 1$. Then $\Sigma_{v_1} Y = \mathrm{Lk}(v_1, \sigma_1) * \mathrm{Lk}(\sigma_1, Y) = \mathbb{S}^{k_1} * \mathrm{Lk}(\sigma_1, Y)$. Similarly, $\Sigma_{v_1} K = \mathrm{Lk}(v_1, \sigma_1) * \mathrm{Lk}(\sigma_1, K) = \mathbb{S}^{k_1} * \mathrm{Lk}(\sigma_1, K)$. Let $K_1 = \mathrm{Lk}(\sigma_1, K)$ and $Y_1 = \mathrm{Lk}(\sigma_1, Y)$. Then they are all-right spherical complexes, K_1 is a subcomplex of Y_1, and Y_1 is CAT(1). Moreover, since $\Sigma_{v_1} K$ is the support set of some top-dimensional homology class in $\Sigma_{v_1} Y$ (Lemma 3-8), so is K_1 in Y_1. As $\gamma^-(t_1) \in \Sigma_{v_1} K = \mathbb{S}^{k_1} * K_1$, we write

$$\gamma^-(t_1) = (\cos\alpha_1)x_1 + (\sin\alpha_1)y_1 \tag{4-3}$$

for $x_1 \in \mathbb{S}^{k_1}$ and $y_1 \in K_1$. By the induction assumption, we can find $y'_1 \in \mathrm{Ant}(y_1, K_1)$ such that $\theta(y'_1) = \theta(y_1)$. Let $x'_1 = \mathrm{Ant}(x_1, \mathbb{S}^{k_1})$. Suppose $\Delta_2 \subset K$ is the unique face ($v_1 \in \Delta_2$) such that $\mathrm{Supp}((\cos\alpha_1)x_1' + (\sin\alpha_1)y_1') = \Sigma_{v_1}\Delta_2$. Let $\overline{v_1v_2} \subset \Delta_2$ be the geodesic segment which starts at v_1, and goes along the direction $(\cos\alpha_1)x_1' + (\sin\alpha_1)y_1'$ until it hits some boundary point v_2 of Δ_2. Note that $\overline{vv_1}$ and $\overline{v_1v_2}$ fit together to form a local geodesic in K.

On the other hand, at $\gamma'(t_1) \in \mathbb{S}^{k'}$, we have $\Sigma_{\gamma'(t_1)}\mathbb{S}^{k'} = \mathrm{Lk}(\gamma'(t_1), \sigma_1') * \mathrm{Lk}(\sigma_1', \mathbb{S}^{k'}) = \mathbb{S}^{k_1} * \mathrm{Lk}(\sigma_1', \mathbb{S}^{k'})$, where $\sigma_1' = \mathrm{Supp}(\gamma'(t_1))$ and k_1 is the same as the previous paragraph. Write $\gamma'^{-}(t_1) = (\cos\alpha_1)u_1 + (\sin\alpha_1)v_1$ for $u_1 \in \mathbb{S}^{k_1}$ and $v_1 \in \mathrm{Lk}(\sigma_1', \mathbb{S}^{k'})$, where we have the same α_1 as (4-3) since Φ is an isometry. Then $\gamma'^{+}(t_1) = (\cos\alpha_1)u_1' + (\sin\alpha_1)v_1'$ for $u_1' = \mathrm{Ant}(u_1, \mathbb{S}^{k_1})$ and $v_1' = \mathrm{Ant}(v_1, \mathrm{Lk}(\sigma_1', \mathbb{S}^{k'}))$. Note that $\theta(v_1') = \theta(v_1) = \theta(y_1) = \theta(y_1')$, so we can extend the isometry Φ to get a map $\Phi'\colon \Delta_1 \cup \Delta_2 \to \Delta_1' \cup \Delta_2'$ such that Φ' is an isometry with respect to the length metric on both sides and $\Phi'(\overline{v_1v_2}) = \gamma([t_1, t_2])$. Thus $d(v_1, v_2) = t_2 - t_1$ and we can define $\gamma\colon [t_1, t_2] \to K$ to be the geodesic segment $\overline{v_1v_2}$. It is clear that $\theta(\gamma(s)) = \theta(\gamma'(s))$ for all $s \in [0, t_2]$. We can repeat this process to define the required local geodesic $\gamma\colon [0, \pi] \to K$. □

Corollary 4-4 *For any $x \in S$ and $v \in \Sigma_x S$, there exists a geodesic ray $\overline{x\xi} \subset S$ which has constant Δ_{mod} direction and $\log_x \xi = v$.*

Proof Since x has a cone neighborhood in S, we can find a short geodesic segment $\overline{xx'}$ in the cone neighborhood such that $\log_x x' = v$. There is a unique closed cube $C_1 \subset S$ such that $\overline{xx'} \subset C_1$ and v is an interior point of $\Sigma_x C_1$. We extend $\overline{xx'}$ in C_1 until it hits the boundary of C_1 at x_1. By Lemma 4-2, there exists a $v_1 \in \mathrm{Ant}(\log_{x_1} x, \Sigma_{x_1} S)$ with $\theta(v_1) = \theta(\log_{x_1} x) = \theta(v)$. Now we choose cube $C_2 \subset S$ and segment $\overline{x_1x_2} \subset C_2$ with $\log_{x_1} x_2 = v_1$ as before. Note that $\overline{xx_1}$ and $\overline{x_1x_2}$ together form a local geodesic segment (hence a geodesic segment). We repeat the previous process to extend the geodesic. Since S is a closed set, the extension can not terminate, which will give us the geodesic ray $\overline{x\xi}$ as required. □

Corollary 4-5 *The set of points in $\partial_T S$ which can be represented by a geodesic ray in S with constant Δ_{mod} direction is dense.*

Proof Fix a base point p, pick some $\xi \in \partial_T S$. For any $\epsilon > 0$, by (3-24), we can find an r_1 such that

$$d_H(\partial_{p,r} S, \partial_T S) < \tfrac{1}{2}\epsilon \tag{4-6}$$

for all $r > r_1$. By Lemma 3-27, we can find r_2 such that for $r > r_2$,

$$\mathrm{diam}(\mathrm{Ant}_\infty(\log_x p, S)) < \tfrac{1}{2}\epsilon \tag{4-7}$$

for any $x \in S \setminus B(p, r)$.

If $r_0 = \max\{r_1, r_2\}$, then we can find $\overline{p\xi_1} \subset B(p, r_0) \cup S$ such that $\angle_T(\xi_1, \xi) \le \frac{\epsilon}{2}$ by (4-6). Pick $x \in \overline{p\xi_1}$ such that $d(p, x) > r_0$. By Corollary 4-4, we can find a geodesic ray $\overline{x\xi_2} \subset S$ of constant Δ_{mod} direction such that $\overline{x\xi_2}$ fits together with $\overline{px}$ to form a geodesic ray $\overline{p\xi_2}$, thus $\angle_T(\xi_1, \xi_2) < \frac{\epsilon}{2}$ by (4-7). Then $\angle_T(\xi, \xi_2) < \epsilon$, which finishes the proof since ϵ and ξ are arbitrary. □

Let l be a geodesic ray of constant Δ_{mod} direction. Then we define $\theta(l)$ to be $\theta(l^{\pm}(t))$. The definition does not depend on the choice of sign $\pm$ and t.

Lemma 4-8 *If $l \subset S$ is a unit-speed geodesic ray of constant Δ_{mod} direction, then there exists a $t_0 < \infty$, which depends on the position of $l(0)$ and $\theta(l)$, such that for any $t > t_0$, $\Sigma_{l(t)} S = \Sigma_{l(t)} l * Y_t$ for some $Y_t \subset \Sigma_{l(t)} S$.*

Proof We apply Theorem 3-32 with $p = l(0)$ and $\epsilon = \chi(\theta(l))$ (see (4-1) for the definition of χ) to get $t_0 < \infty$ such that $\Sigma_{l(t)} S$ is a metric suspension and the suspension point is $\chi(\theta(l))$–close to $l^+(t)$ (or $l^-(t)$) for all $t > t_0$. By Lemma 2-9, $l^+(t)$ and $l^-(t)$ are suspension points, thus $\Sigma_{l(t)} S = \Sigma_{l(t)} l * Y_t$. □

Based on Lemma 4-8, we define a parallel transport of $\Sigma_{l(t)} S$ along l as follows. Let t_0 be as in Lemma 4-8. For any $t > t_0$, $l(t)$ has a product neighborhood in S of form $X_t \times (t - \epsilon, t + \epsilon)$, where X_t is some subset of X with the induced metric. So for any $|t' - t| < \epsilon$, we can identify $\Sigma_{l(t)} S$ and $\Sigma_{l(t')} S$. Moreover, for any $t_1 > t_0$ and $t_2 > t_0$, we can cover the geodesic segment $\overline{l(t_1) l(t_2)}$ by finitely many product neighborhoods, which will induce an identification of $\Sigma_{l(t_1)} S$ and $\Sigma_{l(t_2)} S$. This identification does not depend on the covering we choose.

To see this identification more concretely, take $t > t_0$, a product neighborhood $X_t \times (t - \epsilon, t + \epsilon)$ of $l(t)$ in S, and $v \in \Sigma_{l(t)} S$, we construct a short geodesic $\overline{l(t) x_t} \subset S$ in the cone neighborhood of $l(t)$ going along the direction v. If $|t' - t| < \epsilon$, we can find an isometrically embedded parallelogram in the product neighborhood such that $\overline{l(t) x_t}$ and $\overline{l(t') x_{t'}}$ are opposite sides of the parallelogram and $\overline{l(t) l(t')}$ is one of the remaining sides (we might have to shorten $\overline{l(t) x_t}$ a little).

In general, for any $t' > t_0$, we can cover the geodesic segment $\overline{l(t) l(t')}$ by finitely many product neighborhoods as before and construct a local isometric embedding ϕ from a parallelogram to X such that two opposite sides of the parallelogram are mapped to some geodesic segments $\overline{l(t) x_t}$ and $\overline{l(t') x_{t'}}$ and one of the remaining sides is mapped to $\overline{l(t) l(t')}$. Since X is CAT(0), ϕ is actually an isometric embedding. So we have a well-defined *parallel transport* of $\Sigma_{l(t)} S$ along $l(t)$ for $t > t_0$.

In the construction of the above parallelogram, the length of $\overline{l(t) x_t}$ (or $\overline{l(t') x_{t'}}$) may go to 0 as $|t' - t| \to \infty$. However, in the special case where there exists $s_0 > 0$ such that $\Sigma_{l(t)} l$ is contained in the 0–skeleton of $\Sigma_{l(t)} S$ for all $t > s_0$ (or equivalently $l([s_0, \infty))$ is parallel to some geodesic ray in the 1–skeleton of X), $l([t_0, \infty))$ has a product neighborhood of the form $X' \times [t_0, \infty)$ in S, where X' is a subcomplex of X with the induced metric. Therefore for any $t > t_0$ and a segment $\overline{l(t)x_t} \subset S$ short enough, we can parallel transport $\overline{l(t)x_t}$ along l to infinity, ie there is an isometrically embedded "infinite parallelogram" with one side $\overline{l(t)x_t}$ and one side $l([t, \infty))$.

Lemma 4-9 *If $l\colon [0, \infty) \to S$ is a unit-speed geodesic ray of constant Δ_{mod} direction, then there exists an orthant subcomplex $O \subset X$ satisfying the following conditions:*

(1) $\partial_T l \in \partial_T O$.

(2) *If $\dim(O) = k$ and if $\{l_i\}_{i=1}^k$ are the geodesic rays emanating from the tip of the orthant such that O is the convex hull of $\{l_i\}_{i=1}^k$, then $\partial_T l_i \in \partial_T S$ for all i.*

Proof By the previous lemma, we can choose some $r > 0$ such that for $t > r$, the $l^{\pm}(t)$ are suspension points in $\Sigma_{l(t)} S$. Pick some $t > r$ and let k be the dimension of $\mathrm{Supp}(l^+(t))$. Let $\{v_i^t\}_{i=1}^k$ be vertices of $\mathrm{Supp}(l^+(t))$. Suppose that $\alpha_i = d_{\Sigma_{l(t)} S}(v_i^t, l^+(t))$, where the values of k and α_i are the same for all $t > r$ by the splitting in Lemmas 4-8 and 2-9. Moreover, we would like the labels v_i^t to be consistent under parallel transportation, ie for $t' \neq t$ (with $t' > r$), $v_i^{t'}$ is the parallel transport of v_i^t along l. By Theorem 3-32, we can choose $r' \geq r$ such that if $x \in S \setminus B(l(0), r')$, then $\log_x l(0)$ is ϵ–close to some suspension point in $\Sigma_x S$ for $\epsilon = \min_{1 \leq i \leq k} \bigl\{\frac{1}{2}\bigl(\frac{\pi}{2} - \alpha_i\bigr)\bigr\}$.

Now we pick some $t > r'$, and construct a short geodesic segment $\overline{l(t)x_i} \subset S$ going along the direction of v_i^t in the cone neighborhood of $l(t)$. We choose an arbitrary extension of $\overline{l(t)x_i}$ into S and call the geodesic ray l_i^t for $1 \leq i \leq k$. We claim that for any $y \in l_i^t$ (with $y \neq l(t)$),

$$\Sigma_y S = \Sigma_y l_i^t * Y \tag{4-10}$$

for some $Y \subset \Sigma_y S$, hence the extension is unique and l_i^t is a straight geodesic.

Suppose the claim were not true. Pick the first point $y_i \in l_i^t$ such that (4-10) is not satisfied. Since $\angle_{l(t)}(y_i, l(0)) \geq \pi - \alpha_i > \frac{\pi}{2}$, hence $d(y_i, l(0)) > d(l(0), l(t)) > r'$. By our choice of r', there is a suspension point in $\Sigma_{y_i} S$ which has distance less than $\frac{1}{2}(\frac{\pi}{2} - \alpha_i)$ from $\log_{y_i} l(0)$. Since $\angle_{y_i}(l(0), l(t)) < \alpha_i$, $\log_{y_i} l(t)$ has distance less than $\alpha_i + \frac{1}{2}\bigl(\frac{\pi}{2} - \alpha_i\bigr) < \frac{\pi}{2}$ from a suspension point. Since all points in l_i^t between $l(t)$ and y_i satisfy (4-10), $\log_{y_i} l(t)$ is a vertex in the all-right spherical complex $\Sigma_{y_i} S$. Thus $\log_{y_i} l(t)$ is also a suspension point and (4-10) must hold at $y = y_i$, which is a contradiction.

We claim next that l_i^t is parallel to $l_i^{t'}$ for any $t > r'$ and $t' > r'$. In fact, fixing t, by the discussion before Lemma 4-9 and the uniqueness of $l_i^{t'}$, we know the claim is true for $|t'-t| < \epsilon$, where ϵ depends on t. For the general case, we can apply a covering argument as before.

Fix $t_0 > r'$. By Lemma 2-4, for all i,

$$\angle_T(l_i^{t_0}, l) = \lim_{t\to+\infty} \angle_{l(t)}(l_i^t, l) = \alpha_i < \tfrac{\pi}{2}. \tag{4-11}$$

Thus $\angle_T(l_i^{t_0}, l) = \angle_{l(t_0)}(l_i^{t_0}, l) = \alpha_i$ for all i. It follows that l and $l_i^{t_0}$ bound a flat sector by Lemma 2-5.

We fix a pair i, j with $i \neq j$, and parametrize $l_i^{t_0}$ by arc length. We can assume without loss of generality that $l(t_0)$ is in the 0–skeleton. Let $\{h_m\}_{m=1}^{\infty}$ be the collection of hyperplanes such that $h_m \cap l_i^{t_0} = l_i^{t_0}\big(m+\frac{1}{2}\big)$. Then (4-11) and Lemma 2-4 imply that the CAT(0) projection of l onto $l_i^{t_0}$ is surjective, thus there exists a sequence $\{t_m\}_{m=1}^{\infty}$ such that $l(t_m) \in h_m$. Recalling that $j \neq i$, note that $l_j^{t_m}$ starts at $l(t_m)$, $\angle_{l(t_m)}(l_i^{t_m}, l_j^{t_m}) = \frac{\pi}{2}$ and $l_i^{t_m}$ is orthogonal to h_m, so $l_j^{t_m} \subset h_m$.

By convexity of h_m, we can find a geodesic ray c_m which starts at $l_i^{t_0}\big(m+\frac{1}{2}\big)$, stays inside h_m and is asymptotic to $l_j^{t_m}$ for every m, thus by Lemma 2-4,

$$\angle_T(l_i^{t_0}, l_j^{t_0}) = \lim_{m\to+\infty} \angle_{l_i^{t_0}(m+\frac{1}{2})}(l_i^{t_0}, c_m) = \tfrac{\pi}{2}. \tag{4-12}$$

By (4-12), $\angle_T(l_i^{t_0}, l_j^{t_0}) = \angle_{l(t_0)}(l_i^{t_0}, l_j^{t_0}) = \frac{\pi}{2}$ for $i \neq j$. By Lemma 2-18, we know that the geodesic rays $\{l_i^{t_0}\}_{i=1}^{k}$ span a straight orthant O. Moreover, Lemma 2-8 together with (4-11) imply $\partial_T l \in \partial_T O$. By Remark 2-16, we can replace O by an orthant subcomplex which is Hausdorff close to O. □

4.2 Cycle at infinity

By Lemma 4-9 and Corollary 4-5, there exists a dense subset G of $\partial_T S$ such that for any $v \in G$, there exists an orthant subcomplex $O_v \in X$ such that $v \in \Delta_v = \partial_T O_v$. Denote the vertices of Δ_v by $\mathrm{Fr}(v)$, then $\mathrm{Fr}(v) \subset \partial_T S$ by Lemma 4-9.

It is clear that $G \subset \bigcup_{v\in G} \Delta_v$. We claim $\bigcup_{v\in G} \Delta_v$ is a finite union of all-right spherical simplices. In fact, it suffices to show $\bigcup_{v\in G} \mathrm{Fr}(v)$ is a finite set, which follows from Lemmas 2-17, 3-15 and 4-9 (note that each point in $\bigcup_{v\in G} \mathrm{Fr}(v)$ is represented by a straight geodesic contained in S).

Moreover, $\bigcup_{v\in G} \Delta_v$ has the structure of a finite simplicial complex. Take two simplices Δ_{v_1} and Δ_{v_2}. We know $\Delta_{v_i} = \partial_T O_{v_i}$ for orthant subcomplex O_{v_i}, and Remark 2-13 implies $\Delta_{v_1} \cap \Delta_{v_2}$ is a face of Δ_1 (or Δ_2).

We endow $K = \bigcup_{v \in G} \triangle_v \subset \partial_T Y$ with the angular metric and denote the Euclidean cone over K by CK, which is a subset of $C_T X$.

Lemma 4-13 (1) *K is a topologically embedded finite simplicial complex in $\partial_T X$.*

(2) *CK is linearly contractible.*

Recall that linearly contractible means there exists a constant C such that for any $d > 0$, every cycle of diameter $\le d$ can be filled in by a chain of diameter $\le Cd$.

Proof Let $\angle_T$ be the angular metric on K and d_l be the length metric on K as an all-right spherical complex. Our goal is to show that Id: $(K, \angle_T) \to (K, d_l)$ is a bi-Lipschitz homeomorphism. Let $\{\Delta_i\}$ be the collection of faces of K (each Δ_i is an all-right spherical simplex). Suppose $\{O_i\}_{i=1}^N$ are orthant subcomplexes of X such that $\partial_T O_i = \Delta_i$. If points x and y are in the same Δ_i for some i, then

$$d_l(x, y) = \angle_T(x, y). \tag{4-14}$$

If x and y are not in the same simplex, then we put $\Delta_i = \mathrm{Supp}(x)$, $\Delta_j = \mathrm{Supp}(y)$ and $\Delta_k = \Delta_i \cap \Delta_j$. Assume without loss of generality that $d_l(x, \Delta_k) \ge \frac{1}{2} d_l(x, y)$. Let $(Y_1, Y_2) = \mathcal{I}(O_i, O_j)$. Then $\partial_T Y_1 = \partial_T Y_2 = \Delta_k$. Moreover, it follows from (2-11) and Lemma 2-4 that

$$\angle_T(x, y) \ge 2 \arcsin\bigl(\tfrac{1}{2} A \sin(d_l(x, \Delta_k))\bigr) \ge 2 \arcsin\bigl(\tfrac{1}{2} A \sin\bigl(\tfrac{1}{2} d_l(x, y)\bigr)\bigr), \tag{4-15}$$

where A can be chosen to be independent of i and j since $\{O_i\}_{i=1}^N$ is a finite collection. Equations (4-14) and (4-15) imply Id: $(K, \angle_T) \to (K, d_l)$ is a bi-Lipschitz homeomorphism, thus (1) is true.

To see (2), it suffices to prove $(K, \angle_T)$ is linearly locally contractible, ie there exist $C < \infty$ and $R > 0$ such that for any $d < R$, every cycle of diameter $\le d$ can be filled in by a chain of diameter $\le Cd$. By the above discussion, we only need to prove (K, d_l) is locally linearly contractible.

Since (K, d_l) is compact and can be covered by finitely many cone neighborhoods (see Theorem 2-1), it suffices to show each cone neighborhood is linearly contractible; but any cone neighborhood is isometric to a metric ball in the spherical cone of some lower dimensional finite piecewise spherical complex, thus we can finish the proof by induction on dimension. □

Since G is a dense subset of $\partial_T S$ and K is compact, it follows that $\partial_T S \subset K$ and $C_T S \subset CK \subset C_T X$. We denote the base point of $C_T X$ by o.

Lemma 4-16 *K has the structure of an $(n-1)$–simplicial cycle.*

Proof In the following proof, we will use d to denote the metric on X, and use $\overline{d}$ to denote the metric on $C_T X$.

Pick a base point $p \in X$. By the proof of Lemma 3-16, we know that for any $\epsilon > 0$, there exist a finite collection of constant speed geodesic rays $\{l_i\}_{i=1}^N$ and an $R_\epsilon < \infty$ such that $l_i(t) \in S$ and $\{l_i(t)\}_{i=1}^N$ is a ϵt–net in $B(p,t) \cap S$ for $t \geq R_\epsilon$. Write $\xi_i = \partial_T l_i$ and define $f_\epsilon\colon S \to C_T S \subset CK$ by sending $l_i(t)$ to the point in $\overline{o\xi_i} \subset C_T S$ which has distance $d(l_i(t), p)$ from o ($t \geq R_\epsilon$). For $x \notin \bigcup_{i=1}^N l_i[R_\epsilon, \infty)$, we pick a point $y \in \bigcup_{i=1}^N l_i[R_\epsilon, \infty)$ which is nearest to x and define $f_\epsilon(x) = f_\epsilon(y)$.

It is clear that

$$|d(x,p) - \overline{d}(f_\epsilon(x), o)| \leq \epsilon \max\{d(x,p), R_\epsilon\} \tag{4-17}$$

for any $x \in S$, and

$$|d(x,y) - \overline{d}(f_\epsilon(x), f_\epsilon(y))| \leq \epsilon \max\{d(p,x), d(p,y), R_\epsilon\} \tag{4-18}$$

for any $x \in S$ and $y \in S$. Moreover,

$$\overline{d}_H(f_\epsilon(B(p,r) \cap S), B(o,r) \cap C_T S) \leq \epsilon \max\{r, R_\epsilon\}. \tag{4-19}$$

We might need to pick a larger R_ϵ for (4-19).

We want to approximate f_ϵ by a continuous map. Let us cover S by a collection of open sets $\{B(x, r_x) \cap S\}_{x \in S}$, where $r_x = \epsilon \max\{d(x,p), R_\epsilon\}$. Since S has topological dimension $\leq n$, this covering has a refinement $\{U_i\}_{i=1}^\infty$ of order $\leq n$; see [25, Chapter V]. Note that $\operatorname{diam}(U_i) \leq 2\epsilon \max\{d(p, U_i), R_\epsilon\}$. Denote the nerve of $\{U_i\}_{i=1}^\infty$ by N, which is a simplicial complex of dimension $\leq n$.

Now we define a map $b'\colon N \to CK$ as follows. For any vertex $v_i \in L$, pick $x_i \in U_i$ where U_i is the open set associated with vertex v_i, then set $b'(v_i) = f_\epsilon(x_i)$. Then use the linear contractibility of CK to extend the map skeleton by skeleton to get b'. By choosing a partition of unity subordinate to the covering $\{U_i\}_{i=1}^\infty$, we obtain a barycentric map b from S to the nerve N (see [25, Chapter V]), then the continuous map $b' \circ b\colon S \to CK$ also satisfies (4-17)–(4-19) with ϵ replaced by $L'\epsilon$, where L' is some constant which only depends on the linear contractibility constant of CK. So we can assume without loss of generality that $f_\epsilon\colon S \to CK$ is continuous and (4-17)–(4-19) still hold for f_ϵ.

Recall that S is the support set of some top-dimensional proper homology class $[\sigma]$. We can also view $[\sigma]$ as the fundamental class of S and assume σ is the proper singular cycle representing this class. If $\alpha = f_\epsilon(\sigma)$, then $[\alpha] \in H_n^{\mathrm{p}}(CK)$ since f_ϵ is a proper map by (4-17). Our next goal is to show

$$S_{[\alpha]} = CK \tag{4-20}$$

for ϵ small enough. Since K is a simplicial complex, (4-20) would imply K also has a fundamental class whose support set is exactly K itself, proving Lemma 4-16.

Recall that we have a 1–Lipschitz logarithmic map $\log_p\colon C_T X \to X$ sending base point o to p. By (4-17) and (4-18), there exists a constant $L < \infty$ such that

$$\bar{d}(z, o) = d(\log_p(z), p) \tag{4-21}$$

for all $z \in \operatorname{Im} f_\epsilon$, and

$$|\bar{d}(z, w) - d(\log_p(z), \log_p(w))| \le L\epsilon \max\{\bar{d}(o, z), \bar{d}(o, w), R_\epsilon\} \tag{4-22}$$

for all $z, w \in \operatorname{Im} f_\epsilon$. Moreover,

$$d(\log_p \circ f_\epsilon(x), x) \le L\epsilon \max\{d(x, p), R_\epsilon\} \tag{4-23}$$

for all $x \in S$.

By (4-21), $\log_p$ is proper. Let $\beta = \log_p(\alpha) = \log_p \circ f_\epsilon(\sigma)$. By (4-23), the geodesic homotopy between $\log_p \circ f_\epsilon\colon S \to X$ and the inclusion map $i\colon S \to X$ is proper, thus $[\beta] = [\sigma]$ and $S_{[\beta]} = S_{[\sigma]} = S$. By Lemma 3-2,

$$\log_p(S_{[\alpha]}) \supset S_{[\beta]} = S. \tag{4-24}$$

Equations (4-24), (4-22) and (4-23) imply there exists $L < \infty$ such that

$$\bar{d}_H(B(o, r) \cap S_{[\alpha]}, B(o, r) \cap \operatorname{Im} f_\epsilon) \le L\epsilon \max\{r, R_\epsilon\}. \tag{4-25}$$

This together with (4-19) imply

$$\bar{d}_H(B(o, r) \cap S_{[\alpha]}, B(o, r) \cap C_T S) \le L\epsilon \max\{r, R_\epsilon\}. \tag{4-26}$$

Since K is a simplicial complex, $S_{[\alpha]} = CK'$, where K' is some subcomplex of K. Recall that by the construction of K, the only subcomplex of K that contains $\partial_T S$ is K itself. Now (4-26) implies the Hausdorff distance between $\partial_T S$ and K' is bounded above by $L\epsilon$, thus for ϵ small enough, $K' = K$ and (4-20) holds. We also know $\partial_T S$ is dense in K from this. □

We actually defined a boundary map

$$\partial\colon H^{\mathrm{p}}_{n,n}(X) \to H_{n-1}(\partial_T X) \tag{4-27}$$

in the proof of the above lemma; namely, for ϵ small enough, we send $[\sigma] \in H^{\mathrm{p}}_{n,n}(X)$ to $f_{\epsilon *}[\sigma] \in H^{\mathrm{p}}_n(C_T X)$, which passes to an element in $H_{n-1}(\partial_T X)$ via the map $H^{\mathrm{p}}_n(C_T X) \to H_n(C_T X, C_T X \setminus \{o\}) \cong H_{n-1}(\partial_T X)$.

In the construction of f_ϵ, we have to choose a base point, the geodesic rays $\{l_i(t)\}_{i=1}^N$, the covering $\{U_i\}_{i=1}^\infty$ and the maps b and b'. However, different choices give maps in the same proper homotopy class if the corresponding ϵ is small enough. Also the geodesic homotopy from f_{ϵ_1} to f_{ϵ_2} is proper if ϵ_1 and ϵ_2 are small enough, so the above boundary map is well-defined.

Next we construct a map in the opposite direction as follows. Let η' be a Lipschitz $(n-1)$–cycle in $\partial_T X$. Let α' be the cone over η'. Note that one can cone off elements in $C_{n-1}(\partial_T X)$ to obtain elements in $C_n^{\mathrm{p}}(C_T X)$, which induces a homomorphism $H_{n-1}(\partial_T X) \to H_n^{\mathrm{p}}(C_T X)$. Actually $[\alpha'] \in H_{n,n}^{\mathrm{p}}(C_T X)$ since the cone over a Lipschitz cycle would satisfy the required growth condition. If $\sigma' = \log(\alpha')$, then $[\sigma'] \in H_{n,n}^{\mathrm{p}}(X)$ since log is 1–Lipschitz. Now we define the "coning off" map

$$c\colon H_{n-1}(\partial_T X) \to H_{n,n}^{\mathrm{p}}(X) \tag{4-28}$$

by sending $[\eta']$ to $[\sigma']$. The base point in the definition of log does not matter because different base points give maps which are of bounded distance from each other. It is easy to see that c is a group homomorphism.

For $\epsilon > 0$, pick a finite ϵ–net of $\operatorname{Im}\eta'$ and denote it by $\{\xi_i\}_{i=1}^N$. Suppose $p = \log(o)$ and suppose $\{l_i\}_{i=1}^N$ are the unit-speed geodesic rays emanating from p with $\partial_T l_i = \xi_i$. Pick $R_\epsilon > 0$ such that

$$\left| \frac{d(l_i(t), l_j(t))}{t} - \lim_{t\to+\infty} \frac{d(l_i(t), l_j(t))}{t} \right| < \epsilon \tag{4-29}$$

for $t > R_\epsilon$. Let $I_{\sigma'}$ be the smallest subcomplex of X which contains $\operatorname{Im}\sigma'$. By using the rays $\{l_i\}_{i=1}^N$ as in the proof of Lemma 4-16, we can construct a continuous proper map $g_\epsilon\colon I_{\sigma'} \to C_T X$ skeleton by skeleton so that

$$d(g_\epsilon \circ \log(x), x) \le L\epsilon \max\{d(x,o), R_\epsilon\} \tag{4-30}$$

for $x \in \operatorname{Im}\alpha'$, which implies $g_{\epsilon*}[\sigma'] = [\alpha']$ for ϵ small.

Let $[\sigma'']$ be the fundamental class of $S_{[\sigma']}$ and let $f_{\epsilon*}\colon S_{[\sigma']} \to C_T X$ be the map in Lemma 4-16. We claim that $g_{\epsilon*}[\sigma'] = f_{\epsilon*}[\sigma'']$ for ϵ small, which would imply

$$\partial \circ c = \mathrm{Id}. \tag{4-31}$$

To see the claim, note that $[\sigma'] = [\sigma'']$ in $H_n^{\mathrm{p}}(I_{\sigma'})$. For ϵ small, there is a proper geodesic homotopy between $g_\epsilon|_{S_{[\sigma']}}$ and f_ϵ by (4-23) and (4-30), thus $g_{\epsilon*}[\sigma''] = f_{\epsilon*}[\sigma'']$. Moreover, $g_{\epsilon*}[\sigma''] = g_{\epsilon*}[\sigma']$, so $f_{\epsilon*}[\sigma''] = g_{\epsilon*}[\sigma'] = [\alpha']$.

From (4-23) and the discussion after it we know

$$c \circ \partial = \mathrm{Id}. \tag{4-32}$$

Thus ∂ is also a group homomorphism and we have the following result.

Corollary 4-33 *If X is an n–dimensional* $\mathrm{CAT}(0)$ *cube complex, then:*

(1) $\partial\colon H^{\mathrm{p}}_{n,n}(X) \to H_{n-1}(\partial_T X)$ *is a group isomorphism, and the inverse is given by* $c\colon H_{n-1}(\partial_T X) \to H^{\mathrm{p}}_{n,n}(X)$.

(2) *If $q\colon X \to X'$ is a quasi-isometric embedding from X to another n–dimensional* $\mathrm{CAT}(0)$ *cube complex X', then q induces a monomorphism $q_*\colon H_{n-1}(\partial_T X) \to H_{n-1}(\partial_T X')$. If q is a quasi-isometry, then q_* is an isomorphism.*

Proof We only need to prove (2). Let us approximate q by a Lipschitz quasi-isometric embedding and denote the smallest subcomplex of X' that contains $\operatorname{Im} q$ by I_q. Now we have a homomorphism

$$q_*\colon H^{\mathrm{p}}_{n,n}(X) \to H^{\mathrm{p}}_{n,n}(I_q) \hookrightarrow H^{\mathrm{p}}_{n,n}(X'). \tag{4-34}$$

We can define a continuous map $p\colon I_q \to X$ skeleton by skeleton in such a way that $d(x, p \circ q(x)) < D$ for all $x \in X$ (here D is some positive constant), which induces $p_*\colon H^{\mathrm{p}}_{n,n}(I_q) \to H^{\mathrm{p}}_{n,n}(X)$. It is easy to see $p_* \circ q_* = \mathrm{Id}$ and $q_* \circ p_* = \mathrm{Id}$, so the first map in (4-34) is an isomorphism. Note that the second map in (4-34) is a monomorphism, thus q_* is injective and (2) follows from (1). □

We refer to Theorem A-19 and the remarks after it for generalizations of the above corollary.

Remark 4-35 Though we are working with $\mathbb{Z}/2$ coefficients, it is easy to check that the analogue of Corollary 4-33 for arbitrary coefficients is also true (the same proof goes through).

Remark 4-36 By the above proof and the argument in Lemma 4-16, there exists a positive D', which depends on the quasi-isometry constant of q, such that

$$d_H(q(S_{[\tilde{\sigma}]}), S_{q_*[\tilde{\sigma}]}) < D' \tag{4-37}$$

for any $[\tilde{\sigma}] \in H^{\mathrm{p}}_{n,n}(X)$.

4.3 Cubical coning

Note that the above coning map c does not give us much information about the combinatorial structure of the support set. Now we introduce an alternative coning procedure based on the cubical structure. We can assume, by Lemma 4-16, that $K = \bigcup_{i=1}^{N} \Delta_i$, where each Δ_i is an all-right spherical $(n-1)$–simplex. Let $\{O_i\}_{i=1}^{N}$ be the collection of top-dimensional orthant subcomplexes in X such that $\partial_T O_i = \Delta_i$. By (2-11), we can pass to suborthants and assume $\{O_i\}_{i=1}^{N}$ is a disjoint collection.

The natural quotient map $\bigsqcup_{i=1}^N \Delta_i \to K$ induces a quotient map $Q\colon \bigsqcup_{i=1}^N O_i \to CK$ sending the tip of each O_i to the cone point of CK. We define an inverse map $F\colon CK \to \bigsqcup_{i=1}^N O_i \subset X$ by sending each $x \in CK$ to some point in $Q^{-1}(x)$.

Lemma 4-38 $F\colon CK \to X$ *is a quasi-isometric embedding.*

Recall that CK is endowed with the induced metric from $C_T X$.

Proof Let o_i be the tip of O_i and $L = \max_{i \neq j} d(o_i, o_j)$. For $x \in O_i$ and $y \in O_j$, let $c_{i,x} \subset O_i$ be the constant-speed geodesic ray with $c_{i,x}(0) = o_i$ and $c_{i,x}(1) = x$. We can define $c_{j,y} \subset O_j$ similarly. Let c'_j be the geodesic ray which (1) is asymptotic to $c_{j,y}$; (2) has the same speed as $c_{j,y}$; (3) satisfies $c'_j(0) = o_i$. Then by Lemma 2-4 and convexity of $d(c_i(t), c'_j(t))$,

$$
\begin{aligned}
d(Q(x), Q(y)) &= \lim_{t\to\infty} \frac{d(c_{i,x}(t), c_{j,y}(t))}{t} \\
&= \lim_{t\to\infty} \frac{d(c_{i,x}(t), c'_j(t))}{t} \\
&\geq d(c_{i,x}(1), c'_j(1)) \\
&\geq d(c_{i,x}(1), c_{j,y}(1)) - d(c'_j(1), c_{j,y}(1)) \\
&\geq d(x, y) - d(o_i, o_j) \\
&\geq d(x, y) - L.
\end{aligned}
\tag{4-39}
$$

It follows that

$$d(F(x), F(y)) \leq d(x, y) + L \tag{4-40}$$

for any $x, y \in CK$.

For the other direction, pick $x \in O_i$ and $y \in O_j$, and let us assume without loss of generality that $i \neq j$ and x, y are interior points of O_i and O_j. We extend $\overline{o_i x}$ (or $\overline{o_j y}$) to get a ray $\overline{o_i \xi_1} \subset O_i$ (or $\overline{o_j \xi_2} \subset O_j$). Let $(Y_1, Y_2) = \mathcal{I}(O_i, O_j)$. Since $d(x, y) \leq d(x, Y_1) + d(Y_1, Y_2) + d(y, Y_2) \leq d(x, Y_1) + d(y, Y_2) + L$, we can assume without loss of generality that

$$d(x, Y_1) \geq \tfrac{1}{2}(d(x, y) - L). \tag{4-41}$$

From (4-15), we have

$$
\begin{aligned}
d(F(x), F(y)) \geq d(x, o_i) \sin(\angle_T(\xi_1, \xi_2)) &\geq \tfrac{1}{2} A d(x, o_i) \sin(\angle_T(\xi_1, \partial_T Y_1)) \\
&\geq \tfrac{1}{2} A d(x, Y_1) - L' \\
&\geq \tfrac{1}{4} A d(x, y) - L' - \tfrac{1}{2} L
\end{aligned}
\tag{4-42}
$$

if $\angle_T(\xi_1, \xi_2) < \frac{\pi}{2}$, and

$$d(F(x), F(y)) \geq d(x, o_i) \geq d(x, Y_1) - L' \geq d(x, y) - L' - \tfrac{1}{2}L \tag{4-43}$$

if $\angle_T(\xi_1, \xi_2) \geq \frac{\pi}{2}$. Here A and L' depend on O_i and O_j, but there are finitely many orthants, so we can make A and L' uniform. □

Since X is linearly contractible, we can approximate F by a continuous quasi-isometric embedding F' such that $d(F(x), F'(x)) \leq L$ for some constant L and any $x \in CK$. Let $K^{(n-2)}$ be the $(n-2)$–skeleton of K and define ρ: $CK \to [0,1]$ to be

$$\rho(x) = \begin{cases} 1 & \text{if } d(x, CK^{(n-2)}) \leq 1, \\ 2 - d(x, CK^{(n-2)}) & \text{if } 1 < d(x, CK^{(n-2)}) < 2, \\ 0 & \text{if } d(x, CK^{(n-2)}) \geq 2. \end{cases}$$

Let

$$F_1(x) = \rho(x)F'(x) + (1 - \rho(x))F(x)$$

for $x \in CK$, where $\rho(x)F'(x) + (1 - \rho(x))F(x)$ denotes the point in the geodesic segment $\overline{F'(x)F(x)}$ which has distance $\rho(x)d(F'(x), F(x))$ from $F(x)$. Though F may not be continuous, F_1 is continuous, since the only discontinuity points of F are in the 1–neighborhood of $CK^{(n-2)}$, however inside such a neighborhood we have $F_1 = F'$ by definition. Also note that $d(F(x), F_1(x)) \leq L'$ for all $x \in CK$.

Since $F_1 = F$ outside the 2–neighborhood of $CK^{(n-2)}$, there exists an orthant subcomplex $O_i' \subset O_i$ such that $F_1^{-1}(O_i')$ is an orthant in CK for $1 \leq i \leq N$ and

$$d_H\left(\operatorname{Im} F_1, \bigcup_{i=1}^{N} O_i'\right) < \infty. \tag{4-44}$$

Let $[CK] \in H_n^{\mathrm{p}}(CK)$ be the fundamental class. If $[\tau] = (F_1)_*[CK] \in H_{n,n}^{\mathrm{p}}(X)$, then

$$\bigcup_{i=1}^{N} O_i' \subset S_{[\tau]} \subset \operatorname{Im} F_1. \tag{4-45}$$

The first inclusion follows from the construction of O_i' and the second follows from Lemma 3-2. Equations (4-44) and (4-45) immediately imply:

Lemma 4-46
$$d_H\left(S_{[\tau]}, \bigcup_{i=1}^{N} O_i'\right) < \infty.$$

Now we are ready to prove the main result.

Theorem 4-47 *Let X be a* CAT(0) *cube complex of dimension n. Pick $[\sigma] \in H^{\mathrm{p}}_{n,n}(X)$ and suppose $S = S_{[\sigma]}$. Then there is a finite collection $O_1, \dots, O_k$ of n–dimensional orthant subcomplexes of S such that*

$$d_H\left(S, \bigcup_{i=1}^{k} O_k\right) < \infty.$$

Proof By Lemma 4-46, it suffices to show $[\sigma] = [\tau]$ in $H^{\mathrm{p}}_n(X)$. Note that (4-45) implies $\partial_T S_{[\tau]} = K$, so $\partial([\tau]) = [K] = \partial([\sigma])$, where $[K]$ is the fundamental class of K and ∂ is the map in (4-27). Thus $[\sigma] = [\tau]$ by Corollary 4-33. □

In particular, by Lemma 3-4 and Theorem 4-47, we have:

Theorem 4-48 *If X is a* CAT(0) *cube complex of dimension n, then for every n–quasiflat Q in X, there is a finite collection $O_1, \dots, O_k$ of n–dimensional orthant subcomplexes in X such that*

$$d_H\left(Q, \bigcup_{i=1}^{k} O_k\right) < \infty.$$

5 Preservation of top-dimensional flats

5.1 The lattice generated by top-dimensional quasiflats

We investigate the coarse intersection of the top-dimensional quasiflats in this section.

Let X be a finite-dimensional CAT(0) cube complex. For two subsets A and B, we say they are *coarsely equivalent* (denoted $A \sim B$) if $d_H(A, B) < \infty$. We assume the empty subset is coarsely equivalent to any bounded subset. Denote by $[A]$ the coarse equivalence class which contains A. We say $[A] \subset [B]$ if there exists an $r < \infty$ such that $A \subset N_r(B)$. If $[A] \subset [B]$ and $[A] \neq [B]$, we will write $[A] \subsetneq [B]$. Also we define the union $[A] \cup [B]$ to be $[A \cup B]$, but intersection is not well-defined in general.

The class $[A]$ is *admissible* if it can be represented by a subset which is a finite union of (not necessarily top-dimensional) orthant subcomplexes in X (here A is allowed to be empty). Let $\mathcal{A}(X)$ be the collection of admissible classes of subsets in X. Pick $[A_1], [A_2] \in \mathcal{A}(X)$. We define another two operations between $[A_1]$ and $[A_2]$ as follows.

(1) By Lemma 2-10(4), there exists an $r < \infty$ such that

$$[N_{r_1}(A_1) \cap N_{r_1}(A_2)] = [N_{r_2}(A_1) \cap N_{r_2}(A_2)]$$

for any $r_1 \geq r$ and $r_2 \geq r$. We define the intersection $[A_1] \cap [A_2]$ to be $[N_r(A_1) \cap N_r(A_2)]$, which is also admissible.

(2) By Lemma 2-10(4), there exists an $r < \infty$ such that

$$[A_1 \setminus N_{r_1}(A_2)] = [A_1 \setminus N_{r_2}(A_2)]$$

for any $r_1 \geq r$ and $r_2 \geq r$. Define the subtraction $[A_1] - [A_2]$ to be $[A_1 \setminus N_r(A_2)]$, which is also admissible.

If Y is another CAT(0) cube complex with $\dim(Y) = \dim(X)$ and there is a quasi-isometry $f\colon X \to Y$, then we define $f_\sharp([A])$ to be $[f(A)]$. This is well-defined since $A \sim B$ implies $f(A) \sim f(B)$. Note that:

(1) $f_\sharp([A]) \cup f_\sharp([B]) = f_\sharp([A] \cup [B])$.

(2) If $[A], [B], [f(A)]$ and $[f(B)]$ are all admissible, then

$$f_\sharp([A]) \cap f_\sharp([B]) = f_\sharp([A] \cap [B]) \quad \text{and} \quad f_\sharp([A]) - f_\sharp([B]) = f_\sharp([A] - [B]).$$

We only verify the last equality. Since f is a quasi-isometry, there exist constants $a > 1$, $b > 0$ such that for r large enough, we have

$$f(A) \setminus N_{ar+b}(f(B)) \subset f(A \setminus N_r(B)) \subset f(A) \setminus N_{(r/a)-b}(f(B)).$$

Since $[f(A)]$ and $[f(B)]$ are admissible, the first term and the last term of the above inequality are in the same coarse class for r large enough. This finishes the proof.

Let $\mathcal{Q}(X)$ be the collection of top-dimensional quasiflats in X, modulo the above equivalence relation. Theorem 4-48 implies $\mathcal{Q}(X) \subset \mathcal{A}(X)$. Let $\mathcal{KQ}(X)$ be the smallest subset of $\mathcal{A}(X)$ which contains $\mathcal{Q}(X)$ and is closed under union, intersection and subtraction as defined above. More precisely, each element $\mathcal{KQ}(X)$ can be written as a finite string of elements of $\mathcal{Q}(X)$ with union, intersection or subtraction between adjacent terms and braces which indicate the order of these operations. Let $f\colon X \to Y$ be a quasi-isometry. Then by induction on the length of the string, one can show $[f(A)]$ is admissible and $[f(A)] \in \mathcal{KQ}(Y)$ for each $[A] \in \mathcal{KQ}(X)$. By considering the quasi-isometry inverse of f, we have the following theorem.

Theorem 5-1 *Let X and Y be n–dimensional* CAT(0) *cube complexes. If $f\colon X \to Y$ is a quasi-isometry, then f induces a bijection $f_\sharp\colon \mathcal{KQ}(X) \to \mathcal{KQ}(Y)$. Moreover, for*

$[A], [B] \in \mathcal{KQ}(X)$, *we have:*

$$f_\sharp([A]) \cup f_\sharp([B]) = f_\sharp([A] \cup [B]),$$
$$f_\sharp([A]) \cap f_\sharp([B]) = f_\sharp([A] \cap [B]),$$
$$f_\sharp([A]) - f_\sharp([B]) = f_\sharp([A] - [B]).$$

For $[A]$ admissible, pick a representative in $[A]$ which is a finite union of orthant complexes. Define the *order* of $[A]$, denoted $\|[A]\|$, to be the number of top-dimensional orthant complexes in the representative. By Lemma 2-10, this definition does not depend on the choice of representative. Since every element in $\mathcal{KQ}(X)$ is admissible, we have a map $\mathcal{KQ}(X) \to \{0\} \cup \mathbb{Z}^+$ with the following properties:

(1) $\|[Q]\| \geq 2^{\dim X}$ for $[Q] \in \mathcal{Q}(X)$.

(2) $\|[A] \cup [B]\| = \|[A]\| + \|[B]\| - \|[A] \cap [B]\|$ for $[A], [B] \in \mathcal{KQ}(X)$.

(3) Let f be as in Theorem 5-1. Then $\|[A]\| = 0$ if and only if $|f_\sharp([A])| = 0$ for $[A] \in \mathcal{KQ}(X)$.

The first assertion follows from (3-12).

We say an element $[A] \in \mathcal{KQ}(X)$ is *essential* if $\|[A]\| > 0$. We call $[A]$ a *minimal essential element* if for any $[B] \in \mathcal{KQ}(X)$ with $[B] \subsetneq [A]$, we have $\|[B]\| = 0$. Minimal essential elements have the following properties:

(1) For any $[A] \in \mathcal{KQ}(X)$, there is a decomposition $[A] = \big(\bigcup_{i=1}^N [A_i]\big) \cup [B]$ such that each $[A_i]$ is a minimal essential element and $\|[B]\| = 0$. We also require $[B]$ and each $[A_i]$ to be in $\mathcal{KQ}(X)$.

(2) For two different minimal essential elements $[A_1], [A_2] \in \mathcal{KQ}(X)$, we have $\|[A_1] \cap [A_2]\| = 0$, thus $\|[A_1] \cup [A_2]\| = \|[A_1]\| + \|[A_2]\|$.

(3) Let f be as above. If $[A]$ is a minimal essential element in $\mathcal{KQ}(X)$, then $f_\sharp([A])$ is also a minimal essential element.

We only prove (1). For each top-dimensional orthant subcomplex $[O_i]$ such that $[O_i] \subset [A]$, let $[A_i]$ be the minimal element in $\mathcal{KQ}(X)$ which contains $[O_i]$. We claim that $[A_i]$ is minimal essential. Suppose the contrary true, ie there exists $[A_i'] \in \mathcal{KQ}(X)$ such that $\|[A_i']\| \neq 0$ and $[A_i'] \subsetneq [A_i]$. The minimality of $[A_i]$ implies $[O_i] \subset [A_i']$ does not hold. However, in such a case $[O_i] \subset [A_i] - [A_i'] \subsetneq [A_i]$, which contradicts the minimality of $[A_i]$. We choose $[B] = [A] - \big[\bigcup_{i=1}^N A_i\big]$.

Lemma 5-2 *Let X, Y be n–dimensional* CAT(0) *cube complexes and let $f\colon X \to Y$ be an (L', A')–quasi-isometry. If $|f_\sharp([A])| = \|[A]\|$ for any minimal essential element $[A]$ in $\mathcal{KQ}(X)$, then there exists a constant $C = C(L', A')$ such that for any top-dimensional flat $F \subset X$, there exists a top-dimensional flat $F' \subset Y$ such that $d_H(f(F), F') < C$.*

Proof By Theorem 5-1 and the above discussion, we know $|f_\sharp([A])| = \|[A]\|$ for any $[A] \in \mathcal{KQ}(X)$, in particular $\|[f(F)]\| = \|[F]\| = 2^n$ (here $n = \dim(X) = \dim(Y)$). By Lemma 3-4, let $[\sigma] \in H_n^{\mathrm{p}}(Y)$ be the class such that $d_H(S_{[\sigma]}, f(F)) < \infty$. By Theorem 4-47, $S_{[\sigma]}$ is Hausdorff close to a union of 2^n orthant subcomplexes. Thus $\partial_T S_{[\sigma]}$ is contained in 2^n right-angled spherical simplices of dimension $n-1$. Then $\mathcal{H}^{n-1}(\partial_T S_{[\sigma]}) \le \mathcal{H}^{n-1}(\mathbb{S}^{n-1})$. We pick a base point $p \in S_{[\sigma]}$ and consider the logarithmic map $\log_p\colon C_T Y \to Y$. Lemma 3-6 implies $S_{[\sigma]} \subset \log_p(C_T S_{[\sigma]})$. Thus

$$\frac{\mathcal{H}^n(B(p,r) \cap S_{[\sigma]})}{r^n} \le \frac{\mathcal{H}^n(B(p,r) \cap \log_p(C_T S_{[\sigma]}))}{r^n} \le \frac{\mathcal{H}^n(B(o,r) \cap C_T S_{[\sigma]})}{r^n} \le \omega_n.$$

Here o is the cone point in $C_T Y$ and ω_n is the volume of unit ball in $\mathbb{E}^n$. The second inequality follows from the fact that $\log_p$ is 1–Lipschitz and the third inequality follows from $\mathcal{H}^{n-1}(\partial_T S_{[\sigma]}) \le \mathcal{H}^{n-1}(\mathbb{S}^{n-1})$. By Theorem 3-10(2), $S_{[\sigma]}$ is isometric to $\mathbb{E}^n$. □

5.2 The weakly special cube complexes

It is shown in [8] that the assumption of Lemma 5-2 is satisfied for 2–dimensional RAAGs. Our goal in this section is to find an appropriate class of cube complexes which shares some key properties of the canonical CAT(0) cube complexes of RAAGs such that the assumption of Lemma 5-2 is satisfied. In [20], Haglund and Wise introduced a class of RAAG-like cube complexes, which are called special cube complexes. We adjust their definition for our purposes in the following way.

Definition 5-3 A cube complex K is *weakly special* if:

(1) K is nonpositively curved.

(2) No hyperplane *self-osculates* or *self-intersects*.

The second condition means that for any vertex v and two distinct edges e_1 and e_2 such that $v \in e_1 \cap e_2$, the hyperplanes dual to e_1 and e_2 are different.

If K is compact, then there exists a finite sheet, weakly special cover $\bar{K}$ of K such that every hyperplane in $\bar{K}$ is *two-sided*, ie there exists a small neighborhood of the hyperplane which is a trivial interval bundle over the hyperplane. This follows from the argument in [20, Proposition 3.10].

In the rest of this section, we will denote by W' a compact weakly special cube complex, and W the universal cover of W'. Since we mainly care about W, there is no loss of generality in assuming every hyperplane in W' is two-sided. The goal of this section is to prove the following theorem.

Theorem 5-4 *Let W_1' and W_2' be two compact weakly special cube complexes with $\dim(W_1') = \dim(W_2') = n$. Suppose W_1, W_2 are the universal covers of W_1', W_2' respectively. If $f\colon W_1 \to W_2$ is a (L, A)–quasi-isometry, then there exists a constant $C = C(L, A)$ such that for any top-dimensional flat $F \subset W_1$, there exists a top dimensional flat $F' \subset W_2$ with $d_H(f(F), F') < C$.*

This theorem follows from Lemma 5-2 and the following lemma.

Lemma 5-5 *Let W_1, W_2 and f be as in Theorem 5-4. If $f_\sharp\colon \mathcal{KQ}(W_1) \to \mathcal{KQ}(W_2)$ is the induced bijection in Theorem 5-1, then $|f_\sharp([A])| = \|[A]\|$ for any minimal essential element $[A] \in \mathcal{KQ}(W_1)$.*

In the rest of this section, we will prove Lemma 5-5.

We label the vertices and edges of W' by $\{\overline{v}_i\}_{i=1}^{N_v}$ and $\{\overline{e}_i\}_{i=1}^{N_e}$ such that: (1) different vertices have different labels; (2) two edges have the same label if and only if they are dual to the same hyperplane. We also choose an orientation for each edge such that if two edges are dual to the same hyperplane, then their orientations are consistent with parallelism (this is possible since each hyperplane is two-sided). All the labelings and orientations lift to the universal cover W. The edges in W dual to the same hyperplane also share the same label.

For every edge-path ω in W' or W, define $L(\omega)$ to be the word $\overline{v}_i \overline{e}_{i_1}^{\epsilon_{i_1}} \overline{e}_{i_2}^{\epsilon_{i_2}} \overline{e}_{i_3}^{\epsilon_{i_3}} \cdots$, where $\overline{v}_i$ is the label of the initial vertex of ω, $\overline{e}_{i_j}$ is the label of the j^{th} edge and $\epsilon_{i_j} = \pm 1$ records the orientation of the j^{th} edge.

Definition 5-3 and the way we label W' imply:

(1) For two edges e_1' and e_2' in W' dual to the same hyperplane, e_1' is embedded if and only if e_2' is embedded, ie its end points are distinct.

(2) Pick any vertex $v_i' \in W'$. Then two distinct edges e_1' and e_2' with $v_i' \in e_1' \cap e_2'$ have different labels.

(3) If ω_1' and ω_2' are two edge paths in W' such that $L(\omega_1') = L(\omega_2')$, then $\omega_1' = \omega_2'$. If ω_1 and ω_2 are two edge paths in W such that $L(\omega_1) = L(\omega_2)$, then there exists a unique deck transformation γ such that $\gamma(\omega_1) = \omega_2$.

We will be using the following simple observation repeatedly.

Lemma 5-6 *Pick vertices v_1 and v_2 in W which have the same label. For $i = 1, 2$, let $\{l_{ij}\}_{j=1}^{k}$ be a collection such that each l_{ij} is a geodesic ray, a geodesic segment or a complete geodesic that contains v_i. Suppose that:*

(1) *Each l_{ij} is a subcomplex of W.*

(2) *For each j, there is a graph isomorphism $\phi_j\colon l_{1j} \to l_{2j}$ which preserves the labels of vertices and edges and the orientations of edges, moreover $\phi_j(v_1) = v_2$.*

(3) *The convex hull of $\{l_{1j}\}_{j=1}^{k}$, which we denote by K_1, is a subcomplex isometric to $\prod_{j=1}^{k} l_{1j}$.*

Then the convex hull of $\{l_{2j}\}_{j=1}^{k}$, which we denote by K_2, is a subcomplex isometric to $\prod_{j=1}^{k} l_{2j}$. Moreover, let γ be the deck transformation such that $\gamma(v_1) = v_2$. Then $\gamma(K_1) = K_2$.

Let $\dim(W) = n$ and let O be a top-dimensional orthant subcomplex in W. We now construct a suitable doubling of O which will serve as a basic move to analyze the minimal essential elements in $\mathcal{KQ}(W)$.

Let $\{r_j\}_{i=1}^{n}$ be the geodesic rays emanating from the tip of O such that O is the convex hull of $\{r_j\}_{j=1}^{n}$. We parametrize r_1 by arc length. Since the labeling of W is finite, we can find a sequence of nonnegative integers $\{n_j\}_{j=1}^{\infty}$ with $n_j \to \infty$ such that the label and orientation of the incoming edge at $r_1(n_j)$, the label and orientation of the outgoing edge at $r_1(n_j)$ and the label of $r_1(n_j)$ do not depend on j.

We identify O with $[0,\infty)\times O'$, where O' is an $(n-1)$–dimensional orthant orthogonal to r_1. By our choice of $r_1(n_1)$ and $r_1(n_2)$, we can extend $\overline{r_1(n_2)r_1(n_1)}$ over $r_1(n_1)$ to reach a vertex v such that $L(\overline{r_1(n_1)v}) = L(\overline{r_1(n_2)r_1(n_1)})$. Here v does not need to lie on r_1; see Figure 1.

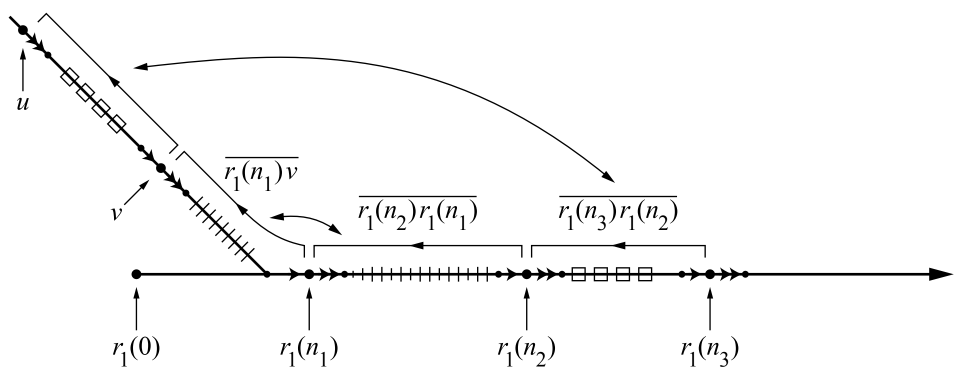


Figure 1

Let K_1 be the convex hull of $\{n_2\} \times O'$ and $\overline{r_1(n_2)r_1(n_1)}$. Then K_1 is of form $K_1 = [n_1, n_2] \times O'$. Note that the parallelism map between $\{n_1\} \times O'$ and $\{n_2\} \times O'$ preserves labeling and orientation of edges. Then it follows from Lemma 5-6 that the convex hull of $\{n_1\} \times O'$ and $\overline{r_1(n_1)v}$ is a subcomplex isometric to $\overline{r_1(n_1)v} \times O'$. (Actually if $\gamma \in \pi_1(W')$ is the deck transformation satisfying $\gamma(r_1(n_2)) = r_1(n_1)$, then $\gamma(K_1)$ is the convex hull of $\{n_1\} \times O'$ and $\overline{r_1(n_1)v}$.) We call this subcomplex the *mirror* of K_1 and denote it by K_1'. Since $K_1 \cap K_1' = \{n_1\} \times O'$, it follows that $K_1' \cup ([n_1, \infty) \times O')$ is again an orthant; see Figure 2.

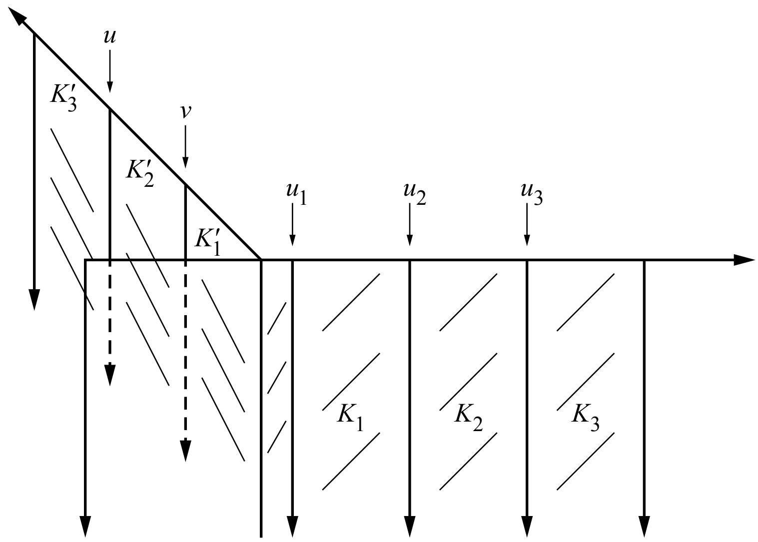


Figure 2

Let $K_2 = [n_2, n_3] \times O'$. We extend $\overline{r_1(n_1)v}$ over v to reach a vertex u such that $L(\overline{vu}) = L(\overline{r_1(n_3)r_1(n_2)})$. Note that the parallelism map between $\{v\} \times O'$ and $\{n_3\} \times O'$ preserves labeling and orientation of edges. Then it follows from Lemma 5-6 that the convex hull of $\{v\} \times O'$ and $\overline{vu}$ is a subcomplex isometric to K_2. (Actually if $\gamma \in \pi_1(W')$ is the deck transformation satisfying $\gamma(r_1(n_3)) = v$, then $\gamma(K_2)$ is the convex hull of $\{v\} \times O'$ and $\overline{vu}$.) This convex hull is called the mirror of K_2, and is denoted by K_2'. Since $\overline{vu}$ and $\overline{r_1(n_1)v}$ fit together to form a geodesic segment, $K_1' \cap K_2' = \{v\} \times O'$. Thus $K_2' \cup K_1' \cup ([n_1, \infty) \times O')$ is again an orthant. We can continue this process, and consecutively construct the mirror of $K_i = [n_i, n_{i+1}] \times O'$ in W (denoted K_i') arranged in the pattern indicated in the above picture. Similarly one can verify that K_i' is isometric to K_i, and $K_i' \cap K_{i+1}'$ is isometric to O'.

Now we obtain a subcomplex $K = \left(\bigcup_{i=1}^{\infty} K_i\right) \cup \left(\bigcup_{i=1}^{\infty} K_i'\right)$. It is clear that $[O] \subset [K]$. The discussion in the previous paragraph implies that $\bigcup_{i=1}^{\infty} K_i'$ is also a top-dimensional orthant. We will call it the *mirror* of O. Moreover, K is isometric to $\mathbb{R} \times (\mathbb{R}_{\geq 0})^{n-1}$. More generally, by the same argument as above and Lemma 5-6, we have the following result.

Lemma 5-7 *If $K \subset W$ is a convex subcomplex isometric to $(\mathbb{R}_{\geq 0})^k \times \mathbb{R}^{n-k}$, then there exists a convex subcomplex K' isometric to $(\mathbb{R}_{\geq 0})^{k-1} \times \mathbb{R}^{n-k+1}$ such that $[K] \subset [K']$.*

Pick a minimal essential element $[A] \in \mathcal{KQ}(W)$. Then there exists a top-dimensional orthant O with $[O] \subset [A]$, where $[O]$ may not be an element in $\mathcal{KQ}(W)$. Using Lemma 5-7, we can double the orthant n times to get a top-dimensional flat F with $[O] \subset [F]$. Since $[A]$ is minimal, $[A] \cap [F] = [A]$, which implies the following result.

Corollary 5-8 *If $[A] \in \mathcal{KQ}(W)$ is a minimal essential element, then there exists a top-dimensional flat $[F]$ such that $[A] \subset [F]$. In particular, $\|[A]\| \leq 2^{\dim(W)} = 2^n$.*

Pick a top-dimensional orthant subcomplex O and denote the $(n-1)$–faces of O by $\{O_i\}_{i=1}^n$. We say that $[O_i]$ is *branched* if there exist top-dimensional orthant subcomplexes O' and O'' such that $[O]$, $[O']$ and $[O'']$ are distinct elements and $[O] \cap [O'] = [O] \cap [O''] = [O_i]$; otherwise $[O_i]$ is called *unbranched.*

Lemma 5-9 *If O and O_i are as above, then $[O_i]$ is branched if and only if there exists a suborthant $O_i' \subset O_i$ and geodesic rays l_1, l_2 and l_3 emanating from the tip of O_i' such that:*

1. *$[O_i'] = [O_i]$.*
2. *$[l_1]$, $[l_2]$ and $[l_3]$ are distinct.*
3. *The convex hull of l_j and O_i' is a top-dimensional orthant for $1 \leq j \leq 3$.*

Proof If O_i is branched, let O' and O'' be the orthant subcomplexes as above. We can assume $O' \cap O = O'' \cap O = \varnothing$. Let $(Y_1, Y_2) = \mathcal{I}(O, O')$. Since Y_1 and Y_2 bound a copy of $Y_1 \times [0, d(O, O')]$ inside W, we have $\dim(Y_1) = \dim(Y_2) \leq n-1$. However, (2-11) implies $[Y_1] = [O] \cap [O'] = [O_i]$, so Y_1 and Y_2 are $(n-1)$–dimensional orthant subcomplexes. We can find a copy of $Y_2 \times [0, \infty)$ inside O' and we claim $Y_1 \times [0, d(O, O')] \cup Y_2 \times [0, \infty)$ is also a top-dimensional orthant subcomplex.

To see this, note that $(Y_1 \times [0, d(O, O')]) \cap (Y_2 \times [0, \infty)) = Y_2$. Pick $y \in Y_2$, and let $\{v_i\}_{i=1}^{n-1}$ be mutually orthogonal directions in $\Sigma_y Y_2$. Moreover, we can assume each v_i is in the 0–skeleton of $\Sigma_y Y_2 \subset \Sigma_y W$. Let $v \in \Sigma_y(Y_1 \times [0, d(O, O')])$ be the direction corresponding to the $[0, d(O, O')]$ factor and let $v' \in \Sigma_y(Y_2 \times [0, \infty))$ be the direction corresponding to the $[0, \infty)$ factor. It is clear that v and v' are distinct points in the 0–skeleton of $\Sigma_y W$. If $d(v, v') = \frac{\pi}{2}$, then v, v' and $\{v_i\}_{i=1}^{n-1}$ would be mutually orthogonal directions, which yields a contradiction with the fact that $\dim(W) = n$.

Thus $d(v,v')=\pi$ and $Y_1\times[0,d(O,O')]\cup Y_2\times[0,\infty)$ is indeed a top-dimensional orthant subcomplex.

Note that the orthant constructed above is the convex hull of Y_1 and some geodesic ray l emanating from the tip of Y_1. We can repeat this argument for O'' to obtain the required suborthants and geodesic rays in the lemma. The other direction of the lemma is trivial. $\square$

Let O, $\{r_j\}_{j=1}^n$, $\{n_i\}_{i=1}^\infty$, K_i and K_i' be as in the discussion before Lemma 5-7. Let $a_j=n_{j+1}-n_1$ for $j\geq 0$. We identify $\big(\bigcup_{i=1}^\infty K_i\big)\cup\big(\bigcup_{i=1}^\infty K_i'\big)$ with $\mathbb{R}\times\prod_{j=2}^n r_j$ such that $K_i=[a_{i-1},a_i]\times\prod_{j=2}^n r_j$. Thus $K_i'=[-a_i,-a_{i-1}]\times\prod_{j=2}^n r_j$. Let l be the unit-speed complete geodesic line in W such that $l(0)=r_1(n_1)$ and it is parallel to the $\mathbb{R}$ factor. For $x\in\mathbb{R}$, we denote the geodesic ray in $\big(\bigcup_{i=1}^\infty K_i\big)\cup\big(\bigcup_{i=1}^\infty K_i'\big)$ that starts at $l(x)$ and goes along the r_j factor by $\{x\}\times r_j$.

Let $\gamma_i\in\pi_1(W')$ be the deck transformation satisfying $\gamma_i(l(a_i))=l(-a_{i-1})$. Then by our construction, $\gamma_i(K_i)=K_i'$. Moreover, under the product decomposition $K_i=[a_{i-1},a_i]\times\prod_{j=2}^n r_j$ and $K_i'=[-a_i,-a_{i-1}]\times\prod_{j=2}^n r_j$, we have that γ_i maps $[a_{i-1},a_i]$ to $[-a_i,-a_{i-1}]$ and fixes the factor $\prod_{j=2}^n r_j$ pointwise.

Let $\widetilde{O}=\bigcup_{i=1}^\infty K_i'$ be the mirror of O. There is an isometry ρ acting on

$$\Big(\bigcup_{i=1}^\infty K_i\Big)\cup\Big(\bigcup_{i=1}^\infty K_i'\Big)=\mathbb{R}\times\prod_{j=2}^n r_j$$

by flipping the $\mathbb{R}$ factor (the other factors are fixed). For $1\leq j\leq n$, let O_j be the $(n-1)$–face of O which is orthogonal to r_j and let $\widetilde{O}_j$ be the $(n-1)$–face of $\widetilde{O}$ such that $[\rho(\widetilde{O}_j)]=[O_j]$. (Recall that $[O]=\big[\bigcup_{i=1}^\infty K_i\big]$.)

Lemma 5-10 *$[O_j]$ is branched if and only if $[\widetilde{O}_j]$ is branched.*

Proof If $j=1$, then $[O_1]=[\widetilde{O}_1]=[O]\cap[\widetilde{O}]$ and the lemma is trivial, so let us assume $j\neq 1$. If $[O_j]$ is branched, then by Lemma 5-9, we can assume without loss of generality (one might need to modify K_i and K_i' by cutting off suitable pieces and replace l by a geodesic in $\mathbb{R}\times\prod_{j=2}^n r_j$ which is parallel to l) that there exist $i_0\geq 0$ and geodesic rays c_1, c_2, c_3 emanating from $l(a_{i_0})$ such that $[c_1]$, $[c_2]$, $[c_3]$ are distinct elements and the convex hull of c_m, $l([a_{i_0},\infty))$ and $\{a_{i_0}\}\times r_k$ (for $k\neq 1,j$), which we denote by H_m, is a top-dimensional orthant subcomplex for $1\leq m\leq 3$.

Let γ be the deck transformation satisfying $\gamma(l(a_{i_0}))=l(-a_{i_0})$. Such a γ exists by the construction of l (in the previous paragraph, we might possibly replace the original l by a geodesic parallel to l, however the same γ works). Let $\widetilde{c}_m=\gamma(c_m)$

for $1 \le m \le 3$. Then $[\widetilde{c}_1]$, $[\widetilde{c}_2]$, $[\widetilde{c}_3]$ are distinct since γ is an isometry. Since γ is label and orientation preserving, c_m and $\widetilde{c}_m$ correspond to the same word for $1 \le m \le 3$, moreover $\gamma(\{a_i\} \times r_k) = \{-a_i\} \times r_k$ for $k \ne 1$. To prove $[\widetilde{O}_j]$ is branched, it suffices to show the convex hull of $\widetilde{c}_m$, $l((-\infty, -a_{i_0}])$ and $\{-a_{i_0}\} \times r_k$ (for $k \ne 1, j$) is a top-dimensional orthant subcomplex.

For $m = 1$, we chop H_1 into pieces so that

$$H_1 = \bigcup_{i=i_0+1}^{\infty} L_i, \quad \text{where} \quad L_i = c_1 \times l([a_{i-1}, a_i]) \times \prod_{k \ne 1, j} r_k.$$

Let γ_i be the deck transformation defined before Lemma 5-10 and let $L'_i = \gamma(L_i)$. We claim that

$$\gamma_i\Bigl(c_1 \times \{a_{i-1}\} \times \prod_{k \ne 1,j} r_k\Bigr) = \gamma_{i+1}\Bigl(c_1 \times \{a_{i+1}\} \times \prod_{k \ne 1,j} r_k\Bigr)$$

for $i \ge i_0 + 1$. This claim follows from the following two observations: (1) both sides of the equality contain $l(-a_i)$; (2) γ_i, γ_{i+1} and the parallelism between $c_1 \times \{a_{i-1}\} \times \prod_{k \ne 1,j} r_k$ and $c_1 \times \{a_{i+1}\} \times \prod_{k \ne 1,j} r_k$ preserve labeling and orientation of edges. It follows from the claim that $H'_1 = \bigcup_{i=i_0+1}^{\infty} L'_i$ is a top-dimensional orthant subcomplex. By a similar argument as before, we know

$$\gamma\Bigl(c_1 \times \{a_i\} \times \prod_{k \ne 1,j} r_k\Bigr) = \gamma_{i_0+1}\Bigl(c_1 \times \{a_{i_0+1}\} \times \prod_{k \ne 1,j} r_k\Bigr),$$

thus H'_1 is the convex hull of $\widetilde{c}_1$, $l((-\infty, -a_i])$ and $\{-a_i\} \times r_k$, for $k \ne 1, j$. Moreover, $[H'_1] \cap [\widetilde{O}] = [\widetilde{O}_j]$. We can repeat this construction for $\widetilde{c}_2$ and $\widetilde{c}_3$, which implies $[\widetilde{O}_j]$ is branched. By the same argument, if $[\widetilde{O}_j]$ is branched, we can prove $[O_j]$ is branched. □

Remark 5-11 It is important that we keep track of information from the labels of O while constructing the mirror of O; in other words, if we construct $\widetilde{O}$ by the pattern indicated in Figure 3, we will not be able to conclude that $[O_j]$ is branched from the fact that $[\widetilde{O}_j]$ is branched.

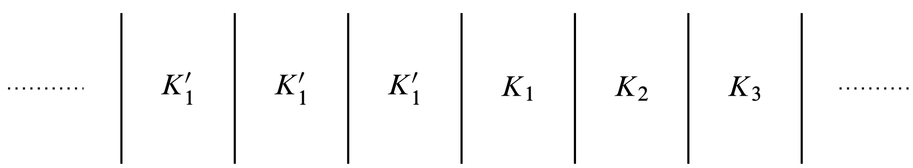


Figure 3

Lemma 5-12 *If $[A] \in \mathcal{KQ}(W)$ is a minimal essential element, then:*

(1) $\|[A]\| = 2^i$ *for some integer i with $1 \le i \le n$.*

(2) *There exists a top-dimensional flat F and another $2^{n-i}-1$ minimal essential elements $\{A_j\}_{j=1}^{2^{n-i}-1}$ with $\|[A_j]\| = \|[A]\|$ such that $[F] = [A] \cup \big(\bigcup_{j=1}^{2^{n-i}-1}[A_j]\big)$.*

Proof We find a top-dimensional orthant subcomplex O such that $[O] \subset [A]$. By the argument before Lemma 5-7, we can double this orthant n times to a get top-dimensional flat F such that $[O] \subset [F]$. Assume without loss of generality that $O \subset F$. Denote by $\{O_i\}_{i=1}^n$ the $(n-1)$–faces of O and let $\rho_i\colon F \to F$ be the isometry that fixes O_i pointwise and flips the direction orthogonal to O_i.

Let G be the group generated by $\{\rho_i\}_{i=1}^n$. Then $G \cong (\mathbb{Z}/2)^n$. We define

$$\Lambda_b = \{1 \le i \le n \mid [O_i] \text{ is branched}\}, \quad \Lambda_u = \{1 \le i \le n \mid [O_i] \text{ is unbranched}\}.$$

Let G_b be the subgroup generated by $\{\rho_i\}_{i\in\Lambda_b}$ and let G_u be the subgroup generated by $\{\rho_i\}_{i\in\Lambda_u}$. We denote by G_i the subgroup generated by $\{\rho_1 \cdots \rho_{i-1}, \rho_{i+1} \cdots \rho_n\}$.

Claim 1 For any $\gamma \in G$, $[O_i]$ is branched if and only if $[\gamma(O_i)]$ is branched.

Proof Writing $\gamma = \rho_{i_1}\rho_{i_2}\cdots\rho_{i_k}$, we prove it by induction on k. The case $k = 0$ is trivial. In general, suppose $[O_i]$ is branched if and only if $[\rho_{i_2}\cdots\rho_{i_k}(O_i)]$ is branched. It follows from the way we construct F that $[\rho_{i_1}\rho_{i_2}\cdots\rho_{i_k}(O)]$ is the mirror of $[\rho_{i_2}\cdots\rho_{i_k}(O)]$. So by Lemma 5-10, $[\rho_{i_1}\rho_{i_2}\cdots\rho_{i_k}(O_i)]$ is branched if and only if $[\rho_{i_2}\cdots\rho_{i_k}(O_i)]$ is branched, thus the claim is true. □

Claim 2
$$[A] \subset \Big[\bigcup_{\gamma\in G_u} \gamma(O)\Big].$$

Proof If $[O_i]$ is branched, by Lemma 5-9 there exists a subcomplex M_i isometric to $(\mathbb{R}_{\ge 0})^{n-1}\times\mathbb{R}$ such that $[M_i]\cap[F]=[O]$. By Lemma 5-7 we can find a top-dimensional flat F_i such that $[M_i]\subset[F_i]$. Since $F_i \cap F \neq \varnothing$, by Lemma 2-10 $[F\cap F_i]=[F]\cap[F_i]$. Note that $F \cap F_i$ is a convex subcomplex of F with $\|[F\cap F_i]\cap[\rho_i(O)]\| = 0$, so $[F\cap F_i] \subset \big[\bigcup_{\gamma\in G_i}\gamma(O)\big]$. Recall that $[A]$ is a minimal essential element, so

$$[A]\subset[F]\cap\Big(\bigcap_{i\in\Lambda_b}[F_i]\Big)=\bigcap_{i\in\Lambda_b}([F_i]\cap[F])\subset\bigcap_{i\in\Lambda_b}\Big[\bigcup_{\gamma\in G_i}\gamma(O)\Big]=\Big[\bigcup_{\gamma\in G_u}\gamma(O)\Big]. \quad □$$

Claim 3
$$\Big[\bigcup_{\gamma\in G_u}\gamma(O)\Big] \subset [A].$$

Proof First we need the following observation. Let $[P_1]$ and $[P_2]$ be two different top-dimensional orthant complexes. Suppose each $[Q] \in \mathcal{Q}(X)$ satisfies the property that either $[P_1] \subset [Q]$ and $[P_2] \subset [Q]$, or $[P_1] \not\subseteq [Q]$ and $[P_2] \not\subseteq [Q]$. Then this property is also true for each element in $\mathcal{KQ}(X)$. To see this, let $\mathcal{A}_{P_1,P_2}(X)$ be the collection of elements in $\mathcal{A}(X)$ which satisfy this property. Then one readily verifies that $\mathcal{A}_{P_1,P_2}(X)$ is closed under union, intersection and subtraction. Moreover, $\mathcal{Q}(X) \subset \mathcal{A}_{P_1,P_2}(X)$. Thus $\mathcal{KQ}(X) \subset \mathcal{A}_{P_1,P_2}(X)$.

Pick an unbranched face $[O_i]$. By Lemma 4-16, Equation (4-34) and Remark 4-36, for every top-dimensional quasiflat Q with $[O] \subset [Q]$, there exists another top-dimensional orthant complex O' such that $[O'] \subset [Q]$ and $\partial_T O' \cap \partial_T O = \partial_T O_i$. This together with Lemma 2-10 (see also Remark 2-13) imply $[O] \cap [O'] = [O_i]$, thus $[O'] = [\rho_i(O)]$ and $[\rho_i(O)] \subset [Q]$ (recall that $[O_i]$ is unbranched). Similarly, one can prove if $[\rho_i(O)] \subset [Q]$ for a top-dimensional quasiflat Q, then $[O] \subset [Q]$. It follows from the above observation that $[\rho_i(O)] \subset [A]$ for $i \in \Lambda_u$.

Let $\gamma \in G_u$. Write $\gamma = \rho_{i_1}\rho_{i_2}\cdots\rho_{i_k}$ with $i_j \in \Lambda_u$ for $1 \le j \le n$. We will prove Claim 3 by induction on k. The case $k = 1$ is already done by the previous paragraph. In general, we assume $[\rho_{i_1}\rho_{i_2}\cdots\rho_{i_{k-1}}(O)] \subset [A]$. Note that $[O] \cap [\rho_{i_k}(O)] = [O_{i_k}]$, where $[O_{i_k}]$ is unbranched, so

$$[\rho_{i_1}\rho_{i_2}\cdots\rho_{i_{k-1}}(O)] \cap [\rho_{i_1}\rho_{i_2}\cdots\rho_{i_k}(O)] = [\rho_{i_1}\rho_{i_2}\cdots\rho_{i_{k-1}}(O_{i_k})].$$

Claim 1 implies $[\rho_{i_1}\rho_{i_2}\cdots\rho_{i_{k-1}}(O_{i_k})]$ is also unbranched, so $[\rho_{i_1}\rho_{i_2}\cdots\rho_{i_k}(O)] \subset [A]$ by the same argument as in the previous paragraph. □

Claim 2 and Claim 3 imply $\left[\bigcup_{\gamma\in G_u}\gamma(O)\right] = [A]$. So $\|[A]\| = |G_u|$, where $|G_u|$ is the order of G_u. Now the first assertion of the lemma follows. Moreover, for any $\gamma \in G$, let $[A_\gamma] \in \mathcal{KQ}(W)$ be the unique minimal essential element such that $[\gamma(O)] \subset [A_\gamma]$. Claim 1 implies $\{[\gamma(O_i)]\}_{i\in\Lambda_b}$ and $\{[\gamma(O_i)]\}_{i\in\Lambda_u}$ are the branched faces and unbranched faces of $[\gamma(O)]$ respectively. By the same argument as in Claim 2 and Claim 3, we can show $[A_\gamma] = \left[\bigcup_{\gamma'\in\gamma G_u}\gamma'(O)\right]$, where γG_u denotes the corresponding coset of G_u. Since there are $|G|/|G_u|$ cosets of G_u, the second assertion of the lemma also follows. □

Proof of Lemma 5-5 If $\|[A]\| = 2^n$, by Corollary 5-8 we know there exists a top-dimensional flat F such that $[A] \subset [F]$, so actually $[A] = [F]$. Then $f(A)$ is a top-dimensional quasiflat, thus $|f_\sharp([A])| \ge 2^n$. However, $f_\sharp([A])$ is also minimal essential, so by Corollary 5-8 we actually have $|f_\sharp([A])| = 2^n = \|[A]\|$. Let g be a quasi-isometry inverse of f. If $[A'] \in \mathcal{KQ}(W_2)$ is a minimal essential element, then by the same argument, we know that $\|[A']\| = 2^n$ implies $|g_\sharp([A'])| = 2^n = \|[A']\|$. So $\|[A]\| = 2^n$ if and only if $|f_\sharp([A])| = 2^n$ for minimal essential element $[A] \in \mathcal{KQ}(W_1)$.

In general, we assume inductively that $\|[A]\| = k$ if and only if $|f_\sharp([A])| = k$ for any $k \geq 2^{n-i+1}$ and any minimal essential element $[A] \in \mathcal{KQ}(W_1)$ (we are doing induction on i). If $[B_1] \in \mathcal{KQ}(W_1)$ is a minimal essential element with $\|[B_1]\| = 2^{n-i}$, then by Lemma 5-12, we can find a top-dimensional flat F and another $2^i - 1$ minimal essential elements $\{[B_j]\}_{j=2}^{2^i}$ such that $\|[B_j]\| = \|[B_1]\|$ and

$$[F] = \bigcup_{j=1}^{2^i} [B_j]. \tag{5-13}$$

Since $f(F)$ is a top-dimensional flat, we have

$$|f_\sharp(F)| = \left| f_\sharp\left(\bigcup_{j=1}^{2^i} [B_j] \right) \right| = \left| \bigcup_{j=1}^{2^i} f_\sharp([B_j]) \right| = \sum_{j=1}^{2^i} |f_\sharp([B_j])| \geq 2^n. \tag{5-14}$$

But our induction assumption implies

$$|f_\sharp([B_j])| < 2^{n-i+1}. \tag{5-15}$$

Since $f_\sharp([B_j])$ is minimal essential element for each j, Equation (5-15) together with assertion (1) of Lemma 5-12 imply

$$|f_\sharp([B_j])| \leq 2^{n-i}. \tag{5-16}$$

Now (5-14) and (5-16) imply

$$\|[B_j]\| = |f_\sharp([B_j])| = 2^{n-i} \tag{5-17}$$

for all j. By considering the quasi-isometry inverse, we know $\|[B]\| = 2^{n-i}$ if and only if $|f_\sharp([B])| = 2^{n-i}$ for minimal essential element $[B] \in \mathcal{KQ}(W_1)$. By Lemma 5-12(1) and our induction assumption, we have actually proved that $\|[B]\| = k$ if and only if $|f_\sharp([B])| = k$ for any $k \geq 2^{n-i}$ and any minimal essential element $[B] \in \mathcal{KQ}(W_1)$. □

5.3 Application to right-angled Coxeter groups and Artin groups

5.3.1 The right-angled Coxeter group case For a finite simplicial graph Γ with vertex set $\{v_i\}_{i \in I}$, there is an associated right-angled Coxeter group (RACG), denoted by $C(\Gamma)$, with the following presentation:

$$\langle \{v_i\}_{i \in I} \mid v_i^2 = 1 \text{ for all } i;\ [v_i, v_j] = 1 \text{ if } v_i \text{ and } v_j \text{ are joined by an edge} \rangle.$$

The group $C(\Gamma)$ has a nice geometric model $D(\Gamma)$, called the *Davis complex*. The 1–skeleton of $D(\Gamma)$ is the Cayley graph of $C(\Gamma)$ with edges corresponding to v_i, v_i^{-1} identified. For $n \geq 2$, the n–skeleton $D^{(n)}(\Gamma)$ of $D(\Gamma)$ is obtained from $D^{(n-1)}(\Gamma)$ by attaching an n–cube whenever one finds a copy of the $(n-1)$–skeleton of an n–cube inside $D^{(n-1)}(\Gamma)$. This process will terminate after finitely many steps and one obtains a CAT(0) cube complex where $C(\Gamma)$ acts properly and cocompactly.

The action of $C(\Gamma)$ on $D(\Gamma)$ is not free; however, $D(\Gamma)$ can be realized as the universal cover of a compact cube complex. The following construction is from [15]. Let $\{e_i\}_{i\in I}$ be the standard basis of $\mathbb{R}^I$ and let $\square^I = [0,1]^I \subset \mathbb{R}^I$ be the unit cube with the standard cubical structure. Let $F(\Gamma)$ be the flag complex of Γ. For each simplex $\Delta \subset F(\Gamma)$, let $\mathbb{R}^\Delta$ be the linear subspace spanned by $\{e_i\}_{v_i\in\Delta}$. Define

$$K(\Gamma) = \bigcup_{\Delta}\{\text{faces of } \square^I \text{ parallel to } \mathbb{R}^\Delta\},$$

where Δ varies among all simplices in $F(\Gamma)$. Then the Davis complex $D(\Gamma)$ is exactly the universal cover of $K(\Gamma)$; see [15, Proposition 3.2.3].

One can verify that $K(\Gamma)$ is weakly special. In order to apply Theorem 5-4 in a nontrivial way, we need the following extra condition:

$(*)$ There is an embedded top-dimensional hyperoctahedron in $F(\Gamma)$.

One can check there exists a top-dimensional flat in $D(\Gamma)$ if and only if $(*)$ is true.

Corollary 5-18 *Let Γ_1 and Γ_2 be two finite simplicial graphs satisfying $(*)$. If $\phi\colon D(\Gamma_1) \to D(\Gamma_2)$ is an (L,A)–quasi-isometry, then* $\dim(D(\Gamma_1)) = \dim(D(\Gamma_2))$. *Moreover there is a constant $D = D(L,A)$ such that for any top-dimensional flat F_1 in $D(\Gamma_1)$, we can find a flat F_2 in $D(\Gamma_2)$ such that*

$$d_H(\phi(F_1), F_2) < D.$$

Proof It suffices to show that $\dim(D(\Gamma_1)) = \dim(D(\Gamma_2))$. The rest follows from Theorem 5-4 and the above discussion.

We can assume the quasi-isometry ϕ is defined on the 0–skeleton of $D(\Gamma_1)$. Since $D(\Gamma_2)$ is CAT(0), we can extend ϕ skeleton by skeleton to obtain a continuous quasi-isometry. Similarly, we assume the quasi-isometry inverse ϕ' is also continuous. Since ϕ and ϕ' are proper, there are induced homomorphisms for the proper homology $\phi_*\colon H^{\mathrm{p}}_*(D(\Gamma_1)) \to H^{\mathrm{p}}_*(D(\Gamma_2))$ and $\phi'_*\colon H^{\mathrm{p}}_*(D(\Gamma_2)) \to H^{\mathrm{p}}_*(D(\Gamma_1))$; see Section 3.1. Note that the geodesic homotopy between $\phi' \circ \phi$ (or $\phi \circ \phi'$) and the identity map is proper, so $\phi_* \circ \phi'_* = \mathrm{Id}$ and $\phi'_* \circ \phi_* = \mathrm{Id}$. Hence ϕ_* is an isomorphism.

By symmetry, it suffices to show $\dim(D(\Gamma_1)) \geq \dim(D(\Gamma_2))$. If $\dim(D(\Gamma_1)) < \dim(D(\Gamma_2))$, then $H_n^{\mathrm{p}}(D(\Gamma_1))$ is trivial ($n = \dim(D(\Gamma_2))$) since there are no n–dimensional cells in $D(\Gamma_1)$. On the other hand, $(*)$ implies there is a top-dimensional flat in $D(\Gamma_2)$, thus $H_n^{\mathrm{p}}(D(\Gamma_2))$ is nontrivial, which yields a contradiction. □

5.3.2 The right-angled Artin group case Recall that for every simplicial graph Γ, there is a corresponding RAAG $G(\Gamma)$. Suppose $\bar{X}(\Gamma)$ is the Salvetti complex of $G(\Gamma)$. Then the 1–cells and 2–cells of $\bar{X}(\Gamma)$ are in 1–1 correspondence with the vertices and edges in Γ receptively. The closure of each k–cell in $\bar{X}(\Gamma)$ is a k–torus, which we call a *standard k–torus*. One can verify that the Salvetti complex $\bar{X}(\Gamma)$ is a weakly special cube complex.

We label the vertices of Γ by distinct letters (they correspond to the generators of $G(\Gamma)$), which induces a labeling of the edges of the Salvetti complex. We choose an orientation for each edge in the Salvetti complex and this gives us a directed labeling of the edges in $X(\Gamma)$. If we specify some base point $v \in X(\Gamma)$ (here v is a vertex), then there is a 1–1 correspondence between words in $G(\Gamma)$ and edge paths in $X(\Gamma)$ which start at v.

A subgraph $\Gamma' \subset \Gamma$ is a *full subgraph* if there does not exist an edge $e \subset \Gamma$ such that the two endpoints of e belong to Γ' but $e \not\subseteq \Gamma'$. In this case, there is an embedding $\bar{X}(\Gamma') \hookrightarrow \bar{X}(\Gamma)$ which is locally isometric. If $p\colon X(\Gamma) \to \bar{X}(\Gamma)$ is the universal cover, then each connected component of $p^{-1}(\bar{X}(\Gamma'))$ is a convex subcomplex isometric to $X(\Gamma')$. Following [8], we call these components *standard subcomplexes associated with* Γ'. Note that there is a 1–1 correspondence between standard subcomplexes associated with Γ' and left cosets of $G(\Gamma')$ in $G(\Gamma)$. A *standard k–flat* is the standard complex associated with a complete subgraph of k vertices. When $k = 1$, we also call it a *standard geodesic*.

Given a subcomplex $K \subset X(\Gamma)$, we denote the collection of labels of edges in K by $\mathrm{label}(K)$ and the corresponding collection of vertices in Γ by $V(K)$.

Let $V \subset \Gamma$ be a set of vertices. We define the *orthogonal complement* of V, denoted by $V^{\perp}$, to be the set $\{w \in \Gamma \mid d(w, v) = 1 \text{ for any } v \in V\}$.

The following theorem follows from Theorem 5-4.

Theorem 5-19 *Let Γ_1, Γ_2 be finite simplicial graphs, and let $\phi\colon X(\Gamma_1) \to X(\Gamma_2)$ be an (L, A)–quasi-isometry. Then $\dim(X(\Gamma_1)) = \dim(X(\Gamma_2))$. Moreover there is a constant $D = D(L, A)$ such that for any top-dimensional flat F_1 in $X(\Gamma_1)$, we can find a flat F_2 in $X(\Gamma_2)$ such that*

$$d_H(\phi(F_1), F_2) < D.$$

One can argue as in Corollary 5-18 or using the invariance of cohomological dimension to show $\dim(X(\Gamma_1)) = \dim(X(\Gamma_2))$.

Using Theorem 5-19, we can set up some immediate quasi-isometry invariants for RAAGs. Let $F(\Gamma)$ be the flag complex of Γ. We will assume $n = \dim(F(\Gamma))$ in the following discussion, then $\dim(X(\Gamma)) = n + 1$.

We construct a family of new graphs $\{\mathcal{G}_d(\Gamma)\}_{d=1}^{n}$, where the vertices of $\mathcal{G}_d(\Gamma)$ are in 1–1 correspondence with the top-dimensional flats in $X(\Gamma)$, and two vertices v_1 and v_2 are joined by an edge if and only if the associated flats F_1 and F_2 satisfy the condition that there exists an $r > 0$ such that $N_r(F_1) \cap N_r(F_2)$ contains a flat of dimension d. Let $\mathcal{G}^s_d(\Gamma)$ be the full subgraph of $\mathcal{G}_d(\Gamma)$ spanned by those vertices representing standard flats of top dimension.

Lemma 2-10 and Theorem 5-19 yield the following result.

Corollary 5-20 *Given a pair of finite simplicial graphs Γ_1, Γ_2 and a quasi-isometry $q\colon X(\Gamma_1) \to X(\Gamma_2)$, there is an induced graph isomorphism $q_*\colon \mathcal{G}_d(\Gamma_1) \to \mathcal{G}_d(\Gamma_2)$ for $1 \le d \le \dim(F(\Gamma_1)) = \dim(F(\Gamma_2))$.*

The relation between $\mathcal{G}_d(\Gamma)$ and Γ is complicated, but several basic properties of $\mathcal{G}_d(\Gamma)$ can be directly read from Γ. We first investigate the connectivity of $\mathcal{G}_d(\Gamma)$.

Lemma 5-21 *Suppose $1 \le d \le n$. Then $\mathcal{G}_d(\Gamma)$ is connected if and only if $\mathcal{G}^s_d(\Gamma)$ is connected.*

Proof For the $\Leftarrow$ direction, it suffices to show every point $v \in \mathcal{G}_d(\Gamma)$ is connected to some point in $\mathcal{G}^s_d(\Gamma)$. Let F_v be the associated top-dimensional flats. Pick a vertex $x \in F_v$ and suppose $\{e_i\}_{i=1}^{n+1}$ are mutually orthogonal edges in F_v emanating from x. Let $e_1^{\perp}$ be the subspace of F_v orthogonal to e_1 and let l_i be the unique standard geodesic such that $e_i \subset l_i$. Then by Lemma 5-6, the convex hull of l_1 and $e_1^{\perp}$ is a top-dimensional flat $F_{v,1}$. By construction, $F_{v,1}$ is adjacent to F_v in $\mathcal{G}_d(\Gamma)$. Now we can replace F_v by $F_{v,1}$, and run the same argument with respect to l_2. After finitely many steps, we will arrive at a standard flat F which is the convex hull of $\{l_i\}_{i=1}^{n+1}$, moreover F is connected to F_v in $\mathcal{G}_d(\Gamma)$. Note that F only depends on the choice of base vertex x and the frame $\{e_i\}_{i=1}^{n+1}$ at x. So we also denote F by $F = F_v(x, e_1, \dots, e_{n+1})$.

Now we prove the other direction. Pick a different base point $x' \in F_v$ and frame $\{e'_i\}_{i=1}^{n+1}$ at x'. We claim $F_v(x, e_1, \dots, e_{n+1})$ and $F_v(x', e'_1, \dots, e'_{n+1})$ are connected in $\mathcal{G}^s_d(\Gamma)$. Note that:

(1) If $x'=x$, $e'_i=e_i$ for $2\le i\le n+1$ and $e'_1=-e_1$, then F and F' are adjacent inside $\mathcal{G}^s_d(\Gamma)$.

(2) If x' is the other end point of e_1, $e'_1=\overrightarrow{x'x}$ and e'_i is parallel to e_i for $2\le i\le n+1$, then $F=F'$.

In general, we can connect x and x' by an edge path $\omega\subset F_v$ and use the previous two properties to induct on the combinatorial length of ω.

Let $\{F_i\}_{i=1}^m$ be a chain of top-dimensional flats representing an edge path in $\mathcal{G}_d(\Gamma)$ such that F_1 and F_m are standard flats. Pick i, let $(Y,Y')=\mathcal{I}(F_i,F_{i+1})$ and let $\phi\colon Y\to Y'$ be the isometry induced by CAT(0) projection as in Lemma 2-10(2). Since Y contains a d–dimensional flat, for vertex $x\in Y$ there are d mutually orthogonal edges $\{e_i\}_{i=1}^d$ such that $x\in e_i\subset Y$. Let $x'=\phi(x)$ and let $e'_i=\phi(e_i)$. We add more edges such that $\{e_i\}_{i=1}^{n+1}$ and $\{e'_i\}_{i=1}^{n+1}$ become bases for F_i and F_{i+1} respectively. Let $F_{i,i}=F_i(x,e_1,\dots,e_{n+1})$ and $F_{i+1,i}=F_{i+1}(x',e'_1,\dots,e'_{n+1})$. By Lemma 2-14, $F_{i,i}$ and $F_{i+1,i}$ are adjacent in $\mathcal{G}^s_d(\Gamma)$ for $1\le i\le m-1$. Moreover, for $2\le i\le m-1$, $F_{i,i}$ and $F_{i,i-1}$ are connected by a path inside $\mathcal{G}^s_d(\Gamma)$ by the previous claim. Thus F_1 and F_m are connected inside $\mathcal{G}^s_d(\Gamma)$. □

Recall that the notion of k–gallery is defined in Definition 1-5.

Lemma 5-22 *$\mathcal{G}^s_d(\Gamma)$ is connected if and only if Γ satisfies the following conditions:*

(1) *For any vertex $v\in F(\Gamma)$, there is a top-dimensional simplex $\Delta\subset F(\Gamma)$ such that $\Delta\cap v^\perp$ contains at least d vertices.*

(2) *Any two top-dimensional simplices Δ_1 and Δ_2 in $F(\Gamma)$ are connected by a $(d-1)$–gallery.*

Proof For the only if part, pick vertex $x\in X(\Gamma)$ and let $\Gamma_{d,x}$ be the full subgraph of $\mathcal{G}^s_d(\Gamma)$ generated by those vertices representing top-dimensional standard flats containing x. Then there is a canonical surjective simplicial map $\phi\colon \mathcal{G}^s_d(\Gamma)\to\Gamma_{d,x}$ by sending any top-dimensional standard flat F to the unique standard flat F' with $x\in F'$ and $\mathrm{label}(F)=\mathrm{label}(F')$. Since $\mathcal{G}^s_d(\Gamma)$ is connected, $\Gamma_{d,x}$ is also connected and (2) is true.

To see (1), suppose there exists a vertex $v\in F(\Gamma)$ such that for any top-dimensional simplex $\Delta\subset F(\Gamma)$, $\Delta\cap v^\perp$ contains less than d vertices. Pick a vertex $x_1\in X(\Gamma)$. If $e\subset X(\Gamma)$ is the unique edge such that $V(e)=v$ and $x_1\in e$, then by our assumption, e is not contained in any top-dimensional standard flat. This is also true for any edge parallel to e. Let h be the hyperplane dual to e. Suppose x_2 is the other endpoint of e.

For $i = 1, 2$, let F_i be the top-dimensional standard flat such that $x_i \in F_i$. Then F_1 and F_2 are separated by h. Since F_1 and F_2 are joined by a chain of top-dimensional standard flats such that each flat in the chain has trivial intersection with h (otherwise some edge parallel to e would be contained in a top-dimensional standard flat), we can find F'_1 and F'_2 in this chain which are adjacent in $\mathcal{G}^s_d(\Gamma)$ and separated by h. Let $(Y_1, Y_2) = \mathcal{I}(F'_1, F'_2)$. Then for a vertex $y \in Y_1$, there are d mutually orthogonal edges $\{e_i\}_{i=1}^d$ such that $y \in e_i \subset Y_1$. Let h_i be the hyperplane dual to e_i. Then $h_i \cap h \neq \varnothing$ for $1 \le i \le d$ by Lemma 2-14, hence in Γ we have $d(V(e_i), V(e)) = d(V(e_i), v) = 1$ for $1 \le i \le d$, which yields a contradiction.

For the other direction, note that (2) implies that $\Gamma_{d,x}$ is connected for any vertex $x \in X(\Gamma)$ and (1) implies that for adjacent vertices $x_1, x_2 \in X(\Gamma)$, there exist $v_i \in \Gamma_{d,x_i}$ for $i = 1, 2$ such that v_1 and v_2 are either adjacent or identical in $\mathcal{G}^s_d(\Gamma)$, thus $\mathcal{G}^s_d(\Gamma)$ is connected. □

The next result follows from Corollary 5-20, Lemma 5-21 and Lemma 5-22.

Theorem 5-23 *Given $G(\Gamma_1)$ and $G(\Gamma_2)$ which are quasi-isometric to each other, for $1 \le d \le \dim(F(\Gamma_1))$, the graph Γ_1 satisfies conditions (1) and (2) in Lemma 5-22 if and only if Γ_2 also satisfies these conditions.*

Now we turn to the diameter of $\mathcal{G}_1(\Gamma)$.

If Γ admits a nontrivial join decomposition $\Gamma = \Gamma_1 \circ \Gamma_2$, then $\operatorname{diam}(\mathcal{G}_1(\Gamma)) \le 2$. To see this, take two arbitrary top-dimensional flats F_1 and F_2 in $X(\Gamma)$, then $F_i = A_i \times B_i$, where A_i and B_i are top-dimensional flats in $X(\Gamma_1)$ and $X(\Gamma_2)$ respectively for $i = 1, 2$; see [30, Lemma 2.3.8]. Let $F = A_1 \times B_2$. Then $\operatorname{diam}(N_r(F_i) \cap N_r(F)) = \infty$ for some $r > 0$ and $i = 1, 2$, thus $\operatorname{diam}(\mathcal{G}_1(\Gamma)) \le 2$. Our next goal is to prove the following converse.

Lemma 5-24 *If $\operatorname{diam}(\mathcal{G}_1(\Gamma)) \le 2$ and if Γ is not one point, then Γ admits a nontrivial join decomposition $\Gamma = \Gamma_1 \circ \Gamma_2$.*

In the first part of the following proof, we will use the argument in [14, Section 4.1].

Proof Following [14, Section 4.2], let Γ^c be the complement graph of Γ. So Γ_c and Γ have the same vertex set, and two vertices are adjacent in Γ_c if and only if they are not adjacent in Γ. It suffices to show Γ_c is disconnected.

We argue by contradiction and suppose Γ_c is connected. Pick a top-dimensional simplex Δ in the flag complex $F(\Gamma)$ of Γ, where we identify Γ with the 1–skeleton of $F(\Gamma)$. Note that Δ corresponds to a top-dimensional standard torus T_Δ in the Salvetti complex. For any vertex $x \in X(\Gamma)$, we denote the unique standard flat in $X(\Gamma)$ which contains x and covers T_Δ by $F_{\Delta,x}$.

If Γ does not contain vertices other than those in Δ, then we are done; otherwise we can find a vertex $\widetilde{v} \in \Gamma$ with

(5-25) $$\widetilde{v} \notin \Delta.$$

Let ω be an edge path of Γ^c which starts at $\widetilde{v}$, ends at $\widetilde{v}$ and travels through every vertex in Γ^c. By recording the labels of consecutive vertices in ω, we obtain a word W. Let W' be the concatenation of eight copies of W.

Pick a vertex $x_1 \in X(\Gamma)$ and let l be the edge path which starts at x_1 and corresponds to the word W'. Let x_2 be the other endpoint of l. Note that l is actually a geodesic segment by our construction of W'. For $i = 1, 2$, let $F_i = F_{\Delta,x_i}$ and let w_i be the vertex in $\mathcal{G}_1(\Gamma)$ corresponding to F_i. We claim $d(w_1, w_2) > 2$.

If $d(w_1, w_2) \le 2$, then there exists a top-dimensional flat F such that

(5-26) $$\operatorname{diam}(N_r(F_i) \cap N_r(F)) = \infty$$

for some $r > 0$ and $i = 1, 2$. Let $(Y_1, Y) = \mathcal{I}(F_1, F)$. By (5-26) and Lemma 2-10, Y_1 is not a point (and neither is Y) and we can identify the convex hull of $Y \cup Y_1$ with $Y_1 \times [0, d(F_1, F)]$. Pick an edge $e_a \subset Y_1$ and let K_1 be the strip $e_a \times [0, d(F_1, F)]$ inside $Y_1 \times [0, d(F_1, F)]$. By considering the pair F and F_2, we can similarly find an edge $e_b \subset F_2$ and a strip K_2 isometric to $e_b \times [0, d(F_2, F)]$ which joins F and F_2. See Figure 4.

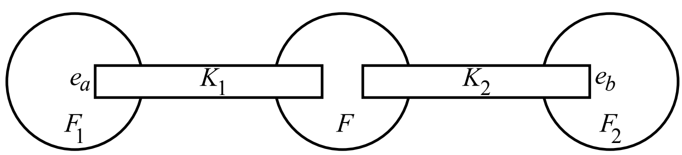


Figure 4

We parametrize the geodesic segment $l = \overline{x_1 x_2}$ by arc length such that $l(0) = v_1$. Assume $l(N) = x_2$. Let h_i be the hyperplane dual to the edge $\overline{l(i-1)l(i)}$ for $1 \le i \le N$. Note that

(5-27) $$h_j \text{ separates } h_i \text{ and } h_k \text{ for } i < j < k.$$

Moreover, each h_i separates F_1 and F_2 by (5-25), hence also separates e_a and e_b. Consider the set $K_1 \cup F \cup K_2$, which is connected and contains e_a and e_b, so each h_i must have nontrivial intersection with at least one of K_1, F and K_2.

We claim each of K_1, F and K_2 could intersect at most $2M$ hyperplanes from the collection $\{h_i\}_{i=1}^N$, where M is the length of word W (and $M > 1$ since Γ contains more than one vertex). This will yield a contradiction since $N = 8M$. We prove the claim for K_1; the case of K_2 is similar.

Let h_a be the hyperplane dual to e_a and let $\Lambda = \{1 \le i \le N \mid h_a \cap h_i \neq \varnothing\}$. Then

$$\{1 \le i \le N \mid K_1 \cap h_i \neq \varnothing\} \subset \Lambda. \tag{5-28}$$

If $h_a = h_{i_0}$ for some i_0, then by (5-27), h_{i_0} is the only hyperplane in $\{h_i\}_{i=1}^N$ intersecting K_1. Hence we are done in this case. Now we assume $h_a \notin \{h_i\}_{i=1}^N$. Let e_i be an edge dual to h_i. Then it follows from $h_a \cap h_i \neq \varnothing$ that for any $i \in \Lambda$,

$$d(V(e_i), V(e_a)) = 1 \tag{5-29}$$

in Γ. If the claim for K_1 is not true, then (5-28) implies Λ has cardinality bigger than $2M$; moreover, it follows from (5-27) that if $i \in \Lambda$ and $j \in \Lambda$, then $k \in \Lambda$ for any $i \le k \le j$. By the construction of the word W, we know every vertex of Γ is contained in the collection $\{V(e_i)\}_{i \in \Lambda}$, which contradicts (5-29).

Now we prove the claim for F. Suppose $F \cap h_i \neq \varnothing$. Then $V(e_i) \in \Delta$. By the construction of W', we know there exist positive integers $a, a' < M$ such that $d(V(e_{i+a}), V(e_i)) \ge 2$ and $d(V(e_{i-a'}), V(e_i)) \ge 2$ in Γ. Then $F \cap h_j = \varnothing$ for $j = i + a$ and $j = i - a'$. By (5-27), $F \cap h_j = \varnothing$ for $j > i + a$ and $j < i - a'$. Thus the claim is true for F. □

Theorem 5-30 *The following are equivalent:*

1. $\operatorname{diam}(\mathcal{G}_1(\Gamma)) < \infty$.
2. $\operatorname{diam}(\mathcal{G}_1(\Gamma)) \le 2$.
3. Γ *admits a nontrivial join decomposition or* Γ *is one point.*

Moreover, these properties are quasi-isometry invariants for right-angled Artin groups.

Note that (1) $\Rightarrow$ (3) follows by considering the concatenation of arbitrarily many copies of W in Lemma 5-24 and applying the same argument.

Remark 5-31 It is shown in [2] and [1] that $G(\Gamma)$ has linear divergence if and only if Γ is a nontrivial join, which also implies that Γ being a nontrivial join is quasi-isometry invariant. Moreover, their results together with [26, Theorem B and Proposition 4.7] implies the following stronger statement.

Given $X = X(\Gamma)$ and $X' = X(\Gamma')$, let $\Gamma = \Delta \circ \Gamma_1 \circ \cdots \circ \Gamma_k$ be the join decomposition such that Δ is the maximal clique factor, and each Γ_i does not allow nontrivial further join decomposition. Similarly, let $\Gamma = \Delta' \circ \Gamma'_1 \circ \cdots \circ \Gamma'_{k'}$. Let $X = \mathbb{R}^n \times \prod_{i=1}^k X(\Gamma_i)$ and let $X' = \mathbb{R}^{n'} \times \prod_{j=1}^{k'} X(\Gamma'_j)$ be the corresponding product decomposition. If $\phi\colon X \to X'$ is an (L, A) quasi-isometry, then $n = n'$, $k = k'$ and there exist constants $L' = L'(L, A)$, $A' = A'(L, A)$, $D = D(L, A)$ such that after re-indexing the factors in X', we have an (L', A') quasi-isometry $\phi_i\colon X(\Gamma_i) \to X(\Gamma'_i)$ so that

$$d\left(p' \circ \phi, \prod_{i=1}^{k} \phi_i \circ p\right) < D,$$

where $p\colon X \to \prod_{i=1}^k X(\Gamma_i)$ and $p'\colon X' \to \prod_{i=1}^k X(\Gamma'_i)$ are the projections.

More generally, let X and X' be locally compact CAT(0) cube complexes which admit a cocompact and essential action. Let $X = \prod_{j=1}^n Z_j \times \prod_{i=1}^k X_i$ be the finest product decomposition of X, where the Z_j are exactly the factors which are quasi-isometric to $\mathbb{R}$. Suppose $Z = \prod_{j=1}^n Z_j$. Then $X = Z \times \prod_{i=1}^k X_i$. Similarly, we decompose X' as $X' = Z' \times \prod_{i=1}^{k'} X'_i$. Then any quasi-isometry between X and X' respects such product decompositions in the sense of the previous paragraph. This is a consequence of [26, Theorem B], [26, Proposition 4.7] and [12, Theorem 6.3].

Appendix: Top-dimensional support sets in spaces of finite geometric dimension, and application to Euclidean buildings

In this section we adjust previous arguments to study the structure of top-dimensional quasiflats in Euclidean buildings and prove the following result.

Theorem A-1 *If Y is a Euclidean building of rank n and $[\sigma] \in H^{\mathrm{p}}_{n,n}(Y)$, then there exist finitely many Weyl cones $\{W_i\}_{i=1}^h$ such that*

$$d_H\left(S_{[\sigma]}, \bigcup_{i=1}^{h} W_i\right) < \infty.$$

Moreover, we can assume $W_i \subset S_{[\sigma]}$ for all i.

For the case of discrete Euclidean buildings, our previous method goes through without much modification. One way to handle the nondiscrete case is to use [30, Lemma 6.2.2], which says the support set of a top-dimensional class locally looks like a polyhedral complex, to reduce to the discrete case. But this lemma relies on the local structure of Euclidean buildings. We introduce another way, based on the generalization of results in Section 3.2 to CAT(0) spaces of finite geometric dimension (or homological dimension), which is of independent interest.

Lemma A-2 *Lemma 3-4 is true under the assumption that Y is a* CAT(0) *space of homological dimension $\leq n$.*

Proof Let $\phi\colon \mathbb{E}^n \to Y$ be a top-dimensional (L,A)–quasiflat. We can assume ϕ is Lipschitz as before since Y is CAT(0). Let $[\mathbb{E}^n]$ be the fundamental class of $\mathbb{E}^n$. Pick $\sigma = \phi_*([\mathbb{E}^n])$ and let $S = S_{[\sigma]}$ be the support set. Pick $\epsilon > 0$, suppose U is the 1–neighborhood of $\operatorname{Im}\phi$ and suppose $\{U_\lambda\}_{\lambda\in\Lambda}$ is a covering of U, where each U_λ is an open subset of U with diameter < 1. Since every metric space is paracompact, we can assume this covering is locally finite and define a continuous map $\varphi\colon U \to \mathbb{E}^n$ via the nerve complex of this covering as in Lemma 4-16, such that there exists a constant C such that

$$d(\varphi\circ\phi(x), x) < C \tag{A-3}$$

for any $x \in \mathbb{E}^n$, thus $\varphi_*([\sigma]) = [\mathbb{E}^n]$. Then we have $S_{[\mathbb{E}^n]} = \mathbb{E}^n \subset \varphi(S_{[\sigma]})$ by Lemma 3-2. It follows that $d_H(S, \operatorname{Im}\phi) < D = D(L,A)$. □

Remark A-4 In the above proof, we need to define φ in an open neighborhood of $\operatorname{Im}\phi$ since $S_{[\sigma],\operatorname{Im}\phi}$ might be strictly smaller than $S_{[\sigma],Y}$. Also we do not need to bound the dimension of the nerve complex of $\{U_\lambda\}_{\lambda\in\Lambda}$ as in Lemma 4-16 since Y is CAT(0), while in Lemma 4-16, the target space CK is linearly contractible with the contractibility constant possibly greater than 1.

Recall that in a polyhedral complex, every top-dimensional homology class can be represented by a cycle with image inside the support of the homology class. However, we do not know if this is still true in the case of an arbitrary metric space of homological dimension n. The following result helps us to get around this point.

Lemma A-5 *Let Y be a metric space of homological dimension $\leq n$ and $[\sigma]\in H_n^{\mathrm{p}}(Y)$. If O is an open neighborhood of $S_{[\sigma]}$, then there exists a proper cycle σ' such that $[\sigma]=[\sigma']$ and $\operatorname{Im}\sigma' \subset O$.*

Proof We first prove a relative version of the above lemma for the usual homology theory. Let $V \subset U$ be open sets in Y. Pick $[\alpha]\in H_n(U,V)$ and let $K = S_{[\alpha]}$. We claim for any open neighborhood $O \supseteq K$, there exist chains β and γ such that $\operatorname{Im}\beta \subset U$, $\operatorname{Im}\gamma \subset V\cup O$ and $\alpha = \partial\beta + \gamma$.

Let $K' = \operatorname{Im}\alpha \setminus (V\cup O)$. For every point $x \in K'$, there exists $\epsilon(x) > 0$ such that $\overline{B}(x, 2\epsilon(x)) \subset U \setminus \operatorname{Im}\ \partial\alpha$ and $[\alpha]$ is trivial in $H_n(U, U\setminus \overline{B}(x,2\epsilon(x)))$. Since K' is compact, we can find finitely many points $\{x_i\}_{i=1}^N$ in K' such that $K' \subset \bigcup_{i=1}^N B(x_i,\epsilon(x_i))$. Suppose $U_{K'} = \bigcup_{i=1}^N B(x_i,\epsilon(x_i))$ and $W = V\cup O\cup U_{K'}$.

Then $\operatorname{Im}\alpha \subset W$ and $[\alpha]$ is trivial in $H_n(U, W)$. Let $W' = W \setminus \bigl(\bigcup_{i=1}^{N} \bar{B}(x_i, 2\epsilon(x_i))\bigr)$. Then $\operatorname{Im}\partial\alpha \subset W' \subset V \cup O$, so it suffices to show $[\alpha]$ is trivial in $H_n(U, W')$, but this follows from the Mayer–Vietoris argument in Lemma 3-2.

Now we turn to the case of $[\sigma] \in H_n^{\mathrm{p}}(Y)$. Pick a base point $p \in Y$, put

$$\begin{aligned} &U_1 := B(p,4), && U_i := B(p, 3^i + 1) \setminus \bar{B}(p, 3^{i-1} - 1) \quad \text{for } i > 1; \\ &U_1' := Y \setminus \bar{B}(p,3), && U_i' := B(p, 3^{i-1}) \cup (Y \setminus \bar{B}(p, 3^i)) \quad \text{for } i > 1. \end{aligned}$$

By barycentric subdivision, we can assume every singular simplex in σ has image of diameter $\le \frac{1}{3}$.

Set $\sigma_0 = \sigma$, $V_0 = Y$ and $V_i = Y \setminus \bar{B}(p, 3^i)$ for $i \ge 1$. Given σ_i with $\operatorname{Im}\sigma_i \subset (O \cup V_i)$ (this is trivially true for $i = 0$), we define σ_{i+1} inductively as follows. First subdivide σ_i to get a proper cycle σ_{i+1}' such that

- $\operatorname{Im}\sigma_i = \operatorname{Im}\sigma_{i+1}'$,
- $\sigma_{i+1}' = \sigma_i + \partial\beta_{i+1}$ with $\operatorname{Im}\beta_{i+1} \subset U_{i+1}$, and
- $\sigma_{i+1}' = \sigma_{i+1,1}' + \sigma_{i+1,2}'$ for $\operatorname{Im}\ \sigma_{i+1,1}' \subset U_{i+1}$ and $\operatorname{Im}\sigma_{i+1,2}' \subset U_{i+1}'$.

It follows that $\operatorname{Im}\partial\sigma_{i+1,1}' \subset U_{i+1} \cap U_{i+1}'$ and $\operatorname{Im}\partial\sigma_{i+1,1}' \subset \operatorname{Im}\sigma_i \subset (O \cup V_i)$. So we can view $[\sigma_{i+1,1}']$ as an element in $H_n(U_{i+1}, U_{i+1} \cap U_{i+1}' \cap (O \cup V_i))$. Then by the previous claim, there exists a chain β_{i+1}' with $\operatorname{Im}\beta_{i+1}' \subset U_{i+1}$ such that $\operatorname{Im}(\sigma_{i+1,1}' + \partial\beta_{i+1}') \subset F$, where

$$\begin{aligned} F = (U_{i+1} \cap U_{i+1}' \cap (O \cup V_i)) \cup (O \cap U_{i+1}) &= (U_{i+1} \cap U_{i+1}' \cap V_i) \cup (O \cap U_{i+1}) \\ &= (U_{i+1} \cap V_{i+1}) \cup (O \cap U_{i+1}) \\ &= (O \cup V_{i+1}) \cap U_{i+1}. \end{aligned}$$

Let $\sigma_{i+1} = \sigma_i + \partial(\beta_{i+1} + \beta_{i+1}')$. Then

$$\operatorname{Im}\sigma_{i+1} \subset (F \cup \operatorname{Im}\ \sigma_{i+1,2}') \subset (F \cup (O \cup V_i)) \subset (O \cup V_{i+1})$$

and the induction goes through.

Let $\sigma' = \sigma + \sum_{i=1}^{\infty} \partial(\beta_i + \beta_i')$. Since $\operatorname{Im}(\beta_i + \beta_i') \subset U_i$, the infinite summation makes sense and σ' is a proper cycle. Also $\operatorname{Im}\sigma' \subset O$ by construction. □

Remark A-6 The above proof also shows that $[\sigma] \in H_n^{\mathrm{p}}(Y)$ is nontrivial if and only if $S_{[\sigma]} \neq \varnothing$. This is not true for lower-dimensional cycles.

Corollary A-7 *Let Z be a* CAT(1) *space of homological dimension n. If $[\sigma] \in H_n(Z)$ is a nontrivial element, then for every point $x \in Z$, there exists a point $y \in S_{[\sigma]}$ such that $d(x, y) = \pi$.*

Proof We argue by contradiction and assume there exists a point $x \in Z$ such that $S_{[\sigma]} \subset B(x,\pi)$. Then by Lemma A-5, there exists a cycle σ' such that $\operatorname{Im}\sigma' \subset B(x,\pi)$ and $[\sigma'] = [\sigma]$. However, $B(x,\pi)$ is contractible and $[\sigma']$ must be trivial, which yields a contradiction. □

Let Y be a CAT(0) space. Pick $p \in Y$ and let T_pY be the tangent cone at p. Denote the base point of T_pY by o. Recall that there are logarithmic maps $\log_p\colon Y \to T_pY$ and $\log_p\colon Y \setminus \{p\} \to \Sigma_p Y$. By [31, Theorem 3.5] (see also [33]), $\log_p\colon Y \setminus \{p\} \to \Sigma_p Y$ is a homotopy equivalence. Thus we get:

Lemma A-8 *The map $(\log_p)_*\colon H_*(Y, Y \setminus \{p\}) \to H_*(T_pY, T_pY \setminus \{o\})$ is an isomorphism.*

We need a simple observation about support sets in cones before we proceed. Let Z be a metric space and let CZ be the Euclidean cone over Z with base point o. Pick a $[\sigma] \in H_i(CZ, CZ \setminus B(o,r))$. Recall that there is an isomorphism

$$\partial\colon H_i(CZ, CZ \setminus B(o,r)) \to H_{i-1}(Z)$$

induced by the boundary map.

Lemma A-9 *Suppose $S = S_{\partial[\sigma],Z}$ and suppose CS is the cone over S inside CZ. Then $S_{[\sigma],CZ,CZ\setminus B(o,r)} = CS \cap B(o,r)$.*

The next lemma is an immediate consequence of [28, Theorem A].

Lemma A-10 *If Z is a CAT(κ) space of homological dimension $\leq n$, then for any $p \in Z$, $\Sigma_p Z$ is of homological dimension $\leq n-1$.*

Now we are ready to prove the geodesic extension property for top-dimensional support sets. The argument is similar to [9, Lemma 3.1].

Lemma A-11 *Let Y be a CAT(0) space of homological dimension n. Pick an element $[\sigma] \in H_n^{\mathrm{p}}(Y)$ and let $S = S_{[\sigma]}$. Then for a geodesic segment $\overline{pq} \subset Y$ with $q \in S$, there exists a geodesic ray $\overline{q\xi} \subset S$ such that $\overline{pq}$ and $\overline{q\xi}$ fit together to form a geodesic ray.*

Proof First we claim that for any $\epsilon > 0$, there exists a point $z \in S \cap S(p,\epsilon)$ such that the concatenation of $\overline{pq}$ and $\overline{qz}$ is a geodesic. Let

$$\log_q\colon (Y, Y \setminus B(q,2\epsilon)) \to (T_qY, T_qY \setminus B(o,2\epsilon)),$$

and let $\alpha = \log_q(\sigma)$. By Lemma A-10, the homological dimension of T_pY is bounded above by n. Then by Lemma 3-2,

$$S_{[\alpha],T_qY,T_qY\setminus B(o,2\epsilon)} \subset f(S_{[\sigma],Y,Y\setminus B(q,2\epsilon)}) = f(S\cap B(q,2\epsilon)). \tag{A-12}$$

Let ∂: $H_n(T_qY, T_qY\setminus B(o,2\epsilon)) \to H_{n-1}(\Sigma_qY)$ be the isomorphism and let $[\beta]=\partial[\alpha]$. Then $[\beta]$ is nontrivial in $H_{n-1}(\Sigma_qY)$ by Lemma A-8 and this commuting diagram:

$$\begin{array}{ccc} H_n(Y, Y\setminus B(q,2\epsilon)) & \xrightarrow{(\log_q)_*} & H_n(T_qY, T_qY\setminus B(o,2\epsilon)) \\ \downarrow & & \downarrow \\ H_n(Y, Y\setminus\{q\}) & \xrightarrow{(\log_q)_*} & H_n(T_qY, T_qY\setminus\{o\}) \end{array}$$

Let $CS_{[\beta]}$ be the Euclidean cone over $S_{[\beta]}\subset \Sigma_qY$ inside T_qY. Then

$$CS_{[\beta]}\cap B(o,2\epsilon) = S_{[\alpha],T_qY,T_qY\setminus B(o,2\epsilon)} \subset f(S\cap B(q,2\epsilon)) \tag{A-13}$$

by (A-12) and Lemma A-9. Moreover, by Corollary A-7 and Lemma A-10, there exists an $x\in S_{[\beta]}$ such that

$$d(x,\log_q(p)) = \pi, \tag{A-14}$$

where $\log_q$: $Y\setminus\{q\}\to\Sigma_qY$. So the claim follows from (A-13).

By repeatedly applying the above claim, for each positive integer n we can obtain a unit-speed geodesic c_n: $[0,\epsilon]\to Y$ such that $c(0)=q$, $c(m\epsilon/2^n)\in S$ for any integer $0\le m\le 2^n$ and $\log_q(c(\epsilon))=x$, where x is the point in (A-14). Note that $S\cap\overline{B}(q,\epsilon)$ is compact, so we assume without loss of generality that $r=\lim_{n\to\infty}c_n(\epsilon)$. If c: $[0,\epsilon]\to Y$ is the unit-speed geodesic joining q and r, then c_n converges uniformly to c, which implies $c([0,\epsilon])\subset S$. Moreover, $\log_q(c(\epsilon))=x$. Thus the concatenation of $\overline{pq}$ and $\overline{qr}$ is a geodesic by (A-14). Now we can repeatedly apply this ϵ–extension procedure to obtain the geodesic ray as required. □

In general, the above set $S\cap\overline{B}(q,\epsilon)$ is not equal to the geodesic cone $C_q(S\cap S(q,\epsilon))$ based at q over $S\cap S(q,\epsilon)$ no matter how small ϵ is. However, we have

$$\begin{aligned}\lim_{\epsilon\to 0} d_{\mathrm{GH}}\big(\tfrac{1}{\epsilon}(C_q(S\cap S(q,\epsilon))), CS_{[\beta]}\cap\overline{B}(o,1)\big)\\ = \lim_{\epsilon\to 0} d_{\mathrm{GH}}\big(\tfrac{1}{\epsilon}(S\cap\overline{B}(q,\epsilon)), CS_{[\beta]}\cap\overline{B}(o,1)\big) = 0.\end{aligned}$$

Thus the tangent cone of S exists for every point in S.

Remark A-15 By the same proof, we know Lemma A-11 is still true if Y is an Alexandrov space which has curvature bounded above and homological dimension $=n$. In this case, $\overline{p\xi}$ is locally geodesic.

Lemma A-16 *Let Z be a* CAT(1) *space of homological dimension $\leq n$ and let $[\sigma] \in \widetilde{H}_n(Z)$ be a nontrivial class. Then the following assertions hold.*

(1) $\mathcal{H}^n(S_{[\sigma]}) \geq \mathcal{H}^n(\mathbb{S}^n)$.

(2) *Let $V(n,r)$ be the volume of r–ball in $\mathbb{S}^n$. Then for any $0 \leq r \leq R < \pi$ and any $p \in S_{[\sigma]}$,*

$$1 \leq \frac{\mathcal{H}^n(B(p,r) \cap S_{[\sigma]})}{V(n,r)} \leq \frac{\mathcal{H}^n(B(p,R) \cap S_{[\sigma]})}{V(n,R)}.$$

(3) *If $\mathcal{H}^n(S_{[\sigma]}) = \mathcal{H}^n(\mathbb{S}^n)$, then $S_{[\sigma]}$ is an isometrically embedded copy of $\mathbb{S}^n$.*

Here $\mathbb{S}^n$ denotes the n–dimensional standard sphere with constant curvature 1.

Proof We claim there exists a 1–Lipschitz map from a subset of $S_{[\sigma]}$ to a full-measure subset of $\mathbb{S}^n$. Let us assume this is true for $i = n-1$. Pick $p \in S_{[\sigma]}$, let $\mathbb{S}^0 * \Sigma_p Z$ be the spherical suspension of $\Sigma_p Z$ and let o be one of the suspension points. Then there is a well-defined 1–Lipschitz map $\log_p\colon B(p,\pi) \to \mathbb{S}^0 * \Sigma_p Z$ sending p to o. Let $[\beta]$ be the image of $[\sigma]$ under the map

$$\widetilde{H}_n(Z) \to H_n(Z, Z \setminus \{p\}) \to H_n(B(p,\pi), B(p,\pi) \setminus \{p\}) \xrightarrow{(\log_p)_*} H_n(B(o,\pi), B(o,\pi) \setminus \{o\}) \to \widetilde{H}_{n-1}(\Sigma_p Z).$$

We can slightly adjust the proof of Lemma A-11 to show that

$$\log_p(S_{[\sigma]} \cap B(p,\pi)) \supseteq (\mathbb{S}^0 * S_{[\beta]}) \cap B(o,\pi). \tag{A-17}$$

The induction assumption implies that there are a subset $K \in S_{[\beta]}$ and a 1–Lipschitz map $f\colon K \to \mathbb{S}^{n-1}$ such that $\mathcal{H}^{n-1}(\mathbb{S}^{n-1} \setminus f(K)) = 0$. Note that f induces a 1–Lipschitz map $\tilde{f}\colon \mathbb{S}^0 * K \to \mathbb{S}^0 * \mathbb{S}^{n-1} = \mathbb{S}^n$ whose image also has full measure, thus by (A-17), there exists $K' \subset S_{[\sigma]}$ such that the image of $\tilde{f} \circ \log_p\colon K' \to \mathbb{S}^n$ has full measure. It follows that $\mathcal{H}^n(S_{[\sigma]}) \geq \mathcal{H}^n(\mathbb{S}^n)$.

The first inequality of (2) follows from (1) and (A-17). The second inequality follows from Remark A-15 and the proof of [9, Corollary 3.3].

Now we prove (3). By Remark A-15, for every point $x \in S_{[\beta]}$, there exists a geodesic segment $l_x \subset S_{[\sigma]}$ emanating from p along the direction x, which has length $= \pi$. Let A be the closure of $\bigcup_{x \in S_{[\beta]}} l_x$. Then $A \subset S_{[\sigma]}$ and $\mathcal{H}^n(A) \geq \mathcal{H}^n(\mathbb{S}^n)$. Then (2) implies that actually $A = S_{[\sigma]}$. Pick arbitrary $q \in S_{[\sigma]} \cap B(p,\pi)$. Then there exists a sequence $\{q_n\}_{n=1}^{\infty} \subset \bigcup_{x \in S_{[\beta]}} l_x$ such that $\lim_{n\to\infty} q_n = q$. Since $\overline{q_n p} \subset S_{[\sigma]}$ by construction, $\overline{qp} \subset S_{[\sigma]}$. It follows that $S_{[\sigma]}$ is π–convex in Y. Then $S_{[\sigma]}$ can be viewed as a compact and geodesically complete CAT(1) space. By [35, Proposition 7.1], $S_{[\sigma]}$ is isometric to $\mathbb{S}^n$. □

We can recover the monotonicity (3-11) and the lower density bound (3-12) from Lemma A-11 and Lemma A-16, then we can define the group $H^{\mathrm{p}}_{n,n}(Y)$ as before when Y is a CAT(0) space of homological dimension $\leq n$ and the rest of the discussion in Section 3.2 goes through without any change. Recall that the homological dimension of a CAT(0) space is equal to its geometric dimension [28, Theorem A], so the following result holds.

Theorem A-18 *Let Y be a* CAT(0) *space of geometric dimension n. Pick an element $[\sigma] \in H^{\mathrm{p}}_{n,n}(Y)$ and let $S = S_{[\sigma]}$. Then:*

(1) **Local property I** *Each point $p \in Y$ has a well-defined tangent cone T_pY.*

(2) **Local property II** *S has the geodesic extension property in the sense of Lemma 3-6.*

(3) **Monotonicity and lower density bound** *For all $0 \leq r \leq R$ and $p \in Y$,*
$$\frac{\mathcal{H}^n(B(p,r)\cap S)}{r^n} \leq \frac{\mathcal{H}^n(B(p,R)\cap S)}{R^n}.$$
If $p \in S$, then
$$\mathcal{H}^n(B(p,r)\cap S) \geq \omega_n r^n,$$
with equality only if $B(p,r)\cap S$ is isometric to an r–ball in $\mathbb{E}^n$. Here ω_n is the volume of an n–dimensional Euclidean ball of radius 1.

(4) **Asymptotic conicality I** *Let $B(o,1)$ be the unit ball in C_TS centered at the cone point o. For any $p \in Y$,*
$$\lim_{r\to+\infty} d_{\mathrm{GH}}\big(\tfrac{1}{r}(B(p,r)\cap S), B(o,1)\big) = 0.$$
Moreover, putting $\partial_{p,r}S := \{\xi \in \partial_T S \mid \overline{p\xi} \subset B(p,r)\cup S\}$, then
$$\lim_{r\to+\infty} d_H(\partial_{p,r}S, \partial_T S) = 0.$$

(5) **Asymptotic conicality II** *For all $\beta > 0$ there is an $r < \infty$ such that if $x \in S \setminus B(p,r)$, then*
$$\mathrm{diam}(\mathrm{Ant}_\infty(\log_x p, S)) < \beta,$$
where the diameter is with respect to the angular metric on $\partial_T Y$.

Now we reinterpret the group $H^{\mathrm{p}}_{n,n}(Y)$. Recall that there is another logarithmic map $\log_p\colon C_TY \to Y$ sending the base point o of C_TY to $p \in Y$. Since $\log_p$ is proper and 1–Lipschitz, it induces a map $H^{\mathrm{p}}_{n,n}(C_TY) \to H^{\mathrm{p}}_{n,n}(Y)$.

Next we define a map in the other direction. Pick $[\sigma] \in H^{\mathrm{p}}_{n,n}(Y)$, let $S = S_{[\sigma]}$ be the support set and let U_S be the 1–neighborhood of S. By Lemma A-5, we can assume $\operatorname{Im}\sigma \subset U_S$. For $\epsilon > 0$, we define the map $f_\epsilon\colon U_S \to C_T S$ as in Lemma 4-16. To approximate f_ϵ by a continuous map, we choose a locally finite covering of U_S by its open subsets which satisfies the diameter condition in Lemma 4-16, then proceed as before to obtain a continuous map $f_\epsilon\colon U_S \to C_T X$. Here the image may not stay inside $C_T S$, however it is sublinearly close to $C_T S$. Now we define $\exp_*([\sigma]) = \lim_{\epsilon\to 0} f_{\epsilon *}([\sigma])$; note that $f_{\epsilon *}([\sigma])$ does not depend on ϵ when it is small. Since (4-23) is still true, $(\log_p)_* \circ \exp_* = \mathrm{Id}$.

To see that $\exp_* \circ (\log_p)_* = \mathrm{Id}$, we follow the proof of (4-30); the only difference is that we need to replace $I_{\sigma'}$ there by the 1–neighborhood of $\operatorname{Im}\sigma'$, then use the nerve complex of a suitable covering to approximate g_ϵ as we did for f_ϵ. So

$$(\log_p)_*\colon H^{\mathrm{p}}_{n,n}(C_T Y) \to H^{\mathrm{p}}_{n,n}(Y)$$

is an isomorphism, with the inverse map $\exp_*$ defined as above.

Let $h_\lambda\colon C_T Y \to C_T Y$ be the homothety map with respect to the base point o and a factor λ. Then h_λ is properly homotopic to h_1 for any $0 < \lambda < \infty$, so for any $[\beta] \in H^{\mathrm{p}}_i(C_T Y)$, we have $h_{\lambda *}([\beta]) = [\beta]$ and $h_\lambda(S_{[\beta]}) = S_{[\beta]}$. It follows that every cycle in $H^{\mathrm{p}}_i(C_T Y)$ is conical. Thus the map

$$j\colon H^{\mathrm{p}}_i(C_T Y) \to H_i(C_T Y, C_T Y \setminus \{o\}) \to H_{i-1}(\partial_T Y)$$

is an isomorphism with inverse given by the "coning off" procedure. It follows that the map defined in (4-27) and (4-28) are isomorphisms, and the analogues of Corollary 4-33 and Remark 4-36 in the case of CAT(0) spaces with finite homological dimension are still true (again, for our argument to go through, we need to replace the set I_q in the proof of Corollary 4-33 by some r–neighborhood of the image of q). This discussion can be summarized as follows.

Theorem A-19 *Let $q\colon Y \to Y'$ be a quasi-isometric embedding, where Y and Y' are* CAT(0) *spaces of geometric dimension $\leq n$. Then:*

1. *The map $\partial := j \circ (\exp_*)\colon H^{\mathrm{p}}_{n,n}(Y) \to H_{n-1}(\partial_T Y)$ is a group isomorphism, with inverse given by the coning off map $c\colon H_{n-1}(\partial_T Y) \to H^{\mathrm{p}}_{n,n}(Y)$; see (4-28).*
2. *The map q induces a monomorphism $q_*\colon H_{n-1}(\partial_T Y) \to H_{n-1}(\partial_T Y')$. If q is a quasi-isometry, then q_* is an isomorphism.*
3. *There exists a $D' > 0$, depending on the quasi-isometry constants of q, such that*

$$d_H(q(S_{[\tilde\sigma]}), S_{q_*[\tilde\sigma]}) < D'$$

for any $[\tilde\sigma] \in H^{\mathrm{p}}_{n,n}(Y)$.

We refer to the work of Kleiner and Lang [29] for a more general version of the above theorem.

Remark A-20 Pick $[\tau] \in H_{n-1}(\partial_T Y)$. Then by Lemma A-8 and Theorem A-19,

$$S_{c([\tau])} = \{y \in Y \mid [\tau] \text{ is nontrivial under } (\log_y)_*\colon H_{n-1}(\partial_T Y) \to H_{n-1}(\Sigma_y Y)\},$$

where c is the coning off map in (4-28).

Now we are ready to prove Theorem A-1. To avoid repetition, we will only sketch the main steps.

Proof of Theorem A-1 If Y is a Euclidean building of rank n, it follows from [30, Corollary 6.1.1] that the homological dimension of Y is less than or equal to n. This also follows from [28, Theorem A] by noticing that $\Sigma_p Y$ is a spherical building of dimension $n-1$ for any $p \in Y$. Let $[\sigma] \in H^{\mathrm{p}}_{n,n}(Z)$.

Step 1 Let $[\alpha] = \exp_*([\sigma]) \in H^{\mathrm{p}}_{n,n}(C_T Y)$. Since $\partial_T Y$ is a spherical building, $S_{[\alpha]}$ is a cone over K, where $K = \bigcup_{i=1}^h C_i$ and each C_i is a chamber in $\partial_T Y$.

Step 2 Let $W_i \subset Y$ be a Weyl cone such that $\partial_T W_i = C_i$. Note that for any $i \neq j$, there is an apartment of $\partial_T Y$ which contains C_i and C_j. Thus we can assume W_i and W_j are contained in a common apartment of Y. So W_i and W_j satisfy inequalities similar to (2-11). The quotient map $\bigsqcup_{i=1}^h C_i \to K$ induces a map $\varphi\colon CK \to Y$ which is a quasi-isometric embedding as in Lemma 4-38. We can assume φ is continuous. Put $[\tau] = \varphi_*([CK])$, where $[CK]$ is the fundamental class of CK. Then it follows from the proof of Lemma 4-46 that $d_H\big(S_{[\tau]}, \bigcup_{i=1}^h W_i\big) < \infty$. Moreover, we can assume that $W_i \subset S_{[\tau]}$.

Step 3 It suffices to show that $[\sigma] = [\tau]$. Pick $p \in Y$. Note that there exists a $D > 0$ such that $d(\log_p(x), \varphi(x)) < D$ for any $x \in CK$. Then

$$[\tau] = \varphi_*([CK]) = (\log_p)_*([\alpha]) = ((\log_p)_* \circ \exp_*)([\sigma]) = [\sigma]. \qquad \square$$

The following result is an immediate consequence of Lemma A-2 and Theorem A-1.

Corollary A-21 *If Y is a Euclidean building of rank n and $Q \subset Y$ is an n–quasiflat, then there exist finitely many Weyl cones $\{W_i\}_{i=1}^h$ such that*

$$d_H\Big(Q, \bigcup_{i=1}^h W_i\Big) < \infty.$$

References

[1] **A Abrams**, **N Brady**, **P Dani**, **M Duchin**, **R Young**, *Pushing fillings in right-angled Artin groups*, J. Lond. Math. Soc. 87 (2013) 663–688 MR

[2] **J Behrstock**, **R Charney**, *Divergence and quasimorphisms of right-angled Artin groups*, Math. Ann. 352 (2012) 339–356 MR

[3] **J Behrstock**, **M F Hagen**, *Cubulated groups: thickness, relative hyperbolicity, and simplicial boundaries*, Groups Geom. Dyn. 10 (2016) 649–707 MR

[4] **J Behrstock**, **M F Hagen**, **A Sisto**, *Hierarchically hyperbolic spaces, I: Curve complexes for cubical groups*, preprint (2014) arXiv To appear in Geom. Topol.

[5] **J A Behrstock**, **T Januszkiewicz**, **W D Neumann**, *Quasi-isometric classification of some high dimensional right-angled Artin groups*, Groups Geom. Dyn. 4 (2010) 681–692 MR

[6] **J Behrstock**, **B Kleiner**, **Y Minsky**, **L Mosher**, *Geometry and rigidity of mapping class groups*, Geom. Topol. 16 (2012) 781–888 MR

[7] **J A Behrstock**, **W D Neumann**, *Quasi-isometric classification of graph manifold groups*, Duke Math. J. 141 (2008) 217–240 MR

[8] **M Bestvina**, **B Kleiner**, **M Sageev**, *The asymptotic geometry of right-angled Artin groups, I*, Geom. Topol. 12 (2008) 1653–1699 MR

[9] **M Bestvina**, **B Kleiner**, **M Sageev**, *Quasiflats in* $\mathrm{CAT}(0)$ 2*–complexes*, Algebr. Geom. Topol. 16 (2016) 2663–2676 MR

[10] **M R Bridson**, *Geodesics and curvature in metric simplicial complexes*, PhD thesis, Cornell Univ. (1991) MR Available at `http://search.proquest.com/docview/303921524`

[11] **M R Bridson**, **A Haefliger**, *Metric spaces of non-positive curvature*, Grundl. Math. Wissen. 319, Springer (1999) MR

[12] **P-E Caprace**, **M Sageev**, *Rank rigidity for CAT(0) cube complexes*, Geom. Funct. Anal. 21 (2011) 851–891 MR

[13] **R Charney**, *An introduction to right-angled Artin groups*, Geom. Dedicata 125 (2007) 141–158 MR

[14] **P Dani**, **A Thomas**, *Divergence in right-angled Coxeter groups*, Trans. Amer. Math. Soc. 367 (2015) 3549–3577 MR

[15] **M W Davis**, **G Moussong**, *Notes on nonpositively curved polyhedra*, from "Low dimensional topology" (K Böröczky, Jr, W Neumann, A Stipsicz, editors), Bolyai Soc. Math. Stud. 8, János Bolyai Math. Soc., Budapest (1999) 11–94 MR

[16] **M J Dunwoody**, *The accessibility of finitely presented groups*, Invent. Math. 81 (1985) 449–457 MR

[17] **A Eskin**, **B Farb**, *Quasi-flats and rigidity in higher rank symmetric spaces*, J. Amer. Math. Soc. 10 (1997) 653–692 MR

[18] **R Frigerio**, **J-F Lafont**, **A Sisto**, *Rigidity of high dimensional graph manifolds*, Astérisque 372, Soc. Math. France, Marseilles (2015) MR

[19] **M Gromov**, *Hyperbolic groups*, from "Essays in group theory" (S M Gersten, editor), Math. Sci. Res. Inst. Publ. 8, Springer (1987) 75–263 MR

[20] **F Haglund**, **D T Wise**, *Special cube complexes*, Geom. Funct. Anal. 17 (2008) 1551–1620 MR

[21] **J Huang**, *Quasi-isometric classification of right-angled Artin groups, I: The finite out case*, preprint (2014) arXiv To appear in Geom. Topol.

[22] **J Huang**, *Quasi-isometry of right-angled artin groups, II: Several infinite out case*, preprint (2015) arXiv

[23] **J Huang**, *Commensurability of groups quasi-isometric to RAAG's*, preprint (2016) arXiv

[24] **J Huang**, **B Kleiner**, *Groups quasi-isometric to RAAG's*, preprint (2016) arXiv

[25] **W Hurewicz**, **H Wallman**, *Dimension theory*, Princeton Mathematical Series 4, Princeton Univ. Press (1941) MR

[26] **M Kapovich**, **B Kleiner**, **B Leeb**, *Quasi-isometries and the de Rham decomposition*, Topology 37 (1998) 1193–1211 MR

[27] **M Kapovich**, **B Leeb**, *Quasi-isometries preserve the geometric decomposition of Haken manifolds*, Invent. Math. 128 (1997) 393–416 MR

[28] **B Kleiner**, *The local structure of length spaces with curvature bounded above*, Math. Z. 231 (1999) 409–456 MR

[29] **B Kleiner**, **U Lang**, *Quasi-minimizing varieties in spaces of higher asymptotic rank*, in preparation

[30] **B Kleiner**, **B Leeb**, *Rigidity of quasi-isometries for symmetric spaces and Euclidean buildings*, C. R. Acad. Sci. Paris Sér. I Math. 324 (1997) 639–643 MR

[31] **L Kramer**, *On the local structure and the homology of* $\mathrm{CAT}(\kappa)$ *spaces and Euclidean buildings*, Adv. Geom. 11 (2011) 347–369 MR

[32] **L Kramer**, **R M Weiss**, *Coarse rigidity of Euclidean buildings*, preprint (2009) arXiv

[33] **A Lytchak**, **K Nagano**, *Geodesically complete spaces with an upper curvature bound*, in preparation

[34] **G D Mostow**, *Strong rigidity of locally symmetric spaces*, Annals of Mathematics Studies 78, Princeton Univ. Press (1973) MR

[35] **K Nagano**, *A volume convergence theorem for Alexandrov spaces with curvature bounded above*, Math. Z. 241 (2002) 127–163 MR

[36] **M Sageev**, *Ends of group pairs and non-positively curved cube complexes*, Proc. London Math. Soc. 71 (1995) 585–617 MR

[37] **K Wortman**, *Quasiflats with holes in reductive groups*, Algebr. Geom. Topol. 6 (2006) 91–117 MR

The Department of Mathematics and Statistics, McGill University
Burnside Hall, Room 1242, 805 Sherbrooke W., Montreal QC H3A 0B9, Canada

jingyin.huang@mcgill.ca

Proposed: Walter Neumann Received: 10 January 2016
Seconded: Bruce Kleiner, Dmitri Burago Revised: 17 May 2016

Geometry & Topology 21 (2017) 2353–2371

A characterisation of alternating knot exteriors

JOSHUA A HOWIE

We give a topological characterisation of alternating knot exteriors based on the presence of special spanning surfaces. This shows that being alternating is a topological property of the knot exterior and not just a property of diagrams, answering an old question of Fox. We also give a characterisation of alternating link exteriors which have marked meridians. We then describe a normal surface algorithm which can decide if a knot is alternating given a triangulation of its exterior as input.

57M25

1 Introduction

Let L be a link in S^3, and let $N(L)$ be a regular open neighbourhood. Then the link exterior $X \cong S^3 \setminus N(L)$ is a compact 3–manifold with torus boundary components. A planar link diagram $\pi(L)$ is a projection

$$\pi\colon S^2 \times I \to S^2,$$

where L has been isotoped to lie in some $S^2 \times I \subset S^3$, together with crossing information. A diagram $\pi(L)$ is alternating if the crossings alternate between over- and under-crossings as we traverse the projection of the link. A nontrivial link is alternating if it admits an alternating diagram. We take the convention that the unknot is not alternating.

A simple Euler characteristic argument shows that if $\pi(L)$ is a planar diagram with n crossings and checkerboard surfaces Σ and Σ', then

$$\chi(\Sigma) + \chi(\Sigma') + n = 2.$$

Furthermore, if $\pi(L)$ is reduced, nonsplit and alternating, then Σ and Σ' are both π_1–essential in X and $2n$ is the difference between the aggregate slopes of Σ and Σ'. If K is a knot, then $2n$ is the difference between their boundary slopes.

Our main result is to prove the converse, where we think of the difference in boundary slopes of Σ and Σ' as the minimal geometric intersection number of $\partial\Sigma$ and $\partial\Sigma'$ on ∂X, which we denote by $i(\partial\Sigma, \partial\Sigma')$.

Published: 19 May 2017 DOI: 10.2140/gt.2017.21.2353

Theorem 3.1 *Let K be a nontrivial knot in S^3 with exterior X. Then K is alternating if and only if there exist a pair of connected spanning surfaces Σ, Σ' in X which satisfy*

$$\chi(\Sigma) + \chi(\Sigma') + \tfrac{1}{2} i(\partial\Sigma, \partial\Sigma') = 2. \tag{$\star$}$$

Notice that all the conditions on X in this characterisation are topological in nature. Theorem 3.1 answers an old question, attributed to Ralph Fox, "What is an alternating knot?" This question has been interpreted as requesting a nondiagrammatic description of alternating knots; see Lickorish [16, page 32].

A similar characterisation of alternating knot exteriors has been independently obtained by Greene [7]. Alternating knot exteriors are characterised by a pair of spanning surfaces which are positive and negative definite with respect to the Gordon–Litherland bilinear form.

By a theorem of Gordon and Luecke [6], a knot exterior has a unique meridian so the concept of a spanning surface is well-defined in X. For link exteriors, this is not true (see Gordon [5]) and there are 3–manifolds which are homeomorphic to the exterior of both alternating and nonalternating links. However, X together with a marked meridian on each boundary component does uniquely determine a link in S^3. With the addition of an extra condition on intersection numbers, we are able to give a characterisation of alternating link exteriors with marked meridians in Theorem 3.2.

For the second half of this article we turn to normal surface theory, and show that given a 3–manifold with connected torus boundary X, it is possible to decide if X is the exterior of an alternating knot in S^3. We also show that given a nonalternating planar diagram of a knot K, we can decide if K is alternating, and if so, we can produce an alternating diagram of either K or its mirror image.

Theorem 5.2 *Let X be the exterior of a knot $K \subset S^3$. Given X, there is a normal surface algorithm to decide if K is alternating.*

In an appendix to [7], Juhász and Lackenby have also found a normal surface algorithm to decide if a knot is prime and alternating based on Greene's characterisation.

Organisation In Section 2, we describe how two spanning surfaces for a link intersect. In Section 3, we prove Theorem 3.1, and give the version for links. In Section 4, we give some background on normal surface theory and the boundary solution space. In Section 5, we detail the algorithm which can decide if a knot manifold is the exterior of an alternating knot.

Acknowledgement The author would like to thank Hyam Rubinstein for many interesting conversations and help with the algorithm.

2 Intersections of spanning surfaces

Let L be a link with m components denoted by L_j for $j = 1, \dots, m$. Denote the boundary components of X by $C_j = \partial N(L_j)$ where each C_j is a torus.

A curve $\mu_j \subset C_j$ is a meridian of X if μ_j bounds an embedded disk in $N(L_j)$ which intersects L_j transversely exactly once. Given X, a set of marked meridians is a set of curves $\{\mu_j\}$ with one μ_j on each C_j such that Dehn filling along each μ_j produces the 3–sphere.

We define a preferred longitude λ_j of C_j to be the unique nontrivial curve on C_j which meets μ_j exactly once and bounds an orientable surface S_j in $S^3 \setminus N(L_j)$. Note that S_j is not necessarily embedded in X since it may intersect other components of L.

Let $\overline{\Sigma}$ be a compact surface embedded in S^3. Then $\overline{\Sigma}$ is a spanning surface for a link L if $\partial\overline{\Sigma} = L$. Let Σ be a surface with boundary, properly embedded in X. Then Σ is a spanning surface for X if each component of $\partial\Sigma$ has minimal geometric intersection number one with the meridian μ_j of C_j. These two notions of spanning surface are related since $\Sigma = X \cap \overline{\Sigma}$ whenever $\overline{\Sigma}$ is in general position with respect to ∂X, and Σ can be extended to $\overline{\Sigma}$ by attaching a small annulus in each component of $N(L)$.

Spanning surfaces Σ and Σ' are in general position in X if they intersect in a set of properly embedded arcs and embedded loops. In particular, there are no triple points or branch points, since each spanning surface is properly embedded.

For $j = 1, \dots, m$, let $\{\sigma_j\}$ be the components of $\partial\Sigma$ and let $\{\sigma'_j\}$ be the components of $\partial\Sigma'$. Fix an orientation on each longitude λ_j and define the orientation of each meridian μ_j so that the (λ_j, μ_j) form a right-handed basis for each torus boundary component C_j. Then $[\sigma_j] = p_j[\mu_j] + [\lambda_j]$ and $[\sigma'_j] = p'_j[\mu_j] + [\lambda_j]$, where $p_j, p'_j \in \mathbb{Z}$. We define the algebraic intersection number of σ_j and σ'_j to be

$$i_a(\sigma_j, \sigma'_j) = p_j - p'_j = -i_a(\sigma'_j, \sigma_j),$$

while the geometric intersection number is

$$i(\sigma_j, \sigma'_j) = |i_a(\sigma_j, \sigma'_j)| = |p_j - p'_j|.$$

Define the geometric intersection number of $\partial\Sigma$ and $\partial\Sigma'$ to be

$$i(\partial\Sigma, \partial\Sigma') = \sum_{j=1}^{m} i(\sigma_j, \sigma'_j).$$

This geometric intersection number measures the difference in aggregate slopes of the two spanning surfaces, as defined in [1]. For a knot this is the difference in boundary slopes.

If we isotope Σ and Σ' so that they realise $i(\partial\Sigma, \partial\Sigma')$, then σ_j and σ'_j form a quadrangulation Q_j of C_j, where each quadrangular face has one pair of nonadjacent vertices on its boundary which are identified. There are $i_j = i(\sigma_j, \sigma'_j)$ vertices, $2i_j$ edges, and i_j faces in Q_j. We refer to $\bigsqcup_{j=1}^{m} Q_j$ as a boundary quadrangulation of ∂X. When forming a quadrangulation on C_j, every intersection of σ_j and σ'_j has the same sign.

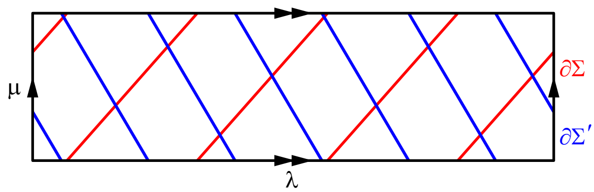


Figure 1: A boundary quadrangulation of ∂X formed by $\partial\Sigma$ and $\partial\Sigma'$ in the case of a knot exterior

An arc of intersection between two spanning surfaces Σ and Σ' in a link exterior X is called a double arc. Let α be a double arc and let $\bar{\alpha}$ be its extension to S^3 such that $\partial\bar{\alpha} \subset L$. There are two types of double arc.

Let W be a regular neighbourhood of $\bar{\alpha}$ in S^3. Let β and β' be the two components of $W \cap L$. We can choose W so that $V = W \cap X$ is a compact handlebody of genus two, and so that $\Sigma \cap V$ and $\Sigma' \cap V$ are both disks.

Fix an orientation on β. This induces orientations on the disks $\Sigma \cap V$ and $\Sigma' \cap V$. If these both induce the same orientation on β', then α is a *parallel* arc. If they induce opposite orientations on β', then α is a *standard* arc.

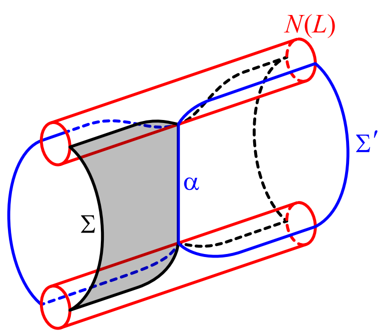


Figure 2: A parallel arc of intersection between two spanning surfaces

This is equivalent to saying $\partial(\Sigma \cap V)$ and $\partial(\Sigma' \cap V)$ have algebraic intersection number zero if α is a parallel arc, and algebraic intersection number two if α is a standard arc. The intersections on ∂X at the endpoints of a parallel arc α have the same sign if α is standard, but opposite signs when α is parallel.

If we collapse a standard arc to a point, then $(\overline{\Sigma} \cup \overline{\Sigma}') \cap W$ collapses to a disk. If we try to collapse a parallel arc to a point, then $(\overline{\Sigma} \cup \overline{\Sigma}') \cap W$ collapses to an object homeomorphic to a neighbourhood of the apex of a double cone.

Lemma 2.1 *Let Σ and Σ' be spanning surfaces for a link L, isotoped so that their boundaries realise the intersection number $i(\partial\Sigma, \partial\Sigma')$ on ∂X. If*

$$i(\partial\Sigma, \partial\Sigma') = \left|\sum_{j=1}^{m} i_a(\sigma_j, \sigma_j')\right|,$$

then every double arc of intersection between Σ and Σ' is standard.

Proof By definition $i(\partial\Sigma, \partial\Sigma') = \sum_{j=1}^{m} i(\sigma_j, \sigma_j')$. If

$$\left|\sum_{j=1}^{m} i_a(\sigma_j, \sigma_j')\right| = \sum_{j=1}^{m} i(\sigma_j, \sigma_j'),$$

then for every $j = 1, \dots, m$, either every intersection between σ_j and σ_j' is positive, or every intersection between σ_j and σ_j' is negative. A parallel arc only occurs when a double arc connects a positive intersection to a negative intersection. □

Note that Lemma 2.1 implies that parallel arcs of intersections can only occur when the double arc runs between different components of ∂X. Hence if K is a knot, then every arc of intersection between two spanning surfaces realising minimal intersection number must be standard.

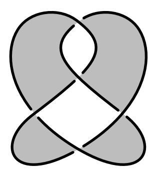


Figure 3: Black and white checkerboard surfaces for a knot

For any nonsplit planar link diagram $\pi(L)$, there is a standard position for the associated checkerboard surfaces. Away from a crossing the checkerboard surfaces are embedded in S^2, but in a small regular neighbourhood of a crossing, we think of the link lying on the surface of a ball U. The ball U intersects S^2 in an equatorial disk, and the over-strand runs over the upper hemisphere, while the under-strand runs under the lower hemisphere. Each checkerboard surface intersects U in a half-twisted band. The ball U is called a bubble and this viewpoint of checkerboard surfaces for planar alternating diagrams was introduced by Menasco [17].

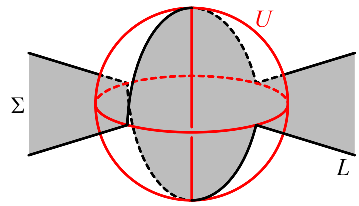


Figure 4: A checkerboard surface in standard position near a bubble U

In standard position, the checkerboard surfaces Σ and Σ' do not intersect in any loops, and intersect only in double arcs corresponding to the north-south axis of each bubble. These double arcs are all standard. The intersection of the corresponding spanning surfaces $\overline{\Sigma}$ and $\overline{\Sigma}'$ is a disjoint union of trivalent graphs and loops consisting of the link L and the collection of vertical axes of the bubbles. If we assume that $\pi(L)$ is nonsplit, then every component of L is involved in a crossing of $\pi(L)$, so that $\overline{\Sigma} \cap \overline{\Sigma}'$ forms a graph Γ', where each connected component is 3–regular.

For a nonsplit alternating planar projection $\pi(L)$ in standard position, as we traverse the image of L_j in $\pi(L)$, it can be seen that σ_j rotates in a positive manner say, with respect to S^2, while σ_j' rotates in a negative manner. Hence the checkerboard surfaces in standard position already realise the minimal geometric intersection number of their boundaries, and their boundaries form a boundary quadrangulation of ∂X.

Thus if $\pi(L)$ is a nonsplit planar alternating diagram of an m–component link L, which has n crossings, then

$$2n = i(\partial\Sigma, \partial\Sigma') = \sum_{j=1}^{m} i(\sigma_j, \sigma_j').$$

A method for calculating the boundary slopes of the checkerboard surfaces associated to a reduced alternating knot diagram is detailed in [3], where it can be seen that $2n$ is the difference between the boundary slopes of the two checkerboard surfaces. If Σ is a spanning surface for a knot and $[\partial\Sigma] = p[\mu] + [\lambda]$, then the boundary slope of Σ is $p \in 2\mathbb{Z}$.

Note that if $\pi(L)$ is a nonalternating planar diagram with n crossings, then somewhere there are two consecutive over-crossings, which forces the boundaries of the associated checkerboard surfaces to create a bigon on some component of C_j. In this case $n > \frac{1}{2} i(\partial\Sigma, \partial\Sigma')$.

The Euler characteristics of the checkerboard surfaces arising from a planar projection are related to the Euler characteristic of the projection sphere by the equation

$$\chi(\Sigma) + \chi(\Sigma') + n = \chi(S^2) = 2.$$

A spanning surface Σ is π_1–essential in a knot exterior X if the induced homomorphism

$$\pi_1(\Sigma) \to \pi_1(X)$$

and the induced map

$$\pi_1(\Sigma, \partial\Sigma) \to \pi_1(X, \partial X)$$

are both injective. This implies that Σ is both incompressible and boundary-incompressible in X. Aumann [2] proved that both checkerboard surfaces associated to a planar reduced alternating projection of a knot K are π_1–essential in X.

3 Characterisation

We now give the proof of the nondiagrammatic characterisation of alternating knot exteriors.

Theorem 3.1 *Let K be a nontrivial knot in S^3 with exterior X. Then K has an alternating projection onto S^2 if and only if there exist a pair of connected spanning surfaces Σ, Σ' for X which satisfy*

$$(\star) \qquad \chi(\Sigma) + \chi(\Sigma') + \tfrac{1}{2} i(\partial\Sigma, \partial\Sigma') = 2.$$

Proof One direction follows from the discussion in Section 2. For the converse, let Σ_0 and Σ_0' be a pair of connected spanning surfaces for X which satisfy $(\star)$.

Since X is not a solid torus, X is boundary-irreducible, so K does not bound a disk in S^3. Hence $\chi(\Sigma_0) + \chi(\Sigma_0') \le 0$, so $i(\partial\Sigma_0, \partial\Sigma_0') \ne 0$ by $(\star)$.

Isotope Σ_0 and Σ_0' in X so that their boundaries realise the minimal geometric intersection number $i(\partial\Sigma_0, \partial\Sigma_0')$. Hence $\partial\Sigma_0$ and $\partial\Sigma_0'$ form a boundary quadrangulation Q, and Q will remain fixed throughout the proof. We may assume that Σ_0 and Σ_0' are in general position, so that they intersect in a set of proper arcs $\mathcal{A}$ and a set of embedded loops $\mathcal{L}_0$.

Recall that $\overline{\Sigma}_0$ is the extension of Σ_0 to S^3, so that $\partial\overline{\Sigma}_0 = K$. Assume that the interiors of $\overline{\Sigma}_0$ and $\overline{\Sigma}_0'$ are in general position, and no loops of intersection have been introduced by the extension process. Let $F_0' = \overline{\Sigma}_0 \cup \overline{\Sigma}_0'$ and let $\Gamma' = (\overline{\Sigma}_0 \cap \overline{\Sigma}_0') \setminus \mathcal{L}_0$, both of which are connected. Let $\overline{\mathcal{A}}$ be the extension of $\mathcal{A}$ to S^3, or in other words, let $\overline{\mathcal{A}}$ be the closure of $\Gamma' \setminus K$.

If we collapse each component of $\overline{\mathcal{A}}$ to a point, then F_0' collapses to an immersed surface F_0 because X is a knot exterior and every arc of $\mathcal{A}$ is standard by Lemma 2.1.

Cutting F'_0, $\overline{\Sigma}_0$ and $\overline{\Sigma}'_0$ along Γ' allows us to calculate that

$$\chi(F'_0) = \chi(\overline{\Sigma}_0) + \chi(\overline{\Sigma}'_0) + \tfrac{1}{2} i(\partial\Sigma_0, \partial\Sigma'_0),$$

from which the equation ($\star$) tells us that $\chi(F'_0) = 2$. The 3–regular graph Γ' collapses to a 4–regular graph Γ. Let

$$f_0\colon S_0 \looparrowright S^3$$

be the immersion of a surface S_0 such that $f_0(S_0) = F_0$. There are no triple points of self-intersection in F_0 since Σ_0 and Σ'_0 are embedded, and the only double loops of self-intersection are precisely the elements of $\mathcal{L}_0$. It follows that

$$\chi(S_0) = \chi(F_0) = \chi(F'_0) = 2,$$

which implies that S_0 is a 2–sphere.

Suppose $\mathcal{L}_0 \neq \varnothing$ and let $\mathcal{B}_0$ be the collection of loops $f_0^{-1}(\mathcal{L}_0)$ on S_0. Since $\mathcal{B}_0$ is the preimage of double loops, we know that $\mathcal{B}_0$ contains an even number of elements. Because S_0 is a 2–sphere, each loop $\beta \in \mathcal{B}_0$ is separating, and $\mathcal{B}_0$ cuts S_0 into a collection of planar surfaces with boundary. Let z_h be the number of planar surfaces in $S_0 \setminus \mathcal{B}_0$ which have h boundary components. Exactly one component of $S_0 \setminus \mathcal{B}_0$ contains the connected graph $f_0^{-1}(\Gamma)$.

Using an Euler characteristic argument, Nowik [19] points out that

$$\sum_{h \geq 1} (2-h) z_h = 2,$$

which in particular implies that z_1, the number of disk regions in $S_0 \setminus \mathcal{B}_0$, is at least 2. Of course, for any collection of disjoint curves on a 2–sphere, there is an innermost one. Thus there is at least one loop in $\mathcal{L}_0$ which bounds a disk in F_0.

Let ℓ_0 be a loop in $\mathcal{L}_0$ which bounds a disk D_0 in either Σ_0 or Σ'_0. Without loss of generality assume $D_0 \subset \Sigma_0$. Let $\{\beta_0, \beta'_0\} = f_0^{-1}(\ell_0)$ where $f_0(N(\beta_0)) \subset \Sigma_0$ and $f_0(N(\beta'_0)) \subset \Sigma'_0$. Notice that f_0^{-1} restricted to the interior of D_0 is a homeomorphism onto the interior of a disk in $S_0 \setminus \mathcal{B}_0$.

Let $A_0 = \ell_0 \times (-1, 1)$ be a regular neighbourhood of ℓ_0 in Σ'_0. We perform surgery on Σ'_0 along D_0, by removing the annulus A_0 from Σ'_0 and gluing in the two disks $D_0 \times \{-1\}$ and $D_0 \times \{1\}$.

Let (Σ_1, Σ'_1) be the result of performing surgery along D_0 on (Σ_0, Σ'_0). Define S_1 to be the result of doing surgery along $f_0^{-1}(D_0)$ in S_0 to remove the loop β'_0, and deleting the curve β_0.

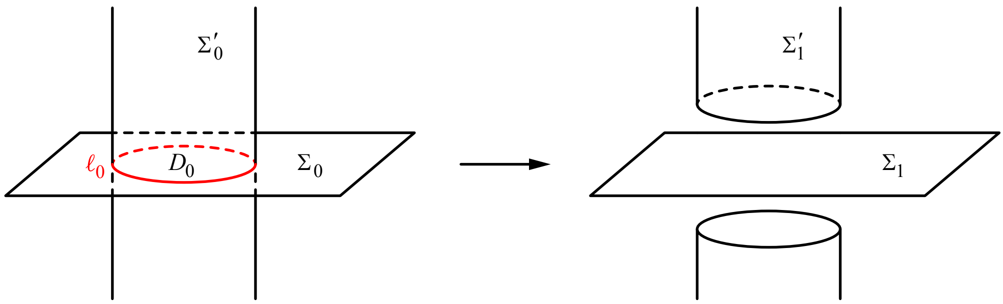


Figure 5: Surgery on Σ_0' along the disk $D_0 \subset \Sigma_0$

Suppose $|\mathcal{L}_0| = k \geq 0$. For $0 \leq j \leq k-1$, inductively define $(\Sigma_{j+1}, \Sigma'_{j+1})$ to be the result of performing surgery along the disk D_j, where $\partial D_j = \ell_j \subset \mathcal{L}_j$ and D_j is a subdisk of either Σ_j or Σ'_j. Let $\mathcal{L}_{j+1} = \mathcal{L}_j \setminus \ell_j$ so that $\mathcal{L}_{j+1}$ is the set of loops of intersection of Σ_{j+1} and Σ'_{j+1}. Let F'_{j+1} be the result of the corresponding surgery on the 2–complex F'_j. Define S_{j+1} to be the result of doing surgery along $f_j^{-1}(D_j)$ in S_j and deleting the curve β'_j or β_j.

A similar calculation to Nowik's shows that

$$\sum_{h \geq 0} (2-h) z_h = \chi(S_j) = 2 + 2j,$$

where S_j is a collection of $j+1$ closed 2–spheres since β_j and β'_j are separating in S_j. Since Γ is connected and disjoint from $\mathcal{L}_0$, $f_j^{-1}(\Gamma)$ is contained in exactly one component of S_j. Hence $z_0 \leq j$, so that $z_1 \geq 2$, and therefore the disk D_j exists.

Continue this inductive surgery process until we have constructed Σ_k and Σ'_k. At this stage $\mathcal{L}_k$ is empty, and S_k consists of $k+1$ 2–spheres. Hence F'_k consists of k unmarked embedded 2–spheres and one 2–complex, denoted F', which contains Γ'.

Define $\overline{\Sigma}$ and $\overline{\Sigma}'$ to be the components of $\overline{\Sigma}_k$ and $\overline{\Sigma}'_k$, respectively, which constitute F'. Then Σ and Σ' are connected spanning surfaces for X which satisfy $(\star)$, and whose intersection is exactly $\mathcal{A}$. Collapsing the arcs of $\overline{\mathcal{A}}$ to points collapses F' and Γ' to F and Γ, respectively. Since $\chi(F') = 2$, it follows that $\chi(F) = 2$, so F is an embedded 2–sphere, which will be our desired projection surface.

Since $\partial\Sigma$ and $\partial\Sigma'$ realise $i(\partial\Sigma, \partial\Sigma') = i(\partial\Sigma_0, \partial\Sigma'_0)$, we can recover the crossing information of $\pi(K)$ from $\Gamma \subset F$. This is because every double arc is standard, so instead of collapsing every arc of $\overline{\mathcal{A}}$ to a point, we could collapse a regular neighbourhood of each $\bar{\alpha}$ to a bubble. The diagram $\pi(K)$ must be alternating since, otherwise, there would be a bigon between $\partial\Sigma$ and $\partial\Sigma'$ on ∂X, which contradicts that the boundary quadrangulation Q has remained fixed. Note that $\pi(K)$ is not necessarily reduced. $\square$

If $\pi(K)$ is a nonalternating planar diagram with checkerboard surfaces Σ and Σ', then there is a bigon on ∂X between $\partial\Sigma$ and $\partial\Sigma'$. Any attempt to isotope the checkerboard surfaces to remove all bigons and obtain a boundary quadrangulation causes Σ and Σ' to create an alternating diagram onto a nonplanar surface F. This surface F could be either embedded or immersed, and in the latter case may even be nonorientable [10, page 146].

Note that Theorem 3.1 is not concerned with primeness. However, a knot is prime if there are no essential annuli properly embedded in X at meridional slope. If a nontrivial knot is prime and alternating, then X is atoroidal, since no prime alternating knot is satellite [17].

In [10, Theorem 4.1], using a different method, the author also proved a variation of Theorem 3.1 which required both spanning surfaces to be π_1–essential in X. We also note that a more complicated characterisation can be obtained as a special case of a theorem proved in [10, Theorem 3.33] which gives a topological characterisation of a class of links which have certain alternating diagrams onto orientable surfaces of higher genus. This will be written up in a forthcoming article with Rubinstein [11].

We have not stated any of the theorems in this section for links. The issue is that, given two spanning surfaces, there could be parallel arcs between different components of ∂X. If the two spanning surfaces have been isotoped to create a boundary quadrangulation and there are parallel arcs of intersection, then the complex F_0' does not collapse to a surface. We note that there exists an example [10, page 139] of a pair of spanning surfaces for a nonsplit 2–component link which intersect in parallel arcs, yet still satisfy $(\star)$.

However, if we assume that all arcs of intersection are standard, then we have the following theorem.

Theorem 3.2 *Let L be a nontrivial nonsplit link in S^3 with exterior X which has a marked meridian on each boundary component. Then L has an alternating projection onto S^2 if and only if there exist a pair of connected spanning surfaces Σ, Σ' for X which satisfy*

$$(\star) \qquad \chi(\Sigma) + \chi(\Sigma') + \tfrac{1}{2} i(\partial\Sigma, \partial\Sigma') = 2,$$

and

$$i(\partial\Sigma, \partial\Sigma') = \left| \sum_{j=1}^{m} i_a(\sigma_j, \sigma_j') \right|.$$

Proof Let $\pi(L)$ be a reduced nonsplit alternating projection of L onto S^2, and let Σ and Σ' be the associated checkerboard surfaces in standard position. Then $\pi(L)$, Σ and Σ' are connected, so every arc of intersection between Σ and Σ' is standard.

This means every algebraic intersection number $i_a(\sigma_j, \sigma_j')$ is either positive or every $i_a(\sigma_j, \sigma_j')$ is negative, which implies the second equation. The first equation follows from the two displayed equations on page 2358.

For the converse, the concept of a spanning surface for X is well-defined since a set of meridians for X are specified. Lemma 2.1 ensures that every arc of intersection is standard. Then the rest of the proof goes through as in Theorem 3.1. Since Σ and Σ' are connected, $\pi(L)$ must be nonsplit. □

4 Normal surface theory

In this section we provide the background material necessary to construct our algorithm. Kneser introduced the concept of a normal surface, before Haken [8] developed normal surface theory into an important tool for algorithmic topology. We will give a brief outline of the theory; for full details the reader is referred to [15].

A knot manifold is a compact irreducible 3–manifold with connected torus boundary. A triangulation $\mathcal{T}$ of a knot manifold M is a collection of t tetrahedra and a set of equations which identify some pairs of faces of the tetrahedra, so that the link of every vertex is either a 2–sphere or a disk, and the unglued faces form the boundary torus ∂M.

A normal surface S is a properly embedded surface in M which is transverse to the 2–skeleton of $\mathcal{T}$, and such that $S \cap \triangle$ is a collection of triangular or quadrilateral disks, where $\triangle$ is any tetrahedron of $\mathcal{T}$, and each disk intersects each edge of $\triangle$ in at most one point. There are seven normal isotopy classes of normal disks, four are triangular and three are quadrilateral, and each of these is known as a disk type.

If we fix an ordering of the disk types $d_1, d_2, \dots, d_{7t}$, then a normal surface S can be represented uniquely up to normal isotopy by a $7t$–tuple of nonnegative integers $\boldsymbol{n}(S) = (x_1, x_2, \dots, x_{7t})$, where x_i is the number of disks of type d_i, and t is the number of tetrahedra in $\mathcal{T}$.

Conversely, given a $7t$–tuple of nonnegative integers $\boldsymbol{n}$, we can impose restrictions on the x_i so that $\boldsymbol{n}$ represents a properly embedded normal surface. We require that at least two of the three quadrilateral disk types are not present in each tetrahedra. This ensures that the surface is embedded. We also need to make sure that the disk types match up with the disk types in neighbouring tetrahedra.

An arc type is the normal isotopy class of the intersection of a normal surface with a face of a tetrahedron. There are three arc types in each face of each tetrahedron, and each arc type is contributed to by two different disk types, one triangular, the other quadrilateral. We require that the number of each arc type in each face agrees with the number of arcs of the corresponding type in the face of the tetrahedron which is glued to it. This condition can be described by a linear equation for each arc type. Together they are called the matching equations for the normal surface S, and in a one-vertex triangulation of a knot manifold M there are $6t-3$ matching equations.

The set of nonnegative integer solutions to the normal surface equations lie within an infinite linear cone $\mathcal{S}_{\mathcal{T}} \subset \mathbb{R}^{7t}$. The linear cone $\mathcal{S}_{\mathcal{T}}$ is called the solution space.

The additional condition that

$$\sum_{i=1}^{7t} x_i = 1$$

turns the solution space into a compact, convex, linear cell $\mathcal{P}_{\mathcal{T}} \subset \mathcal{S}_{\mathcal{T}}$. We call $\mathcal{P}_{\mathcal{T}}$ the projective solution space, and we let $\hat{\boldsymbol{n}}(S)$ represent the projective class of the normal surface S. The carrier of a normal surface S, denoted $\mathcal{C}_{\mathcal{T}}(S)$, is defined to be the unique minimal face of $\mathcal{P}_{\mathcal{T}}$ which contains $\hat{\boldsymbol{n}}(S)$.

Let S be a properly embedded surface in a 3-manifold M with triangulation $\mathcal{T}$. Haken [8] showed that after a series of isotopies, compressions, boundary-compressions and the removal of trivial 2–spheres and disks, S' can be represented as the union of properly embedded normal surfaces with respect to $\mathcal{T}$. In particular, if S is π_1–essential in M, then S can be isotoped to be normal with respect to $\mathcal{T}$.

Two normal surfaces S and S' are compatible if, for each tetrahedron $\triangle$ of $\mathcal{T}$, S and S' do not contain quadrilateral disks of different types. If S and S' are compatible, then we can form the Haken sum of S and S', which we denote $S \oplus S'$. The Haken sum is a geometric sum along each arc and loop of intersection between S and S', which is uniquely defined by the requirement that $S \oplus S'$ is a normal surface. Any other choice of geometric sums would produce a surface with folds. If $\boldsymbol{n}(S) = (x_1, x_2, \dots, x_{7t})$ and $\boldsymbol{n}(S') = (x'_1, x'_2, \dots, x'_{7t})$ are representatives of compatible normal surfaces S and S' in a triangulation $\mathcal{T}$ of a 3–manifold M, then $\boldsymbol{n}(S \oplus S') = \boldsymbol{n}(S) + \boldsymbol{n}(S') = (x_1 + x'_1, x_2 + x'_2, \dots, x_{7t} + x'_{7t})$. Also, $\chi(S \oplus S') = \chi(S) + \chi(S')$.

A normal surface S is called a vertex surface if $\hat{\boldsymbol{n}}(S)$ lies at a vertex of the projective solution space. This means that whenever some multiple of S can be written as a Haken sum of two surfaces, both the summands are also multiples of S.

A normal surface S is called a fundamental surface if $\boldsymbol{n}(S)$ cannot be written as the sum of two solutions to the normal surface equations. Every vertex surface is a fundamental surface, but there exist fundamental surfaces which are not vertex surfaces.

All normal surfaces can be written as a finite sum of fundamental surfaces. There are a finite number of fundamental surfaces. They can be found algorithmically and Haken used this fact to construct his algorithms. Many of these algorithms have been subsequently improved so that they use vertex solutions rather than fundamental solutions, which makes the algorithms more efficient.

A triangulation is 0–efficient if the only normal disks or normal spheres are vertex-linking. Jaco and Rubinstein [12] showed that every compact orientable irreducible and boundary-irreducible 3–manifold with nonempty boundary admits a 0–efficient triangulation. Since the solid torus admits a one-vertex triangulation, it then follows that every knot exterior admits a one-vertex triangulation.

Let $\mathcal{T}$ be a one-vertex triangulation of a knot manifold M. Then there is an induced one-vertex triangulation $\mathcal{T}_\partial$ of ∂M. The boundary triangulation $\mathcal{T}_\partial$ consists of one vertex, three edges and two faces. There are six normal arc types; however, a normal curve is determined by just three of these arc types. Every curve on ∂M has a unique normal representative. This means that isotopy classes of curves on ∂M correspond to normal isotopy classes of curves on $\mathcal{T}_\partial$.

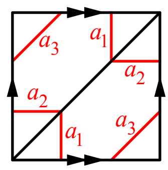


Figure 6: Normal arc types in the boundary triangulation

Fix an ordering of the disk types in $\mathcal{T}$ such that $d_1, \dots, d_7$ represent the disk types in one of the tetrahedra which meets ∂M in a face ϕ. Furthermore, let $d_1, \dots, d_4$ represent triangular disk types, and d_5, d_6, d_7 represent quadrilateral disks, such that d_i and d_{i+4} meet ϕ in the same arc type a_i for $i = 1, 2, 3$. Here d_4 is the triangular disk type which is disjoint from ϕ.

Let y_i be the number of arcs of type a_i in ϕ. It follows that

$$y_i = x_i + x_{i+4}$$

for $i = 1, 2, 3$. Jaco and Sedgwick [14, Theorem 3.6] showed that y_1, y_2, y_3 and the matching equations for normal curves determine the number of arcs of each type in the other 2–simplex of $\mathcal{T}_\partial$. We define the boundary solution space of ∂M to be

$$\mathcal{S}_{\mathcal{T}_\partial} = \{(y_1, y_2, y_3) \mid y_i \in \mathbb{N}_0\} \subset \mathbb{R}^3,$$

where $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$. If ∂S is the boundary of a properly embedded normal surface, then ∂S is represented by $\boldsymbol{n}(\partial S) = (y_1, y_2, y_3)$ in $\mathcal{S}_{\mathcal{T}_\partial}$.

If each coordinate of $\boldsymbol{n}(\partial S)$ is nonzero, then ∂S contains a trivial curve. Hence if S is an incompressible surface, then at least one of the coordinates of $\boldsymbol{n}(\partial S)$ is zero.

Jaco and Sedgwick [14, Section 3.3] proved that if S is a properly embedded π_1–essential normal surface with boundary in a compact 3–manifold M with triangulation $\mathcal{T}$, then every surface in $\mathcal{C}_{\mathcal{T}}(S)$ is either closed, or has the same slope as S. This means that if p/q is a boundary slope of X, then there is a vertex surface S which has slope p/q. Hence it is only necessary to check the vertices of $\mathcal{P}_{\mathcal{T}}$ in order to list all boundary slopes of X. In proving this theorem, Jaco and Sedgwick have given another proof of a theorem of Hatcher [9] that there are only a finite number of slopes bounding π_1–essential surfaces in any knot exterior. Recall that the set of slopes of a link exterior which bound π_1–essential surfaces is not necessarily finite, so our algorithm is only designed to work for knots.

Jaco and Sedgwick [14, Theorem 6.4] also gave an algorithm to decide if a knot manifold M is a knot exterior in S^3, which is also an algorithm to find the unique meridian μ of a knot exterior X. This algorithm makes use of the Rubinstein–Thompson algorithm [21; 22] which can decide if a 3–manifold is homeomorphic to the 3–sphere. There is also an algorithm which can decide if X is a solid torus [8], which is equivalent to deciding if K is the unknot, and it is now known that some spanning disk for K can be found as a vertex solution if K is the unknot.

The boundary triangulation $\mathcal{T}_\partial$ consists of two 2–simplices and three edges. We can modify the triangulation $\mathcal{T}$ by gluing two faces of a tetrahedron $\triangle$ to $\mathcal{T}_\partial$. The resulting triangulation $\mathcal{T}' = \mathcal{T} \cup \triangle$ is another one-vertex triangulation of M, and $\mathcal{T}'$ is called a layered triangulation. In effect, this is a $(2,2)$–Pachner move on $\mathcal{T}_\partial$. The other two faces of $\triangle$ form the boundary triangulation $\mathcal{T}'_\partial$.

Layering a tetrahedron changes the slope of one of the edges in the boundary triangulation. It is always possible to layer a triangulation with a sequence of tetrahedra so that the edges of $\mathcal{T}'_\partial$ have slopes $\infty, k, k+1$, for some $k \in \mathbb{Z}$. We will choose to do this so that $(1,0,0) \in \mathcal{S}_{\mathcal{T}_\partial}$ represents the meridian μ.

5 Alternating algorithm

We now describe an algorithm to decide if a knot is alternating on S^2. The input is a triangulation $\mathcal{T}$ of a knot exterior X. If instead we are given a nonalternating planar diagram $\pi(K)$, then there is a method of Petronio [20] to construct a spine of the knot complement $S^3 \setminus K$ from the diagram $\pi(K)$. Dual to this spine is an ideal triangulation of $S^3 \setminus K$. We can then use an inflation of Jaco and Rubinstein [13] to construct a one-vertex triangulation of the knot exterior X from the ideal triangulation.

First we need one more result on spanning surfaces.

Lemma 5.1 *Suppose that Σ and Σ' are spanning surfaces for a knot $K \subset S^3$. If*

$$\chi(\Sigma) + \chi(\Sigma') + \tfrac{1}{2} i(\partial\Sigma, \partial\Sigma') > 2,$$

then at least one of Σ or Σ' is not connected.

Proof Isotope Σ and Σ' so that they are in general position and realise $i(\partial\Sigma, \partial\Sigma')$ on ∂X. As in the proof for Theorem 3.1, we form the pseudo-2–complex $F' = \overline{\Sigma} \cup \overline{\Sigma}'$ and collapse the arcs of intersection to points to obtain an immersed surface F, where $\chi(F) = \chi(\Sigma) + \chi(\Sigma') + \frac{1}{2} i(\partial\Sigma, \partial\Sigma') > 2$. Let $f \colon S \looparrowright S^3$ be an immersion of a closed surface S such that $f(S) = F$. The only possible self-intersections of $f(S)$ are loops, so $\chi(S) = \chi(F) > 2$, and therefore S is not connected. Hence either Σ or Σ' is not connected. □

Theorem 5.2 *Let X be the exterior of a knot $K \subset S^3$. Given X, there is a normal surface algorithm to decide if K is alternating.*

Proof Let $\mathcal{T}'$ be a one-vertex triangulation of X. As shown in [14], any other triangulation of X can be modified to a one-vertex triangulation.

Use the Jaco–Sedgwick algorithm to find the unique meridian μ of X. Included in this process is a check whether X is a solid torus. If X is a solid torus, then K is the unknot, which by our convention is not alternating.

Layer the triangulation until one of the edges in the boundary is parallel to μ. Then the other edges in the boundary are parallel to $\lambda + k\mu$ and $\lambda + (k+1)\mu$ for some $k \in \mathbb{Z}$. Call this triangulation $\mathcal{T}$.

Let $\triangle$ be one of the two tetrahedra that meets the boundary and let ϕ be a face of $\triangle$ which lies in the boundary. Let $(x_1, x_2, x_3, x_4, x_5, x_6, x_7)$ describe the normal coordinates of $S \cap \triangle$ where S is a properly embedded surface with boundary in X. Let (y_1, y_2, y_3) describe the normal coordinates of $\partial S \cap \phi$. As described in Section 4, we label the arc and disc types so that $y_i = x_i + x_{i+4}$ for each $i = 1, 2, 3$, and so that $(1, 0, 0)$ represents μ in ∂X. Let $(0, 1, 0)$ represent $\lambda + k\mu$ so that $(0, 0, 1)$ represents $\lambda + (k+1)\mu$ for some $k \in \mathbb{Z}$.

Any spanning surface for X meets μ exactly once. It follows that if (y_1, y_2, y_3) represents a spanning surface, then $y_2 + y_3 = 1$. So there are two types of coordinates in $\mathcal{S}_{\mathcal{T}_\partial}$ which can represent spanning surfaces: $(y, 1, 0)$ and $(y, 0, 1)$ for some $y \in \mathbb{N}_0$.

Let Σ and Σ' be normal spanning surfaces in X. We can read off the minimal geometric intersection number of their boundaries from their coordinates in $\mathcal{S}_{\mathcal{T}_\partial}$. Let $y, y' \in \mathbb{N}_0$. There are three cases:

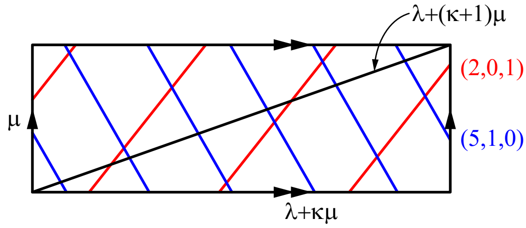


Figure 7: Normal curves in the boundary triangulation after layering

(1) If $\partial\Sigma$ and $\partial\Sigma'$ are represented by $(y,1,0)$ and $(y',1,0)$, respectively, then $i(\partial\Sigma,\partial\Sigma') = |y-y'|$.
(2) If $\partial\Sigma$ and $\partial\Sigma'$ are represented by $(y,0,1)$ and $(y',0,1)$, respectively, then $i(\partial\Sigma,\partial\Sigma') = |y-y'|$.
(3) If $\partial\Sigma$ and $\partial\Sigma'$ are represented by $(y,1,0)$ and $(y',0,1)$, respectively, then $i(\partial\Sigma,\partial\Sigma') = y+y'+1$. See Figure 7 for an example of this case.

Note that we could continue layering the triangulation until $k=0$, which would require detection of a Seifert surface, but this is not necessary since we are only interested in the differences of spanning slopes, and not the boundary slopes themselves.

Theorem 3.1 tells us that we need to find a pair of connected spanning surfaces at even boundary slope, which satisfy

$$(\star) \qquad \chi(\Sigma)+\chi(\Sigma')+\tfrac{1}{2}i(\partial\Sigma,\partial\Sigma') = 2.$$

The checkerboard surfaces Σ and Σ' associated to a reduced alternating diagram of K are one such pair of surfaces. Aumann [2] showed that they are both π_1–essential in X, so we know that both Σ and Σ' must have normal representatives in their isotopy classes.

Let Σ and Σ' be a pair of connected π_1–essential normal spanning surfaces which satisfy $(\star)$. Suppose that $\boldsymbol{n}(\Sigma)$ is not a fundamental solution. Then Σ can be written as a Haken sum of fundamental surfaces,

$$\Sigma = \Sigma_1 \oplus \cdots \oplus \Sigma_a \oplus S_1 \oplus \cdots \oplus S_b,$$

where each Σ_i is a properly embedded compact surface with boundary, and each S_j is a properly embedded closed surface.

Since Σ is π_1–essential, it follows from Jaco and Sedgwick [14, Section 3.3] that that each Σ_i must have the same slope as Σ. Since Σ is a spanning surface, then a must equal 1, and Σ_1 is also a spanning surface at the same slope as Σ. In fact, Σ_1 must be fundamental.

Since X is irreducible and embedded in S^3, it follows that $\chi(S_j) \leq 0$ for each $j = 1, \ldots, b$. Suppose that for some k, we have $\chi(S_k) \leq -2$. But then $\chi(\Sigma) < \chi(\Sigma_1)$ which implies that

$$\chi(\Sigma_1) + \chi(\Sigma') + \tfrac{1}{2} i(\partial\Sigma_1, \partial\Sigma') > 2.$$

Hence, by Lemma 5.1, at least one of Σ_1 or Σ' must be disconnected. Every fundamental surface is connected, so Σ' must be disconnected, which gives a contradiction.

Hence $\chi(S_i) = 0$ for each i, and every S_i is an embedded torus. Moreover, $\chi(\Sigma) = \chi(\Sigma_1)$. Similarly,

$$\Sigma' = \Sigma'_1 \oplus S'_1 \oplus \cdots \oplus S'_c,$$

where Σ'_1 is fundamental spanning surface with the same slope and Euler characteristic as Σ', and S'_j is an embedded torus for each $j = 1, \ldots, c$. Therefore Σ_1 and Σ'_1 are fundamental spanning surfaces which satisfy ($\star$).

Let $\mathcal{F}$ be the set of all fundamental spanning surfaces in X. For each pair of surfaces $\Sigma, \Sigma' \in \mathcal{F}$, calculate the intersection number $i(\partial\Sigma, \partial\Sigma')$, and calculate $\chi(\Sigma)$ and $\chi(\Sigma')$. There is an algorithm to compute the Euler characteristic of a properly embedded normal surface described in [15, Algorithm 9.1]. If Σ and Σ' satisfy ($\star$), then K is alternating by Theorem 3.1. If no pair of surfaces from $\mathcal{F}$ satisfy ($\star$), then K is not alternating. □

Let $\pi(K)$ be an alternating diagram of the prime knot K with associated checkerboard surfaces Σ and Σ'. Let $\pi_*(K)$ be a different alternating diagram of K with associated checkerboard surfaces Σ_* and Σ'_*. If $\pi(K)$ and $\pi_*(K)$ are both reduced, then we know from a theorem of Menasco and Thistlethwaite [18] that $\pi(K)$ and $\pi_*(K)$ are related by a sequence of flypes. In that case, Σ and Σ_* are homeomorphic and have the same boundary slope, but Σ and Σ_* may not be isotopic in X. The same is true for Σ' and Σ'_*.

However, every checkerboard surface for a reduced alternating diagram is π_1–essential, and thus will appear amongst our collection of fundamental spanning surfaces $\mathcal{F}$. The collection $\mathcal{F}$ may also contain some pairs of surfaces which correspond to an alternating diagram which is not reduced. In this case, at least one of the checkerboard surfaces fails to be π_1–essential.

Let Σ and Σ' have minimal intersection number amongst all surfaces from $\mathcal{F}$ which satisfy ($\star$). Place an orientation on $\partial\Sigma$, and label the vertices of $\partial\Sigma \cap \partial\Sigma'$ in the order they are encountered as one traverses $\partial\Sigma$ by $1, \ldots, i$, where $i = i(\partial\Sigma, \partial\Sigma')$. Then each arc of intersection between Σ and Σ' is labelled by two numbers, one even and one odd. These pairs of numbers, listed as a sequence of even positive integers in the order of their paired odd numbers, correspond to the Dowker–Thistlethwaite notation [4] of a planar alternating diagram of K or its mirror image.

Therefore, given a nonalternating planar diagram of a knot K, there is an algorithm to decide if K is alternating, and if so, there is an algorithm to produce an alternating diagram of K up to chirality.

References

[1] **C Adams**, **T Kindred**, *A classification of spanning surfaces for alternating links*, Algebr. Geom. Topol. 13 (2013) 2967–3007 MR

[2] **R J Aumann**, *Asphericity of alternating knots*, Ann. of Math. 64 (1956) 374–392 MR

[3] **C L Curtis**, **S J Taylor**, *The Jones polynomial and boundary slopes of alternating knots*, J. Knot Theory Ramifications 20 (2011) 1345–1354 MR

[4] **C H Dowker**, **M B Thistlethwaite**, *Classification of knot projections*, Topology Appl. 16 (1983) 19–31 MR

[5] **C M Gordon**, *Links and their complements*, from "Topology and geometry: commemorating SISTAG" (A J Berrick, M C Leung, X Xu, editors), Contemp. Math. 314, Amer. Math. Soc., Providence, RI (2002) 71–82 MR

[6] **C M Gordon**, **J Luecke**, *Knots are determined by their complements*, J. Amer. Math. Soc. 2 (1989) 371–415 MR

[7] **J E Greene**, *Alternating links and definite surfaces*, preprint (2015) arXiv To appear in Duke Math. J.

[8] **W Haken**, *Theorie der Normalflächen*, Acta Math. 105 (1961) 245–375 MR

[9] **A E Hatcher**, *On the boundary curves of incompressible surfaces*, Pacific J. Math. 99 (1982) 373–377 MR

[10] **J A Howie**, *Surface-alternating knots and links*, PhD thesis, University of Melbourne (2015) Available at `http://hdl.handle.net/11343/58991`

[11] **J A Howie**, **J H Rubinstein**, *Weakly generalised alternating links* In preparation

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

[13] **W Jaco**, **J H Rubinstein**, *Inflations of ideal triangulations*, Adv. Math. 267 (2014) 176–224 MR

[14] **W Jaco**, **E Sedgwick**, *Decision problems in the space of Dehn fillings*, Topology 42 (2003) 845–906 MR

[15] **W Jaco**, **J L Tollefson**, *Algorithms for the complete decomposition of a closed 3–manifold*, Illinois J. Math. 39 (1995) 358–406 MR

[16] **W B R Lickorish**, *An introduction to knot theory*, Graduate Texts in Mathematics 175, Springer, New York (1997) MR

[17] **W Menasco**, *Closed incompressible surfaces in alternating knot and link complements*, Topology 23 (1984) 37–44 MR

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

[19] **T Nowik**, *Dissecting the 2–sphere by immersions*, Geom. Dedicata 127 (2007) 37–41 MR

[20] **C Petronio**, *Ideal triangulations of link complements and hyperbolicity equations*, Geom. Dedicata 66 (1997) 27–50 MR

[21] **J H Rubinstein**, *An algorithm to recognize the 3–sphere*, from "Proceedings of the International Congress of Mathematicians" (S D Chatterji, editor), volume 1, Birkhäuser, Basel (1995) 601–611 MR

[22] **A Thompson**, *Thin position and the recognition problem for* S^3, Math. Res. Lett. 1 (1994) 613–630 MR

School of Mathematics and Statistics, University of Melbourne
Melbourne VIC 3010, Australia

`josh.howie@monash.edu`

Proposed: Cameron Gordon Received: 5 February 2016
Seconded: Ciprian Manolescu, András I. Stipsicz Accepted: 19 September 2016

Geometry & Topology 21 (2017) 2373–2418

Homology of FI-modules

THOMAS CHURCH
JORDAN S ELLENBERG

We prove an explicit and sharp upper bound for the Castelnuovo–Mumford regularity of an FI-module in terms of the degrees of its generators and relations. We use this to refine a result of Putman on the stability of homology of congruence subgroups, extending his theorem to previously excluded small characteristics and to integral homology while maintaining explicit bounds for the stable range.

18G10, 20C30

1 Introduction

In recent years, there has been swift development in the study of various abelian categories related, in one way or another, to stable representation theory; see Church, Ellenberg and Farb [4], Church, Ellenberg, Farb and Nagpal [5], Sam and Snowden [14] and Wiltshire-Gordon [16]. The simplest of these is the category of *FI-modules* introduced in [4], which can be seen as a category of modules for a certain twisted commutative algebra. A critical question about these categories is whether they are *noetherian*; that is, whether a subobject of a finitely generated object is itself finitely generated.[1]

The category of FI-modules over $\mathbb{Z}$ is noetherian [5, Theorem A], so any finitely generated FI-module V can be resolved by finitely generated projectives. One can ask for more — in the spirit of the notion of Castelnuovo–Mumford regularity from commutative algebra, one can ask for a resolution of V whose terms have explicitly bounded degree. Castelnuovo–Mumford regularity has proven to be a very useful invariant in commutative algebra, and we expect the same to be the case in this twisted commutative setting. In the present paper, we prove a strong bound for the Castelnuovo–Mumford regularity of FI-modules, and explain how this regularity theorem allows us to refine a result of Putman [12] on the homology of congruence subgroups. Although much of the paper is homological-algebraic in nature, the heart of the main results is Theorem E; this is a basic structure theorem for FI-modules, whose proof at the core

[1] In some contexts, such abelian categories are called "locally noetherian", the term "noetherian" being reserved for categories where *every* object is noetherian. We use "noetherian" here in the broader sense, but we acknowledge that not every FI-module is finitely generated.

Published: 19 May 2017 DOI: 10.2140/gt.2017.21.2373

boils down to a combinatorial argument on injections from $[d]$ to $[n]$ involving certain sets of integers enumerated by the Catalan numbers.

The theorems we obtain with these combinatorial methods naturally hold for FI-modules with coefficients in $\mathbb{Z}$. This is in contrast with earlier representation-theoretic approaches, which tend to apply only to FI-modules with coefficients in a field, usually required to have characteristic 0. On the other hand, the approach via representation theory provides a very beautiful theory unifying the study of many different categories (see eg Sam and Snowden [13]), while the arguments of the present paper are quite specific to FI-modules. It would be very interesting to understand the extent to which the combinatorics in Section 3 can be generalized beyond FI-modules to the family of stable representation categories considered by Sam and Snowden.

Notation FI is the category of finite sets and injections; an *FI-module* W is a functor $W\colon \mathrm{FI} \to \mathbb{Z}$–Mod. Given a finite set T, we write W_T for $W(T)$. For every $n \in \mathbb{N} = \{0, 1, 2, \dots\}$, we set $[n] := \{1, \dots, n\}$, and we write W_n for $W_{[n]} = W([n])$.

When W is an FI-module, we write $\deg W$ for the largest $k \in \mathbb{N}$ such that $W_k \neq 0$. To include edge cases such as $W = 0$, we formally define $\deg W \in \{-\infty\} \cup \mathbb{N} \cup \{\infty\}$ by

$$\deg W := \inf\bigl\{k \in \{-\infty\} \cup \mathbb{N} \cup \{\infty\} \bigm| W_n = 0 \text{ for all } n > k \in \mathbb{N}\bigr\}.$$

FI-homology The functor $H_0\colon$ FI-Mod $\to$ FI-Mod captures the notion of "minimal generators" for an FI-module. Given an FI-module W, the FI-module $H_0(W)$ is the quotient of W defined by

$$H_0(W)_T := W_T \,/\, \mathrm{span}(\mathrm{im}\, f_*\colon W_S \to W_T \mid f\colon S \hookrightarrow T,\ |S| < |T|).$$

This is the largest FI-module quotient of W such that all maps $f_*\colon H_0(W)_S \to H_0(W)_T$ with $|S| < |T|$ are zero. An FI-module W is *generated in degree at most m* if $\deg H_0(W) \leq m$.

The functor H_0 is right exact, and we define $H_p\colon$ FI-Mod $\to$ FI-Mod to be its p^{th} left-derived functor. One can think of $H_p(W)$ as giving minimal generators for the "p^{th} syzygy" of the FI-module W. Our first main theorem bounds $H_p(W)$ in terms of $H_0(W)$ and $H_1(W)$.

Theorem A *Let W be an FI-module with $\deg H_0(W) \leq k$ and $\deg H_1(W) \leq d$. Then W has regularity at most $k + d - 1$: that is, for all $p > 0$, we have*

$$\deg H_p(W) \leq p + k + d - 1.$$

It is natural to shift our indexing by writing $d_p(W) := \deg H_p(W) - p$; with this indexing, Theorem A states simply that $d_p(W) \leq d_0(W) + d_1(W)$.

We will define in Section 2.1 the notion of a *free FI-module*, and we will see that $H_p(W)$ can be computed explicitly from a free resolution of W. For now, we record one corollary:

Corollary B *Let M be a free FI-module generated in degree at most k, and let V be an arbitrary FI-module generated in degree at most d. For any homomorphism $V \to M$, the kernel $\ker(V \to M)$ is generated in degree at most $k+d+1$.*

Uniform description of an FI-module Our next main result is the following theorem, which gives a uniform description of an FI-module in terms of a explicit finite amount of data.

Theorem C *Let W be an arbitrary FI-module, and define*

$$N := \max(\deg H_0(W), \deg H_1(W)).$$

Then for any finite set T,

$$W_T = \operatorname*{colim}_{\substack{S\subset T \\ |S|\le N}} W_S. \tag{1}$$

Moreover, N is the smallest integer such that (1) *holds for all finite sets.*

We deduce Theorem C from [5, Corollary 2.24] by showing that the complex $\widetilde{S}_{-*}W$ we introduced there computes the FI-homology $H_*(W)$. An alternate proof of Theorem C has recently been given by Gan and Li [8]; in contrast with our approach via FI-homology, they prove directly that an FI-module that is presented in finite degree admits a description as in (1).

Homology of congruence subgroups As an application of these theorems, we have the following result on the homology of congruence subgroups, which strengthens a recent theorem of Putman [12]. For $L \neq 0 \in \mathbb{Z}$, let $\Gamma_n(L)$ be the level-L principal congruence subgroup

$$\Gamma_n(L) := \ker\big(\mathrm{GL}_n(\mathbb{Z}) \to \mathrm{GL}_n(\mathbb{Z}/L\mathbb{Z})\big).$$

For $S \subset [n]$, let $\Gamma_S(L) \subset \Gamma_n(L)$ be the subgroup

$$\Gamma_S(L) := \{M \in \Gamma_n(L) \mid M_{ij} = \delta_{ij} \text{ if } i \notin S \text{ or } j \notin S\}.$$

Notice that if $|S| = m$, the subgroup $\Gamma_S(L)$ is isomorphic to $\Gamma_m(L)$.

Theorem D *For all $L \neq 0 \in \mathbb{Z}$, all $n \geq 0$, and all $k \geq 0$,*

$$H_k(\Gamma_n(L);\mathbb{Z}) = \operatorname*{colim}_{\substack{S\subset[n] \\ |S|<11\cdot 2^{k-2}}} H_k(\Gamma_S(L);\mathbb{Z}).$$

In fact, we prove a version of Theorem D for any ring satisfying one of Bass's stable range conditions; see Theorem D$'$ in Section 5.2. This theorem has already been used by Calegari and Emerton [2, Section 5] to prove stability for the completed homology of arithmetic groups.

The conclusion of Theorem D is based on the main result of Putman in [12] on "central stability" for $H_k(\Gamma_n(M);\mathbb{Z})$, but its formulation here is a combination of [12, Theorem B] and our earlier theorem with Farb and Nagpal [5, Theorem 1.6]. Our main improvement over Putman is that Theorem D applies to homology with integral coefficients (or any other coefficients), while [12] only applied to coefficients in a field of characteristic at least $2^{k-2}\cdot 18-3$. This limitation was removed in Church, Ellenberg, Farb and Nagpal [5], but at the cost of losing any hope of an explicit stable range. The methods of the present paper maintain the applicability to arbitrary coefficients while recovering Putman's stable range.

Ingredients of Theorem D In light of Theorem C, in order to obtain the conclusion of Theorem D, we must bound the degree of H_0 and H_1 for the FI-module $\mathcal{H}_k$ satisfying $(\mathcal{H}_k)_n = H_k(\Gamma_n(L);\mathbb{Z})$. The key technical ingredients are Theorem A and a theorem of Charney on a congruence version of the complex of partial bases. We obtain in Proposition 5.13 a spectral sequence with $E^2_{pq} = H_p(\mathcal{H}_q)$. Charney's theorem tells us that this spectral sequence converges to zero in an appropriate sense, and Theorem A then lets us work backward to conclude that E^2_{pq} vanishes outside the corresponding range, giving the desired bound on the degree of $H_0(\mathcal{H}_q)$ and $H_1(\mathcal{H}_q)$.

Remark The argument of Theorem D bears an interesting resemblance to that of the second author with Venkatesh and Westerland in [6]. In that paper, one proves a stability theorem for the cohomology of Hurwitz spaces, using the fact that this cohomology carries the structure of module for a certain graded $\mathbb{Q}$–algebra R. As in the present paper (indeed most stable cohomology theorems), the topological side of the argument requires proving that a certain complex, carrying an action of the group whose cohomology we wish to control, is approximately contractible. The algebraic piece of [6] involves showing that $\deg \operatorname{Tor}_i^R(M,\mathbb{Q})$ can be bounded in terms of $\deg \operatorname{Tor}_0^R(M,\mathbb{Q})$ and $\deg \operatorname{Tor}_1^R(M,\mathbb{Q})$ [6, Proposition 4.10]. Exactly as in the proof of Theorem D, it is these bounds that allow us to carry out an induction in the spectral sequence arising from the quotient of the highly connected complex by the group of interest.

Combinatorial structure of FI-modules Our last theorem is a basic structural property of FI-modules; this structural theorem provides the technical foundation for our other results, and is also of independent interest in its own right.

An FI-module M is *torsion-free* if for every injection $f\colon S \hookrightarrow T$ between finite sets, the map $f_*\colon M_S \to M_T$ is injective. In this case, for any subset $S \subset T$, we may regard M_S as a submodule of M_T by identifying it with its image under the canonical inclusion.

Theorem E *Let M be a torsion-free FI-module generated in degree at most k, and let $V \subset M$ be a sub-FI-module generated in degree at most d. Then for all $n > \min(k,d)+d$ and any $a \le n$,*

$$V_n \cap \bigl(M_{[n]-\{1\}} + \cdots + M_{[n]-\{a\}}\bigr) = V_{[n]-\{1\}} + \cdots + V_{[n]-\{a\}}.$$

Theorem E holds for any $d \ge 0$ and $k \ge 0$. However, in the cases of primary interest, we will have $k < d$, so in practice, the threshold for Theorem E will be $n > k+d$. We note also that Theorem E is trivially true for $a > d$: the inclusion

$$V_{[n]-\{1\}} + \cdots + V_{[n]-\{a\}} \subset V_n \cap \bigl(M_{[n]-\{1\}} + \cdots + M_{[n]-\{a\}}\bigr)$$

always holds, and it is easy to show that $V_{[n]-\{1\}} + \cdots + V_{[n]-\{a\}} = V_n$ when $a > d$.

Stating Theorem E without M Although Theorem E seems to be a theorem about the relation between the FI-module M and its submodule V, actually the key object is V; the role of M is somewhat auxiliary. In fact, in Section 3.1, we will give a more general formulation in Theorem E$'$ that makes no reference to M; in place of the intersection $V_n \cap (M_{[n]-\{1\}} + \cdots + M_{[n]-\{a\}})$, we use the subspace of V_n annihilated by the operator $\prod_{i=1}^{a}(\mathrm{id} - (i\ \ n{+}i)) \in \mathbb{Z}[S_{n+a}]$; see Section 3.1 for more details. Theorem E$'$ is stronger than Theorem E and has content even in the case corresponding to $V = M$, when Theorem E says nothing. Aside from their application to FI-homology in this paper, these results are fundamental properties of the structure of FI-modules, and should be of interest on their own.

Theorem E and homology We will show that if M is a free FI-module, Theorem E has a natural homological interpretation as a bound on the degree of vanishing of a certain derived functor applied to M/V; see Remark 2.7 and Corollary 4.5 for details. It is this interpretation that allows us to connect Theorem E with the bounds on regularity in Theorem A.

Bounds on torsion The conclusion of Theorem E can be phrased as a statement about the quotient FI-module M/V, and in the case $a = 1$, this conclusion becomes particularly simple: it states that the map $(M/V)_{[n]-\{1\}} \to (M/V)_{[n]}$ is injective when $n > \min(k,d)+d$. This yields the following corollary of Theorem E. In general, an FI-module W is *torsion-free in degrees at least m* if the maps $f_*\colon M_S \to M_T$ are injective whenever $|S| \ge m$.

Corollary F *If M is torsion-free and generated in degree at most k, and its sub-FI-module V is generated in degree at most d, then the quotient M/V is torsion-free in degrees at least* $\min(k,d)+d$.

Alternate proofs of Theorem A In the time since this paper was first posted, alternate proofs of Theorem A have been given by Li [10] (based partly on Li and Yu [11]) and Gan [7]. The structure of those proofs is different from ours. In this paper, we prove Theorem E in a self-contained way and then deduce Theorem A as a direct consequence. By contrast, both Li and Gan use Corollary F as a stepping stone (replacing the need for the full strength of Theorem E); they prove both Theorem A and Corollary F together, using an induction on k that bounces back and forth between those two results.

Sharpness of Theorem E Before moving on, we give a simple example showing that the bound of $\min(k,d)+d$ in Theorem E and Corollary F is sharp.

Example Fix any $k \geq 0$ and any $d > k$. Let M be the FI-module over $\mathbb{Q}$ such that M_T is freely spanned by the k–element subsets of T. The FI-module M is torsion-free and generated in degree k by $M_k \simeq \mathbb{Q}$.

For any d–element set U, consider the element $v_U := \sum_{S\subset U,\,|S|=k} e_S$. Let $V \subset M$ be the sub-FI-module such that V_T is spanned by the elements $v_U \in M_T$ for all d–element subsets $U \subset T$. The FI-module V is generated by $v_{[d]} \in V_d \simeq \mathbb{Q}$, so Corollary F asserts that the quotient $W := M/V$ should be torsion-free in degrees at least $k+d$.

In fact, we have $W_n \neq 0$ for $n < k+d$ and $W_n = 0$ for $n \geq k+d$, which we can verify as follows. By definition, V_n is spanned by the $\binom{n}{d}$ elements v_U as U ranges over the d–element subsets $U \subset [n]$, so the dimension of V_n is at most $\binom{n}{d}$. When $n < k+d$, we have $\dim V_n \leq \binom{n}{d} < \binom{n}{k} = \dim M_n$ so $V_n \neq M_n$, verifying the first claim. On the other hand, with a bit of work, one can check directly that $V_{k+d} = M_{k+d}$, which then implies $V_n = M_n$ for all $n \geq k+d$, verifying the second claim.

Since $W_n = 0$ for $n \geq k+d$, we see that W is torsion-free in degrees at least $k+d$ as guaranteed by Corollary F; however, the fact that $W_{k+d-1} \neq 0$ shows that this bound cannot be improved.

Castelnuovo–Mumford regularity What we prove in Theorem A is that

$$\deg H_p(V) \leq c_V + p$$

for some constant c_V depending on V. By analogy with commutative algebra, this statement could be thought of as saying that the *Castelnuovo–Mumford regularity* of V

is at most c_V. For FI-modules over fields of characteristic 0, that all finitely generated FI-modules have finite Castelnuovo–Mumford regularity in this sense is a recent result of Sam and Snowden [14, Corollary 6.3.4].

We emphasize that Theorem A gives an explicit description of the regularity of V which depends only on the *degrees* of generators and relations for V. This is much stronger than the bounds for the regularity of finitely generated modules M over polynomial rings $\mathbb{C}[x_1, \dots, x_r]$, which depend on the *number* of generators of M. We take this strong bound on regularity as support for the point of view that the category of FI-modules is in some sense akin to the category of graded modules for a univariate polynomial ring $\mathbb{C}[T]$. (See Table 1 for more details of this analogy.) Of course, in the latter context, the fact that the regularity is bounded by the degree of generators and relations is a triviality because $H_p(V) = 0$ for all $p > 1$; by contrast, the category of FI-modules has infinite global dimension.

Despite these analogies, we would like to emphasize one surprising feature of the bound

$$\deg H_p(V) - p \leq \deg H_0(V) + \deg H_1(V) - 1$$

we obtain for FI-modules: one cannot expect a bound of this form to hold for graded modules over a general graded ring, for the simple reason that the bound is not invariant under shifts in grading.[2] The existence of such a bound for FI-modules reflects the fact that, although a version of the grading shift does exist for FI-modules (see Section 2.2), its effect on generators and relations is considerably more complicated. In particular, this shift is not *invertible* for FI-modules.

Infinitely generated FI-modules One striking feature of Theorem A, and another contrast with polynomial rings, is that its application is not restricted to finitely generated FI-modules: Theorem A bounds the regularity of *any* FI-module which is presented in finite degree. This is critical for the applications to homology of congruence subgroups in Section 5.2: for congruence subgroups such as

$$\Gamma_n(t) = \ker\big(\mathrm{GL}_n(\mathbb{C}[t]) \to \mathrm{GL}_n(\mathbb{C})\big),$$

the FI-modules arising from the homology of $\Gamma_n(t)$ are not even countably generated! Nevertheless, the bounds in Theorem D$'$ below apply equally well to this case.

[2] Indeed, since $\deg H_i(V[k]) = \deg H_i(V) - k$ for graded modules, applying this inequality to $V[k]$ rather than V would shift the left side by $-k$ but the right side by $-2k$, leading to the absurd conclusion that $\deg H_p(V) - p \leq \deg H_0(V) + \deg H_1(V) - 1 - k$ for any k. This is impossible unless $\deg H_p(V) = -\infty$ for $p > 1$, meaning the ring has homological dimension 1 (as we saw for $\mathbb{C}[T]$ above).

Acknowledgements The authors gratefully acknowledge support from the National Science Foundation. Church's work was supported in part by NSF grants DMS-1103807 and DMS-1350138, and by the Alfred P Sloan Foundation. Ellenberg's work was supported in part by NSF grants DMS-1101267 and DMS-1402620, a Romnes Faculty Fellowship, and a John Simon Guggenheim Fellowship.

The authors are very grateful to the University of Copenhagen for its hospitality in August 2013, where the first version of the main results of this paper were completed. We thank Jennifer Wilson for many helpful comments on an earlier draft of this paper, and Eric Ramos, Jens Reinhold and Emily Riehl for helpful conversations. We thank Wee Liang Gan and Liping Li for conversations regarding their results in [8] and [9]. We are very grateful to two anonymous referees for their careful reading and thoughtful suggestions, which improved this paper greatly, and especially for comments which led us to the statement of Theorem E$'$ in Section 3.1.

2 Summary of FI-modules

In this introductory section, we record the basic definitions and properties of FI-modules that we will use in this paper. Experts who are already familiar with FI-modules can likely skip Section 2 on a first reading (with the exception of Lemma 2.6 and Remark 2.7, which are less standard and play a key role in later sections).

As we mentioned in the introduction, there is a productive analogy between FI-modules and graded $\mathbb{C}[T]$–modules. For the benefit of readers unfamiliar with FI-modules, in Table 1 we have listed all the constructions for FI-modules described in this section, along with the analogous construction for $\mathbb{C}[T]$–modules. These analogies are not intended as precise mathematical assertions, only as signposts to help the reader orient themself in the world of FI-modules.

(Those readers used to the six-functors formalism may prefer to dualize the right side of Table 1, thinking of $\varphi\colon \mathrm{FB} \hookrightarrow \mathrm{FI}$ as analogous to the structure map $f\colon \operatorname{Spec}\mathbb{C}[T] \to \operatorname{Spec}\mathbb{C}$, so that the adjoint functors $M \leftrightarrows \varphi^*$ correspond to $f^{-1} \leftrightarrows f_*$. Similarly, $\pi\colon \mathbb{Z}\,\mathrm{FI} \twoheadrightarrow \mathbb{Z}\,\mathrm{FB}$ is analogous to the closed inclusion $i\colon \operatorname{Spec}\mathbb{C} \to \operatorname{Spec}\mathbb{C}[T]$, and the adjunctions $\lambda \leftrightarrows \pi^* \leftrightarrows \rho$ correspond to $i^{-1} \leftrightarrows i_* = i_! \leftrightarrows i^!$.)

2.1 Free FI-modules and generation

FB-modules Just as FI denotes the category of finite sets and injections, FB denotes the category of finite sets and bijections. An FB-module W is an element of FB-Mod, the abelian category of functors $W\colon \mathrm{FB} \to \mathbb{Z}$–Mod. An FB-module W is just a sequence W_n of $\mathbb{Z}[S_n]$–modules, with no additional structure.

FI	a category	$\mathbb{C}[T]$	an algebra over
FB $\subset$ FI	its subcat. of isos	$\mathbb{C} \subset \mathbb{C}[T]$	a field
φ: FB $\hookrightarrow$ FI		$\mathbb{C} \hookrightarrow \mathbb{C}[T]$	
φ^*: FI-Mod $\to$ FB-Mod	forget action of nonisos	$\mathbb{C}[T]$–Mod $\to \mathbb{C}$–Mod	take underlying vector space
M: FB-Mod $\to$ FI-Mod	left adjoint of φ^* = "free FI-mod on W"	$\mathbb{C}$–Mod $\to \mathbb{C}[T]$–Mod	$V \mapsto \mathbb{C}[T] \otimes_{\mathbb{C}} V$ = free $\mathbb{C}[T]$–Mod on V
π: $\mathbb{Z}$ FI $\twoheadrightarrow \mathbb{Z}$ FB	nonisos $\mapsto 0$	$\mathbb{C}[T] \twoheadrightarrow \mathbb{C}$	$T \mapsto 0$
π^*: FB-Mod $\to$ FI-Mod	make nonisos act by 0	$\mathbb{C}$–Mod $\to \mathbb{C}[T]$–Mod	make T act by 0
λ: FI-Mod $\to$ FB-Mod	left adjoint of π^* (extend scalars by π)	$\mathbb{C}[T]$–Mod $\to \mathbb{C}$–Mod	$M \mapsto M/TM$ $= M \otimes_{\mathbb{C}[T]} \mathbb{C}$
H_0: FI-Mod $\to$ FI-Mod	$H_0 := \pi^* \circ \lambda$	M/TM considered as $\mathbb{C}[T]$–Mod	
ρ: FI-Mod $\to$ FB-Mod	right adjoint of π^*	$\mathbb{C}[T]$–Mod $\to \mathbb{C}$–Mod	$M \mapsto M[T]$ $= \ker(M \xrightarrow{T} M[1])$
S: FI-Mod $\to$ FI-Mod	$(SW)_T := W_{T \sqcup \{\star\}}$	grading shift $M \mapsto M[1]$	
D: FI-Mod $\to$ FI-Mod	$\operatorname{coker}(W \to SW)$	$\operatorname{coker}(M \xrightarrow{T} M[1])$	
K: FI-Mod $\to$ FI-Mod	$\ker(W \to SW)$ $= \pi^* \circ \rho$	$\ker(M \xrightarrow{T} M[1])$	

Table 1: Analogies between FI-modules and graded $\mathbb{C}[T]$–modules

Free FI-modules FB is the subcategory of FI consisting of all the isomorphisms (its maximal subgroupoid). From the inclusion φ: FB $\hookrightarrow$ FI, we obtain a natural forgetful functor φ^* from FI-Mod to FB-Mod that simply forgets about the action of all nonisomorphisms. Its left adjoint M: FB-Mod $\to$ FI-Mod takes an FB-module W to the "free FI-module $M(W)$ on W". We call any FI-module of the form $M(W)$ a *free FI-module*.

We recall from [4, Definition 2.2.2] an explicit formula for $M(W)$, from which we can see that M is exact:

$$M(W)_T = \bigoplus_{S \subset T} W_S. \tag{2}$$

For notational convenience, for $m \in \mathbb{N}$ we write $M(m) := M(\mathbb{Z}[S_m])$. These FI-modules have the defining property that $\mathrm{Hom}_{\text{FI-Mod}}(M(m), V) \simeq V_m$, since we can write $M(m) \simeq \mathbb{Z}\big[\mathrm{Hom}_{\mathrm{FI}}([m], -)\big]$.

As a consequence, $M(m)$ is a projective FI-module (they are the "principal projective" FI-modules). In general, an FI-module is projective if and only if it is a summand of some $\bigoplus_{i \in I} M(m_i)$.

We point out that despite the name, free FI-modules need not be projective (since nonprojective $\mathbb{Z}[S_n]$–modules are in abundance!).[3] Nevertheless, for our purposes free FI-modules will be just as good as projective FI-modules (see Lemma 2.3 and Corollary 4.5), so this discrepancy will not bother us.

Generation in degree at most k Every FI-module V has a natural increasing filtration

$$V_{\langle \leq 0\rangle} \subset V_{\langle \leq 1\rangle} \subset \cdots \subset V_{\langle \leq m\rangle} \subset \cdots \subset V = \bigcup_{m\geq 0} V_{\langle \leq m\rangle},$$

where $V_{\langle \leq m\rangle}$ is the sub-FI-module of V "generated by elements in degree at most m". This filtration, which is respected by all maps of FI-modules, can be defined as follows.

Given an FI-module V, by a slight abuse of notation we write $M(V)$ for the free FI-module on the FB-module $\varphi^* V$ underlying V. From the adjunction $M \leftrightarrows \varphi^*$ we have a canonical map $M(V) \twoheadrightarrow V$, which is always surjective. We modify this slightly to define the filtration $V_{\langle \leq m\rangle}$.

Definition 2.1 Let $V_{\leq m}$ be the FB-module defined by $(V_{\leq m})_T = V_T$ if $|T| \leq m$ and $(V_{\leq m})_T = 0$ if $|T| > m$. Then the natural inclusion of FB-modules $V_{\leq m} \hookrightarrow \varphi^* V$ induces a map of FI-modules $M(V_{\leq m}) \to V$.

We define $V_{\langle \leq m\rangle} \subset V$ to be the image of the canonical map $M(V_{\leq m}) \to V$. Equivalently, $V_{\langle \leq m\rangle}$ is the smallest sub-FI-module $U \subset V$ satisfying $U_n = V_n$ for all $n \leq m$. We sometimes write $V_{\langle < m\rangle}$ as an abbreviation for $V_{\langle \leq m-1\rangle}$.

In the introduction we said that an FI-module W is *generated in degree at most m* if $\deg H_0(W) \leq m$, but there are many equivalent ways to formulate this definition.

Lemma 2.2 *Let V be an FI-module, and fix $m \geq 0$. The following are equivalent:*

(i) *V is generated in degree at most m.*

(ii) $\deg H_0(V) \leq m$.

(iii) $V = V_{\langle \leq m\rangle}$.

(iv) *V admits a surjection from $\bigoplus_{i\in I} M(m_i)$ with all $m_i \leq m$.*

(v) *The natural map $M(W_m) \to W$ is surjective in degrees at least m.*

[3]For example, consider the FI-module V for which $V_T := \mathbb{Z}[e_{\{i,j\}}]_{i\neq j\in T}$, the free abelian group on the 2–element subsets of T. This FI-module V is free on the FB-module W having $W_2 \simeq \mathbb{Z}$ (the trivial $\mathbb{Z}[S_2]$–module) and $W_n = 0$ for $n \neq 2$. However, V is not projective, since W_2 is not a projective $\mathbb{Z}[S_2]$–module.

The functor H_0 In the other direction, we do not quite have a projection from FI to FB, because the noninvertible morphisms in FI have nowhere to go. One wants to map them to zero, but there's no "zero morphism" in FB. This problem can be solved by passing to the $\mathbb{Z}$–enriched versions $\mathbb{Z}\,\mathrm{FI}$ and $\mathbb{Z}\,\mathrm{FB}$ of these categories: we now have a functor $\pi\colon \mathbb{Z}\,\mathrm{FI} \to \mathbb{Z}\,\mathrm{FB}$ which sends all noninvertible morphisms to zero and is the identity on isomorphisms.

This induces the "extension by zero" functor $\pi^*\colon$ FB-Mod $\to$ FI-Mod which takes an FB-module W and simply regards it as an FI-module by defining $f_* = 0$ for all noninvertible $f\colon S \to T$. This functor is exact and has both a left adjoint λ and right adjoint ρ.

In this section, we consider the left adjoint λ: FI-Mod $\to$ FB-Mod. Since every noninvertible map $f\colon S \hookrightarrow T$ increases cardinality, we have the formula $(\lambda V)_n = (V/V_{\langle <n\rangle})_n$. This is almost exactly the definition of H_0: FI-Mod $\to$ FI-Mod given in the introduction; the only difference is that λV is an FB-module whereas $H_0(V)$ is the same thing regarded as an FI-module, ie $H_0 = \pi^* \circ \lambda$.

We adopt the convention in this paper that if F is a right-exact functor, H_p^F denotes its p^{th} left-derived functor. As we explained in the introduction, we write $H_p(V)$ for the derived functors $H_p^{H_0}(V)$ of H_0, and call these the *FI-homology* of V.

Lemma 2.3 *Free FI-modules are H_0–acyclic.*

Proof Our goal is to prove that $H_p(M(W)) = 0$ for $p > 0$. Since M is exact and takes projectives to projectives, there is an isomorphism $H_p(M(W)) \simeq H_p^{H_0\circ M}(W)$. However, the composition $H_0 \circ M$ is just the exact functor π^*. Indeed, the composition $\pi \circ \varphi$ is the identity, so $\varphi^* \circ \pi^* = \mathrm{id}$. It follows that its left adjoint $\lambda \circ M$ is the identity as well. Since $H_0 = \pi^* \circ \lambda$, we have $H_0 \circ M = \pi^* \circ \lambda \circ M = \pi^*$ as claimed. We conclude that $H_p(M(W)) \simeq H_p^{H_0\circ M}(W) = H_p^{\pi^*}(W)$, which vanishes for $p > 0$ since π^* is exact. □

We can now explain how Corollary B follows from Theorem A.

Proof of Corollary B If $M = 0$, the corollary is trivial, so assume that $M \neq 0$. Let $K = \ker(V \to M)$ and $W = \operatorname{coker}(V \to M)$. Thanks to the equivalences in Lemma 2.2, the statement of the corollary is that $\deg H_0(K) \le \deg H_0(M) + \deg H_0(V) + 1$.

From the exact sequence $0 \to K \to V \to M \to W \to 0$, we obtain the inequalities

$$\begin{aligned} &\deg H_0(W) \le \deg H_0(M), \\ &\deg H_1(W) \le \max\bigl(\deg H_0(V),\ \deg H_1(M)\bigr), \\ &\deg H_0(K) \le \max\bigl(\deg H_0(V),\ \deg H_1(M),\ \deg H_2(W)\bigr). \end{aligned}$$

Therefore, to prove the corollary, it suffices to show that the degrees of $H_0(V)$, $H_1(M)$ and $H_2(W)$ are bounded by $\deg H_0(M) + \deg H_0(V) + 1$. For $H_0(V)$, this is trivial, since $\deg H_0(M) \geq 0$. Since M is free, $H_1(M) = 0$ by Lemma 2.3, so $\deg H_1(M) = -\infty$; this also shows $\deg H_1(W) \leq \deg H_0(V)$. Finally, applying Theorem A to W shows that, as desired,

$$\deg H_2(W) \leq \deg H_0(W) + \deg H_1(W) + 1 \leq \deg H_0(M) + \deg H_0(V) + 1. \qquad \square$$

2.2 Shifts and derivatives of FI-modules

The shift functor S Fix a one-element set $\{\star\}$. Let $\sqcup\colon \mathrm{Sets}\times\mathrm{Sets} \to \mathrm{Sets}$ be the coproduct, ie the disjoint union of sets. This must be formalized in some fixed functorial way such as $S \sqcup T := (S \times \{0\}) \cup (T \times \{1\})$; but since the coproduct is unique up to canonical isomorphism, the choice of formalization is irrelevant.

The disjoint union with $\{\star\}$ defines a functor $\sigma\colon \mathrm{FI} \to \mathrm{FI}$ by $T \mapsto T \sqcup \{\star\}$. The *shift functor* $S\colon$ FI-Mod $\to$ FI-Mod is given by precomposition with σ: the FI-module SV is the composition $SV\colon \mathrm{FI} \xrightarrow{\sigma} \mathrm{FI} \xrightarrow{V} \mathbb{Z}$–Mod. Concretely, for any finite set T we have $(SV)_T = V_{T\sqcup\{\star\}}$. The functor S is evidently exact.

The kernel functor K and derivative functor D The inclusion of S into $S \sqcup \{\star\}$ defines a natural transformation from $\mathrm{id}_{\mathrm{FI}}$ to σ. From this we obtain a natural transformation ι from $\mathrm{id}_{\text{FI-Mod}}$ to S. Concretely, this is a natural map of FI-modules $\iota\colon V \to SV$ which, for every finite set T, sends V_T to $(SV)_T = V_{T\sqcup\{\star\}}$ via the map corresponding to the inclusion i_T of T into $T \sqcup \{\star\}$.

The functor $D\colon$ FI-Mod $\to$ FI-Mod, the *derivative*, is defined to be the cokernel of this map:

$$DV := \operatorname{coker}(V \xrightarrow{\iota} SV).$$

We similarly define $K\colon$ FI-Mod $\to$ FI-Mod to be the kernel $KV := \ker(V \xrightarrow{\iota} SV)$. For any FI-module, we have a natural exact sequence

$$0 \to KV \to V \to SV \to DV \to 0.$$

Since id and S are exact functors, D is right exact and K is left exact. Concretely, we have

$$(DV)_T \simeq V_{T\sqcup\{\star\}}/\operatorname{im}(V_T \to V_{T\sqcup\{\star\}}) \quad \text{and} \quad (KV)_T = \{v \in V_T \mid i(v) = 0 \in V_{T\sqcup\{\star\}}\}.$$

From this formula for KV, one can check that the functor K essentially coincides with the right adjoint $\rho\colon$ FI-Mod $\to$ FB-Mod of π^*; as we saw with H_0, the only difference is that KV is ρV considered as an FI-module, ie $K = \pi^* \circ \rho$.

Remark 2.4 Readers paying attention to the analogies in Table 1 might object that although H_0 and D are very different functors of FI-modules, the table indicates that both correspond to the functor $M \mapsto M/TM$ of graded $\mathbb{C}[T]$–modules. But this is not quite right, and by being careful with gradings we can see the distinction: H_0 corresponds to $\operatorname{coker}(M[-1] \xrightarrow{T} M)$ whereas D corresponds to $\operatorname{coker}(M \xrightarrow{T} M[1])$. In the graded case we rarely need to worry about the distinction, since grading shifts are invertible. But for FI-modules this is not true, and the distinction is important. (Lemma 4.4 below may clarify the behavior of D.)

Lemma 2.5 *An FI-module V is torsion-free if and only if $KV = 0$.*

Proof Recall that an FI-module V is *torsion-free* if for any injection $f\colon S \hookrightarrow T$ of finite sets, the map $f_*\colon V_S \to V_T$ is injective. By a simple induction, this holds if and only if f_* is injective for all $f\colon S \hookrightarrow T$ with $|T| = |S|+1$. However, such an inclusion can be factored as $f = g \circ i_S$ for some bijection $g\colon S \sqcup \{\star\} \simeq T$. Since g_* is necessarily injective, we see that V is torsion-free if and only if $\iota_S = (i_S)_*\colon V_S \to V_{S \sqcup \{\star\}}$ is injective for all finite sets S, ie if $KV = 0$. □

Iterates of shift and derivative We can iterate the shift functor S, obtaining FI-modules $S^b V$ for any $b \geq 0$. To avoid the notational confusion of writing $(S^2 V)_T \simeq V_{T \sqcup \{\star\} \sqcup \{\star\}}$, we adopt the notation that $[\star_b]$ denotes a fixed b–element set $[\star_b] := \{\star_1, \dots, \star_b\}$. We can then naturally identify $(S^2 V)_T \simeq V_{T \sqcup [\star_2]}$, and so on.

The iterates D^a are also right exact and can be described quite explicitly. For every FI-module V and every finite set T, we have

$$(D^a V)_T \simeq \frac{V_{T \sqcup [\star_a]}}{\sum_{j=1}^{a} \operatorname{im}(V_{T \sqcup [\star_a] - \{\star_j\}})}, \tag{3}$$

where $\operatorname{im}(V_{T \sqcup [\star_a] - \{\star_j\}})$ denotes the image of the natural map $V_{T \sqcup [\star_a] - \{\star_j\}} \to V_{T \sqcup [\star_a]}$ induced by the inclusion $T \sqcup [\star_a] - \{\star_j\} \subset T \sqcup [\star_a]$. We remark that D^a is the left adjoint of the functor B^a of [5, Definition 2.16].

For any submodule $V \subset M$ the inclusion induces a map $D^a V \to D^a M$.

Lemma 2.6 *If M is a torsion-free FI-module and $V \subset M$ a submodule,*

$$\ker(D^a V \to D^a M)_{n-a} = 0 \iff$$
$$V_n \cap \big(M_{[n]-\{1\}} + \dots + M_{[n]-\{a\}}\big) = V_{[n]-\{1\}} + \dots + V_{[n]-\{a\}}.$$

Proof Since M is torsion-free we can identify M_S with its image in $M_{S \sqcup \{\star\}}$ and so on, so

$$\ker(D^a V \to D^a M)_T \simeq \ker\left(\frac{V_{T \sqcup [\star_a]}}{\sum_{j=1}^{a} V_{T \sqcup [\star_a] - \{\star_j\}}} \to \frac{M_{T \sqcup [\star_a]}}{\sum_{j=1}^{a} M_{T \sqcup [\star_a] - \{\star_j\}}} \right).$$

In other words,

$$\ker(D^a V \to D^a M)_T = 0 \iff V_{T \sqcup [\star_a]} \cap \left(\sum_{j=1}^{a} M_{T \sqcup [\star_a] - \{\star_j\}} \right) = \sum_{j=1}^{a} V_{T \sqcup [\star_a] - \{\star_j\}}.$$

Setting $T = \{a+1, \dots, n\}$ and identifying $[\star_a]$ with $\{1, \dots, a\}$, we obtain the desired expression. $\square$

Remark 2.7 Notice that the right side of Lemma 2.6 is precisely the conclusion of Theorem E. Therefore, we can restate Theorem E as saying: if M is torsion-free and generated in degree at most k, and $V \subset M$ is generated in degree at most d, then $\ker(D^a V \to D^a M)_{n-a}$ vanishes for $n > d + \min(k,d)$; in other words, $\deg \ker(D^a V \to D^a M) \le d + \min(k,d) - a$. This observation will be used in Section 4, and specifically in the proof of Theorem 4.8 to obtain bounds on $H_p^{D^a}$.

3 Combinatorics of finite injections and FI-modules

The goal of this section is to prove Theorem E. In Section 3.1 we generalize Theorem E to Theorem E′ which does not refer to the ambient FI-module M, and is of independent interest. In Section 3.2 we establish the combinatorial properties of $\mathbb{Z}\big[\mathrm{Hom}_{\mathrm{FI}}([d],[n])\big]$ that make our proof possible; throughout that section we do not mention FI-modules at all. In Section 3.3 we apply these properties to prove Theorem E′. But before moving to the combinatorics, we begin by motivating the connections with FI-modules.

3.1 The ideal I_m and Theorem E′

The ideal I_m For each pair of distinct elements $i \ne j$ in $[n]$, we write $(i\ \ j)$ for the transposition in S_n interchanging i and j, and we define $J_j^i := \mathrm{id} - (i\ \ j) \in \mathbb{Z}[S_n]$. Note that $J_j^i = J_i^j$, and that J_j^i and J_l^k commute when their four indices are distinct (since the transpositions $(i\ \ j)$ and $(k\ \ l)$ commute in this case).

For $m \in \mathbb{N}$, define $I_m \subset \mathbb{Z}[S_n]$ to be the two-sided ideal generated by products of the form

$$J_{j_1}^{i_1} J_{j_2}^{i_2} \cdots J_{j_m}^{i_m},$$

where $i_1, j_1, \dots, i_m, j_m$ are $2m$ *distinct* elements of $[n]$. (In particular, the terms of the product commute.) Multiplying out such a product, we have

$$J_{j_1}^{i_1} J_{j_2}^{i_2} \cdots J_{j_m}^{i_m} = \sum_{K \subset [m]} (-1)^{|K|} \prod_{k \in K} (i_k\ \ j_k) = \sum_{\sigma \in (\mathbb{Z}/2)^m} (-1)^{\sigma} \sigma, \tag{4}$$

where $(\mathbb{Z}/2)^m$ denotes the subgroup generated by the commuting transpositions $(i_k\ \ j_k)$, and $(-1)^\sigma$ denotes the image of σ under the sign homomorphism $S_n \to \pm 1$.

Although the ideals I_m will play multiple different roles in the proof of Theorem E, the following property provides a simple illustration of why we consider these ideals. Recall that the group ring $\mathbb{Z}[S_n]$ acts on W_n for any FI-module W.

Proposition 3.1 *Let M be an FI-module generated in degree at most k. Then $I_{k+1} \cdot M_n = 0$ for all $n \geq 0$.*

Proof We prove first that I_m annihilates the free module $M(a)$ if $a < m$, meaning that $I_m \cdot M(a)_n = 0$ for all n. For any a, we have a basis for $M(a)_n$ given by injections $f\colon [a] \hookrightarrow [n]$. The key observation is that given $f\colon [a] \hookrightarrow [n]$ and a generator $J^{i_1}_{j_1} J^{i_2}_{j_2} \cdots J^{i_m}_{j_m} \in I_m$,

$$\text{if } \operatorname{im} f \cap \{i_\ell, j_\ell\} = \varnothing \text{ for some } \ell \in [m], \quad \text{then } J^{i_1}_{j_1} J^{i_2}_{j_2} \cdots J^{i_m}_{j_m} \cdot f = 0. \tag{5}$$

Indeed the assumption implies $(i_\ell\ \ j_\ell) \circ f = f$, so $J^{i_\ell}_{j_\ell} = \mathrm{id} - (i_\ell\ \ j_\ell)$ satisfies $J^{i_\ell}_{j_\ell} \cdot f = 0$. Since the terms of the product commute, it follows that $J^{i_1}_{j_1} J^{i_2}_{j_2} \cdots J^{i_m}_{j_m} \cdot f = 0$.

However, when $a < m$, for *every* $f\colon [a] \hookrightarrow [m]$ and $J = J^{i_1}_{j_1} J^{i_2}_{j_2} \cdots J^{i_m}_{j_m} \in I_m$ there exists some ℓ for which $\operatorname{im} f \cap \{i_\ell, j_\ell\} = \varnothing$. Therefore, $J \cdot f = 0$ for all basis elements $f \in M(a)_n$ and all generators $J \in I_m$, proving that $I_m \cdot M(a) = 0$ as claimed.

Returning to the general claim, let M be an FI-module generated in degree at most k. By Lemma 2.2(iv), M is a quotient of a sum of free modules $M(a)$ generated in degrees $a \leq k$. We have just proved that I_{k+1} annihilates any such free module, so it annihilates the quotient M as well. □

Generalizing Theorem E by removing M The statement of Theorem E can be generalized by removing M from its statement. Recall that Theorem E states that if M is a torsion-free FI-module and $V \subset M$ is a submodule, then

$$V_n \cap \left(M_{[n]-\{1\}} + \cdots + M_{[n]-\{a\}}\right) = V_{[n]-\{1\}} + \cdots + V_{[n]-\{a\}} \tag{6}$$

for sufficiently large n. Though it may not be obvious, the central object in this statement is V. In fact, we can remove the FI-module M from the statement entirely, and at the same time strengthen the theorem.

Consider the case $a = 1$, when our goal (6) is that $V_n \cap M_{[n]-\{1\}}$ coincides with $V_{[n]-\{1\}}$ for large enough n. When M is free, the submodule $M_{[n]-\{1\}} \subset M_{[n]}$ can be cut out as

$$M_{[n]-\{1\}} = \{m \in M_{[n]} \subset M_{[n]\sqcup[\star_1]} \mid (1\ \ \star_1) \cdot m = m\}.$$

In other words, recalling that $\iota\colon [n] \hookrightarrow [n] \sqcup [\star_1]$ denotes the standard inclusion, the element $\widetilde{J}_{[1]} := J_1^{\star_1} \circ \iota \in \mathbb{Z}\big[\mathrm{Hom}_{\mathrm{FI}}([n], [n] \sqcup [\star_1])\big]$ has the property that

$$M_{[n]-\{1\}} = \ker \widetilde{J}_{[1]}|_{M_{[n]}}$$

when M is free. For general M we need *not* have equality here, but we do always have the containment $M_{[n]-\{1\}} \subset \ker \widetilde{J}_{[1]}|_{M_{[n]}}$. Intersecting with V, we always have the containments

$$V_{[n]-\{1\}} \subset V_n \cap M_{[n]-\{1\}} \subset \ker \widetilde{J}_{[1]}|_{V_{[n]}}.$$

The statement of Theorem E is that the *first* containment is an equality for large enough n. But we can actually prove the stronger statement that *both* are equalities: $V_{[n]-\{1\}} = \ker \widetilde{J}_{[1]}|_{V_{[n]}}$ for large enough n. Notice that this statement no longer makes reference to M!

For larger a, we consider the element $\widetilde{J}_{[a]} := J_1^{\star_1} \cdots J_a^{\star_a} \circ \iota \in \mathbb{Z}\big[\mathrm{Hom}_{\mathrm{FI}}([n], [n] \sqcup [\star_a])\big]$. We saw in the previous paragraph that $M_{[n]-\{i\}} \subset \ker J_i^{\star_i}$ (identifying $M_{[n]-\{i\}}$ with its image). Since all the operators $J_i^{\star_i}$ commute, this implies that $M_{[n]-\{i\}} \subset \ker \widetilde{J}_{[a]}$ for any $i \in [a]$, so

$$M_{[n]-\{1\}} + \cdots + M_{[n]-\{a\}} \subset \ker \widetilde{J}_{[a]}|_{M_{[n]}}.$$

This means that

$$V_{[n]-\{1\}} + \cdots + V_{[n]-\{a\}} \subset V_n \cap \big(M_{[n]-\{1\}} + \cdots + M_{[n]-\{a\}}\big) \subset \ker \widetilde{J}_{[a]}|_{V_{[n]}} \tag{7}$$

for any $n \geq a$. This leads us to the following generalization of Theorem E.

Theorem E$'$ *Let V be a torsion-free FI-module generated in degree at most d satisfying $I_{K+1} \cdot V = 0$. Then for all $n > K + d$ and any $a \leq n$,*

$$V_{[n]-\{1\}} + \cdots + V_{[n]-\{a\}} = \ker \widetilde{J}_{[a]}.$$

Theorem E$'$ is proved in Section 3.3 below, but we first verify here that it implies Theorem E.

Proof that Theorem E$'$ implies Theorem E We begin in the setup of Theorem E, so let M be a torsion-free FI-module generated in degree at most k, and let V be a submodule of M generated in degree at most d.

Proposition 3.1 states that $I_{k+1} \cdot M = 0$, so the same is true of its submodule V. Applying Proposition 3.1 to V directly shows that $I_{d+1} \cdot V = 0$. Therefore, if we set $K = \min(k, d)$, we have $I_{K+1} \cdot V = 0$.

Applying Theorem E$'$, we conclude that $V_{[n]-\{1\}} + \cdots + V_{[n]-\{a\}} = \ker \widetilde{J}_{[a]}|_{V_{[n]}}$ for all $n > K + d$. In light of (7), this implies that $V_{[n]-\{1\}} + \cdots + V_{[n]-\{a\}} = V_n \cap \big(M_{[n]-\{1\}} + \cdots + M_{[n]-\{a\}}\big)$ for $n > K + d$, as desired. $\square$

In fact, Theorem E$'$ is strictly stronger than Theorem E. To see this, notice that Theorem E says nothing when $V = M$, while Theorem E$'$ implies the following structural statement (by taking $a = K + 1$ and noting that $\widetilde{J}_{[K+1]} \in I_{K+1}$), which is nontrivial whenever $K < d$.

Corollary 3.2 *Let V be a torsion-free FI-module generated in degree at most d satisfying $I_{K+1} \cdot V = 0$. (For example, this holds if V can be embedded into some FI-module generated in degree at most K.) Then for all $n > K + d$,*

$$V_n = V_{[n]-\{1\}} + \cdots + V_{[n]-\{K\}} + V_{[n]-\{K+1\}}.$$

3.2 The combinatorics of $\mathbb{Z}\big[\mathrm{Hom}_{\mathrm{FI}}([d], [n])\big]$

The discussion above did not depend on any ordering on $[n]$ (essentially treating it as an arbitrary finite set). By contrast, throughout the rest of this section we rely heavily on the ordering on $[n]$. This is inconsistent with the philosophy of FI-modules, so throughout Section 3.2 we will not mention the category FI at all.

Definition 3.3 (the collection $\Sigma(b)$) For $b \in \mathbb{N}$, let $\Sigma(b)$ denote the set of b–element subsets $S \subset [2b]$ satisfying the following property:

$(**)$ The i^{th} largest element of S is at most $2i - 1$.

For $a \in \mathbb{N}$ with $1 \le a \le b$, let $\Sigma(a, b) \subset \Sigma(b)$ consist of all those $S \in \Sigma(b)$ containing $[a]$:

$$\Sigma(a, b) := \{S \in \Sigma(b) \mid [a] \subset S \subset [2b]\}. \tag{8}$$

For example, it follows from $(**)$ that $1 \in S$ for any $S \in \Sigma(b)$, so for any $b \in \mathbb{N}$ we have $\Sigma(1, b) = \Sigma(b)$. At the other extreme, we have $\Sigma(b, b) = \{[b]\}$. The subsets $\Sigma(a, b)$ interpolate between $\Sigma(1, b) = \Sigma(b)$ and $\Sigma(b, b) = \{[b]\}$; for example,

$$\Sigma(1,4) = 1234, 1235, 1236, 1237, 1245, 1246, 1247, 1256, 1257, \\ 1345, 1346, 1347, 1356, 1357;$$

$$\Sigma(2,4) = 1234, 1235, 1236, 1237, 1245, 1246, 1247, 1256, 1257;$$

$$\Sigma(3,4) = 1234, 1235, 1236, 1237;$$

$$\Sigma(4,4) = 1234.$$

We have written the elements of $\Sigma(a, b)$ in lexicographic order, which ordering we denote by $\preccurlyeq$. We denote by $\overline{S}$ the complement $\overline{S} := [2b] \setminus S$. We will only use this notation for b–element subsets $S \subset [2b]$, so the notation is unambiguous; in particular, $\overline{S}$ is always a b–element subset of $[2b]$ as well.

Remark 3.4 We record some relations between the different collections $\Sigma(a,b)$:

(a) For any $S \in \Sigma(b)$ and any $m \leq 2b+1$ with $m \notin S$, the union $S \cup \{m\}$ belongs to $\Sigma(b+1)$. In particular, this holds if $m \in \overline{S}$. If $S \in \Sigma(a,b)$, then $S \cup \{m\} \in \Sigma(a,b+1)$.

(b) For any $S \in \Sigma(b)$ and any $c \leq b$, if $R \subset S$ is the c–element subset consisting of the c smallest elements, then $R \in \Sigma(c)$. If $S \in \Sigma(a,b)$ for $a \leq c \leq b$, then $R \in \Sigma(a,c)$ as well.

(c) If $S, T \in \Sigma(b)$ satisfy $T \preccurlyeq S$ and $S \in \Sigma(a,b)$, then $T \in \Sigma(a,b)$ as well. In other words, $\Sigma(a,b)$ is an "initial segment" of $\Sigma(b)$ (this is immediately visible in the description of $\Sigma(a,4)$ above).

Descendants The condition $(**)$ gives one way to define the Catalan numbers: the n^{th} Catalan number is $|\Sigma(n)| = (1/(n+1))\binom{2n}{n}$. This is not a coincidence; our interest in $\Sigma(b)$ comes from the following characterization of the sets $S \in \Sigma(b)$, which is related to another definition of the Catalan numbers.

Given any b–element subset $S \subset [2b]$, write the elements of S in increasing order as $s_1, \dots, s_b$ and the elements of $\overline{S}$ in increasing order as $t_1, \dots, t_b$. Let $(\mathbb{Z}/2)^S$ denote the subgroup of S_{2b} generated by the commuting transpositions $(s_k\ \ t_k) \in S_{2b}$. If we define $J_S \in I_b$ as

$$J_S := \prod_i J_{s_i}^{t_i}, \tag{9}$$

by (4) we have $J_S = \sum_{\sigma \in (\mathbb{Z}/2)^S} (-1)^\sigma \sigma$. In these terms, the defining property $(**)$ of $\Sigma(b)$ has the following formulation:

$$S \in \Sigma(b) \iff S \text{ is lexicographically first among } \{\sigma \cdot S \mid \sigma \in (\mathbb{Z}/2)^S\}. \tag{10}$$

Given $S \in \Sigma(b)$, we refer to the subsets $\{\sigma \cdot S \mid \sigma \in (\mathbb{Z}/2)^S\}$ as the *descendants* of S; by (10), S lexicographically precedes all of its descendants.[4] In fact, we will use the following generalization. For any subset $U \subset [n]$ with $S \subset U$, and any b distinct elements $u_1 < \dots < u_b$ of $[n] \setminus U$, we can consider the subgroup $(\mathbb{Z}/2)^b$ generated by the disjoint transpositions $(s_i\ \ u_i)$. By comparison with (10), it is straightforward to conclude:

Lemma 3.5 *$S \in \Sigma(b)$ implies U is lexicographically first among $\{\sigma \cdot U \mid \sigma \in (\mathbb{Z}/2)^b\}$.*

---

[4] A set S and its descendant $\sigma \cdot S$ need not determine the same subgroup $(\mathbb{Z}/2)^S \neq (\mathbb{Z}/2)^{\sigma \cdot S}$, so the relation of being a descendant is neither symmetric nor transitive. For example, if $S = \{1,2\} \subset [4]$, then $S' = \{1,4\}$ is a descendant of S, but $(\mathbb{Z}/2)^S = \langle (1\ 3), (2\ 4) \rangle$ whereas $(\mathbb{Z}/2)^{S'} = \langle (1\ 2), (3\ 4) \rangle$. The descendants of S are $S = 12$, $S' = 14$, 23 and 34 whereas the descendants of S' are $S' = 14$, 23, 13 and 24.

The S_n–module F and subgroups Fix $d \in \mathbb{N}$ and $n \in \mathbb{N}$ for the remainder of Section 3. Let F denote the $\mathbb{Z}[S_n]$–module associated to the permutation action on the set of injections $f\colon [d] \hookrightarrow [n]$. (In other words, as an S_n–module F is isomorphic to $\mathbb{Z}\big[\mathrm{Hom}_{\mathrm{FI}}([d],[n])\big]$; however, we will wait until the next section to explore this connection with the category FI.)

Definition 3.6 We define certain subgroups of the free abelian group F corresponding to particular subsets of the basis $\{f\colon [d] \hookrightarrow [n]\}$. In these definitions, S is a b–element subset $S \in \Sigma(b)$:

$$\begin{aligned}
F^{\neq S} &:= \langle f\colon [d] \hookrightarrow [n] \mid S \not\subset \operatorname{im} f \rangle, \\
F^{b} &:= \langle f\colon [d] \hookrightarrow [n] \mid \forall S \in \Sigma(b),\ S \not\subset \operatorname{im} f \rangle = \bigcap_{S\in\Sigma(b)} F^{\neq S}, \\
F^{a,b} &:= \langle f\colon [d] \hookrightarrow [n] \mid \forall S \in \Sigma(a,b),\ S \not\subset \operatorname{im} f \rangle = \bigcap_{S\in\Sigma(a,b)} F^{\neq S}, \\
F_{=S} &:= \langle f\colon [d] \hookrightarrow [n] \mid \operatorname{im} f \cap [2b] = S \rangle.
\end{aligned}$$

In general none of these subgroups are preserved by the action of S_n on F.

We emphasize the contrast between $F^{\neq S}$ and $F_{=S}$: for fixed b, a given injection $f\colon [d] \hookrightarrow [n]$ may lie in $F^{\neq S}$ for many different $S \in \Sigma(b)$; in contrast, f lies in $F_{=S}$ for at most one $S \in \Sigma(b)$ (namely $S = \operatorname{im} f \cap [2b]$, if this subset happens to belong to $\Sigma(b)$).

Since $\Sigma(b) = \Sigma(1,b) \supset \cdots \supset \Sigma(b,b)$, we have $F^b = F^{1,b} \subset F^{2,b} \subset \cdots \subset F^{b,b}$. Similarly, from Remark 3.4(b) we have $F^{a,a} \subset \cdots \subset F^{a,b} \subset \cdots$. In other words, if $a \le a'$ and $b \le b'$, then $F^{a,b} \subset F^{a',b'}$. Note that, since $\Sigma(a,a)$ consists of the single set $S = [a]$, the subgroup $F^{a,a}$ is spanned by injections $f\colon [d] \hookrightarrow [n]$ with $i \notin \operatorname{im} f$ for some $i \in [a]$.

We make no assumptions whatsoever on d, n or b in this section, although in some cases the definitions become rather trivial. (For example, when $b > d$, we have $F = F^b$; when $d > n$, we have $F = 0$; when $2b > n$, we have $I_b = 0$.)

Proposition 3.7 *For any b such that $n \ge b + d$, we have*

$$F = I_b \cdot F + F^b.$$

Proof It is vacuous that $I_b \cdot F + F^b \subset F$, so we must prove that $F \subset I_b \cdot F + F^b$. Assume otherwise; then some basis element f does not lie in $I_b \cdot F + F^b$. Choose f so that $\operatorname{im} f$ is lexicographically last among all such f. Since $f \notin F^b$, there exists some $S \in \Sigma(b)$ with $S \subset \operatorname{im} f$. Since $n \geq b+d$, we may choose b distinct elements $u_1 < \cdots < u_b$ from $[n] \setminus \operatorname{im} f$. Let $J = J_{s_1}^{u_1} \cdots J_{s_b}^{u_b}$, and consider the element

$$(J - \mathrm{id}) \cdot f = \sum_{\sigma \neq 1 \in (\mathbb{Z}/2)^b} (-1)^\sigma \sigma \cdot f.$$

By Lemma 3.5 we have $\operatorname{im}(\sigma \cdot f) = \sigma \cdot \operatorname{im} f \succ \operatorname{im} f$ for all $\sigma \neq 1$. By our definition of f (that its image was lexicographically last), $\sigma \cdot f$ is contained in $I_b \cdot F + F^b$ for all $\sigma \neq 1$, so $(J - \mathrm{id}) \cdot f \in I_b \cdot F + F^b$. However, $J \cdot f \in I_b \cdot F$ by definition, so this implies that $J \cdot f - (J - \mathrm{id}) \cdot f = f$ lies in $I_b \cdot F + F^b$, contradicting our assumption. $\square$

Decomposing F in terms of the subgroups $J_S F_{=S}$ We will also need, for a different purpose, a more specific version of Proposition 3.7. For each $S \in \Sigma(b)$, we have defined in (9) the operator $J_S \in \mathbb{Z}[S_{2b}]$. For any $n \geq 2b$ we may consider this as an operator in $\mathbb{Z}[S_n]$, which we also denote by J_S.

Proposition 3.8 *For any $a \leq b$ such that $2b \leq n$,*

$$F^{a,b+1} \subset F^{a,b} + \sum_{S \in \Sigma(a,b)} J_S F_{=S}.$$

Proof For this proof only, define

$$F^{(a,b)} := F^{a,b} + \sum_{S \in \Sigma(a,b)} F_{=S} \tag{11}$$

$$= \big\langle f \colon [d] \hookrightarrow [n] \,\big|\, \nexists S \in \Sigma(a,b) \text{ s.t. } S \subsetneq \operatorname{im} f \cap [2b] \big\rangle.$$

In words, $F^{(a,b)}$ is spanned by those injections $f \colon [d] \hookrightarrow [n]$ such that $\operatorname{im} f \cap [2b]$ does not *properly* contain any element of $\Sigma(a,b)$ (but $\operatorname{im} f \cap [2b]$ is allowed to be *equal* to some $S \in \Sigma(a,b)$).

We begin by showing that $F^{a,b+1} \subset F^{(a,b)}$. Consider a basis element f which does not lie in $F^{(a,b)}$. By definition, there exists $S \in \Sigma(a,b)$ such that $S \subsetneq \operatorname{im} f \cap [2b]$. Choose $m \in \operatorname{im} f \cap [2b]$ with $m \notin S$, and define $T = S \cup \{m\}$. We have $T \in \Sigma(a,b+1)$ by Remark 3.4(a), so $f \notin F^{a,b+1}$ as desired.

We now show that for any $S \in \Sigma(a,b)$, we have

$$F_{=S} \ \subset\ J_S F_{=S} + F^{a,b} + \sum_{\substack{S' \in \Sigma(a,b) \\ S' \succ S}} F_{=S'}. \tag{12}$$

Consider a basis element $f \in F_{=S}$ and the associated element

$$(J_S - \mathrm{id}) \cdot f = \sum_{\sigma \neq 1 \in (\mathbb{Z}/2)^S} (-1)^\sigma \sigma \cdot f.$$

By assumption, $\operatorname{im} f \cap [2b] = S$, so $\operatorname{im}(\sigma \cdot f) \cap [2b] = \sigma \cdot \operatorname{im} f \cap [2b] = \sigma \cdot S$ is a descendant of S. By (10), the fact that $S \in \Sigma(a,b)$ means that $\sigma \cdot S \succ S$ for all $\sigma \neq 1 \in (\mathbb{Z}/2)^S$. Thus for each σ, there are two possibilities for the b–element subset $\sigma \cdot S$: either $\sigma \cdot S$ does not belong to $\Sigma(a,b)$, in which case $\sigma \cdot f \in F^{a,b}$; or $\sigma \cdot S \in \Sigma(a,b)$ but $\sigma \cdot S \succ S$, in which case $\sigma \cdot f \in F_{=\sigma \cdot S}$. In other words,

$$(\mathrm{id} - J_S) \cdot f \in F^{a,b} + \sum_{\substack{S' \in \Sigma(a,b) \\ S' \succ S}} F_{=S'}.$$

Writing $f = J_S \cdot f - (J_S - \mathrm{id}) \cdot f$, this demonstrates (12).

Beginning with (11), we apply (12) to each $S \in \Sigma(a,b)$ in lexicographic order to obtain the desired

$$F^{a,b+1} \subset F^{(a,b)} = F^{a,b} + \sum_{S \in \Sigma(a,b)} F_{=S} \subset \sum_{S \in \Sigma(a,b)} J_S F_{=S} + F^{a,b}. \qquad \square$$

3.3 Proof of Theorem E′

We are now ready to apply the combinatorial apparatus above to FI-modules and prove Theorem E′.

Proof of Theorem E′ We continue with the notation of Section 3.2, so F denotes the S_n–module $\mathbb{Z}\big[\mathrm{Hom}_{\mathrm{FI}}([d],[n])\big]$, and F^b and $F^{a,b}$ are the subgroups of F defined in Definition 3.6. Define subgroups $V^b \subset V_n$ and $V^{a,b} \subset V_n$ by $V^b := \operatorname{im}(F^b \otimes V_d \to V_n)$ and $V^{a,b} := \operatorname{im}(F^{a,b} \otimes V_d \to V_n)$. From the containments following Definition 3.6 we see that $V^b = V^{1,b} \subset V^{2,b} \subset \cdots \subset V^{b,b}$.

Let us understand these subgroups $V^{a,b}$ more concretely. To say that V is generated in degree at most d means that V_n is spanned by its subgroups V_T as T ranges over subsets $T \subset [n]$ with $|T| = d$. (Throughout this proof, T will always denote a subset $T \subset [n]$ with $|T| = d$.)

By definition, V^b is the subgroup of V_n spanned by $f_*(V_d)$ where $f\colon [d] \hookrightarrow [n]$ ranges over injections for which $\operatorname{im} f$ does not contain any $S \in \Sigma(b)$. In other words,

$$V^b = \operatorname{span}\big\{V_T \,\big|\, T \subset [n],\ |T| = d \text{ s.t. } T \text{ does not contain any } S \in \Sigma(b)\big\}.$$

Similarly, $V^{b,b}$ is by definition the subgroup of V_n spanned by those V_T for which $[b] \not\subset T$ and $|T| = d$. Since $V_{[n]-\{i\}}$ is the subgroup spanned by those V_T where $i \notin T$, we see that

$$V^{b,b} = V_{[n]-\{1\}} + \cdots + V_{[n]-\{b\}}. \tag{13}$$

Fix some $n > K + d$ and some $a \le n$. According to (13), the desired conclusion of the theorem states that $\ker \widetilde{J}_{[a]} = V^{a,a}$ when $n > K + d$. From (7), we know that $V^{a,a} \subset \ker \widetilde{J}_{[a]}$ for *all* n, so what we need to prove is that $\ker \widetilde{J}_{[a]} \subset V^{a,a}$ when $n > K + d$. We accomplish this by proving by reverse induction on b that $\ker \widetilde{J}_{[a]} \subset V^{a,b}$ for all $b \ge a$.

Our base case is $b = K + 1$. In this case we will prove something much stronger than the inductive hypothesis; we will prove Corollary 3.2 by showing that it is a direct consequence of Proposition 3.7. Recall that we always have the containments $V^{K+1} \subset V^{a,K+1} \subset V^{K+1,K+1} \subset V_n$. The statement of Proposition 3.7 for $b = K+1$ is that $F = I_{K+1} \cdot F + F^{\ell}$, and the hypothesis is satisfied since $n \ge (K + 1) + d$. Therefore,

$$V_n = \mathrm{im}(I_{K+1} \cdot F \otimes V_d) + V^{K+1} = I_{K+1} \cdot V_n + V^{K+1}.$$

Since $I_{K+1} \cdot V_n = 0$ by assumption, we conclude that

$$V_n = V^{K+1} = V^{a,K+1} = V^{K+1,K+1}.$$

Notice that $V_n = V^{K+1,K+1}$ is precisely the conclusion of Corollary 3.2, as mentioned above. This concludes the base case.

For the inductive step, The key is to show that for all $a \le b \le K$ we have

$$V^{a,b+1} \cap \ker \widetilde{J}_{[a]} \subset V^{a,b}. \tag{14}$$

Given this, if we assume $\ker \widetilde{J}_{[a]} \subset V^{a,b+1}$ by induction, (14) implies $\ker \widetilde{J}_{[a]} = V^{a,b+1} \cap \ker \widetilde{J}_{[a]} \subset V^{a,b}$, which is the desired inductive hypothesis. The remainder of the argument thus consists of the proof of (14).

For convenience, we would like to assume that $K \le d$. If $K > d$, replacing K by d in the statement of Theorem E$'$ makes the conclusion stronger, while the hypothesis is still satisfied because $I_{d+1} \cdot V = 0$ by Proposition 3.1. Therefore, making this replacement if necessary, we may assume that $K \le d$. Our assumption on n thus implies $n > K + d \ge 2K \ge 2b$. Therefore, we may apply Proposition 3.8, which states that $F^{a,b+1} \subset F^{a,b} + \sum_{S \in \Sigma(a,b)} J_S F_{=S}$. We conclude that every $v \in V^{a,b+1}$ can be written as

$$v = v^{a,b} + \sum_{S \in \Sigma(a,b)} v_S, \quad \text{where } v^{a,b} \in V^{a,b},\ v_S \in J_S F_{=S} \cdot V_d. \tag{15}$$

It will suffice to show that if an element v as in (15) lies in $\ker \widetilde{J}_{[a]}$, then in fact each term v_S is zero, which implies (14).

Assume that $v \in V^{a,b+1} \cap \ker \widetilde{J}_{[a]}$, and suppose for a contradiction that $v_S \neq 0$ for some $S \in \Sigma(a,b)$. Let S be the lexicographically first such element of $\Sigma(a,b)$. We may thus write

$$v = v^{a,b} + v_S + \sum_{\substack{T \in \Sigma(a,b) \\ S \prec T}} v_T. \tag{16}$$

For any $S \in \Sigma(a,b)$, write the elements of S in order as $s_1 < \cdots < s_b$, and define

$$\widetilde{J}_S := J_{s_1}^{\star_1} \cdots J_{s_b}^{\star_b} \circ \iota \in \mathbb{Z}\big[\mathrm{Hom}_{\mathrm{FI}}([n], [n] \sqcup [\star_b])\big].$$

We will establish a series of claims about $\widetilde{J}_S$, which hold for any $S \in \Sigma(a,b)$.

Claim 3.9 $\widetilde{J}_S \cdot (\ker \widetilde{J}_{[a]}) = 0.$

Proof To say that $S \in \Sigma(a,b)$ means that $[a] \subset S$, so the elements of S are necessarily $1 < 2 < \cdots < a < s_{a+1} < \cdots < s_b$. Therefore,

$$\widetilde{J}_S = J_1^{\star_1} \cdots J_a^{\star_a} J_{s_{a+1}}^{\star_{a+1}} \cdots J_{s_b}^{\star_b} = \widetilde{J}_{[a]} \cdot X.$$

Since $\widetilde{J}_S = \widetilde{J}_{[a]} \cdot X = X \cdot \widetilde{J}_{[a]}$, we have $\ker \widetilde{J}_{[a]} \subset \ker \widetilde{J}_S$ as claimed. □

By (5), we have

$$\widetilde{J}_S \cdot f = 0 \quad \text{for any } f\colon [d] \hookrightarrow [n] \text{ with } S \not\subset \operatorname{im} f \tag{17}$$

since for any such f there exists $s_i \notin \operatorname{im} f$, so $\{s_i, \star_i\} \cap \operatorname{im} f = \varnothing$. This has the following consequences.

Claim 3.10 $\widetilde{J}_S \cdot F^{a,b} = 0.$

Proof By definition, any $f \in F^{a,b}$ has $S \not\subset \operatorname{im} f$, so $\widetilde{J}_S \cdot f = 0$ by (17). □

Claim 3.11 *$\widetilde{J}_S \cdot J_T F_{=T} = 0$ for any $T \in \Sigma(a,b)$ such that $S \prec T$.*

Proof Given a generator $f \in F_{=T}$ we know that $\operatorname{im} f \cap [2b] = T$. As in the proof of Proposition 3.8, the terms of $J_T \cdot f$ consist of $\sigma \cdot f$ for $\sigma \in (\mathbb{Z}/2)^T$. The intersections $\operatorname{im}(\sigma \cdot f) \cap [2b]$ are precisely the descendants $\sigma \cdot T$. Since $T \in \Sigma(a,b)$ we have $\sigma \cdot T \succcurlyeq T$. In particular, since $S \prec T \preccurlyeq \sigma \cdot T$, every term satisfies $S \not\subset \operatorname{im}(\sigma \cdot f)$. By (17), $\widetilde{J}_S \cdot J_T F_{=T} = 0$ as desired. □

We now apply these consequences to the decomposition (16). By Claim 3.9, our assumption that $v \in \ker \widetilde{J}_{[a]}$ implies that $\widetilde{J}_S \cdot v = 0$. Claims 3.10 and 3.11 show that $\widetilde{J}_S \cdot v^{a,b} = 0$ and $\widetilde{J}_S \cdot v_T = 0$. We conclude that $\widetilde{J}_S \cdot v_S = \widetilde{J}_S \cdot v = 0$; it remains to show that this implies $v_S = 0 \in V_n$.

We show this using the following two claims, which we prove in turn. Define $\tau \in \mathrm{End}_{\mathrm{FI}}([n] \sqcup [\star_b])$ to be the involution $\tau := (t_1 \ \ \star_1) \cdots (t_b \ \ \star_b)$, where $(t_1, \dots, t_b)$ denotes the complement of S in $[2b]$ as before.

Claim 3.12 $\widetilde{J}_S \cdot J_S = \widetilde{J}_S$ *when restricted to* $F_{=S}$.

Proof As in Claim 3.11, given $f \in F_{=S}$ with $\operatorname{im} f \cap [2b] = S$, the terms of $J_S \cdot f$ consist of f together with $\sigma \cdot f$ for $\sigma \neq 1 \in (\mathbb{Z}/2)^S$. Each of the latter has $\operatorname{im}(\sigma \cdot f) \cap [2b] = \sigma \cdot S \succ S$. Therefore, $S \not\subset \sigma \cdot f$ for $\sigma \neq 1$, so $\widetilde{J}_S \cdot \sigma \cdot f = 0$ by (17). We conclude that $\widetilde{J}_S \cdot J_S \cdot f = \widetilde{J}_S \cdot (f + \sum (-1)^\sigma \sigma \cdot f) = \widetilde{J}_S \cdot f$, as claimed. □

Claim 3.13 $\tau \widetilde{J}_s = \iota \circ J_S$ *when restricted to* $F_{=S}$.

Proof Note that $\tau (J_{s_1}^{\star_1} \cdots J_{s_b}^{\star_b}) \tau^{-1} = J_{s_1}^{t_1} \cdots J_{s_b}^{t_b}$. Therefore,

$$\tau \widetilde{J}_S = J_{s_1}^{t_1} \cdots J_{s_b}^{t_b} \circ \tau \circ \iota \in \mathbb{Z}\big[\mathrm{Hom}_{\mathrm{FI}}([n], [n] \sqcup [\star_b])\big].$$

By definition, the image of a map $f \in F_{=S}$ does not contain t_i, so when restricted to $F_{=S}$, we have $\tau \circ \iota = \iota$. We conclude that $\tau \widetilde{J}_S = J_{s_1}^{t_1} \cdots J_{s_b}^{t_b} \circ \iota = \iota \circ J_S$, as claimed. □

We now complete the proof. Write $v_S = J_S \cdot w_S$ for $w_S \in F_{=S} \cdot V_d \subset V_n$. Claim 3.12 implies that $\widetilde{J}_S \cdot v_S = \widetilde{J}_S \cdot J_S \cdot w_S = \widetilde{J}_S \cdot w_S$. Thus $\widetilde{J}_S \cdot w_S = 0$, so certainly $\tau \widetilde{J}_S w_S = 0$. Claim 3.13 implies that $\tau \widetilde{J}_S w_S = \iota(J_S w_S) = \iota(v_S)$. Combining these, we see that $\iota(v_S) = 0$. Since V is torsion-free, ι is injective, so this proves that $v_S = 0$.

This contradicts our assumption that $v_S \neq 0$, so we conclude from (15) that $v \in V^{a,b}$. This concludes the proof of the containment (14); as we explained following (14), this completes the proof of the inductive hypothesis and thus concludes the proof of the theorem. □

4 Bounds on the homology of FI-modules

An outline of the proof of Theorem A Before launching into the proof of Theorem A, we outline the steps that we will take. Recall that Theorem A states that for an FI-module W, the degree of the FI-homology $H_p(W)$ can be bounded in terms of certain invariants of W. In this outline, whenever we speak of a "bound on" a particular FI-module, we mean a bound on its degree:

(1) We prove that a bound on $D^a X$ can be converted to a bound on $H_0(X)$ (Proposition 4.6).

(2) We show that Theorem E gives a bound on the degree of $H_1^{D^a}(W)$ for all a.

(3) Using homological properties of the functor D, we show that this bound on $H_1^{D^a}(W)$ implies a bound on $H_p^{D^a}(W)$ for all p and all a.

(4) If X_p is the p^{th} syzygy of W,[5] it is almost true that $H_p(W) = H_0(X_p)$; specifically, we have $H_p(W) = H_1(X_{p-1})$ and $H_p(W) \hookrightarrow H_0(X_p)$. Similarly, it is almost true that $H_p^{D^a}(W) = D^a(X_p)$, and we prove that for sufficiently large a this *is* true. Therefore, by using step (1), we can convert our bound on $H_p^{D^a}(W)$ to the desired bound on $H_p(W)$.

4.1 Relations and H_1

Our main theorems will be proved in terms of a presentation of the FI-module in question. We saw in Lemma 2.2 that W is generated in degree at most k if and only if $\deg H_0(V) \leq k$. The existence of a presentation for W with relations in degree at most d is very close to the condition $\deg H_1(W) \leq d$, but they are not quite equivalent.[6] Therefore, we distinguish these in our terminology as follows.

Definition 4.1 We say that an FI-module W is generated in degree at most k and related in degree at most d if there exists a short exact sequence

$$0 \to V \to M \to W \to 0,$$

where M is a free FI-module generated in degree at most k and V is generated in degree at most d.

Proposition 4.2 *Any FI-module W is generated in degree at most* $\deg H_0(W)$ *and related in degree at most* $\max(\deg H_0(W), \deg H_1(W))$.

Proof Set $M := M(W_{\langle \leq \deg H_0(W) \rangle})$. By Lemma 2.2, the natural map $M \twoheadrightarrow W$ is surjective.

[5] Here we consider syzygies relative to a free resolution of W that is *minimal* in the sense that all maps become 0 after applying H_0.

[6] For instance, a FI-module W admitting a finite-length filtration whose graded pieces are free has $H_1(W) = 0$, but such a W need not itself be free (recall that free FI-modules need not be projective). If we could always find a surjection $M(H_0(W)) \twoheadrightarrow W$ lifting the isomorphism on H_0, there would be no problem, but such a surjection does not always exist. For example, it can happen that $H_0(W)_n \simeq \mathbb{Z}/2\mathbb{Z}$ while W_n is a free abelian group, in which case there is no map $H_0(W)_n \to W_n$ at all.

Let V be its kernel, so that $0 \to V \to M \to W \to 0$ is a presentation of W as in Definition 4.1. By Lemma 2.3 M is H_0–acyclic, so we have the exact sequence

$$0 \to H_1(W) \to H_0(V) \to H_0(M).$$

From this, we conclude that $\deg H_0(V)$ is bounded by the degrees of the other two terms. Since $H_0(M) = W_{\leq \deg H_0(W)}$, we see in particular that $\deg H_0(M) = \deg H_0(W)$. Thus V is generated in degree at most $\max(\deg H_0(W), \deg H_1(W))$, as desired. □

From this proposition, we see that relations will indeed behave as we would expect, as long as $\deg H_0(W) \leq \deg H_1(W)$. We will reduce to this case in the proof of Theorem A using the following proposition, whose proof was explained to us by Eric Ramos; we are grateful to the referee for suggesting the current statement. Similar arguments appear in Li and Yu [11] in the proof of Corollary 3.4 and the second proof of Lemma 3.3.

Given an FI-module W and some $m \geq 0$, consider the FI-module $Z = W_{\langle \leq m\rangle} / W_{\langle < m\rangle}$. Note that $H_0(Z)$ vanishes except in degree m, where $H_0(Z)_m = Z_m = H_0(W)_m$, so we have a surjection $M(Z_m) \twoheadrightarrow Z$. In terms of the original FI-module W, we have a natural surjection from $M(H_0(W)_m)$ to $W_{\langle \leq m\rangle} / W_{\langle < m\rangle}$ which is an isomorphism in degree m.

Proposition 4.3 *Let W be an FI-module with $\deg H_0(W) < \infty$. Then the natural surjection*

$$M(H_0(W)_m) \twoheadrightarrow W_{\langle \leq m\rangle} / W_{\langle < m\rangle}$$

is an isomorphism whenever $m \geq \deg H_1(W)$ or $m > \deg H_0(W)$. In particular, the inclusion $W_{\langle < \deg H_1(W)\rangle} \hookrightarrow W$ induces an isomorphism on H_i for all $i > 0$.

Proof We proceed by reverse induction on m, showing both that $M(H_0(W)_m) \simeq W_{\langle \leq m\rangle} / W_{\langle < m\rangle}$ *and* that the inclusion $W_{\langle < m\rangle} \hookrightarrow W$ induces isomorphisms on H_i for all $i > 0$. Our base case consists of all $m > \deg H_0(W)$, when both claims are essentially tautological: in this case $M(H_0(W))_m = 0$ and $W_{\langle < m\rangle} = W_{\langle \leq m\rangle} = W$, so both sides of the claimed isomorphism vanish, proving the first claim. Similarly $W_{\langle < m\rangle} = W$ if $m > \deg H_0(W)$, so the second claim is automatic.

For "usual" FI-modules with $\deg H_0(W) < \deg H_1(W)$ there is nothing left to prove; it remains to handle FI-modules with $\deg H_1(W) \leq \deg H_0(W)$.

For the inductive step, write Z for the quotient $Z := W_{\langle \leq m\rangle} / W_{\langle < m\rangle}$, and let A be the kernel of the surjection $M(Z_m) \twoheadrightarrow Z$, so that $0 \to A \to M(Z_m) \to Z \to 0$. The

FI-module A vanishes in degrees at most m: in degree m the map $M(Z_m)_m \to Z_m$ is an isomorphism, and $M(Z_m)$ itself is zero in degrees less than m. Since A vanishes in degrees at most m, $H_0(A)$ also vanishes in degrees at most m; since $H_0(M(Z_m))$ vanishes in degrees greater than m, the map $H_0(A) \to H_0(M(Z_m))$ is zero. Since $M(Z_m)$ is H_0–acyclic, we conclude that there is an isomorphism $H_1(Z) \simeq H_0(A)$.

Now consider the long exact sequence

$$\cdots \to H_1(W_{\langle \leq m \rangle}) \to H_1(Z) \to H_0(W_{\langle < m \rangle}) \to \cdots .$$

By induction, we know that $\deg H_1(W_{\langle \leq m \rangle}) = \deg H_1(W) \leq m$, and by definition $\deg H_0(W_{\langle < m \rangle}) < m$. Therefore, $\deg H_1(Z) \leq m$.

We showed above that $H_1(Z)$ vanishes in degrees greater than m, while $H_0(A)$ vanishes in degrees at most m, so $H_1(Z) = H_0(A) = 0$. Therefore, $A = 0$, and the natural map $M(H_0(W)_m) = M(Z_m) \to Z$ is an isomorphism, as claimed. Since free FI-modules are H_0–acyclic, we conclude that the inclusion $W_{\langle < m \rangle} \hookrightarrow W_{\langle \leq m \rangle}$ induces an isomorphism on H_i for all $i > 0$; the inclusion $W_{\langle \leq m \rangle} \hookrightarrow W$ induces an isomorphism on H_i for all $i > 0$ by induction, so we have proved the inductive hypothesis. □

4.2 Homological properties of the derivative

Considering FB-Mod as a full subcategory of FI-Mod, the functor S restricts to a functor S: FB-Mod → FB-Mod.

Lemma 4.4 *There is a natural isomorphism of functors*

$$D \circ M = M \circ S\colon \text{FB-Mod} \to \text{FI-Mod}.$$

Proof There is automatically a natural transformation $M \circ S \to D \circ M$. It would suffice to check that this is an isomorphism on free FB-modules, but it will be no more difficult to check this on arbitrary FB-modules W. From the formula (2) for $M(W)$ we see that $(SM(W))_T = \bigoplus_{S \subset T \sqcup \{\star\}} W_S$, with $\iota\colon M(W) \to SM(W)$ the inclusion of those summands with $\star \notin S$. (Incidentally, this shows that free FI-modules are torsion-free.) It follows that

$$(DM(W))_T = \bigoplus_{\substack{S \subset T \sqcup \{\star\} \\ \star \in S}} W_S = \bigoplus_{R \subset T} W_{R \sqcup \{\star\}} = \bigoplus_{R \subset T} (SW)_R = M(SW)_T,$$

as claimed (for the second equality we reindex by $S = R \sqcup \{\star\}$). It is straightforward to check that this identification agrees on morphisms as well. □

Corollary 4.5 *Free FI-modules are D^a–acyclic for all $a \geq 1$.*

Proof Just as in the proof of Lemma 2.3, we have $H_p^{D^a}(M(W)) \simeq H_p^{D^a \circ M}(W)$. However, Lemma 4.4 implies that $D^a \circ M = M \circ S^a$. This is exact since both M and S are, so $H_p^{M \circ S^a} = 0$ for all $p > 0$. □

Proposition 4.6 *If V is an FI-module generated in degree at most k, then $D^a V = 0$ for all $a > k$. On the other hand, if $\deg D^a V \leq m$ for some $m \geq -1$, then V is generated in degree at most $m + a$.*

Proof If V is generated in degree at most k, there is a surjection $M(V_{\leq k}) \twoheadrightarrow V$. Since D^a is right exact, we have a surjection $D^a M(V_{\leq k}) \twoheadrightarrow D^a V$ for any a. By Lemma 4.4, $D^a M(V_{\leq k}) \simeq M(S^a V_{\leq k})$. However, $S^a V_{\leq k} = 0$ when $a > k$, since $(S^a V_{\leq k})_R = (V_{\leq k})_{R \sqcup [\star_a]} = 0$. Therefore, $D^a V = 0$ for $a > k$.

For the second claim, to say that $\deg D^a V \leq m$ means that $(D^a V)_T = 0$ whenever $|T| > m$. The formula (3) for $(D^a V)_T$ shows that the defining surjection $V_{T \sqcup [\star_a]} \twoheadrightarrow H_0(V)_{T \sqcup [\star_a]}$ factors through $(D^a V)_T \twoheadrightarrow H_0(V)_{T \sqcup [\star_a]}$, so it follows that $H_0(V)_R = 0$ whenever $|R| > m + a$. In other words, V is generated in degree at most $m + a$. □

The derived functors of D We can now establish the basic properties of the derived functors H_p^D of the derivative D.

Lemma 4.7 *Let W be an FI-module.*

(i) *The derived functor H_1^D coincides with K, so there is a natural exact sequence*
$$0 \to H_1^D(W) \to W \to SW \to DW \to 0.$$
(ii) *W is torsion-free if and only if $H_1^D(W) = 0$.*

(iii) *$H_p^D = 0$ for all $p > 1$.*

(iv) *D takes projective FI-modules to projective FI-modules.*

(v) *If Y is an FI-module of finite degree, then we have $\deg DY \leq \deg Y - 1$ and $\deg H_1^D(Y) \leq \deg Y$.*

Proof Given W, let M be a free FI-module with $M \twoheadrightarrow W$; for instance, we may take the universal $M = M(W) \twoheadrightarrow W$. Let V be the kernel of this surjection, so we have $0 \to V \to M \to W \to 0$. Since M is free, $H_1^D(M) = 0$ by Corollary 4.5, so we have an isomorphism $H_1^D(W) \simeq \ker(DV \to DM)$.

(i) The key properties are that S is exact and that $M \xrightarrow{\iota} SM$ is injective, ie that free FI-modules are torsion-free (which we saw in the proof of Lemma 4.4). Thus we have a diagram:

$$\begin{tikzcd}
 & V \arrow[r] \arrow[d] & SV \arrow[r] \arrow[d] & DV \arrow[r] \arrow[d] & 0 \\
0 \arrow[r] & M \arrow[r] & SM \arrow[r] & DM \arrow[r] & 0
\end{tikzcd}$$

Applying the snake lemma, we obtain the desired exact sequence

$$\ker(SV \to SM) = 0 \to H_1^D(W) \to W \to SW \to DW \to 0.$$

In particular, this identifies $H_1^D(W)$ with $KW = \ker(W \to SW)$.

(ii) Given that $H_1^D = K$, this is the statement of Lemma 2.5.

(iii) Since M is D–acyclic, we have $H_2^D(W) \simeq H_1^D(V)$. The FI-module V is torsion-free, being a submodule of M, so $H_1^D(V) = 0$ by (ii). Since W was arbitrary, this proves that $H_2^D = 0$, which implies that $H_p^D = 0$ for all $p > 1$.

(iv) Since projective FI-modules are summands of $\bigoplus M(m_i)$, it suffices to prove this for $M(m) = M(\mathbb{Z}[S_m])$. Lemma 4.4 states that

$$DM(m) = DM(\mathbb{Z}[S_m]) \simeq M(S\mathbb{Z}[S_m]) \simeq M\left(\bigoplus_{i=1}^{m} \mathbb{Z}[S_{m-1}]\right) = \bigoplus_{i=1}^{m} M(m-1),$$

which is indeed projective.

(v) It is clear that $\deg SY = \deg Y - 1$, since $(SY)_n = Y_{[n]\sqcup[\star]} \simeq Y_{n+1}$ (unless $\deg Y = 0$, when $\deg SY = -\infty$). Both claims now follow from (i), the first from the surjection $SY \twoheadrightarrow DY$ and the second from the injection $H_1^D(Y) \hookrightarrow Y$. □

4.3 Proof of Theorem A

We now have in place all the tools we need to prove our main theorems bounding the degree of homology of FI-modules. The key technical result is Theorem E, together with Lemma 2.6 and Remark 2.7 establishing a connection between its conclusion and D^a.

Theorem 4.8 *Let W be an FI-module generated in degree at most k and related in degree at most d, and let $N := d + \min(k, d) - 1$. For all $a \geq 1$ and all $p \geq 1$,*

$$(*_p^a) \qquad \deg H_p^{D^a}(W) \leq N - a + p.$$

Proof We will reduce by induction to the case when $a=1$ or $p=1$. To accomplish this reduction, we prove that $(*_p^{a-1})+(*_{p-1}^{a-1}) \implies (*_p^a)$ for any $a \geq 2$ and $p \geq 2$.

Fix $a \geq 2$ and $p \geq 2$. By Lemma 4.7(iv), D^a takes projective FI-modules to projective FI-modules, so we may compute the left derived functors of D^a by means of the Grothendieck spectral sequence applied to the composition $D \circ D^{a-1}$. Thanks to the vanishing of H_p^D for $p>1$ from Lemma 4.7(iii), this spectral sequence has only two nonzero columns, so it degenerates to the short exact sequences

$$0 \to DH_p^{D^{a-1}}(W) \to H_p^{D^a}(W) \to H_1^D(H_{p-1}^{D^{a-1}}W) \to 0. \tag{18}$$

The assertions $(*_p^{a-1})$ and $(*_{p-1}^{a-1})$ state respectively that

$$\begin{aligned}\deg H_p^{D^{a-1}}(W) &\leq N-(a-1)+p = N-a+p+1,\\ \deg H_{p-1}^{D^{a-1}}(W) &\leq N-(a-1)+(p-1) = N-a+p.\end{aligned}$$

Lemma 4.7(v) tells us that $\deg DY \leq \deg Y - 1$ and $\deg H_1^D(Y) \leq \deg Y$, so these bounds imply

$$\begin{aligned}\deg DH_p^{D^{a-1}}(W) &\leq N-a+p,\\ \deg H_1^D(H_{p-1}^{D^{a-1}}(W)) &\leq N-a+p.\end{aligned}$$

The short exact sequence (18) now implies

$$\deg H_p^{D^a}(W) \leq N-a+p,$$

which is precisely the assertion $(*_p^a)$. This establishes that $(*_p^{a-1})+(*_{p-1}^{a-1}) \implies (*_p^a)$ for any $a \geq 2$ and $p \geq 2$.

Given this implication, it suffices to prove directly that $(*_p^a)$ holds when either $a=1$ or $p=1$, since all remaining cases with $a \geq 2$ and $p \geq 2$ then follow by induction. When $a=1$ and $p \geq 2$, we have $H_p^D(W)=0$ by Lemma 4.7(iii), so $\deg H_p^D(W) = -\infty$ and the bound $(*_p^a)$ certainly holds. What remains as the unavoidable core of the problem is the bound $(*_p^a)$ when $p=1$, namely that $\deg H_1^{D^a}(W) \leq N-a+1$ for all $a \geq 1$.

To compute $H_1^{D^a}(W)$, consider a presentation $0 \to V \to M \to W \to 0$ as in Definition 4.1, with M free and generated in degree at most k and V generated in degree at most d. Since M is free, it is D^a–acyclic by Corollary 4.5, so

$$H_1^{D^a}(W) \simeq \ker(D^a V \to D^a M).$$

But recall from Remark 2.7 that the conclusion of Theorem E can be restated as a claim about the map $D^a V \to D^a M$ and its kernel! Specifically, the conclusion of

Theorem E says for any $a \geq 1$, that

$$\ker(D^a V \to D^a M)_{n-a} = 0 \quad \text{for all } n > d + \min(k, d),$$

or in other words, that

$$\deg \ker(D^a V \to D^a M) \leq d + \min(k, d) - a = N + 1 - a.$$

This means that the conclusion of Theorem E applied to the relations $V \subset M$ is precisely the claim $(*^a_p)$ for $p = 1$ and all $a \geq 1$. As explained above, all other cases now follow by induction. $\square$

Proof of Theorem A Fix $k' \geq 0$ and $d \geq 0$, and let U be an FI-module with $\deg H_0(U) \leq k'$ and $\deg H_1(U) \leq d$. Our goal is to prove that $\deg H_p(U) - p \leq k' + d - 1$ for all $p > 0$.

We first reduce to the case when $k' < d$. Let $k := \min(k', d-1)$ and define W to be the submodule $W := U_{\langle \leq k \rangle}$. In the most common case when $k' < d$, this has no effect: we have $k = k'$ and $W = U$. In the other case when $k' \geq d$, we have $\deg H_1(U) \leq d = k+1$, so Proposition 4.3 states that $H_p(W) \simeq H_p(U)$ for all $p > 0$. Since $k \leq k'$ in either case, to prove the theorem it suffices to prove that

$$\deg H_p(W) - p \leq k + d - 1 \quad \text{for all } p > 0.$$

For the rest of the proof, we discard the FI-module U and work only with W, which has $\deg H_0(W) \leq k$ and $\deg H_1(W) \leq d$ with $k < d$.

Given these bounds, Proposition 4.2 tells us that W is generated in degree at most k and related in degree at most $\max(k, d) = d$. Therefore, there exists a surjection $M \twoheadrightarrow W$ from a free FI-module M generated in degree at most k, whose kernel is generated in degree at most d. Set $M_0 := M$ and extend this to a resolution of W by free FI-modules:

$$\cdots \to M_2 \to M_1 \to M_0 \to W \to 0.$$

For each $p > 0$, let X_p be the p^{th} syzygy of W, namely $X_p := \operatorname{im}(M_p \to M_{p-1}) \simeq \ker(M_{p-1} \to M_{p-2})$. Let us assume that this resolution is *minimal* in the very weak sense that $\deg H_0(X_p) = \deg H_0(M_p)$ for all $p > 0$. (The existence of such a resolution is a consequence of the fact that every FI-module V generated in degree at most k admits a surjection from a free FI-module generated in degree at most k, namely $M(V_{\leq k})$ as discussed in Proposition 4.6.) Set $X_0 := W$.

For all $p \geq 1$, we have an exact sequence

$$0 \to X_p \to M_{p-1} \to X_{p-1} \to 0. \tag{19}$$

Since the M_i are H_0–acyclic by Corollary 4.5, applying H_0 to (19) gives $H_i(X_p) \simeq H_{i+1}(X_{p-1})$ for all $i \geq 1$; iterating, we obtain $H_p(W) \simeq H_1(X_{p-1})$. Similarly, $H_p^{D^a}(W) \simeq H_1^{D^a}(X_{p-1})$ for any $a \geq 1$.

Let us write

$$N := d + k - 1;$$

so our eventual goal is to prove that $\deg H_p(W) \leq N + p$ for all $p \geq 1$. Before that, we prove that for all $p \geq 1$,

$$\deg H_0(X_p) \leq N + p. \tag{20}$$

By construction $X_1 = \ker(M \to W)$; by our hypothesis, X_1 is generated in degree at most d, so $\deg H_0(X_1) \leq d \leq d + k = N + 1$. This proves (20) for $p = 1$; we proceed by induction on p.

Fix $p \geq 2$, and assume by induction that (20) holds for $p-1$, ie that $\deg H_0(X_{p-1}) \leq N + p - 1$. By minimality of the resolution, $\deg H_0(M_{p-1}) = \deg H_0(X_{p-1})$, so M_{p-1} is generated in degree at most $N + p - 1$. By Proposition 4.6, this implies that $D^{N+p} M_{p-1} = 0$. Then applying D^{N+p} to (19) yields a long exact sequence containing the segment

$$H_1^{D^{N+p}}(X_{p-1}) \to D^{N+p} X_p \to 0 = D^{N+p} M_{p-1}.$$

This shows that $D^{N+p} X_p$ is a quotient of $H_1^{D^{N+p}}(X_{p-1}) \simeq H_p^{D^{N+p}}(W)$. We proved in Theorem 4.8 that

$$\deg H_p^{D^{N+p}}(W) \leq N - (N + p) + p = 0,$$

so $\deg D^{N+p} X_p \leq 0$. (The statement of Theorem 4.8 has $d + \min(k, d) - 1$, but this coincides with $N = d + k - 1$ since $k < d$.) By Proposition 4.6, this implies that X_p is generated in degree at most $N + p$, which is the result to be proved. This concludes the proof of (20).

We saw above that (19) implies $H_i(X_p) \simeq H_{i+1}(X_{p-1})$ for $i \geq 1$. To complete the proof of the theorem, we consider the segment of the long exact sequence involving $i = 0$:

$$0 = H_1(M_{p-1}) \to H_1(X_{p-1}) \to H_0(X_p) \to H_0(M_{p-1}) \to H_0(X_{p-1}).$$

This shows that $H_p(W) \simeq H_1(X_{p-1})$ injects into $H_0(X_p)$ for all $p > 0$. We proved in (20) that $\deg H_0(X_p) \leq N + p$, so we conclude that $\deg H_p(W) \leq N + p$ for all $p > 0$, as desired. □

5 Application to homology of congruence subgroups

5.1 A complex computing $H_i(V)$

For any category $\mathcal{C}$, let $\mathcal{C}$–Mod denote the category of functors $\mathcal{C} \to \mathbb{Z}$–Mod. Given $V \in \mathcal{C}$–Mod and $W \in \mathcal{C}^{\mathrm{op}}$–Mod, their tensor product over $\mathcal{C}$ is an abelian group $V \otimes_{\mathcal{C}} W$. It can be defined as the largest quotient of

$$\bigoplus_{X \in \mathrm{Ob}\,\mathcal{C}} V(X) \otimes_{\mathbb{Z}} W(X)$$

in which

$$v_X \otimes f^*(w_Y) \in V(X) \otimes W(X) \;\equiv\; f_*(v_X) \otimes w_Y \in V(Y) \otimes W(Y)$$

for all $X, Y \in \mathrm{Ob}\,\mathcal{C}$, $v_X \in V(X)$, $w_Y \in W(Y)$ and $f \in \mathrm{Hom}_{\mathcal{C}}(X, Y)$.[7]

In this paper we will be interested in the tensor product of an FI-module V and co-FI-module W. This can be described explicitly as follows.

Definition 5.1 Given $V \in$ FI-Mod and $W \in \mathrm{FI}^{\mathrm{op}}$-Mod, the abelian group $V \otimes_{\mathrm{FI}} W$ is defined by

$$\begin{aligned} V \otimes_{\mathrm{FI}} W \\ &= \Big(\bigoplus_{T \in \mathrm{Ob}\,\mathrm{FI}} V_T \otimes_{\mathbb{Z}} W_T\Big) \Big/ \big\langle f_*(v_S) \otimes w_T \equiv v_S \otimes f^*(w_T) \;\big|\; f\colon S \hookrightarrow T\big\rangle \\ &= \Big(\bigoplus_{n \geq 0} V_n \otimes_{\mathbb{Z}S_n} W_n\Big) \Big/ \big\langle f_*(v_n) \otimes w_{n+1} \equiv v_n \otimes f^*(w_{n+1}) \;\big|\; f\colon [n] \hookrightarrow [n+1]\big\rangle. \end{aligned}$$

We think of an FI-module $V \in$ FI-Mod as a "right module over FI", and a co-FI-module $W \in \mathrm{FI}^{\mathrm{op}}$-Mod as a "left module over FI". This is consistent with our notation $V \otimes_{\mathrm{FI}} W$ for the tensor. Moreover, if W is an $\mathrm{FI}^{\mathrm{op}} \times \mathrm{FI}$-module, we will say that W is an *FI-bimodule*; in this case $V \otimes_{\mathrm{FI}} W$ is not just an abelian group, but in fact an FI-module. This is familiar from the analogous situation with R–modules: the tensor of a right R–module with an R–bimodule is a right R–module. To verify the claim in this setting, just note that

$$(\mathrm{FI}^{\mathrm{op}} \times \mathrm{FI})\text{-Mod} = [\mathrm{FI}^{\mathrm{op}} \times \mathrm{FI}, \mathbb{Z}\text{–Mod}] = [\mathrm{FI}, [\mathrm{FI}^{\mathrm{op}}, \mathbb{Z}\text{–Mod}]] = [\mathrm{FI}, \mathrm{FI}^{\mathrm{op}}\text{-Mod}].$$

[7] The reader may recognize this as an example of a *coend*: given V and W we can define a functor $V \boxtimes W\colon \mathcal{C} \times \mathcal{C}^{\mathrm{op}} \to \mathbb{Z}$–Mod; then $V \otimes_{\mathcal{C}} W$ coincides with the coend $\int^{C} V \boxtimes W$, and the quotient construction above is just the standard coequalizer formula for a coend.

In other words, we can think of an FI-bimodule W as a functor from FI to $\mathrm{FI}^{\mathrm{op}}$-Mod; after tensoring with $V \in$ FI-Mod, we are left with a functor from FI to $\mathbb{Z}$–Mod, which is just an FI-module.

Definition 5.2 The $\mathrm{FI}^{\mathrm{op}} \times \mathrm{FI}$-module K is defined on objects by

$$K(S,T) = \mathbb{Z}[\mathrm{Bij}(S,T)];$$

in particular, $K(S,T) = 0$ if $|S| \neq |T|$. Given a morphism

$$(f\colon S' \hookrightarrow S,\ g\colon T \hookrightarrow T') \quad \text{in } \mathrm{Hom}_{\mathrm{FI}^{\mathrm{op}} \times \mathrm{FI}}\big((S,T),(S',T')\big),$$

we consider two cases. If f and g are both bijective, $K_{(f,g)}\colon K(S,T) \to K(S',T')$ is the map defined by $\mathrm{Bij}(S,T) \ni \varphi \mapsto g \circ \varphi \circ f \in \mathrm{Bij}(S',T')$. If either f or g is not bijective, $K_{(f,g)} = 0$.

Since K is an FI-bimodule, the tensor $V \otimes_{\mathrm{FI}} K$ is itself an FI-module. In fact, this FI-module is already familiar to us! To avoid confusion, in the remainder of the paper we will write $H_i^{\mathrm{FI}}(V)$ for the FI-homology of V, which was denoted simply by $H_i(V)$ in previous sections.

Proposition 5.3 *Given $V \in$ FI-Mod, the FI-module $V \otimes_{\mathrm{FI}} K$ is isomorphic to the FI-module $H_0^{\mathrm{FI}}(V)$ defined in the introduction. As a consequence,*

$$H_i^{\mathrm{FI}}(V) = \mathrm{Tor}_i^{\mathrm{FI}}(V,K) \quad \text{for any } i \geq 0.$$

Proof Definition 5.1 presents $V \otimes_{\mathrm{FI}} K$ as a quotient of

$$\bigoplus_{n \geq 0} V_n \otimes_{\mathbb{Z}S_n} K_n,$$

so we first identify the FI-module $V_n \otimes_{\mathbb{Z}S_n} K_n$. Since K is not only a co-FI-module but an FI-bimodule, K_n is an $S_n \times \mathrm{FI}$-module: as an FI-module K_n sends a set T to $\mathbb{Z}[\mathrm{Bij}([n],T)]$, and the action of S_n by precomposition commutes with this FI-module structure. Thus the FI-module $V_n \otimes_{\mathbb{Z}S_n} K_n$ sends T to V_T if $|T| = n$, and to 0 if $|T| \neq n$. Passing to the direct sum, we find that $\bigoplus_{n \geq 0} V_n \otimes_{\mathbb{Z}S_n} K_n$ sends T to V_T for any finite set T of any cardinality; in other words, the FI-module $\bigoplus_{n \geq 0} V_n \otimes_{\mathbb{Z}S_n} K_n$ can be identified with V itself.

We now consider the relations: Definition 5.1 states that $V \otimes_{\mathrm{FI}} K$ is the quotient of $V \simeq \bigoplus V_n \otimes_{\mathbb{Z}S_n} K_n$ by the relations

$$f_*(v_n) \otimes k_{n+1} \equiv v_n \otimes f^*(k_{n+1}) \quad \text{for all } f\colon [n] \hookrightarrow [n+1].$$

However, by definition, f^* acts as 0 on K whenever f is not bijective. Therefore, these relations reduce to $f_*(v_n) \equiv 0$ for all $v_n \in V_n$ and $f\colon [n] \hookrightarrow [n+1]$. The quotient of $\bigoplus_n V_n$ by these relations is precisely $H_0^{\mathrm{FI}}(V)$ as we defined it in the introduction. The assertion that $H_i^{\mathrm{FI}}(V) = \mathrm{Tor}_i^{\mathrm{FI}}(V, K)$ is then tautological (but see Remarks 5.4 and 5.5 for further discussion). □

Remark 5.4 The notation $\mathrm{Tor}_*^{\mathrm{FI}}(V, W)$ requires some justification, since this could denote the left-derived functors of $V \otimes_{\mathrm{FI}} -$ or of $- \otimes_{\mathrm{FI}} W$. Fortunately, the tensor product functor

$$- \otimes_{\mathrm{FI}} -\colon \text{FI-Mod} \times \mathrm{FI}^{\mathrm{op}}\text{-Mod} \to \mathbb{Z}\text{–Mod}$$

is a left-balanced functor in the sense of [15, Definition 2.7.7], so by [15, Exercise 2.7.4] its left-derived functors in the first variable and in the second variable coincide. In other words, these derived functors $\mathrm{Tor}_*^{\mathrm{FI}}(V, W)$ can be computed either from a resolution $V_\bullet$ of V by projective FI-modules, or from a resolution $W_\bullet$ of W by projective $\mathrm{FI}^{\mathrm{op}}$-modules, as we would expect.

Remark 5.5 When W is an FI-bimodule, $V \otimes_{\mathrm{FI}} W$ and thus $\mathrm{Tor}_*^{\mathrm{FI}}(V, W)$ are FI-modules, but there is one important point to make. We can compute the FI-module $\mathrm{Tor}_i^{\mathrm{FI}}(V, W)$ from a projective resolution $W_\bullet \to W$ of FI-bimodules, but in fact something much weaker suffices. We do not need the terms W_i of this resolution to be *projective* FI-bimodules; it suffices that each FI-bimodule W_i be "$\mathrm{FI}^{\mathrm{op}}$-projective", meaning that for each finite set $T \in \mathrm{Ob}\,\mathrm{FI}$ the $\mathrm{FI}^{\mathrm{op}}$-module $(W_i)_T$ is a projective $\mathrm{FI}^{\mathrm{op}}$-module.

This is familiar from the situation of R–modules: if M is a right R–module and N is an R-S–bimodule, then to compute the S–modules $\mathrm{Tor}_*^R(M, N)$ from a resolution $N_\bullet \to N$ by R-S–bimodules, it suffices that each N_i be projective (or even flat) as an R–module. The reason is that such an R-S–bimodule is acyclic for the functor $M \otimes_R -\colon R\text{-}S\text{–Mod} \to S\text{–Mod}$; the situation for FI-modules is the same.

The only projective $\mathrm{FI}^{\mathrm{op}}$-modules we will need to consider are the corepresentable functors $\mathbb{Z}[\mathrm{Inj}(-, U)] = \mathbb{Z}[\mathrm{Hom}_{\mathrm{FI}^{\mathrm{op}}}(U, -)]$ for a fixed finite set U (such corepresentable functors are always projective).

We may therefore describe $H_i^{\mathrm{FI}}(V)$ in a uniform way that applies to all FI-modules V by finding an appropriate resolution $C_\bullet \to K$ of $\mathrm{FI}^{\mathrm{op}}$-projective FI-bimodules.

A uniform construction of FI-complexes We will make use of the same construction in multiple places below, so we begin by describing this construction in a general context; we are grateful to the referee for suggesting this.

Definition 5.6 We denote by $\mathrm{FI}^{\leadsto}$ the twisted arrow category whose objects are pairs (T, U) where T is a finite set and $U \subset T$ is a subset, and where a morphism from (T, U) to (T', U') is an injection $f \colon T \hookrightarrow T'$ such that $f(U) \supseteq U'$.

Given an $\mathrm{FI}^{\leadsto}$-module F, we will construct two chain complexes of FI-modules. In fact, for any functor F from $\mathrm{FI}^{\leadsto}$ to any abelian category $\mathcal{A}$, we construct two chain complexes $C_\bullet^F$ and $\widetilde{C}_\bullet^F$ taking values in $[\mathrm{FI}, \mathcal{A}]$.

Construction 5.7 (the complexes $C_\bullet^F$ and $\widetilde{C}_\bullet^F$) Given a functor $F \colon \mathrm{FI}^{\leadsto} \to \mathcal{A}$, for each $k \geq 0$, define $\widetilde{C}_k^F \colon \mathrm{FI} \to \mathcal{A}$ by

$$\widetilde{C}_k^F(T) = \bigoplus_{f \colon [k] \hookrightarrow T} F(T, \operatorname{im} f).$$

An FI-morphism $g \colon T \hookrightarrow T'$ defines for each $f \colon [k] \hookrightarrow T$ an $\mathrm{FI}^{\leadsto}$-morphism $g \colon (T, \operatorname{im} f) \to (T', \operatorname{im} g \circ f)$, and $g_* \colon \widetilde{C}_k^F(T) \to \widetilde{C}_k^F(T')$ is given by the induced maps.

Next, we define the boundary map $\partial \colon \widetilde{C}_k^F \to \widetilde{C}_{k-1}^F$. For $k \geq 1$ and $1 \leq i \leq k$, let $\delta_i \colon [k-1] \hookrightarrow [k]$ be the ordered injection whose image does not contain i. For any $f \colon [k] \hookrightarrow T$, the identity id_T defines an $\mathrm{FI}^{\leadsto}$-morphism from $(T, \operatorname{im} f)$ to $(T, \operatorname{im} f \circ \delta_i)$. Let $d_i \colon \widetilde{C}_k^F \to \widetilde{C}_{k-1}^F$ be the map induced on each factor by $\mathrm{id}_T \colon (T, \operatorname{im} f) \to (T, \operatorname{im} f \circ \delta_i)$; note that this commutes with the FI-action g_* defined above.

We define $\partial \colon \widetilde{C}_k^F \to \widetilde{C}_{k-1}^F$ by $\partial := \sum (-1)^i d_i$. The familiar formula $\delta_i \circ \delta_j = \delta_{j+1} \circ \delta_i$ for $i \leq j$ implies that $d_j \circ d_i = d_i \circ d_{j+1}$ by the functoriality of F, so $\partial^2 = 0$. Therefore, the differential ∂ makes $\widetilde{C}_\bullet^F$ a chain complex with values in $[\mathrm{FI}, \mathcal{A}]$.

We define the complex $C_\bullet^F$ as the quotient of $\widetilde{C}_\bullet^F$ by the following relations. The permutations $\sigma \in S_k$ act on $\widetilde{C}_k^F$ by precomposition, and breaking up into orbits we have

$$\widetilde{C}_k^F(T) = \bigoplus_{U \subset T, |U| = k} \; \bigoplus_{f \colon [k] \simeq U} F(T, U).$$

We define C_k^F to be the quotient of $\widetilde{C}_k^F$ by the relations $\sigma_* = (-1)^\sigma$ for all $\sigma \in S_k$; in other words, we pass to the quotient where S_k acts by the sign representation. The functoriality of F guarantees that C_k^F is still a functor $\mathrm{FI} \to \mathcal{A}$.

The individual homomorphisms d_i do not respect these relations, so they do not descend to C_k^F. However, the alternating sum $\partial = \sum (-1)^i d_i$ does descend to a differential $\partial \colon C_k^F \to C_{k-1}^F$, and so we obtain a chain complex

$$C_\bullet^F = \quad \cdots \to C_k^F \to C_{k-1}^F \to \cdots \to C_0^F$$

with values in $[\mathrm{FI}, \mathcal{A}]$. Note that on objects we have

$$(C_k^F)_T = \bigoplus_{U \subset T, |U|=k} F(T,U),$$

where FI-morphisms act with a factor of ± 1 coming from the orientation of the subset U.

Remark 5.8 When the finite set T is fixed, the following standard argument shows that the chain complex $C_\bullet^F(T)$ is a summand of $\widetilde{C}_\bullet^F(T)$. Choosing an ordering of T, let $\widetilde{C}_\bullet^{\mathrm{ord}}(T)$ be the subcomplex of $\widetilde{C}_\bullet^F(T)$ spanned by those summands where $f\colon [k] \hookrightarrow T$ is order preserving. The differential ∂ preserves this subcomplex, and the projection $\widetilde{C}_\bullet^F(T) \to C_\bullet^F(T)$ restricts to an isomorphism $\widetilde{C}_\bullet^{\mathrm{ord}}(T) \simeq C_\bullet^F(T)$. However, we emphasize that $C_\bullet^F$ is *not* a summand of $\widetilde{C}_\bullet^F$ when these are considered as complexes of FI-modules.

We now use this construction to define a complex $C_\bullet \to K$ of FI-bimodules, which will give us our resolution of K.

Definition 5.9 (the complex $C_\bullet$) Given $U \subset T$, let $F(T,U)$ be the $\mathrm{FI}^{\mathrm{op}}$-module defined by $F(T,U)_S = \mathbb{Z}[f\colon S \hookrightarrow T \setminus U]$. One easily checks that this defines a functor $F\colon \mathrm{FI}^{\leadsto} \to \mathrm{FI}^{\mathrm{op}}\text{-Mod}$, so Construction 5.7 defines a chain complex $C_\bullet := C_\bullet^F$ with values in $[\mathrm{FI}, \mathrm{FI}^{\mathrm{op}}\text{-Mod}]$, ie a complex of FI-bimodules. Concretely, $C_k(S,T)$ is the free abelian group on pairs $(U \subset T, f\colon S \hookrightarrow T)$ where $|U| = k$ and $\operatorname{im} f$ is disjoint from U.

Remark See [5, Equation (10)] and the surrounding section for more discussion of this complex. A caution: we could similarly have defined a complex $\widetilde{C}_\bullet^F$ of FI-bimodules, but be warned that the $\mathrm{FI}^{\mathrm{op}} \times \mathrm{FI}$-module B_V discussed following [5, Corollary 2.18] is *not* isomorphic to $\widetilde{C}_\bullet^F$, although they contain much the same information.

The resolution $C_\bullet \to K$ We consider the augmentation map $\partial\colon C_0 \to K$ defined by

$$C_0(S,T) \ni (\varnothing, f\colon S \hookrightarrow T) \mapsto \begin{cases} f \in \mathrm{Bij}(S,T) & \text{if } |S| = |T|, \\ 0 & \text{if } |S| < |T|, \end{cases} \in K(S,T).$$

Since $C_1(S,T)$ has basis $(\{u\} \subset T, f\colon S \hookrightarrow T \setminus \{u\})$, the composition $\partial^2\colon C_1 \to C_0 \to K$ is 0. Therefore, this augmentation extends $C_\bullet$ to a complex

$$\cdots \to C_1 \to C_0 \to K \to 0.$$

Proposition 5.10 *The complex $C_\bullet \to K$ is a resolution of K by FI$^{\mathrm{op}}$-projective FI-bimodules. As a consequence, given any FI-module V, the FI-homology of V is computed by the FI-chain complex $V \otimes_{\mathrm{FI}} C_\bullet$:*

$$H_i^{\mathrm{FI}}(V) = H_i(V \otimes_{\mathrm{FI}} C_\bullet).$$

Proof We first verify that $C_\bullet \to K$ is a resolution, ie that $H_0(C_\bullet) \simeq K$ and $H_*(C_\bullet) = 0$ for $* > 0$. It suffices to check this pointwise, so fix finite sets S and T and consider the chain complex of abelian groups $C_\bullet(S,T)$.

For each $h\colon S \hookrightarrow T$, let $C_k^h(S,T)$ be the summand of $C_k(S,T)$ spanned by the elements of the form (U,h). The differential ∂ preserves this summand, so we have a direct sum decomposition $C_\bullet(S,T) = \bigoplus_{h\colon S\hookrightarrow T} C_\bullet^h(S,T)$. Similarly, let $K^h(S,T)$ be the corresponding summand of $K(S,T)$; concretely, this summand is isomorphic to $\mathbb{Z}$ if h is bijective and 0 otherwise. It therefore suffices to show for fixed $h\colon S \hookrightarrow T$ that $C_\bullet^h(S,T)$ is a resolution of $K^h(S,T)$.

Let $\Delta^{T-h(S)}$ be the $(|T-h(S)|-1)$–dimensional simplex with vertex set $T-h(S)$, and let $\widetilde{C}_\bullet(\Delta^{T-h(S)})$ be its reduced cellular chain complex. A basis for $C_k^h(S,T)$ is given by the k–element subsets U of $T-h(S)$, oriented appropriately. In other words, we can identify $C_k^h(S,T) \simeq \widetilde{C}_{k-1}(\Delta^{T-h(S)})$, and this extends to an isomorphism of chain complexes $C_\bullet^h(S,T) \simeq \widetilde{C}_{\bullet-1}(\Delta^{T-h(S)})$.

If $T-h(S)$ is nonempty, the simplex $\Delta^{T-h(S)}$ is contractible, so $H_*(C_\bullet^h(S,T)) \simeq \widetilde{H}_{*-1}(\Delta^{T-h(S)}) = 0$ for all $* \geq 0$. Since $K^h(S,T) = 0$ when h is not bijective, this is as desired. In the remaining case when h is a bijection and $\Delta^{T-h(S)}$ is empty, the only nonzero term of this resolution is $C_0^h(S,T) \simeq \widetilde{C}_{-1}(\varnothing) \simeq \mathbb{Z} \simeq K^h(S,T)$, which again is as desired.

We next verify that the FI-bimodules C_k are FI$^{\mathrm{op}}$-projective, meaning that for each finite set T the FI$^{\mathrm{op}}$-module $C_k(-,T)$ is a projective FI$^{\mathrm{op}}$-module. For a fixed k–element subset $U \subset T$, let $C_k^U(S,T)$ be the summand of $C_k(S,T)$ spanned by elements $(U, f\colon S \hookrightarrow T\setminus U)$. These summands are preserved by FI$^{\mathrm{op}}$-morphisms, so this defines a summand $C_k^U(-,T)$ of the FI$^{\mathrm{op}}$-module $C_k(-,T)$. This summand $C_k^U(-,T)$ is isomorphic to the corepresentable functor $\mathbb{Z}[\mathrm{Inj}(-, T\setminus U)] = \mathbb{Z}[\mathrm{Hom}_{\mathrm{FI}^{\mathrm{op}}}(T\setminus U, -)]$. Since $C_k(-,T) = \bigoplus C_k^U(-,T)$, this shows that $C_k(-,T)$ is a projective FI$^{\mathrm{op}}$-module, as desired.

It now follows from Proposition 5.3 and Remark 5.5 that $H_i^{\mathrm{FI}}(V) = H_i(V \otimes_{\mathrm{FI}} C_\bullet)$. □

Remark 5.11 A result essentially equivalent to the conclusion of Proposition 5.10 has been proved independently in a recent preprint of Gan and Li [9, Theorem 1].

Remark 5.12 It is possible to interpret $C_\bullet$ as the "Koszul resolution of FI over K", thinking of $f \in \mathrm{Hom}_{\mathrm{FI}}(S,T)$ as graded by $|T|-|S| = |T - f(S)|$. Moreover, under Schur–Weyl duality $C_\bullet$ corresponds to the classical Koszul resolution of $\mathrm{Sym}^* V$ by $\bigwedge^* V^\vee \otimes \mathrm{Sym}^* V$. For reasons of space we will not pursue this further here; see [14, Section 6] for more details, including strong theorems regarding this Koszul duality for FI-modules over $\mathbb{C}$.

We can now prove Theorem C.

Proof of Theorem C The desired result states for a particular integer N (namely the maximum of $\deg H_0^{\mathrm{FI}}(V)$ and $\deg H_1^{\mathrm{FI}}(V)$), that

$$\operatorname*{colim}_{S\subset T,\, |S|\le N} V_S = V_T \quad \text{for all finite sets } T. \tag{21}$$

We introduced in [5, Definition 2.19] a certain complex of FI-modules $\widetilde{S}_{-\bullet}V$, and combining our earlier results [5, Theorem C, Corollary 2.24] shows that (21) holds if and only if $H_0(\widetilde{S}_{-\bullet}V)_n = 0$ and $H_1(\widetilde{S}_{-\bullet}V)_n = 0$ for all $n > N$.

Our main goal will be therefore to prove that $V \otimes_{\mathrm{FI}} C_\bullet \simeq \widetilde{S}_{-\bullet}(V)$. Given this, we know that

$$H_i(\widetilde{S}_{-\bullet}V) \simeq H_i(V \otimes_{\mathrm{FI}} C_\bullet) \simeq \mathrm{Tor}_i^{\mathrm{FI}}(V,K) \simeq H_i^{\mathrm{FI}}(V),$$

where the second isomorphism holds by Proposition 5.10 and the third isomorphism holds by Proposition 5.3. Therefore, (21) holds if and only if $H_0^{\mathrm{FI}}(V)_n = 0$ and $H_1^{\mathrm{FI}}(V)_n = 0$ for all $n > N$. In other words, the desired condition (21) holds exactly when $\deg H_0^{\mathrm{FI}}(V) \le K$ and $\deg H_1^{\mathrm{FI}}(V) \le N$, which is precisely what the theorem claims.

Recall from Definition 5.6 the category $\mathrm{FI}^{\rightsquigarrow}$. For any FI-module V, we can define an $\mathrm{FI}^{\rightsquigarrow}$-module F_V by $F_V(T,U) = V_{T\setminus U}$, since an $\mathrm{FI}^{\rightsquigarrow}$-morphism $(T,U) \to (T',U')$ restricts to an inclusion $T\setminus U \hookrightarrow T'\setminus U'$. We first show that the complex of FI-modules $V \otimes_{\mathrm{FI}} C_\bullet$ coincides with the complex $C_\bullet^{F_V}$ of Construction 5.7.

We saw in the proof of Proposition 5.10 that $C_k(-,T) = \bigoplus_{|U|=k} C_k^U(-,T)$ where $C_k^U(-,T)$ is the corepresentable functor $\mathbb{Z}[\mathrm{Hom}_{\mathrm{FI}^{\mathrm{op}}}(T\setminus U, -)]$. By the Yoneda lemma, the tensor of V with a functor corepresented by R is simply V_R. Therefore, as abelian groups we have an isomorphism

$$(V \otimes_{\mathrm{FI}} C_k)_T \simeq \bigoplus_{|U|=k} V_{T\setminus U} \simeq C_k^{F_V}(T).$$

Checking the morphisms and differential, we see that $V \otimes_{\mathrm{FI}} C_\bullet$ and $C_\bullet^{F_V}$ coincide as chain complexes of FI-modules.

We conclude by showing that $C_\bullet^{F_V}$ coincides with $\widetilde{S}_{-\bullet}(V)$. We will in fact show that $\widetilde{C}_\bullet^{F_V}$ coincides with the S_n–complex of FI-modules $B_\bullet(V)$ of [5, Equation (10)]. As an abelian group

$$\widetilde{C}_k^{F_V}(T) = \bigoplus_{f\colon [k]\hookrightarrow T} V_{T\setminus \operatorname{im} f},$$

and $B_k(V)_T$ is defined by the same formula [5, Definition 2.9]. Given an injection $g\colon T \hookrightarrow T'$, unwinding Construction 5.7 shows that the map $g_*\colon \widetilde{C}_k^{F_V}(T) \to \widetilde{C}_k^{F_V}(T')$ sends the summand labeled by f to the summand labeled by $g \circ f\colon [k] \hookrightarrow T'$ by the map $(g|_{T\setminus \operatorname{im} f})_*\colon V_{T\setminus \operatorname{im} f} \to V_{T'\setminus \operatorname{im} g\circ f}$. This is precisely the FI-structure on $B_k(V)$. Finally, the maps d_i of Construction 5.7 agree with those defined just before [5, Equation (10)], so the resulting differentials $\partial = \sum (-1)^i d_i$ agree as well.

The S_k–actions on $\widetilde{C}_k^{F_V}$ and on $B_k(V)$ agree, and $C_k^{F_V}$ and $\widetilde{S}_{-k}(V)$ are respectively obtained from these by tensoring over S_k with the sign representation. So we conclude that $V \otimes_{\mathrm{FI}} \widetilde{C}_\bullet \simeq C_\bullet^{F_V}$ is isomorphic to $\widetilde{S}_{-\bullet}(V)$ as chain complexes of FI-modules, as desired. □

5.2 Homology of congruence subgroups

In this section, we state and prove Theorem D′, a more general version of Theorem D from the introduction.

Let R be a commutative ring satisfying Bass's stable range condition SR_{d+2}, and fix a proper ideal $\mathfrak{p} \subsetneq R$. (We use Bass's indexing convention, under which a field satisfies SR_2, and any noetherian d–dimensional ring satisfies SR_{d+2}.) Let $\Gamma_n(\mathfrak{p})$ be the congruence subgroup defined by the exact sequence of groups:

$$1 \to \Gamma_n(\mathfrak{p}) \to \mathrm{GL}_n(R) \to \mathrm{GL}_n(R/\mathfrak{p}).$$

As explained in [5, Section 3], these groups form an FI-group $\Gamma(\mathfrak{p})$ (a functor $\mathrm{FI} \to$ Groups satisfying $\Gamma(\mathfrak{p})_T \simeq \Gamma_{|T|}(\mathfrak{p})$), and thus their integral homology forms an FI-module:

$$\mathcal{H}_k := H_k(\Gamma(\mathfrak{p}); \mathbb{Z}).$$

Theorem D′ *Let R be a commutative ring satisfying Bass's stable range condition SR_{d+2}, and let $\mathfrak{p} \subsetneq R$ be a proper ideal. Then for all $k \geq 2$,*

$$\deg H_0^{\mathrm{FI}}(\mathcal{H}_k) \leq 2^{k-2}(2d+9) - 2 \quad \text{and} \quad \deg H_1^{\mathrm{FI}}(\mathcal{H}_k) \leq 2^{k-2}(2d+9) - 1.$$

In particular, for all $n \geq 0$ and all $k \geq 0$, we have

$$H_k(\Gamma_n(\mathfrak{p}); \mathbb{Z}) = \operatorname*{colim}_{\substack{S \subset [n] \\ |S| < 2^{k-2}(2d+9)}} H_k(\Gamma_S(\mathfrak{p}); \mathbb{Z}). \tag{22}$$

Theorem D is the special case of Theorem D$'$ when $R = \mathbb{Z}$. Indeed, any Dedekind domain R satisfies Bass's condition SR_3 (ie SR_{d+2} for $d = 1$), yielding the bound $|S| < 11 \cdot 2^{k-2}$ in Theorem D. Note that although we take group homology with integer coefficients in the statement of Theorem D$'$, these coefficients could be replaced by any other abelian group; the proof applies unchanged.

By the *stable range*, we mean the range $n \geq 2^{k-2}(2d + 9)$ where the description (22) is not vacuous. Our stable range is slightly better than that of [12], where Putman obtained the range $n \geq 2^{k-2}(2d + 16) - 3$. For example, [12, Theorem B] gives for a Dedekind domain R the stable range $n \geq 18 \cdot 2^{k-2} - 3$, while Theorem D$'$ gives the stable range $n \geq 11 \cdot 2^{k-2}$.

Proof of Theorem D$'$ To avoid confusion with the homology of a chain complex, in this section we write $H_p^{\mathrm{FI}}(W)$ for the FI-homology of an FI-module W (which in previous sections was denoted simply by $H_p(W)$).

An action of an FI-group Γ on an FI-module M is a collection of actions of Γ_T on M_T that are consistent with the FI-structure. Given such an action, the coinvariants form an FI-module $\mathbb{Z} \otimes_\Gamma M$, whose components are simply $\mathbb{Z} \otimes_{\Gamma_T} M_T$. The left-derived functors $H_i(\Gamma; M)$ are simply the FI-modules defined by $H_i(\Gamma; M)_T := H_i(\Gamma_T; M_T)$. In the special case when $M = M(0)$ and the action is trivial, we write $\mathcal{H}_i(\Gamma)$; this is the group homology, considered as an FI-module $\mathcal{H}_i(\Gamma)_T := H_i(\Gamma_T; \mathbb{Z})$.

We will need the following proposition, which constructs for any FI-group a spectral sequence based on the FI-homology of its group homology.

Proposition 5.13 *To any FI-group Γ there is naturally associated an explicit FI-chain complex $X_\bullet^\Gamma$ on which Γ acts, for which we have a spectral sequence:*

$$E^2_{pq} = H_p^{\mathrm{FI}}(\mathcal{H}_q(\Gamma)) \Longrightarrow H_{p+q}(\Gamma; X_\bullet^\Gamma).$$

Proof Recall from Definition 5.6 the category $\mathrm{FI}^{\rightsquigarrow}$ used in Construction 5.7. Define the $\mathrm{FI}^{\rightsquigarrow}$-module A by $A(T, U) = \mathbb{Z}[\Gamma_T / \Gamma_{T \setminus U}]$. An $\mathrm{FI}^{\rightsquigarrow}$-morphism $f\colon (T, U) \to (T', U')$ has $f(U) \supseteq U'$, so the induced map $f_*\colon \Gamma_T \to \Gamma_{T'}$ satisfies $f_*(\Gamma_{T \setminus U}) \subset \Gamma_{T' \setminus U'}$, verifying that A is indeed an $\mathrm{FI}^{\rightsquigarrow}$-module.

The FI-chain complex $X_\bullet = X_\bullet^\Gamma$ we are interested in will be the FI-chain complex $X_\bullet := C_\bullet^A$ arising from A via Construction 5.7:

$$X_k(T) = \bigoplus_{U \subset T, |U| = k} \mathbb{Z}[\Gamma_T / \Gamma_{T \setminus U}].$$

For each T the obvious action of Γ_T on $\mathbb{Z}[\Gamma_T/\Gamma_{T\setminus U}]$ induces an action of Γ_T on $X_k(T)$. The FI-module structure on X_k is induced by the FI-structure maps $\Gamma_T\to\Gamma_{T'}$, and the differential ∂ descends from the *identity* on Γ_T. Therefore, the action of Γ_T on $X_k(T)$ is compatible with both, giving an action of the FI-group Γ on the FI-chain complex $X_\bullet$.

From this action, we obtain two spectral sequences converging to the homology $H_*(\Gamma; X_\bullet)$ of the complex $X_\bullet$:

$$\begin{aligned}\bar{E}^2_{pq} &= H_p(\Gamma; H_q(X_\bullet)) \Longrightarrow H_{p+q}(\Gamma; X_\bullet),\\ E^1_{pq} &= H_q(\Gamma; X_p) \Longrightarrow H_{p+q}(\Gamma; X_\bullet).\end{aligned}$$

The desired spectral sequence mentioned in the proposition is the second one (though we will use the first spectral sequence later). It remains to identify E^2_{pq} with $H_p^{\mathrm{FI}}(\mathcal{H}_q(\Gamma))$, so let us compute $E^1_{pq}=H_q(\Gamma; X_p)$.

By definition $X_p(T)$ is a direct sum of factors $\mathbb{Z}[\Gamma_T/\Gamma_{T\setminus U}]$. By Shapiro's lemma, the contribution of such a factor to $H_q(\Gamma_T; X_p(T))$ is precisely $H_q(\Gamma_{T\setminus U})=\mathcal{H}_q(\Gamma)_{T\setminus U}$. We find that

$$H_q(\Gamma; X_p)_T = H_q(\Gamma_T; X_p(T)) = \bigoplus_{U\subset T,\,|U|=p} \mathcal{H}_q(\Gamma)_{T\setminus U} = (\mathcal{H}_q(\Gamma)\otimes_{\mathrm{FI}} C_p)_T,$$

where the last equality comes from the proof of Theorem C. We conclude that

$$E^1_{pq} = H_q(\Gamma; X_p) \simeq \mathcal{H}_q(\Gamma)\otimes_{\mathrm{FI}} C_p.$$

Moreover, the differential $d^1\colon H_q(\Gamma;X_p)\to H_q(\Gamma;X_{p-1})$ is induced by $\partial\colon X_p\to X_{p-1}$, and comparing the definitions of $X_\bullet$ and $C_\bullet$ shows that $(E^1_{pq}, d^1)=(\mathcal{H}_q(\Gamma)\otimes_{\mathrm{FI}} C_\bullet, \partial)$. By Proposition 5.10 we conclude that, as claimed,

$$E^2_{pq} = \mathrm{Tor}_p^{\mathrm{FI}}(\mathcal{H}_q(\Gamma), K) = H_p^{\mathrm{FI}}(\mathcal{H}_q(\Gamma)) \Longrightarrow H_{p+q}(\Gamma; X_\bullet). \qquad \square$$

We now continue with the proof of Theorem D$'$. Returning to the notation of that theorem, let Γ be the congruence FI-group $\Gamma(\mathfrak{p})$, and $\mathcal{H}_k=\mathcal{H}_k(\Gamma(\mathfrak{p}))$ its group homology. We would like to apply Proposition 5.13, but to do this we need to bound the equivariant homology $H_{p+q}(\Gamma; X_\bullet)$. We can do this using the other spectral sequence $\bar{E}^2_{pq}=H_p(\Gamma; H_q(X_\bullet))\Longrightarrow H_{p+q}(\Gamma; X_\bullet)$ if we can bound $H_q(X_\bullet)$. And fortunately, this complex $X_\bullet$ (or a complex quite close to it) has already been considered by Charney!

In Proposition 5.13 we defined $X_\bullet=C^A_\bullet$ based on the functor $A(T,U)=\mathbb{Z}[\Gamma_T/\Gamma_{T\setminus U}]$. Let $\widetilde{X}_\bullet := \widetilde{C}^A_\bullet$ be the ordered version of this complex; concretely, we can write

$$\widetilde{X}_k(T) = \bigoplus_{(t_1,\dots,t_k)\subset T} \mathbb{Z}[\Gamma_T/\Gamma_{T-\{t_1,\dots,t_k\}}].$$

In the foundational paper [3], Charney considered (in the case $T = \{1, \dots, n\}$) a complex $Y_\bullet(T)$ that is similar to $\widetilde{X}_\bullet(T)$ but somewhat larger. Her key technical result in that paper is that $Y_\bullet(T)$ is q–acyclic if $|T| \geq 2q + d + 1$. Moreover, [3, Proposition 3.2] implies that $Y_q(T)$ coincides with $\widetilde{X}_q(T)$ as long as $|T| \geq q + d$, so Charney's result[8] implies that $\widetilde{X}_\bullet(T)$ is q–acyclic if $|T| \geq 2q + d + 1$. By Remark 5.8 we know that $X_\bullet(T)$ is a summand of $\widetilde{X}_\bullet(T)$, so $X_\bullet(T)$ is q–acyclic in the same range. Said differently, $H_q(X_\bullet)_T = 0$ for $|T| > 2q + d$; that is, $\deg H_q(X_\bullet) \leq 2q + d$.

Any FI-module M with $\deg M \leq N$ automatically has $\deg H_i(\Gamma; M) \leq N$ for all i, since $H_*(\Gamma; M) = H_*(\Gamma_T; 0) = 0$ when $|T| > N$. Therefore, Charney's bound $\deg H_q(X_\bullet) \leq 2q + d$ implies, for all p,

$$\deg \bar{E}^2_{pq} = \deg H_p(\Gamma; H_q(X_\bullet)) \leq 2q + d.$$

Since this spectral sequence converges to $\bar{E}^2_{pq} \Longrightarrow H_{p+q}(\Gamma; X_\bullet)$, we conclude that

$$\deg H_k(\Gamma; X_\bullet) \leq 2k + d. \tag{23}$$

The bound (23) marks the end of the input from topology in this proof. The remainder of the proof is just careful bookkeeping and repeatedly applying Theorem A to our spectral sequence of FI-modules

$$E^2_{pq} = H^{\mathrm{FI}}_p(\mathcal{H}_q) \Longrightarrow H_{p+q}(\Gamma; X_\bullet).$$

In fact, this bookkeeping can be formulated as the following completely general statement:

Claim 5.14 *Consider a spectral sequence of FI-modules $E^2_{pq} \Longrightarrow V_{p+q}$ converging to FI-modules V_k satisfying $\deg V_k \leq 2k + d$ for some integer $d \geq 0$. Suppose that for all q, we know that*

$$\deg E^2_{pq} \leq \deg E^2_{0q} + \deg E^2_{1q} - 1 + p, \tag{α}$$

and suppose for simplicity that $E^2_{p0} = 0$ for $p > 0$. Then for all $k \geq 2$, we have

$$\deg E^2_{0k} \leq 2^{k-2}(2d + 9) - 2 \quad \textit{and} \quad \deg E^2_{1k} \leq 2^{k-2}(2d + 9) - 1. \tag{24}$$

Proof We would like to prove this claim (24) by induction on k for all $k \geq 2$, but we need to modify it slightly so it holds in the base cases $k \in \{0, 1\}$ as well. Therefore, we will prove along the way that

$$\forall p \geq 2,\ \deg E^2_{pk} \leq 2^{k-1}(2d + 9) - 4 + p \tag{25}$$

[8] Note that our indexing differs from Charney's in that her complex has $\widetilde{X}_q(T)$ in degree $q - 1$; this is why we have $2q + d + 1$ and $q + d$ in place of her $2q + d + 3$ and $q + d + 1$, respectively.

holds for all $k \geq 0$. Notice that (24) + (α) $\implies$ (25), so this only requires additional work in the base cases when $k \in \{0,1\}$. We first prove (25) in these base cases, and then prove by induction on k that both (24) and (25) hold for all $k \geq 2$.

Case $k = 0$ Our assumption that $E^2_{p0} = 0$ for all $p \geq 1$ implies $\deg E^2_{p0} = -\infty$, so (25) holds.

Case $k = 1$ Since $E^2_{3,0} = 0$ and $E^2_{4,0} = 0$, the spectral sequence degenerates at E^2 for $E^2_{0,1}$ and $E^2_{1,1}$, yielding $E^2_{0,1} = E^\infty_{0,1} = V^1$ and $E^2_{1,1} = E^\infty_{1,1} \subset V^2$. Since $\deg V^1 \leq 2 + d$ and $\deg V^2 \leq 4 + d$, we conclude that $\deg E^2_{0,1} \leq d + 2$ and $\deg E^2_{1,1} \leq d + 4$. Applying the assumption (α), we conclude that $\deg E^2_{p,1} \leq 2d + 5 + p$ for all $p \geq 2$; this is precisely the bound (25) in the case $k = 1$.

General case Let $N_{p,m} := 2^{m-1}(2d+9) - 4 + p$ be the bound occurring in (25). Fix $k \geq 2$, and assume by induction that (25) holds for all $m < k$; that is, $\deg E^2_{p,m} \leq N_{p,m}$ for all $p \geq 2$ and all $m < k$.

Now consider the entry $E^2_{0,k}$. Since $E^\infty_{0,k}$ is a constituent of V^k, we have $\deg E^\infty_{0,k} \leq \deg V^k \leq 2d + k$. No nontrivial differential has source $E^r_{0,k}$, but we have differentials $d^r\colon E^r_{r,k-r+1} \to E^r_{0,k}$. The maximum of $N_{r,k-r+1}$ over $r \geq 2$ occurs when $r = 2$, when we have $N_{2,k-1} = 2^{k-2}(2d+9) - 2$. Therefore, for all $r \geq 2$ the sources of these differentials satisfy $\deg E^r_{r,k-r+1} \leq 2^{k-2}(2d+9) - 2$. Since $\deg E^\infty_{0,k} \leq 2d + k < 2^{k-2}(2d+9) - 2$, we conclude that $\deg E^2_{0,k} \leq 2^{k-2}(2d+9) - 2$, as claimed in (24). Similarly, the degrees of the sources of the differentials $d^r\colon E^r_{1+r,k-r+1} \to E^r_{1,k}$ are bounded above by $N_{3,k-1} = 2^{k-2}(2d+9) - 1$. Since

$$\deg E^\infty_{1,k} \leq \deg V^{k+1} \leq 2d + k + 1 < 2^{k-2}(2d+9) - 1,$$

we conclude that $\deg E^2_{1,k} \leq 2^{k-2}(2d+9) - 1$, as claimed in (24).

Now applying the assumption (α) to (24), we conclude that (25) holds for k as well. This concludes the proof of the claim. □

We now finish the proof of Theorem D′ by applying this claim to the spectral sequence $E^2_{pq} = H^{\mathrm{FI}}_p(\mathcal{H}_q) \implies H_{p+q}(\Gamma; X_\bullet)$ of Proposition 5.13. The hypothesis $E^2_{p0} = 0$ of the claim is satisfied because $\mathcal{H}_0$ is the free FI-module $\mathcal{H}_0 \simeq M(0)$, so $E^2_{p0} = H^{\mathrm{FI}}_p(\mathcal{H}_0) = 0$ for $p > 0$. The assumption (α) is precisely the statement of Theorem A, and the bound $\deg H_k(\Gamma; X_\bullet) \leq 2k + d$ was obtained in (23) above.

The description (22) for $k \geq 2$ follows from (24) by Theorem C. The only thing that remains is some arithmetic to check that (22) holds for $k = 0$ and $k = 1$ as well.

For $k=0$ this is trivial, since $\mathcal{H}_0 = M(0)$ is free: this means $\deg H_0^{\mathrm{FI}}(\mathcal{H}_0) = 0$ and $\deg H_1^{\mathrm{FI}}(\mathcal{H}_0) = -\infty$, so Theorem C then gives an identification as in (22) over $|S| \le 0$. Since $d \ge 0$, we have $2^{0-2}(2d+9) \ge \frac{9}{4} > 1$, so the bound in (22) holds.

Similarly, for $k=1$ we saw in the proof above that $\deg H_0^{\mathrm{FI}}(\mathcal{H}_1) = \deg E^2_{0,1} \le 2+d$ and $\deg H_1^{\mathrm{FI}}(\mathcal{H}_1) = \deg E^2_{1,1} \le 4+d$, so Theorem C gives an identification as in (22) over $|S| \le 4+d$. For integer m the conditions $m < 2^{1-2}(2d+9) = d + \frac{9}{2}$ and $m \le d+4$ are equivalent, so again the bound in (22) follows. □

We close with a variant of Theorem D$'$ which has been used by Calegari and Emerton [2] in their study of completed homology. An inclusion of ideals $\mathfrak{q} \subset \mathfrak{p}$ induces an inclusion $\Gamma_n(\mathfrak{q}) \subset \Gamma_n(\mathfrak{p})$, so given an inverse system of ideals such as $\cdots \subset \mathfrak{p}^i \subset \cdots \subset \mathfrak{p}^2 \subset \mathfrak{p}$, we can consider the inverse limit $\varprojlim H_k(\Gamma_n(\mathfrak{p}^i))$ of the homology of the corresponding congruence subgroups.

Theorem D$''$ *Let R be the ring of integers in a number field, and let $(\mathfrak{p}_i)_{i\in I}$ be an inverse system of proper ideals in R. Fix $N > 1$. Then for all $n \ge 0$ and all $k \ge 0$, we have*

$$\varprojlim_{i\in I} H_k(\Gamma_n(\mathfrak{p}_i); \mathbb{Z}/N) = \operatorname*{colim}_{\substack{S\subset[n]\\ |S|<11\cdot 2^{k-2}}} \varprojlim_{i\in I} H_k(\Gamma_S(\mathfrak{p}_i); \mathbb{Z}/N).$$

Proof Any number ring R is a Dedekind domain, so R satisfies Bass's stable range condition SR_3. Therefore, for any $n \ge 0$ and any $k \ge 0$, we can deduce from Theorem D$'$ that

$$\varprojlim_{i\in I} H_k(\Gamma_n(\mathfrak{p}_i); \mathbb{Z}/N) = \varprojlim_{i\in I} \operatorname*{colim}_{\substack{S\subset[n]\\ |S|<11\cdot 2^{k-2}}} H_k(\Gamma_S(\mathfrak{p}_i); \mathbb{Z}/N).$$

It remains to check that we can exchange the limit and colimit. This is of course not true in general, but we can verify that it is true in this case as follows. The existence of the Borel–Serre compactification [1] implies that $H_k(\Gamma_n(\mathfrak{p}); \mathbb{Z}/N)$ is a finitely generated $\mathbb{Z}/N$–module for any $\mathfrak{p} \subset R$. This is enough to give the desired result: since this colimit is over a finite poset, it can therefore be written as a coequalizer of finitely generated $\mathbb{Z}/N$–modules. The limit of the coequalizers is the coequalizer of the limits (any inverse system of finite abelian groups satisfies the Mittag-Leffler condition, so the $\lim^1$ term vanishes), which is to say that the limit and colimit can be exchanged as desired. □

References

[1] **A Borel**, **J-P Serre**, *Corners and arithmetic groups*, Comment. Math. Helv. 48 (1973) 436–491 MR

[2] **F Calegari**, **M Emerton**, *Homological stability for completed homology*, Math. Ann. 364 (2016) 1025–1041 MR

[3] **R Charney**, *On the problem of homology stability for congruence subgroups*, Comm. Algebra 12 (1984) 2081–2123 MR

[4] **T Church**, **J S Ellenberg**, **B Farb**, *FI-modules and stability for representations of symmetric groups*, Duke Math. J. 164 (2015) 1833–1910 MR

[5] **T Church**, **J S Ellenberg**, **B Farb**, **R Nagpal**, *FI-modules over Noetherian rings*, Geom. Topol. 18 (2014) 2951–2984 MR

[6] **J S Ellenberg**, **A Venkatesh**, **C Westerland**, *Homological stability for Hurwitz spaces and the Cohen–Lenstra conjecture over function fields*, Ann. of Math. 183 (2016) 729–786 MR

[7] **W L Gan**, *A long exact sequence for homology of FI-modules*, preprint (2016) arXiv

[8] **W L Gan**, **L Li**, *On central stability*, preprint (2016) arXiv

[9] **W L Gan**, **L Li**, *A remark on FI-module homology*, Michigan Math. J. 65 (2016) 855–861 MR

[10] **L Li**, *Upper bounds of homological invariants of FI_G–modules*, preprint (2016) arXiv

[11] **L Li**, **N Yu**, *Filtrations and homological degrees of FI-modules*, J. Algebra 472 (2017) 369–398 MR

[12] **A Putman**, *Stability in the homology of congruence subgroups*, Invent. Math. 202 (2015) 987–1027 MR

[13] **S V Sam**, **A Snowden**, *Stability patterns in representation theory*, Forum Math. Sigma 3 (2015) e11 (108 pages) MR

[14] **S V Sam**, **A Snowden**, *GL-equivariant modules over polynomial rings in infinitely many variables*, Trans. Amer. Math. Soc. 368 (2016) 1097–1158 MR

[15] **C A Weibel**, *An introduction to homological algebra*, Cambridge Studies in Advanced Mathematics 38, Cambridge Univ. Press (1994) MR

[16] **J D Wiltshire-Gordon**, *Uniformly presented vector spaces*, preprint (2014) arXiv

Department of Mathematics, Stanford University
450 Serra Mall, Stanford, CA 94305, United States

Department of Mathematics, University of Wisconsin
480 Lincoln Drive, Madison, WI 53706, United States

tfchurch@stanford.edu, ellenber@math.wisc.edu

http://math.stanford.edu/~church/, http://www.math.wisc.edu/~ellenber/

Proposed: Benson Farb Received: 22 February 2016
Seconded: Haynes Miller, Jesper Grodal Revised: 31 August 2016

Geometry & Topology 21 (2017) 2419–2460

Hodge modules on complex tori and generic vanishing for compact Kähler manifolds

GIUSEPPE PARESCHI
MIHNEA POPA
CHRISTIAN SCHNELL

We extend the results of generic vanishing theory to polarizable real Hodge modules on compact complex tori, and from there to arbitrary compact Kähler manifolds. As applications, we obtain a bimeromorphic characterization of compact complex tori (among compact Kähler manifolds), semipositivity results and a description of the Leray filtration for maps to tori.

14C30; 14F17

A Introduction

The term "generic vanishing" refers to a collection of theorems about the cohomology of line bundles with trivial first Chern class. The first results of this type were obtained by Green and Lazarsfeld in the late 1980s [13; 14]; they were proved using classical Hodge theory and are therefore valid on arbitrary compact Kähler manifolds. About ten years ago, Hacon [15] found a more algebraic approach, using vanishing theorems and the Fourier–Mukai transform, that has led to many additional results in the projective case; see also Chen and Jiang [9], Pareschi and Popa [23] and Popa and Schnell [26]. The purpose of this paper is to show that the newer results are in fact also valid on arbitrary compact Kähler manifolds.

Besides Hacon [15], our motivation also comes from a paper by Chen and Jiang [9], in which they prove, roughly speaking, that the direct image of the canonical bundle under a generically finite morphism to an abelian variety is semiample. Before we can state more precise results, recall the following definitions (see Section 13 for more details).

Definition Given a coherent $\mathscr{O}_T$–module $\mathscr{F}$ on a compact complex torus T, define

$$S^i(T, \mathscr{F}) = \{L \in \mathrm{Pic}^0(T) \mid H^i(T, \mathscr{F} \otimes L) \neq 0\}.$$

We say that $\mathscr{F}$ is a *GV-sheaf* if $\operatorname{codim} S^i(T, \mathscr{F}) \geq i$ for every $i \geq 0$; we say that $\mathscr{F}$ is *M–regular* if $\operatorname{codim} S^i(T, \mathscr{F}) \geq i+1$ for every $i \geq 1$.

Published: 19 May 2017 DOI: 10.2140/gt.2017.21.2419

Hacon [15, Section 4] showed that if $f\colon X \to A$ is a morphism from a smooth projective variety to an abelian variety, then the higher direct image sheaves $R^j f_* \omega_X$ are GV-sheaves on A; in the special case where f is generically finite over its image, Chen and Jiang [9, Theorem 1.2] proved the much stronger result that $f_* \omega_X$ is, up to tensoring by line bundles in $\mathrm{Pic}^0(A)$, the direct sum of pullbacks of M–regular sheaves from quotients of A. Since GV-sheaves are nef, whereas M–regular sheaves are ample, one should think of this as saying that $f_* \omega_X$ is not only nef but actually semiample. One of our main results is the following generalization of this fact:

Theorem A *Let $f\colon X \to T$ be a holomorphic mapping from a compact Kähler manifold to a compact complex torus. Then, for $j \geq 0$, one has a decomposition*

$$R^j f_* \omega_X \simeq \bigoplus_{k=1}^{n} (q_k^* \mathscr{F}_k \otimes L_k),$$

where each $\mathscr{F}_k$ is an M–regular (hence ample) coherent sheaf with projective support on the compact complex torus T_k, each $q_k\colon T \to T_k$ is a surjective morphism with connected fibers, and each $L_k \in \mathrm{Pic}^0(T)$ has finite order. In particular, $R^j f_* \omega_X$ is a GV-sheaf on T.

This leads to strong positivity properties for higher direct images of canonical bundles under maps to tori. For instance, if f is a surjective map that is submersive away from a divisor with simple normal crossings, then $R^j f_* \omega_X$ is a semipositive vector bundle on T. See Section 20 for more on this circle of ideas.

One application of Theorem A is the following effective criterion for a compact Kähler manifold to be bimeromorphically equivalent to a torus; this generalizes a well-known theorem of Chen and Hacon in the projective case [6].

Theorem B *A compact Kähler manifold X is bimeromorphic to a compact complex torus if and only if $\dim H^1(X, \mathbb{C}) = 2 \dim X$ and $P_1(X) = P_2(X) = 1$.*

The proof is inspired by the approach to the Chen–Hacon theorem given by Pareschi [20]; even in the projective case, however, the result in Corollary 16.2 greatly simplifies the existing proof. In Theorem 19.1, we deduce that the Albanese map of a compact Kähler manifold with $P_1(X) = P_2(X) = 1$ is surjective with connected fibers; in the projective case, this was first proved by Jiang [16], as an effective version of Kawamata's theorem about projective varieties of Kodaira dimension zero. It is likely that the present methods can also be applied to the classification of compact Kähler manifolds with $\dim H^1(X, \mathbb{C}) = 2 \dim X$ and small plurigenera; for the projective case, see for instance Chen and Hacon [8].

In a different direction, Theorem A combined with results of Lazarsfeld, Popa and Schnell [18] leads to a concrete description of the Leray filtration on the cohomology of ω_X, associated with a holomorphic mapping $f\colon X \to T$ as above. Recall that, for each $k \geq 0$, the Leray filtration is a decreasing filtration $L^\bullet H^k(X, \omega_X)$ with the property that

$$\mathrm{gr}_L^i\, H^k(X, \omega_X) = H^i(T, R^{k-i} f_* \omega_X).$$

One can also define a natural decreasing filtration $F^\bullet H^k(X, \omega_X)$ induced by the cup product action of $H^1(T, \mathscr{O}_T)$, namely

$$F^i H^k(X, \omega_X) = \mathrm{Im}\big(\textstyle\bigwedge^i H^1(T, \mathscr{O}_T) \otimes H^{k-i}(X, \omega_X) \to H^k(X, \omega_X)\big).$$

Theorem C *The filtrations $L^\bullet H^k(X, \omega_X)$ and $F^\bullet H^k(X, \omega_X)$ coincide.*

We give a dual description of the filtration on global holomorphic forms in Corollary 21.3. Despite the elementary nature of the statement, we do not know how to prove Theorem C using only methods from classical Hodge theory; finding a more elementary proof is an interesting problem.

Our approach to Theorem A is to address generic vanishing for a larger class of objects of Hodge-theoretic origin, namely polarizable real Hodge modules on compact complex tori. This is not just a matter of higher generality; we do not know how to prove Theorem A using methods of classical Hodge theory in the spirit of Green and Lazarsfeld [13]. This is precisely due to the lack of an a priori description of the Leray filtration on $H^k(X, \omega_X)$ as in Theorem C.

The starting point for our proof of Theorem A is a result by Saito [29], which says that the coherent $\mathscr{O}_T$–module $R^j f_* \omega_X$ is part of a polarizable real Hodge module $M = (\mathcal{M}, F_\bullet \mathcal{M}, M_\mathbb{R}) \in \mathrm{HM}_\mathbb{R}(T, \dim X + j)$ on the torus T; more precisely,

$$R^j f_* \omega_X \simeq \omega_T \otimes F_{p(M)} \mathcal{M}$$

is the first nontrivial piece in the Hodge filtration $F_\bullet \mathcal{M}$ of the underlying regular holonomic $\mathscr{D}$–module $\mathcal{M}$. (Please see Section 1 for some background on Hodge modules.) Note that M is supported on the image $f(X)$, and that its restriction to the smooth locus of f is the polarizable variation of Hodge structure on the $(\dim f + j)^{\text{th}}$ cohomology of the fibers. The reason for working with real coefficients is that the polarization is induced by a choice of Kähler form in $H^2(X, \mathbb{R}) \cap H^{1,1}(X)$; the variation of Hodge structure itself is of course defined over $\mathbb{Z}$.

In light of the above identity, Theorem A is a consequence of the following general statement about polarizable real Hodge modules on compact complex tori:

Theorem D *Let $M = (\mathcal{M}, F_\bullet\mathcal{M}, M_\mathbb{R}) \in \mathrm{HM}_\mathbb{R}(T, w)$ be a polarizable real Hodge module on a compact complex torus T. Then, for each $k \in \mathbb{Z}$, the coherent $\mathscr{O}_T$–module $\mathrm{gr}_k^F \mathcal{M}$ decomposes as*

$$\mathrm{gr}_k^F \mathcal{M} \simeq \bigoplus_{j=1}^{n} (q_j^* \mathscr{F}_j \otimes_{\mathscr{O}_T} L_j),$$

where $q_j \colon T \to T_j$ is a surjective map with connected fibers to a complex torus, $\mathscr{F}_j$ is an M–regular coherent sheaf on T_j with projective support and $L_j \in \mathrm{Pic}^0(T)$. If M admits an integral structure, then each L_j has finite order.

Let us briefly describe the most important elements in the proof. In Popa and Schnell [26] we already exploited the relationship between generic vanishing and Hodge modules on abelian varieties, but the proofs relied on vanishing theorems. What allows us to go further is a beautiful idea by Botong Wang [41], namely that, up to taking direct summands and tensoring by unitary local systems, every polarizable real Hodge module on a complex torus actually comes from an abelian variety. (Wang showed this for Hodge modules of geometric origin.) This is a version with coefficients of Ueno's result [39] that every irreducible subvariety of T is a torus bundle over a projective variety, and is proved by combining this geometric fact with some arguments about variations of Hodge structure.

The existence of the decomposition in Theorem D is due to the fact that the regular holonomic $\mathscr{D}$–module $\mathcal{M}$ is semisimple, hence isomorphic to a direct sum of simple regular holonomic $\mathscr{D}$–modules. This follows from a theorem by Deligne and Nori (see Deligne [11]), which says that the local system underlying a polarizable real variation of Hodge structure on a Zariski-open subset of a compact Kähler manifold is semisimple. It turns out that the decomposition of $\mathcal{M}$ into simple summands is compatible with the Hodge filtration $F_\bullet\mathcal{M}$; in order to prove this, we introduce the category of "polarizable complex Hodge modules" (which are polarizable real Hodge modules together with an endomorphism whose square is minus the identity), and show that every simple summand of $\mathcal{M}$ underlies a polarizable complex Hodge module in this sense.

Note Our ad hoc definition of complex Hodge modules is good enough for the purposes of this paper, but is certainly not the final word. A more satisfactory treatment, in terms of $\mathscr{D}$–modules and distribution-valued pairings, is currently being developed by Claude Sabbah and the third author [27].

The M–regularity of the individual summands in Theorem D turns out to be closely related to the Euler characteristic of the corresponding $\mathscr{D}$–modules. The results in [26] show that when $(\mathcal{M}, F_\bullet\mathcal{M})$ underlies a polarizable complex Hodge module on an abelian variety A, the Euler characteristic satisfies $\chi(A, \mathcal{M}) \geq 0$, and each coherent $\mathscr{O}_A$–module $\operatorname{gr}_k^F \mathcal{M}$ is a GV-sheaf. The new result (in Lemma 15.1) is that each $\operatorname{gr}_k^F \mathcal{M}$ is actually M–regular provided that $\chi(A, \mathcal{M}) > 0$. That we can always get into the situation where the Euler characteristic is positive follows from some general results about simple holonomic $\mathscr{D}$–modules from Schnell [34].

Theorem D implies that each graded quotient $\operatorname{gr}_k^F \mathcal{M}$ with respect to the Hodge filtration is a GV-sheaf, the Kähler analogue of a result in [26]. However, the stronger formulation above is new even in the case of smooth projective varieties, and has further useful consequences. One such is the following: for a holomorphic mapping $f\colon X \to T$ that is generically finite onto its image, the locus

$$S^0(T, f_*\omega_X) = \{L \in \operatorname{Pic}^0(T) \mid H^i(T, f_*\omega_X \otimes_{\mathscr{O}_T} L) \neq 0\}$$

is preserved by the involution $L \mapsto L^{-1}$ on $\operatorname{Pic}^0(T)$; see Corollary 16.2. This is a crucial ingredient in the proof of Theorem B.

Going back to Wang's paper [41], its main purpose was to prove Beauville's conjecture, namely that, on a compact Kähler manifold X, every irreducible component of every $\Sigma^k(X) = \{\rho \in \operatorname{Char}(X) \mid H^k(X, \mathbb{C}_\rho) \neq 0\}$ contains characters of finite order. In the projective case, this is of course a famous theorem by Simpson [37]. Combining the structural Theorem 7.1 with known results about Hodge modules on abelian varieties (Schnell [35]) allows us to prove the following generalization of Wang's theorem (which dealt with Hodge modules of geometric origin):

Theorem E *If a polarizable real Hodge module $M \in \mathrm{HM}_{\mathbb{R}}(T, w)$ on a compact complex torus admits an integral structure, then the sets*

$$S_m^i(T, M) = \{\rho \in \operatorname{Char}(T) \mid \dim H^i(T, M_{\mathbb{R}} \otimes_{\mathbb{R}} \mathbb{C}_\rho) \geq m\}$$

are finite unions of translates of linear subvarieties by points of finite order.

The idea is to use Kronecker's theorem (about algebraic integers all of whose conjugates have absolute value one) to prove that certain characters have finite order. Roughly speaking, the characters in question are unitary because of the existence of a polarization on M, and they take values in the group of algebraic integers because of the existence of an integral structure on M.

Projectivity questions

We conclude by noting that many of the results in this paper can be placed in the broader context of the following problem: how far are natural geometric or sheaf-theoretic constructions on compact Kähler manifolds in general, and on compact complex tori in particular, from being determined by similar constructions on projective manifolds? Theorems A and D provide the answer on tori in the case of Hodge-theoretic constructions. We thank János Kollár for suggesting this point of view, and also the statements of the problems in the paragraph below.

Further structural results could provide a general machine for reducing certain questions about Kähler manifolds to the algebraic setting. For instance, by analogy with positivity conjectures in the algebraic case, one hopes for the following result in the case of varying families: if X and Y are compact Kähler manifolds and $f\colon X \to Y$ is a fiber space of maximal variation, ie such that the general fiber is bimeromorphic to at most countably many other fibers, then Y is projective. More generally, for an arbitrary such f, is there a mapping $g\colon Y \to Z$, with Z projective, such that the fibers of f are bimeromorphically isotrivial over those of Y?

A slightly more refined version in the case when $Y = T$ is a torus, which is essentially a combination of Iitaka fibrations and Ueno's conjecture, is this: there should exist a morphism $h\colon X \to Z$, where Z is a variety of general type generating an abelian quotient $g\colon T \to A$, such that the fibers of h have Kodaira dimension 0 and are bimeromorphically isotrivial over the fibers of g.

B Real and complex Hodge modules

1 Real Hodge modules

In this paper, we work with polarizable real Hodge modules on complex manifolds. This is the natural setting for studying compact Kähler manifolds, because the polarizations induced by Kähler forms are defined over $\mathbb{R}$ (but usually not over $\mathbb{Q}$, as in the projective case). Saito originally developed the theory of Hodge modules with rational coefficients, but as explained in [29], everything works just as well with real coefficients and with the following weaker assumption on the local monodromy: the eigenvalues of the monodromy operator on the nearby cycles are allowed to be arbitrary complex numbers of absolute value one, rather than just roots of unity. This has already been observed several times in the literature [33]; the point is that Saito's theory rests on certain results about polarizable variations of Hodge structure [32; 43; 5], which hold in this generality.

Let X be a complex manifold. We first recall some terminology.

Definition 1.1 We denote by $\mathrm{HM}_{\mathbb{R}}(X, w)$ the category of polarizable real Hodge modules of weight w; this is a semisimple $\mathbb{R}$–linear abelian category, endowed with a faithful functor to the category of real perverse sheaves.

Saito constructs $\mathrm{HM}_{\mathbb{R}}(X, w)$ as a full subcategory of the category of all filtered regular holonomic $\mathscr{D}$–modules with real structure, in several stages. To begin with, recall that a *filtered regular holonomic $\mathscr{D}$–module with real structure* on X consists of the following four pieces of data: (1) a regular holonomic left $\mathscr{D}_X$–module $\mathcal{M}$; (2) a good filtration $F_\bullet\mathcal{M}$ by coherent $\mathscr{O}_X$–modules; (3) a perverse sheaf $M_\mathbb{R}$ with coefficients in $\mathbb{R}$; (4) an isomorphism $M_\mathbb{R} \otimes_\mathbb{R} \mathbb{C} \simeq \mathrm{DR}(\mathcal{M})$. Although the isomorphism is part of the data, we usually suppress it from the notation and simply write $M = (\mathcal{M}, F_\bullet\mathcal{M}, M_\mathbb{R})$. The *support* $\operatorname{Supp} M$ is defined to be the support of the underlying perverse sheaf $M_\mathbb{R}$; one says that M has *strict support* if $\operatorname{Supp} M$ is irreducible and if M has no nontrivial subobjects or quotient objects that are supported on a proper subset of $\operatorname{Supp} M$.

Now M is called a *real Hodge module of weight* w if it satisfies several additional conditions that are imposed by recursion on the dimension of $\operatorname{Supp} M$. Although they are not quite stated in this way in [28], the essence of these conditions is that (1) every Hodge module decomposes into a sum of Hodge modules with strict support, and (2) every Hodge module with strict support is generically a real variation of Hodge structure, which uniquely determines the Hodge module. Given $k \in \mathbb{Z}$, set $\mathbb{R}(k) = (2\pi i)^k\mathbb{R} \subseteq \mathbb{C}$; then one has the *Tate twist*

$$M(k) = (\mathcal{M}, F_{\bullet-k}\mathcal{M}, M_\mathbb{R} \otimes_\mathbb{R} \mathbb{R}(k)) \in \mathrm{HM}_\mathbb{R}(X, w-2k).$$

Every real Hodge module of weight w has a well-defined *dual* $\boldsymbol{D}M$, which is a real Hodge module of weight $-w$ whose underlying perverse sheaf is the Verdier dual $\boldsymbol{D}M_\mathbb{R}$. A *polarization* is an isomorphism of real Hodge modules $\boldsymbol{D}M \simeq M(w)$, subject to certain conditions that are again imposed recursively; one says that M is *polarizable* if it admits at least one polarization.

Example 1.2 Every polarizable real variation of Hodge structure of weight w on X gives rise to an object of $\mathrm{HM}_\mathbb{R}(X, w + \dim X)$. If $\mathcal{H}$ is such a variation, we denote the underlying real local system by $\mathcal{H}_\mathbb{R}$, its complexification by $\mathcal{H}_\mathbb{C} = \mathcal{H}_\mathbb{R} \otimes_\mathbb{R} \mathbb{C}$ and the corresponding flat bundle by $(\mathcal{H}, \nabla)$; then $\mathcal{H} \simeq \mathcal{H}_\mathbb{C} \otimes_\mathbb{C} \mathscr{O}_X$. The flat connection makes $\mathcal{H}$ into a regular holonomic left $\mathscr{D}$–module, filtered by $F_\bullet\mathcal{H} = F^{-\bullet}\mathcal{H}$; the real structure is given by the real perverse sheaf $\mathcal{H}_\mathbb{R}[\dim X]$.

We list a few useful properties of polarizable real Hodge modules. By definition, every object $M \in \mathrm{HM}_\mathbb{R}(X, w)$ admits a locally finite *decomposition by strict support*; when X is compact, this is a finite decomposition

$$M \simeq \bigoplus_{j=1}^{n} M_j,$$

where each $M_j \in \mathrm{HM}_\mathbb{R}(X, w)$ has strict support equal to an irreducible analytic subvariety $Z_j \subseteq X$. There are no nontrivial morphisms between Hodge modules with different strict support; if we assume that $Z_1, \dots, Z_n$ are distinct, the decomposition by strict support is therefore unique. Since the category $\mathrm{HM}_{\mathbb{R}}(X, w)$ is semisimple, it follows that every polarizable real Hodge module of weight w is isomorphic to a direct sum of simple objects with strict support.

One of Saito's most important results is the following structure theorem, relating polarizable real Hodge modules and polarizable real variations of Hodge structure.

Theorem 1.3 (Saito) *The category of polarizable real Hodge modules of weight w with strict support $Z \subseteq X$ is equivalent to the category of generically defined polarizable real variations of Hodge structure of weight $w - \dim Z$ on Z.*

In other words, for any $M \in \mathrm{HM}_{\mathbb{R}}(X, w)$ with strict support Z, there is a dense Zariski-open subset of the smooth locus of Z over which it restricts to a polarizable real variation of Hodge structure; conversely, every such variation extends uniquely to a Hodge module with strict support Z. The proof in [30, Theorem 3.21] carries over to the case of real coefficients; see [29] for further discussion.

Lemma 1.4 *The support of $M \in \mathrm{HM}_{\mathbb{R}}(X, w)$ lies in a submanifold $i\colon Y \hookrightarrow X$ if and only if M belongs to the image of the functor $i_*\colon \mathrm{HM}_{\mathbb{R}}(Y, w) \to \mathrm{HM}_{\mathbb{R}}(X, w)$.*

This result is often called *Kashiwara's equivalence*, because Kashiwara proved the same thing for arbitrary coherent $\mathscr{D}$–modules. In the case of Hodge modules, the point is that the coherent $\mathscr{O}_X$–modules $F_k\mathcal{M}/F_{k-1}\mathcal{M}$ are in fact $\mathscr{O}_Y$–modules.

2 Compact Kähler manifolds and semisimplicity

In this section, we prove some results about the underlying regular holonomic $\mathscr{D}$–modules of polarizable real Hodge modules on compact Kähler manifolds. Our starting point is the following semisimplicity theorem:

Theorem 2.1 (Deligne, Nori) *Let X be a compact Kähler manifold. If*

$$M = (\mathcal{M}, F_\bullet\mathcal{M}, M_{\mathbb{R}}) \in \mathrm{HM}_{\mathbb{R}}(X, w),$$

then the perverse sheaf $M_{\mathbb{R}}$ and the $\mathscr{D}$–module $\mathcal{M}$ are semisimple.

Proof Since the category $\mathrm{HM}_{\mathbb{R}}(X, w)$ is semisimple, we may assume without loss of generality that M is simple, with strict support an irreducible analytic subvariety $Z \subseteq X$. By Saito's Theorem 1.3, M restricts to a polarizable real variation of Hodge structure $\mathcal{H}$ of weight $w - \dim Z$ on a Zariski-open subset of the smooth locus of Z; note that $\mathcal{H}$ is a simple object in the category of real variations of Hodge structure. Now

$M_{\mathbb{R}}$ is the intersection complex of $\mathcal{H}_{\mathbb{R}}$, and so it suffices to prove that $\mathcal{H}_{\mathbb{R}}$ is semisimple. After resolving singularities, we can assume that $\mathcal{H}$ is defined on a Zariski-open subset of a compact Kähler manifold; in that case, Deligne and Nori have shown that $\mathcal{H}_{\mathbb{R}}$ is semisimple [11, Section 1.12]. It follows that the complexification $M_{\mathbb{R}} \otimes_{\mathbb{R}} \mathbb{C}$ of the perverse sheaf is semisimple as well; by the Riemann–Hilbert correspondence, the same is true for the underlying regular holonomic $\mathscr{D}$–module $\mathcal{M}$. $\square$

A priori, there is no reason why the decomposition of the regular holonomic $\mathscr{D}$–module $\mathcal{M}$ into simple factors should lift to a decomposition in the category $\mathrm{HM}_{\mathbb{R}}(X, w)$. Nevertheless, it turns out that we can always chose the decomposition in such a way that it is compatible with the filtration $F_\bullet \mathcal{M}$.

Proposition 2.2 *Let $M \in \mathrm{HM}_{\mathbb{R}}(X, w)$ be a simple polarizable real Hodge module on a compact Kähler manifold. Then one of the following two statements is true:*

(1) *The underlying perverse sheaf $M_{\mathbb{R}} \otimes_{\mathbb{R}} \mathbb{C}$ is simple.*

(2) *There is an endomorphism $J \in \mathrm{End}(M)$ with $J^2 = -\mathrm{id}$ such that*

$$(\mathcal{M}, F_\bullet \mathcal{M}, M_{\mathbb{R}} \otimes_{\mathbb{R}} \mathbb{C}) = \ker(J - i \cdot \mathrm{id}) \oplus \ker(J + i \cdot \mathrm{id}),$$

and the perverse sheaves underlying $\ker(J \pm i \cdot \mathrm{id})$ are simple.

We begin by proving the following key lemma:

Lemma 2.3 *Let $\mathcal{H}$ be a polarizable real variation of Hodge structure on a Zariski-open subset of a compact Kähler manifold. If $\mathcal{H}$ is simple, then*

(a) *either the underlying complex local system $\mathcal{H}_{\mathbb{C}}$ is also simple,*

(b) *or there is an endomorphism $J \in \mathrm{End}(\mathcal{H})$ with $J^2 = -\mathrm{id}$ such that*

$$\mathcal{H}_{\mathbb{C}} = \ker(J_{\mathbb{C}} - i \cdot \mathrm{id}) \oplus \ker(J_{\mathbb{C}} + i \cdot \mathrm{id})$$

is the sum of two (possibly isomorphic) simple local systems.

Proof Since X is a Zariski-open subset of a compact Kähler manifold, the theorem of the fixed part holds on X, and the local system $\mathcal{H}_{\mathbb{C}}$ is semisimple [11, Section 1.12]. Choose a base point $x_0 \in X$, and write $H_{\mathbb{R}}$ for the fiber of the local system $\mathcal{H}_{\mathbb{R}}$ at the point x_0; it carries a polarizable Hodge structure

$$H_{\mathbb{C}} = H_{\mathbb{R}} \otimes_{\mathbb{R}} \mathbb{C} = \bigoplus_{p+q=w} H^{p,q},$$

say of weight w. The fundamental group $\Gamma = \pi_1(X, x_0)$ acts on $H_{\mathbb{R}}$, and, as we remarked above, $H_{\mathbb{C}}$ decomposes into a sum of simple Γ–modules. The proof of

[11, Proposition 1.13] shows that there is a nontrivial simple Γ–module $V \subseteq H_{\mathbb{C}}$ compatible with the Hodge decomposition, meaning that

$$V = \bigoplus_{p+q=w} V \cap H^{p,q}.$$

Let $\overline{V} \subseteq H_{\mathbb{C}}$ denote the conjugate of V with respect to the real structure $H_{\mathbb{R}}$; it is another nontrivial simple Γ–module with

$$\overline{V} = \bigoplus_{p+q=w} \overline{V} \cap H^{p,q}.$$

The intersection $(V + \overline{V}) \cap H_{\mathbb{R}}$ is therefore a Γ–invariant real sub-Hodge structure of $H_{\mathbb{R}}$. By the theorem of the fixed part, it extends to a real subvariation of $\mathcal{H}$; since $\mathcal{H}$ is simple, this means that $H_{\mathbb{C}} = V + \overline{V}$. Now there are two possibilities:

(1) If $V = \overline{V}$, then $H_{\mathbb{C}} = V$, and $\mathcal{H}_{\mathbb{C}}$ is a simple local system.

(2) If $V \neq \overline{V}$, then $H_{\mathbb{C}} = V \oplus \overline{V}$, and $\mathcal{H}_{\mathbb{C}}$ is the sum of two (possibly isomorphic) simple local systems.

The endomorphism algebra $\operatorname{End}(\mathcal{H}_{\mathbb{R}})$ coincides with the subalgebra of Γ–invariants in $\operatorname{End}(H_{\mathbb{R}})$; by the theorem of the fixed part, it is also a real sub-Hodge structure. Let $p \in \operatorname{End}(H_{\mathbb{C}})$ and $\overline{p} \in \operatorname{End}(H_{\mathbb{C}})$ denote the projections to the two subspaces V and $\overline{V}$; both preserve the Hodge decomposition, and are therefore of type $(0,0)$. This shows that the element $J = i(p - \overline{p}) \in \operatorname{End}(H_{\mathbb{C}})$ is a real Hodge class of type $(0,0)$ with $J^2 = -\operatorname{id}$; by the theorem of the fixed part, J is the restriction to x_0 of an endomorphism of the variation of Hodge structure $\mathcal{H}$. This completes the proof because V and $\overline{V}$ are exactly the $\pm i$–eigenspaces of J. □

Proof of Proposition 2.2 Since M is simple, it has strict support equal to an irreducible analytic subvariety $Z \subseteq X$; by Theorem 1.3, M is obtained from a polarizable real variation of Hodge structure $\mathcal{H}$ of weight $w - \dim Z$ on a dense Zariski-open subset of the smooth locus of Z. Let $\mathcal{H}_{\mathbb{R}}$ denote the underlying real local system; then $M_{\mathbb{R}}$ is isomorphic to the intersection complex of $\mathcal{H}_{\mathbb{R}}$. Since we can resolve the singularities of Z by blowing up along submanifolds of X, Lemma 2.3 applies to this situation; it shows that $\mathcal{H}_{\mathbb{C}} = \mathcal{H}_{\mathbb{R}} \otimes_{\mathbb{R}} \mathbb{C}$ has at most two simple factors. The same is true for $M_{\mathbb{R}} \otimes_{\mathbb{R}} \mathbb{C}$ and, by the Riemann–Hilbert correspondence, for $\mathcal{M}$.

Now we have to consider two cases. If $\mathcal{H}_{\mathbb{C}}$ is simple, then $\mathcal{M}$ is also simple, and we are done. If $\mathcal{H}_{\mathbb{C}}$ is not simple, then by Lemma 2.3 there is an endomorphism $J \in \operatorname{End}(\mathcal{H})$ with $J^2 = -\operatorname{id}$ such that the two simple factors are the $\pm i$–eigenspaces of J. By Theorem 1.3, it extends uniquely to an endomorphism of $J \in \operatorname{End}(M)$ in the category $\mathrm{HM}_{\mathbb{R}}(X, w)$; in particular, we obtain an induced endomorphism

$$J\colon \mathcal{M} \to \mathcal{M}$$

that is strictly compatible with the filtration $F_\bullet \mathcal{M}$ by [28, Proposition 5.1.14]. Now the $\pm i$–eigenspaces of J give us the desired decomposition

$$(\mathcal{M}, F_\bullet \mathcal{M}) = (\mathcal{M}', F_\bullet \mathcal{M}') \oplus (\mathcal{M}'', F_\bullet \mathcal{M}'');$$

note that the two regular holonomic $\mathscr{D}$–modules $\mathcal{M}'$ and $\mathcal{M}''$ are simple because the corresponding perverse sheaves are the intersection complexes of the simple complex local systems $\ker(J_{\mathbb{C}} \pm i \cdot \mathrm{id})$, where $J_{\mathbb{C}}$ stands for the induced endomorphism of the complexification $M_{\mathbb{R}} \otimes_{\mathbb{R}} \mathbb{C}$. □

3 Complex Hodge modules

In Saito's recursive definition of the category of polarizable Hodge modules, the existence of a real structure is crucial: to say that a given filtration on a complex vector space is a Hodge structure of a certain weight, or that a given bilinear form is a polarization, one needs to have complex conjugation. This explains why there is as yet no general theory of "polarizable complex Hodge modules" — although it seems likely that such a theory can be constructed within the framework of twistor $\mathscr{D}$–modules developed by Sabbah and Mochizuki. We now explain a workaround for this problem, suggested by Proposition 2.2.

Definition 3.1 A *polarizable complex Hodge module* on a complex manifold X is a pair (M, J), consisting of a polarizable real Hodge module $M \in \mathrm{HM}_{\mathbb{R}}(X, w)$ and an endomorphism $J \in \operatorname{End}(M)$ with $J^2 = -\mathrm{id}$.

The space of morphisms between two polarizable complex Hodge modules (M_1, J_1) and (M_2, J_2) is defined in the obvious way:

$$\operatorname{Hom}((M_1, J_1), (M_2, J_2)) = \{f \in \operatorname{Hom}(M_1, M_2) \mid f \circ J_1 = J_2 \circ f\}.$$

Note that composition with J_1 (or equivalently, J_2) puts a natural complex structure on this real vector space.

Definition 3.2 We denote by $\mathrm{HM}_{\mathbb{C}}(X, w)$ the category of polarizable complex Hodge modules of weight w; it is $\mathbb{C}$–linear and abelian.

From a polarizable complex Hodge module (M, J), we obtain a filtered regular holonomic $\mathscr{D}$–module as well as a complex perverse sheaf, as follows. Denote by

$$\mathcal{M} = \mathcal{M}' \oplus \mathcal{M}'' = \ker(J - i \cdot \mathrm{id}) \oplus \ker(J + i \cdot \mathrm{id})$$

the induced decomposition of the regular holonomic $\mathscr{D}$–module underlying M, and observe that $J \in \mathrm{End}(\mathcal{M})$ is strictly compatible with the Hodge filtration $F_\bullet\mathcal{M}$. This means that we have a decomposition

$$(\mathcal{M}, F_\bullet\mathcal{M}) = (\mathcal{M}', F_\bullet\mathcal{M}') \oplus (\mathcal{M}'', F_\bullet\mathcal{M}'')$$

in the category of filtered $\mathscr{D}$–modules. Similarly, let $J_{\mathbb{C}} \in \mathrm{End}(M_{\mathbb{C}})$ denote the induced endomorphism of the complex perverse sheaf underlying M; then

$$M_{\mathbb{C}} = M_{\mathbb{R}} \otimes_{\mathbb{R}} \mathbb{C} = \ker(J_{\mathbb{C}} - i \cdot \mathrm{id}) \oplus \ker(J_{\mathbb{C}} + i \cdot \mathrm{id}),$$

and the two summands correspond to $\mathcal{M}'$ and $\mathcal{M}''$ under the Riemann–Hilbert correspondence. Note that they are isomorphic as *real* perverse sheaves; the only difference is in the $\mathbb{C}$–action. We obtain a functor

$$(M, J) \mapsto \ker(J_{\mathbb{C}} - i \cdot \mathrm{id})$$

from $\mathrm{HM}_{\mathbb{C}}(X, w)$ to the category of complex perverse sheaves on X; it is faithful, but depends on the choice of i.

Definition 3.3 Given $(M, J) \in \mathrm{HM}_{\mathbb{C}}(X, w)$, we call

$$\ker(J_{\mathbb{C}} - i \cdot \mathrm{id}) \subseteq M_{\mathbb{C}}$$

the *underlying complex perverse sheaf*, and

$$(\mathcal{M}', F_\bullet\mathcal{M}') = \ker(J - i \cdot \mathrm{id}) \subseteq (\mathcal{M}, F_\bullet\mathcal{M})$$

the *underlying filtered regular holonomic $\mathscr{D}$–module*.

There is also an obvious functor from polarizable real Hodge modules to polarizable complex Hodge modules: it takes $M \in \mathrm{HM}_{\mathbb{R}}(X, w)$ to the pair

$$(M \oplus M, J_M), \quad J_M(m_1, m_2) = (-m_2, m_1).$$

Not surprisingly, the underlying complex perverse sheaf is isomorphic to $M_{\mathbb{R}} \otimes_{\mathbb{R}} \mathbb{C}$, and the underlying filtered regular holonomic $\mathscr{D}$–module to $(\mathcal{M}, F_\bullet\mathcal{M})$. The proof of the following lemma is left as an easy exercise.

Lemma 3.4 *A polarized complex Hodge module $(M, J) \in \mathrm{HM}_{\mathbb{C}}(X, w)$ belongs to the image of $\mathrm{HM}_{\mathbb{R}}(X, w)$ if and only if there exists $r \in \mathrm{End}(M)$ with*

$$r \circ J = -J \circ r \quad \textit{and} \quad r^2 = \mathrm{id}.$$

In particular, (M, J) should be isomorphic to its *complex conjugate* $(M, -J)$, but this in itself does not guarantee the existence of a real structure — for example when M is simple and $\mathrm{End}(M)$ is isomorphic to the quaternions $\mathbb{H}$.

Proposition 3.5 *The category* $\mathrm{HM}_{\mathbb{C}}(X, w)$ *is semisimple, and the simple objects are of the following two types:*

(i) $(M \oplus M, J_M)$, *where* $M \in \mathrm{HM}_{\mathbb{R}}(X, w)$ *is simple and* $\mathrm{End}(M) = \mathbb{R}$.

(ii) (M, J), *where* $M \in \mathrm{HM}_{\mathbb{R}}(X, w)$ *is simple and* $\mathrm{End}(M) \in \{\mathbb{C}, \mathbb{H}\}$.

Proof Since $\mathrm{HM}_{\mathbb{R}}(X, w)$ is semisimple, every object of $\mathrm{HM}_{\mathbb{C}}(X, w)$ is isomorphic to a direct sum of polarizable complex Hodge modules of the form

$$(M^{\oplus n}, J), \tag{3.6}$$

where $M \in \mathrm{HM}_{\mathbb{R}}(X, w)$ is simple and J is an $n \times n$ matrix with entries in $\mathrm{End}(M)$ such that $J^2 = -\mathrm{id}$. By Schur's lemma and the classification of real division algebras, the endomorphism algebra of a simple polarizable real Hodge module is one of $\mathbb{R}$, $\mathbb{C}$ or $\mathbb{H}$. If $\mathrm{End}(M) = \mathbb{R}$, elementary linear algebra shows that n must be even and that (3.6) is isomorphic to the direct sum of $\frac{n}{2}$ copies of (i). If $\mathrm{End}(M) = \mathbb{C}$, one can diagonalize the matrix J; this means that (3.6) is isomorphic to a direct sum of n objects of type (ii). If $\mathrm{End}(M) = \mathbb{H}$, it is still possible to diagonalize J, but this needs some nontrivial results about matrices with entries in the quaternions [42]. Write $J \in M_n(\mathbb{H})$ in the form $J = J_1 + J_2 j$, with $J_1, J_2 \in M_n(\mathbb{C})$, and consider the "adjoint matrix"

$$\chi_J = \begin{pmatrix} J_1 & J_2 \\ -\bar{J}_2 & \bar{J}_1 \end{pmatrix} \in M_{2n}(\mathbb{C}).$$

Since $J^2 = -\mathrm{id}$, one also has $\chi_J^2 = -\mathrm{id}$, and so the matrix J is normal by [42, Theorem 4.2]. According to [42, Corollary 6.2], this implies the existence of a unitary matrix $U \in M_n(\mathbb{H})$ such that $U^{-1} A U = i \cdot \mathrm{id}$; here unitary means that $U^{-1} = U^*$ is equal to the conjugate transpose of U. The consequence is that (3.6) is again isomorphic to a direct sum of n objects of type (ii). Since it is straightforward to prove that both types of objects are indeed simple, this concludes the proof. □

Note The three possible values for the endomorphism algebra of a simple object $M \in \mathrm{HM}_{\mathbb{R}}(X, w)$ reflect the splitting behavior of its complexification $(M \oplus M, J_M)$ in $\mathrm{HM}_{\mathbb{C}}(X, w)$: if $\mathrm{End}(M) = \mathbb{R}$, it remains irreducible; if $\mathrm{End}(M) = \mathbb{C}$, it splits into two nonisomorphic simple factors; if $\mathrm{End}(M) = \mathbb{H}$, it splits into two isomorphic simple factors. Note that the endomorphism ring of a simple polarizable complex Hodge module is always isomorphic to $\mathbb{C}$, in accordance with Schur's lemma.

Our ad hoc definition of the category $\mathrm{HM}_{\mathbb{C}}(X, w)$ has the advantage that every result about polarizable real Hodge modules that does not explicitly mention the real structure extends to polarizable complex Hodge modules. For example, each $(M, J) \in \mathrm{HM}_{\mathbb{C}}(X, w)$ admits a unique decomposition by strict support: M admits such a decomposition, and since there are no nontrivial morphisms between objects with different strict support, J is automatically compatible with the decomposition. For much the same reason, Kashiwara's equivalence (in Lemma 1.4) holds also for polarizable complex Hodge modules.

Another result that immediately carries over is Saito's direct image theorem. The strictness of the direct image complex is one of the crucial properties of polarizable Hodge modules; in the special case of the morphism from a projective variety X to a point, it is equivalent to the E_1–degeneration of the spectral sequence

$$E_1^{p,q} = H^{p+q}(X, \mathrm{gr}_p^F \mathrm{DR}(\mathcal{M}')) \Rightarrow H^{p+q}(X, \mathrm{DR}(\mathcal{M}')),$$

a familiar result in classical Hodge theory when $\mathcal{M}' = \mathscr{O}_X$.

Theorem 3.7 *Let $f\colon X \to Y$ be a projective morphism between complex manifolds.*

(a) *If $(M, J) \in \mathrm{HM}_{\mathbb{C}}(X, w)$, then for each $k \in \mathbb{Z}$, the pair*

$$\mathcal{H}^k f_*(M, J) = (\mathcal{H}^k f_* M, \mathcal{H}^k f_* J) \in \mathrm{HM}_{\mathbb{C}}(Y, w + k)$$

is again a polarizable complex Hodge module.

(b) *The direct image complex $f_+(\mathcal{M}', F_\bullet\mathcal{M}')$ is strict, and $\mathcal{H}^k f_+(\mathcal{M}', F_\bullet\mathcal{M}')$ is the filtered regular holonomic $\mathscr{D}$–module underlying $\mathcal{H}^k f_*(M, J)$.*

Proof Since $M \in \mathrm{HM}_{\mathbb{R}}(X, w)$ is a polarizable real Hodge module, $\mathcal{H}^k f_* M$ is in $\mathrm{HM}_{\mathbb{R}}(Y, w+k)$ by Saito's direct image theorem [28, Théorème 5.3.1]. Now it suffices to note that $J \in \mathrm{End}(M)$ induces an endomorphism $\mathcal{H}^k f_* J \in \mathrm{End}(\mathcal{H}^k f_* M)$ whose square is equal to minus the identity. Since

$$(\mathcal{M}, F_\bullet\mathcal{M}) = (\mathcal{M}', F_\bullet\mathcal{M}') \oplus (\mathcal{M}'', F_\bullet\mathcal{M}''),$$

the strictness of the complex $f_+(\mathcal{M}', F_\bullet\mathcal{M}')$ follows from that of $f_+(\mathcal{M}, F_\bullet\mathcal{M})$, which is part of the above-cited theorem by Saito. □

On compact Kähler manifolds, the semisimplicity results from the previous section can be summarized as follows:

Proposition 3.8 *Let X be a compact Kähler manifold.*

(a) *A polarizable complex Hodge module $(M, J) \in \mathrm{HM}_{\mathbb{C}}(X, w)$ is simple if and only if the underlying complex perverse sheaf*

$$\ker\bigl(J_{\mathbb{C}} - i \cdot \mathrm{id}\colon M_{\mathbb{R}} \otimes_{\mathbb{R}} \mathbb{C} \to M_{\mathbb{R}} \otimes_{\mathbb{R}} \mathbb{C}\bigr)$$

is simple.

(b) *If $M \in \mathrm{HM}_{\mathbb{R}}(X, w)$, then every simple factor of the complex perverse sheaf $M_{\mathbb{R}} \otimes_{\mathbb{R}} \mathbb{C}$ underlies a polarizable complex Hodge module.*

Proof This is a restatement of Proposition 2.2. □

4 Complex variations of Hodge structure

In this section, we discuss the relation between polarizable complex Hodge modules and polarizable complex variations of Hodge structure.

Definition 4.1 A *polarizable complex variation of Hodge structure* is a pair $(\mathcal{H}, J)$, where $\mathcal{H}$ is a polarizable real variation of Hodge structure and $J \in \operatorname{End}(\mathcal{H})$ is an endomorphism with $J^2 = -\mathrm{id}$.

As before, the *complexification* of a real variation $\mathcal{H}$ is defined as

$$(\mathcal{H} \oplus \mathcal{H}, J_{\mathcal{H}}), \quad J_{\mathcal{H}}(h_1, h_2) = (-h_2, h_1),$$

and a complex variation $(\mathcal{H}, J)$ is real if and only if there is an endomorphism $r \in \operatorname{End}(\mathcal{H})$ with $r \circ J = -J \circ r$ and $r^2 = \mathrm{id}$. Note that the direct sum of $(\mathcal{H}, J)$ with its *complex conjugate* $(\mathcal{H}, -J)$ has an obvious real structure.

The definition above is convenient for our purposes; it is also not hard to show that it is equivalent to the one in [11, Section 1], up to the choice of weight. (Deligne only considers complex variations of weight zero.)

Example 4.2 Let $\rho \in \operatorname{Char}(X)$ be a unitary character of the fundamental group, and denote by $\mathbb{C}_\rho$ the resulting unitary local system. It determines a polarizable complex variation of Hodge structure in the following manner. The underlying real local system is $\mathbb{R}^2$, with monodromy acting by

$$\begin{pmatrix} \operatorname{Re}\rho & -\operatorname{Im}\rho \\ \operatorname{Im}\rho & \operatorname{Re}\rho \end{pmatrix};$$

the standard inner product on $\mathbb{R}^2$ makes this into a polarizable real variation of Hodge structure $\mathcal{H}_\rho$ of weight zero, with $J_\rho \in \operatorname{End}(\mathcal{H}_\rho)$ acting as $J_\rho(x, y) = (-y, x)$; for simplicity, we continue to denote the pair $(\mathcal{H}_\rho, J_\rho)$ by the symbol $\mathbb{C}_\rho$.

We have the following criterion for deciding whether a polarizable complex Hodge module is *smooth*, meaning induced by a complex variation of Hodge structure.

Lemma 4.3 *Given $(M, J) \in \mathrm{HM}_{\mathbb{C}}(X, w)$, let us denote by*

$$\mathcal{M} = \mathcal{M}' \oplus \mathcal{M}'' = \ker(J - i \cdot \mathrm{id}) \oplus \ker(J + i \cdot \mathrm{id})$$

the induced decomposition of the regular holonomic $\mathscr{D}$–module underlying M. If $\mathcal{M}'$ is coherent as an $\mathscr{O}_X$–module, then M is smooth.

Proof Let $M_{\mathbb{C}} = \ker(J_{\mathbb{C}} - i \cdot \mathrm{id}) \oplus \ker(J_{\mathbb{C}} + i \cdot \mathrm{id})$ be the analogous decomposition of the underlying perverse sheaf. Since $\mathcal{M}'$ is $\mathscr{O}_X$–coherent, it is a vector bundle with flat connection; by the Riemann–Hilbert correspondence, the first factor is therefore (up to a shift in degree by $\dim X$) a complex local system. Since it is isomorphic to $M_{\mathbb{R}}$ as a real perverse sheaf, it follows that $M_{\mathbb{R}}$ is also a local system; but then M is smooth by [28, Lemme 5.1.10]. □

In general, the relationship between complex Hodge modules and complex variations of Hodge structure is governed by the following theorem; it is of course an immediate consequence of Saito's results (see Theorem 1.3).

Theorem 4.4 *The category of polarizable complex Hodge modules of weight w with strict support $Z \subseteq X$ is equivalent to the category of generically defined polarizable complex variations of Hodge structure of weight $w - \dim Z$ on Z.*

5 Integral structures on Hodge modules

By working with polarizable real (or complex) Hodge modules, we lose certain arithmetic information about the monodromy of the underlying perverse sheaves, such as the fact that the monodromy eigenvalues are roots of unity. One can recover some of this information by asking for the existence of an "integral structure" [35, Definition 1.9], which is just a constructible complex of sheaves of $\mathbb{Z}$–modules that becomes isomorphic to the perverse sheaf underlying the Hodge module after tensoring by $\mathbb{R}$.

Definition 5.1 An *integral structure* on a polarizable real Hodge module M in $\mathrm{HM}_{\mathbb{R}}(X, w)$ is a constructible complex $E \in \mathrm{D}_c^b(\mathbb{Z}_X)$ such that $M_{\mathbb{R}} \simeq E \otimes_{\mathbb{Z}} \mathbb{R}$.

As explained in [35, Section 1.2.2], the existence of an integral structure is preserved by many of the standard operations on (mixed) Hodge modules, such as direct and inverse images or duality. Note that even though it makes sense to ask whether a given (mixed) Hodge module admits an integral structure, there appears to be no good functorial theory of "polarizable integral Hodge modules".

Lemma 5.2 *If $M \in \mathrm{HM}_{\mathbb{R}}(X, w)$ admits an integral structure, then the same is true for every summand in the decomposition of M by strict support.*

Proof Consider the decomposition

$$M = \bigoplus_{j=1}^{n} M_j$$

by strict support, with $Z_1, \dots, Z_n \subseteq X$ distinct irreducible analytic subvarieties. Each M_j is a polarizable real Hodge module with strict support Z_j, and therefore comes from a polarizable real variation of Hodge structure $\mathcal{H}_j$ on a dense Zariski-open subset of Z_j. What we must prove is that each $\mathcal{H}_j$ can be defined over $\mathbb{Z}$. Let $M_{\mathbb{R}}$ be the underlying real perverse sheaf, and set $d_j = \dim Z_j$. According to [2, Proposition 2.1.17], Z_j is an irreducible component in the support of the $(-d_j)^{\text{th}}$ cohomology sheaf of $M_{\mathbb{R}}$ and $\mathcal{H}_{j,\mathbb{R}}$ is the restriction of that constructible sheaf to a Zariski-open subset of Z_j. Since $M_{\mathbb{R}} \simeq E \otimes_{\mathbb{Z}} \mathbb{R}$, it follows that $\mathcal{H}_j$ is defined over $\mathbb{Z}$. □

6 Operations on Hodge modules

In this section, we recall three useful operations for polarizable real (and complex) Hodge modules. If $\operatorname{Supp} M$ is compact, we define the *Euler characteristic* of $M = (\mathcal{M}, F_\bullet\mathcal{M}, M_{\mathbb{R}}) \in \mathrm{HM}_{\mathbb{R}}(X, w)$ by the formula

$$\chi(X, M) = \sum_{i \in \mathbb{Z}} (-1)^i \dim_{\mathbb{R}} H^i(X, M_{\mathbb{R}}) = \sum_{i \in \mathbb{Z}} (-1)^i \dim_{\mathbb{C}} H^i(X, \mathrm{DR}(\mathcal{M})).$$

For $(M, J) \in \mathrm{HM}_{\mathbb{C}}(X, w)$, we let $\mathcal{M} = \mathcal{M}' \oplus \mathcal{M}'' = \ker(J - i \cdot \mathrm{id}) \oplus \ker(J + i \cdot \mathrm{id})$ be the decomposition into eigenspaces, and define

$$\chi(X, M, J) = \sum_{i \in \mathbb{Z}} (-1)^i \dim_{\mathbb{C}} H^i(X, \mathrm{DR}(\mathcal{M}')).$$

With this definition, one has $\chi(X, M) = \chi(X, M, J) + \chi(X, M, -J)$.

Given a smooth morphism $f\colon Y \to X$ of relative dimension $\dim f = \dim Y - \dim X$, we define the *naive inverse image*

$$f^{-1}M = (f^*\mathcal{M}, f^*F_\bullet\mathcal{M}, f^{-1}M_{\mathbb{R}}).$$

One can show that $f^{-1}M \in \mathrm{HM}_{\mathbb{R}}(Y, w + \dim f)$; see [36, Section 9] for more details. The same is true for polarizable complex Hodge modules: if $(M, J) \in \mathrm{HM}_{\mathbb{C}}(X, w)$, then one obviously has

$$f^{-1}(M, J) = (f^{-1}M, f^{-1}J) \in \mathrm{HM}_{\mathbb{C}}(Y, w + \dim f).$$

One can also twist a polarizable complex Hodge module by a unitary character.

Lemma 6.1 *For any unitary character $\rho \in \mathrm{Char}(X)$, there is an object*

$$(M, J) \otimes_{\mathbb{C}} \mathbb{C}_\rho \in \mathrm{HM}_{\mathbb{C}}(X, w)$$

whose associated complex perverse sheaf is $\ker(J_{\mathbb{C}} - i \cdot \mathrm{id}) \otimes_{\mathbb{C}} \mathbb{C}_\rho$.

Proof In the notation of Example 4.2, consider the tensor product

$$M \otimes_{\mathbb{R}} \mathcal{H}_\rho \in \mathrm{HM}_{\mathbb{R}}(X, w);$$

it is again a polarizable real Hodge module of weight w because $\mathcal{H}_\rho$ is a polarizable real variation of Hodge structure of weight zero. The square of the endomorphism $J \otimes J_\rho$ is the identity, and so

$$N = \ker(J \otimes J_\rho + \mathrm{id}) \subseteq M \otimes_{\mathbb{R}} \mathcal{H}_\rho$$

is again a polarizable real Hodge module of weight w. Now $K = J \otimes \mathrm{id} \in \mathrm{End}(N)$ satisfies $K^2 = -\mathrm{id}$, which means that the pair (N, K) is a polarizable complex Hodge module of weight w. On the associated complex perverse sheaf

$$\ker(K_{\mathbb{C}} - i \cdot \mathrm{id}) \subseteq M_{\mathbb{C}} \otimes_{\mathbb{C}} \mathcal{H}_{\rho,\mathbb{C}},$$

both $J_{\mathbb{C}} \otimes \mathrm{id}$ and $\mathrm{id} \otimes J_{\rho,\mathbb{C}}$ act as multiplication by i, which means that

$$\ker(K_{\mathbb{C}} - i \cdot \mathrm{id}) = \ker(J_{\mathbb{C}} - i \cdot \mathrm{id}) \otimes_{\mathbb{C}} \mathbb{C}_{\rho}.$$

The corresponding regular holonomic $\mathscr{D}$–module is obviously

$$\mathcal{N}' = \mathcal{M}' \otimes_{\mathscr{O}_X} (L, \nabla),$$

with the filtration induced by $F_{\bullet}\mathcal{M}'$; here (L, ∇) denotes the flat bundle corresponding to the complex local system $\mathbb{C}_{\rho}$, and $\mathcal{M} = \mathcal{M}' \oplus \mathcal{M}''$ as above. □

Note The proof shows that

$$N_{\mathbb{C}} = \bigl(\ker(J_{\mathbb{C}} - i \cdot \mathrm{id}) \otimes_{\mathbb{C}} \mathbb{C}_{\rho}\bigr) \oplus \bigl(\ker(J_{\mathbb{C}} + i \cdot \mathrm{id}) \otimes_{\mathbb{C}} \mathbb{C}_{\overline{\rho}}\bigr),$$
$$\mathcal{N} = \bigl(\mathcal{M}' \otimes_{\mathscr{O}_X} (L, \nabla)\bigr) \oplus \bigl(\mathcal{M}'' \otimes_{\mathscr{O}_X} (L, \nabla)^{-1}\bigr),$$

where $\overline{\rho}$ is the complex conjugate of the character $\rho \in \mathrm{Char}(X)$.

C Hodge modules on complex tori

7 Main result

The paper [26] contains several results about Hodge modules of geometric origin on abelian varieties. In this chapter, we generalize these results to arbitrary polarizable complex Hodge modules on compact complex tori. To do so, we develop a beautiful idea due to Wang [41], namely that, up to direct sums and character twists, every such object actually comes from an abelian variety.

Theorem 7.1 *Let $(M, J) \in \mathrm{HM}_{\mathbb{C}}(T, w)$ be a polarizable complex Hodge module on a compact complex torus T. Then there is a decomposition*

$$(M, J) \simeq \bigoplus_{j=1}^{n} q_j^{-1}(N_j, J_j) \otimes_{\mathbb{C}} \mathbb{C}_{\rho_j}, \tag{7.2}$$

where $q_j \colon T \to T_j$ is a surjective morphism with connected fibers, $\rho_j \in \mathrm{Char}(T)$ is a unitary character and $(N_j, J_j) \in \mathrm{HM}_{\mathbb{C}}(T_j, w - \dim q_j)$ is a simple polarizable complex Hodge module with $\mathrm{Supp}\, N_j$ projective and $\chi(T_j, N_j, J_j) > 0$.

For Hodge modules of geometric origin, a less precise result was proved by Wang [41]. His proof makes use of the decomposition theorem, which in the setting of arbitrary compact Kähler manifolds is only known for Hodge modules of geometric origin [29]. This technical issue can be circumvented by putting everything in terms of generically defined variations of Hodge structure.

To get a result for a polarizable real Hodge module $M \in \mathrm{HM}_{\mathbb{R}}(T, w)$, we simply apply Theorem 7.1 to its complexification $(M \oplus M, J_M) \in \mathrm{HM}_{\mathbb{C}}(T, w)$. One could say more about the terms in the decomposition below, but the following version is enough for our purposes.

Corollary 7.3 *Let $M \in \mathrm{HM}_{\mathbb{R}}(T, w)$ be a polarizable real Hodge module on a compact complex torus T. Then, in the notation of Theorem 7.1, one has*

$$(M \oplus M, J_M) \simeq \bigoplus_{j=1}^{n} q_j^{-1}(N_j, J_j) \otimes_{\mathbb{C}} \mathbb{C}_{\rho_j}.$$

If M admits an integral structure, then each $\rho_j \in \mathrm{Char}(T)$ has finite order.

The proof of these results takes up the rest of the chapter.

8 Subvarieties of complex tori

This section contains a structure theorem for subvarieties of compact complex tori. The statement is contained in [41, Propositions 2.3 and 2.4], but we give a simpler argument below.

Proposition 8.1 *Let X be an irreducible analytic subvariety of a compact complex torus T. Then there is a subtorus $S \subseteq T$ with the following two properties:*

(a) *$S + X = X$ and the quotient $Y = X/S$ is projective.*

(b) *If $D \subseteq X$ is an irreducible analytic subvariety with $\dim D = \dim X - 1$, then $S + D = D$.*

In particular, every divisor on X is the preimage of a divisor on Y.

Proof It is well known that the algebraic reduction of T is an abelian variety. More precisely, there is a subtorus $S \subseteq T$ such that $A = T/S$ is an abelian variety, and every other subtorus with this property contains S; see eg [4, Chapter 2, Section 6].

Now let $X \subseteq T$ be an irreducible analytic subvariety of T; without loss of generality, we may assume that $0 \in X$ and that X is not contained in any proper subtorus of T. By a theorem of Ueno [39, Theorem 10.9], there is a subtorus $S' \subseteq T$ with $S' + X \subseteq X$

and such that $X/S' \subseteq T/S'$ is of general type. In particular, X/S' is projective; but then T/S' must also be projective, which means that $S \subseteq S'$. Setting $Y = X/S$, we get a cartesian diagram

$$\begin{tikzcd} X \arrow[r, hook] \arrow[d] & T \arrow[d] \\ Y \arrow[r, hook] & A \end{tikzcd}$$

with Y projective. Now it remains to show that every divisor on X is the pullback of a divisor from Y.

Let $D \subseteq X$ be an irreducible analytic subvariety with $\dim D = \dim X - 1$; as before, we may assume that $0 \in D$. For dimension reasons, either $S + D = D$ or $S + D = X$; let us suppose that $S + D = X$ and see how this leads to a contradiction. Define $T_D \subseteq T$ to be the smallest subtorus of T containing D; then $S + T_D = T$. If $T_D = T$, then the same reasoning as above would show that $S + D = D$; therefore $T_D \neq T$, and $\dim(T_D \cap S) \leq \dim S - 1$. Now

$$D \cap S \subseteq T_D \cap S \subseteq S,$$

and, because $\dim(D \cap S) = \dim S - 1$, it follows that $D \cap S = T_D \cap S$ consists of a subtorus S'' and finitely many of its translates. After dividing out by S'', we may assume that $\dim S = 1$ and that $D \cap S = T_D \cap S$ is a finite set; in particular, D is finite over Y, and therefore also projective. Now consider the addition morphism

$$S \times D \to T.$$

Since $S + D = X$, its image is equal to X; because S and D are both projective, it follows that X is projective, and hence that T is projective. But this contradicts our choice of S. The conclusion is that $S + D = D$, as asserted. □

Note It is possible for S to be itself an abelian variety; this is why the proof that $S + D \neq X$ requires some care.

9 Simple Hodge modules and abelian varieties

We begin by proving a structure theorem for *simple* polarizable complex Hodge modules on a compact complex torus T; this is evidently the most important case, because every polarizable complex Hodge module is isomorphic to a direct sum of simple ones. Fix a simple polarizable complex Hodge module $(M, J) \in \mathrm{HM}_{\mathbb{C}}(T, w)$. By Proposition 3.5, the polarizable real Hodge module $M \in \mathrm{HM}_{\mathbb{R}}(X, w)$ has strict support equal to an irreducible analytic subvariety; we assume in addition that $\operatorname{Supp} M$ is not contained in any proper subtorus of T.

Theorem 9.1 *There is an abelian variety A, a surjective morphism $q\colon T \to A$ with connected fibers, a simple $(N, K) \in \mathrm{HM}_{\mathbb{C}}(A, w - \dim q)$ with $\chi(A, N, K) > 0$, and a unitary character $\rho \in \mathrm{Char}(T)$, such that*

$$(M, J) \simeq q^{-1}(N, K) \otimes_{\mathbb{C}} \mathbb{C}_{\rho}. \tag{9.2}$$

In particular, $\mathrm{Supp}\, M = q^{-1}(\mathrm{Supp}\, N)$ *is covered by translates of* $\ker q$.

Let $X = \mathrm{Supp}\, M$. By Proposition 8.1, there is a subtorus $S \subseteq T$ such that $S + X = X$ and such that $Y = X/S$ is projective. Since Y is not contained in any proper subtorus, it follows that $A = T/S$ is an abelian variety. Let $q\colon T \to A$ be the quotient mapping, which is proper and smooth of relative dimension $\dim q = \dim S$. This will not be our final choice for Theorem 9.1, but it does have almost all the properties that we want (except for the lower bound on the Euler characteristic).

Proposition 9.3 *There is a simple $(N, K) \in \mathrm{HM}_{\mathbb{C}}(A, w - \dim q)$ with strict support Y and a unitary character $\rho \in \mathrm{Char}(T)$ for which* (9.2) *holds.*

By Theorem 4.4, (M, J) corresponds to a polarizable complex variation of Hodge structure of weight $w - \dim X$ on a dense Zariski-open subset of X. The crucial observation, due to Wang, is that we can choose this set to be of the form $q^{-1}(U)$, where U is a dense Zariski-open subset of the smooth locus of Y.

Lemma 9.4 *There is a dense Zariski-open subset $U \subseteq Y$, contained in the smooth locus of Y, and a polarizable complex variation of Hodge structure $(\mathcal{H}, J)$ of weight $w - \dim X$ on $q^{-1}(U)$ such that (M, J) is the polarizable complex Hodge module corresponding to $(\mathcal{H}, J)$ in Theorem 4.4.*

Proof Let $Z \subseteq X$ be the union of the singular locus of X and the singular locus of M. Then Z is an analytic subset of X, and according to Theorem 1.3, the restriction of M to $X \setminus Z$ is a polarizable real variation of Hodge structure $\mathcal{H}$ of weight $w - \dim X$. By Proposition 8.1, no irreducible component of Z of dimension $\dim X - 1$ dominates Y; we can therefore find a Zariski-open subset $U \subseteq Y$, contained in the smooth locus of Y, such that the intersection $q^{-1}(U) \cap Z$ has codimension ≥ 2 in $q^{-1}(U)$. Now $\mathcal{H}$ extends uniquely to a polarizable real variation of Hodge structure on the entire complex manifold $q^{-1}(U)$, see [32, Proposition 4.1]. The assertion about J follows easily. □

For any $y \in U$, the restriction of $(\mathcal{H}, J)$ to the fiber $q^{-1}(y)$ is a polarizable complex variation of Hodge structure on a translate of the compact complex torus $\ker q$. By Lemma 11.1, the restriction to $q^{-1}(y)$ of the underlying local system

$$\ker(J_{\mathbb{C}} - i \cdot \mathrm{id}\colon \mathcal{H}_{\mathbb{C}} \to \mathcal{H}_{\mathbb{C}})$$

is the direct sum of local systems of the form $\mathbb{C}_\rho$ for $\rho \in \mathrm{Char}(T)$ unitary; when M admits an integral structure, ρ has finite order in the group $\mathrm{Char}(T)$.

Proof of Proposition 9.3 Let $\rho \in \mathrm{Char}(T)$ be one of the unitary characters in question, and let $\overline{\rho} \in \mathrm{Char}(T)$ denote its complex conjugate. The tensor product $(\mathcal{H}, J) \otimes_{\mathbb{C}} \mathbb{C}_{\overline{\rho}}$ is a polarizable complex variation of Hodge structure of weight $w - \dim X$ on the open subset $q^{-1}(U)$. Since all fibers of $q\colon q^{-1}(U) \to U$ are translates of the compact complex torus $\ker q$, classical Hodge theory for compact Kähler manifolds [43, Theorem 2.9] implies that

$$q_*((\mathcal{H}, J) \otimes_{\mathbb{C}} \mathbb{C}_{\overline{\rho}}) \tag{9.5}$$

is a polarizable complex variation of Hodge structure of weight $w - \dim X$ on U; in particular, it is again semisimple. By our choice of ρ, the adjunction morphism

$$q^{-1}q_*((\mathcal{H}, J) \otimes_{\mathbb{C}} \mathbb{C}_{\overline{\rho}}) \to (\mathcal{H}, J) \otimes_{\mathbb{C}} \mathbb{C}_{\overline{\rho}}$$

is nontrivial. Consequently, (9.5) must have at least one simple summand $(\mathcal{H}_U, K)$ in the category of polarizable complex variations of Hodge structure of weight $w - \dim X$ for which the induced morphism $q^{-1}(\mathcal{H}_U, K) \to (\mathcal{H}, J) \otimes_{\mathbb{C}} \mathbb{C}_{\overline{\rho}}$ is nontrivial. Both sides being simple, the morphism is an isomorphism; consequently,

$$q^{-1}(\mathcal{H}_U, K) \otimes_{\mathbb{C}} \mathbb{C}_\rho \simeq (\mathcal{H}, J). \tag{9.6}$$

Now let $(N, K) \in \mathrm{HM}_{\mathbb{C}}(A, w - \dim q)$ be the polarizable complex Hodge module on A corresponding to $(\mathcal{H}_U, K)$; by construction, (N, K) is simple with strict support Y. Arguing as in [34, Lemma 20.2], one proves that the naive pullback $q^{-1}(N, K) \in \mathrm{HM}_{\mathbb{C}}(T, w)$ is simple with strict support X. By (9.6), this means that (M, J) is isomorphic to $q^{-1}(N, K) \otimes_{\mathbb{C}} \mathbb{C}_\rho$ in the category $\mathrm{HM}_{\mathbb{C}}(T, w)$. □

We have thus proved Theorem 9.1, except for the inequality $\chi(A, N, K) > 0$. Let $\mathcal{N}$ denote the regular holonomic $\mathscr{D}$–module underlying N; then

$$\mathcal{N} = \mathcal{N}' \oplus \mathcal{N}'' = \ker(K - i \cdot \mathrm{id}) \oplus \ker(K + i \cdot \mathrm{id}),$$

where $K \in \mathrm{End}(\mathcal{N})$ refers to the induced endomorphism. By Proposition 3.8, both $\mathcal{N}'$ and $\mathcal{N}''$ are simple with strict support Y. Since A is an abelian variety, one has, for example by [34, Section 5], that

$$\chi(A, N, K) = \sum_{i \in \mathbb{Z}} (-1)^i \dim H^i(A, \mathrm{DR}(\mathcal{N}')) \geq 0.$$

Now the point is that a simple holonomic $\mathscr{D}$–module with vanishing Euler characteristic is always (up to a twist by a line bundle with flat connection) the pullback from a lower-dimensional abelian variety [34, Section 20].

Proof of Theorem 9.1 Keeping the notation from Proposition 9.3, we have a surjective morphism $q\colon T \to A$ with connected fibers, a simple polarizable complex Hodge module $(N, K) \in \mathrm{HM}_{\mathbb{C}}(Y, w - \dim q)$ with strict support $Y = q(X)$, and a unitary character $\rho \in \mathrm{Char}(T)$ such that

$$(M, J) \simeq q^{-1}(N, K) \otimes_{\mathbb{C}} \mathbb{C}_\rho.$$

If (N, K) has positive Euler characteristic, we are done, so let us assume from now on that $\chi(A, N, K) = 0$. This means that $\mathcal{N}'$ is a simple regular holonomic $\mathscr{D}$–module with strict support Y and Euler characteristic zero.

By [34, Corollary 5.2], there is a surjective morphism $f\colon A \to B$ with connected fibers from A to a lower-dimensional abelian variety B, such that $\mathcal{N}'$ is (up to a twist by a line bundle with flat connection) the pullback of a simple regular holonomic $\mathscr{D}$–module with positive Euler characteristic. Setting

$$\mathcal{M} = \mathcal{M}' \oplus \mathcal{M}'' = \ker(J - i \cdot \mathrm{id}) \oplus \ker(J + i \cdot \mathrm{id}),$$

it follows that $\mathcal{M}'$ is (again up to a twist by a line bundle with flat connection) the pullback by $f \circ q$ of a simple regular holonomic $\mathscr{D}$–module on B. Consequently, there is a dense Zariski-open subset $U \subseteq f(Y)$ such that the restriction of $\mathcal{M}'$ to $(f \circ q)^{-1}(U)$ is coherent as an $\mathscr{O}$–module. By Lemma 4.3, the restriction of (M, J) to this open set is therefore a polarizable complex variation of Hodge structure of weight $w - \dim X$. After replacing our original morphism $q\colon T \to A$ by the composition $f \circ q\colon T \to B$, we can argue as in the proof of Proposition 9.3 to show that (9.2) is still satisfied (for a different choice of $\rho \in \mathrm{Char}(T)$, perhaps).

With some additional work, one can prove that now $\chi(A, N, K) > 0$. Alternatively, the same result can be obtained by the following more indirect method: as long as $\chi(A, N, K) = 0$, we can repeat the argument above; since the dimension of A goes down each time, we must eventually get to the point where $\chi(A, N, K) > 0$. This completes the proof of Theorem 9.1. □

10 Proof of the main result

As in Theorem 7.1, let $(M, J) \in \mathrm{HM}_{\mathbb{C}}(T, w)$ be a polarizable complex Hodge module on a compact complex torus T. Using the decomposition by strict support, we can assume without loss of generality that (M, J) has strict support equal to an irreducible analytic subvariety $X \subseteq T$. After translation, we may assume moreover that $0 \in X$. Let $T' \subseteq T$ be the smallest subtorus of T containing X; by Kashiwara's equivalence, we have $(M, J) = i_*(M', J')$ for some $(M', J') \in \mathrm{HM}_{\mathbb{C}}(T', w)$, where $i\colon T' \hookrightarrow T$ is the inclusion. Now Theorem 9.1 gives us a morphism $q'\colon T' \to A'$ such that (M', J') is isomorphic to the direct sum of pullbacks of polarizable complex Hodge modules twisted by unitary local systems. Since i^{-1}: $\mathrm{Char}(T) \to \mathrm{Char}(T')$ is surjective, the same is then true for (M, J) with respect to the quotient mapping $q\colon T \to T/\ker q'$. This proves Theorem 7.1.

Proof of Corollary 7.3 By considering the complexification

$$(M \oplus M, J_M) \in \mathrm{HM}_{\mathbb{C}}(T, w),$$

we reduce the problem to the situation of Theorem 7.1. It remains to show that all the characters in (7.2) have finite order in $\mathrm{Char}(T)$ if M admits an integral structure. By Lemma 5.2, every summand in the decomposition of M by strict support still admits an integral structure, and so we may assume without loss of generality that M has strict support equal to $X \subseteq T$ and that $0 \in X$. As before, we have $(M, J) = i_*(M', J')$, where $i\colon T' \hookrightarrow T$ is the smallest subtorus of T containing X; it is easy to see that M' again admits an integral structure. Now we apply the same argument as in the proof of Theorem 7.1 to the finitely many simple factors of (M, J), noting that the characters $\rho \in \mathrm{Char}(T)$ that come up always have finite order by Lemma 11.1 below. □

Note As in the proof of Lemma 6.1, it follows that $M \oplus M$ is isomorphic to the direct sum of the polarizable real Hodge modules

$$\ker(q_j^{-1} J_j \otimes J_{\rho_j} + \mathrm{id}) \subseteq q_j^{-1} N_j \otimes_{\mathbb{R}} \mathcal{H}_{\rho_j}. \tag{10.1}$$

Furthermore, one can show that, for each $j = 1, \dots, n$, exactly one of two things happens:

(1) Either the object in (10.1) is simple, and therefore occurs among the simple factors of M; in this case, the underlying regular holonomic $\mathscr{D}$–module $\mathcal{M}$ will contain the two simple factors

$$(q_j^* \mathcal{N}_j' \otimes_{\mathscr{O}_T} (L_j, \nabla_j)) \oplus (q_j^* \mathcal{N}_j'' \otimes_{\mathscr{O}_T} (L_j, \nabla_j)^{-1}).$$

(2) Or the object in (10.1) splits into two copies of a simple polarizable real Hodge module, which also has to occur among the simple factors of M. In this case, one can actually arrange that (N_j, J_j) is real and that the character ρ_j takes values in $\{-1, +1\}$. The simple object in question is the twist of (N_j, J_j) by the polarizable real variation of Hodge structure of rank one determined by ρ_j; moreover, $\mathcal{M}$ will contain $q_j^* \mathcal{N}_j' \otimes_{\mathscr{O}_T} (L_j, \nabla_j) \simeq q_j^* \mathcal{N}_j'' \otimes_{\mathscr{O}_T} (L_j, \nabla_j)^{-1}$ as a simple factor.

11 A lemma about variations of Hodge structure

The fundamental group of a compact complex torus is abelian, and so every polarizable complex variation of Hodge structure is a direct sum of unitary local systems of rank one; this is the content of the following elementary lemma [35, Lemma 1.8]:

Lemma 11.1 *Let $(\mathcal{H}, J)$ be a polarizable complex variation of Hodge structure on a compact complex torus T. Then the local system $\mathcal{H}_{\mathbb{C}} = \mathcal{H}_{\mathbb{R}} \otimes_{\mathbb{R}} \mathbb{C}$ is isomorphic to a direct sum of unitary local systems of rank one. If $\mathcal{H}$ admits an integral structure, then each of these local systems of rank one has finite order.*

Proof According to [11, Section 1.12], the underlying local system of a polarizable complex variation of Hodge structure on a compact Kähler manifold is semisimple; in the case of a compact complex torus, it is therefore a direct sum of rank-one local systems. The existence of a polarization implies that the individual local systems are unitary [11, Proposition 1.13]. Now suppose that $\mathcal{H}$ admits an integral structure, and let $\mu\colon \pi_1(A, 0) \to \mathrm{GL}_n(\mathbb{Z})$ be the monodromy representation. We already know that the complexification of μ is a direct sum of unitary characters. Since μ is defined over $\mathbb{Z}$, the values of each character are algebraic integers of absolute value one; by Kronecker's theorem, they must be roots of unity. □

12 Integral structure and points of finite order

One can combine the decomposition in Corollary 7.3 with known results about Hodge modules on abelian varieties [35] to prove the following generalization of Wang's theorem:

Corollary 12.1 *If $M \in \mathrm{HM}_{\mathbb{R}}(T, w)$ admits an integral structure, then the sets*

$$S^i_m(T, M) = \{\rho \in \mathrm{Char}(T) \mid \dim H^i(T, M_{\mathbb{R}} \otimes_{\mathbb{R}} \mathbb{C}_\rho) \geq m\}$$

are finite unions of translates of linear subvarieties by points of finite order.

Proof The result in question is known for abelian varieties: if $M \in \mathrm{HM}_{\mathbb{R}}(A, w)$ is a polarizable real Hodge module on an abelian variety, and if M admits an integral structure, then the sets $S^i_m(A, M)$ are finite unions of "arithmetic subvarieties" (namely translates of linear subvarieties by points of finite order). This is proved in [35, Theorem 1.4] for polarizable rational Hodge modules, but the proof carries over unchanged to the case of real coefficients. The same argument shows more generally that if the underlying perverse sheaf $M_{\mathbb{C}}$ of a polarizable real Hodge module $M \in \mathrm{HM}_{\mathbb{R}}(A, w)$ is isomorphic to a direct factor in the complexification of some $E \in \mathrm{D}^b_c(\mathbb{Z}_A)$, then each $S^i_m(A, M)$ is a finite union of arithmetic subvarieties.

Now let us see how to extend this result to compact complex tori. Passing to the underlying complex perverse sheaves in Corollary 7.3, we get

$$M_{\mathbb{C}} \simeq \bigoplus_{j=1}^{n} (q_j^{-1} N_{j,\mathbb{C}} \otimes_{\mathbb{C}} \mathbb{C}_{\rho_j});$$

recall that $\operatorname{Supp} N_j$ is a projective subvariety of the complex torus T_j, and that $\rho_j \in \operatorname{Char}(T)$ has finite order. In light of this decomposition and the comments above, it is therefore enough to prove that each $N_{j,\mathbb{C}}$ is isomorphic to a direct factor in the complexification of some object of $\mathrm{D}_c^b(\mathbb{Z}_{T_j})$.

Let $E \in \mathrm{D}_c^b(\mathbb{Z}_T)$ be some choice of integral structure on the real Hodge module M; obviously $M_{\mathbb{C}} \simeq E \otimes_{\mathbb{Z}} \mathbb{C}$. Let $r \geq 1$ be the order of the point $\rho_j \in \operatorname{Char}(T)$, and denote by $[r]\colon T \to T$ the finite morphism given by multiplication by r. We define

$$E' = \boldsymbol{R}[r]_*([r]^{-1} E) \in \mathrm{D}_c^b(\mathbb{Z}_T)$$

and observe that the complexification of E' is isomorphic to the direct sum of $E \otimes_{\mathbb{Z}} \mathbb{C}_\rho$, where $\rho \in \operatorname{Char}(T)$ runs over the finite set of characters whose order divides r. This set includes ρ_j^{-1}, and so $q_j^{-1} N_{j,\mathbb{C}}$ is isomorphic to a direct factor of $E' \otimes_{\mathbb{Z}} \mathbb{C}$. Because $q_j\colon T \to T_j$ has connected fibers, this implies that

$$N_{j,\mathbb{C}} \simeq \mathcal{H}^{-\dim q_j} q_{j*}(q_j^{-1} N_{j,\mathbb{C}})$$

is isomorphic to a direct factor of

$$\mathcal{H}^{-\dim q_j} q_{j*}(E' \otimes_{\mathbb{Z}} \mathbb{C}).$$

As explained in [35, Section 1.2.2], this is again the complexification of a constructible complex in $\mathrm{D}_c^b(\mathbb{Z}_{T_j})$, and so the proof is complete. □

D Generic vanishing theory

Let X be a compact Kähler manifold, and let $f\colon X \to T$ be a holomorphic mapping to a compact complex torus. The main purpose of this chapter is to show that the higher direct image sheaves $R^j f_* \omega_X$ have the same properties as in the projective case (such as being GV-sheaves). As explained in the introduction, we do not know how to obtain this using classical Hodge theory; this forces us to prove a more general result for arbitrary polarizable complex Hodge modules.

13 GV-sheaves and M–regular sheaves

We begin by reviewing a few basic definitions. Let T be a compact complex torus, $\hat{T} = \operatorname{Pic}^0(T)$ its dual, and P the normalized Poincaré bundle on the product $T \times \hat{T}$. It induces an integral transform

$$\boldsymbol{R}\Phi_P\colon \mathrm{D}_{\mathrm{coh}}^b(\mathscr{O}_T) \to \mathrm{D}_{\mathrm{coh}}^b(\mathscr{O}_{\hat{T}}), \quad \boldsymbol{R}\Phi_P(\mathscr{F}) = \boldsymbol{R} p_{2*}(p_1^* \mathscr{F} \otimes P),$$

where $\mathrm{D}_{\mathrm{coh}}^b(\mathscr{O}_T)$ is the derived category of cohomologically bounded and coherent complexes of $\mathscr{O}_T$–modules. Likewise, we have $\boldsymbol{R}\Psi_P\colon \mathrm{D}_{\mathrm{coh}}^b(\mathscr{O}_{\hat{T}}) \to \mathrm{D}_{\mathrm{coh}}^b(\mathscr{O}_T)$ going in the opposite direction. An argument analogous to Mukai's for abelian varieties shows that the Fourier–Mukai equivalence holds in this case as well [3, Theorem 2.1].

Theorem 13.1 *With the notations above, $\boldsymbol{R}\Phi_P$ and $\boldsymbol{R}\Psi_P$ are equivalences of derived categories. More precisely, one has*

$$\boldsymbol{R}\Psi_P \circ \boldsymbol{R}\Phi_P \simeq (-1)_T^*[-\dim T] \quad \textit{and} \quad \boldsymbol{R}\Phi_P \circ \boldsymbol{R}\Psi_P \simeq (-1)_{\hat{T}}^*[-\dim T].$$

Given a coherent $\mathscr{O}_T$–module $\mathscr{F}$ and an integer $m \geq 1$, we define

$$S_m^i(T, \mathscr{F}) = \{L \in \mathrm{Pic}^0(T) \mid \dim \ H^i(T, \mathscr{F} \otimes_{\mathscr{O}_T} L) \geq m\}.$$

It is customary to denote

$$S^i(T, \mathscr{F}) = S_1^i(T, \mathscr{F}) = \{L \in \mathrm{Pic}^0(T) \mid H^i(T, \mathscr{F} \otimes_{\mathscr{O}_T} L) \neq 0\}.$$

Recall the following definitions, from [23] and [21], respectively.

Definition 13.2 A coherent $\mathscr{O}_T$–module $\mathscr{F}$ is called a *GV-sheaf* if the inequality

$$\mathrm{codim}_{\mathrm{Pic}^0(T)} S^i(T, \mathscr{F}) \geq i$$

is satisfied for every integer $i \geq 0$. It is called *M–regular* if the inequality

$$\mathrm{codim}_{\mathrm{Pic}^0(T)} S^i(T, \mathscr{F}) \geq i + 1$$

is satisfied for every integer $i \geq 1$.

A number of local properties of integral transforms for complex manifolds, based only on commutative algebra results, were proved in [22; 25]. For instance, the following is a special case of [22, Theorem 2.2]:

Theorem 13.3 *Let $\mathscr{F}$ be a coherent sheaf on a compact complex torus T. Then the following statements are equivalent:*

(i) *$\mathscr{F}$ is a GV-sheaf.*

(ii) *$R^i\Phi_P(\boldsymbol{R}\Delta\mathscr{F}) = 0$ for $i \neq \dim T$, where $\boldsymbol{R}\Delta\mathscr{F} := \boldsymbol{R}\mathcal{H}om(\mathscr{F}, \mathscr{O}_T)$.*

Note that this statement was inspired by work of Hacon [15] in the projective setting. In the course of the proof of Theorem 13.3, and also for some of the results below, the following consequence of Grothendieck duality for compact complex manifolds is needed:

$$\boldsymbol{R}\Phi_P(\mathscr{F}) \simeq \boldsymbol{R}\Delta(\boldsymbol{R}\Phi_{P^{-1}}(\boldsymbol{R}\Delta\mathscr{F})[\dim T]); \tag{13.4}$$

see the proof of [22, Theorem 2.2], and especially the references there. In particular, if $\mathscr{F}$ is a GV-sheaf, then if we let $\widehat{\mathscr{F}} := R^{\dim T}\Phi_{P^{-1}}(\boldsymbol{R}\Delta\mathscr{F})$, Theorem 13.3 and (13.4) imply that

$$R\Phi_P(\mathscr{F}) \simeq R\mathcal{H}om(\widehat{\mathscr{F}}, \mathscr{O}_{\widehat{A}}). \tag{13.5}$$

As in [24, Proposition 2.8], $\mathscr{F}$ is M–regular if and only if $\widehat{\mathscr{F}}$ is torsion-free.

The fact that Theorems 13.1 and 13.3 and (13.5) hold for arbitrary compact complex tori allows us to deduce important properties of GV-sheaves in this setting. Besides these statements, the proofs only rely on local commutative algebra and base change, and so are completely analogous to those for abelian varieties; we will thus only indicate references for that case.

Proposition 13.6 *Let $\mathscr{F}$ be a GV-sheaf on T.*

(a) *One has $S^{\dim T}(T, \mathscr{F}) \subseteq \cdots \subseteq S^1(T, \mathscr{F}) \subseteq S^0(T, \mathscr{F}) \subseteq \hat{T}$.*

(b) *If $S^0(T, \mathscr{F})$ is empty, then $\mathscr{F} = 0$.*

(c) *If an irreducible component $Z \subseteq S^0(T, \mathscr{F})$ has codimension k in $\operatorname{Pic}^0(X)$, then $Z \subseteq S^k(T, \mathscr{F})$, and hence $\dim \operatorname{Supp} \mathscr{F} \geq k$.*

Proof For (a), see [23, Proposition 3.14]; for (b), see [20, Lemma 1.12]; for (c), see [20, Lemma 1.8]. □

14 Higher direct images of dualizing sheaves

Saito [29] and Takegoshi [38] have extended to Kähler manifolds many of the fundamental theorems on higher direct images of canonical bundles proved by Kollár for smooth projective varieties. The following theorem summarizes some of the results in [38, pages 390–391] in the special case that is needed for our purposes.

Theorem 14.1 (Takegoshi) *Let $f: X \to Y$ be a proper holomorphic mapping from a compact Kähler manifold to a reduced and irreducible analytic space, and let $L \in \operatorname{Pic}^0(X)$ be a holomorphic line bundle with trivial first Chern class.*

(a) *The Leray spectral sequence*

$$E_2^{p,q} = H^p(Y, R^q f_*(\omega_X \otimes L)) \Rightarrow H^{p+q}(X, \omega_X \otimes L)$$

degenerates at E_2.

(b) *If f is surjective, then $R^q f_*(\omega_X \otimes L)$ is torsion-free for every $q \geq 0$; in particular, it vanishes for $q > \dim X - \dim Y$.*

Saito [29] obtained the same results in much greater generality, using the theory of Hodge modules. In fact, his method also gives the splitting of the complex $R f_* \omega_X$

in the derived category, thus extending the main result of [17] to all compact Kähler manifolds.

Theorem 14.2 (Saito) *Keeping the assumptions of the previous theorem, one has*

$$\boldsymbol{R} f_* \omega_X \simeq \bigoplus_j (R^j f_* \omega_X)[-j]$$

in the derived category $\mathrm{D}^b_{\mathrm{coh}}(\mathscr{O}_Y)$.

Proof Given [29], the proof in [31] goes through under the assumption that X is a compact Kähler manifold. □

15 Euler characteristic and M–regularity

In this section, we relate the Euler characteristic of a simple polarizable complex Hodge module on a compact complex torus T to the M–regularity of the associated graded object.

Lemma 15.1 *Let* $(M, J) \in \mathrm{HM}_{\mathbb{C}}(T, w)$ *be a simple polarizable complex Hodge module on a compact complex torus. If* $\operatorname{Supp} M$ *is projective and* $\chi(T, M, J) > 0$, *then the coherent* $\mathscr{O}_T$*–module* $\mathrm{gr}_k^F \mathcal{M}'$ *is* M*–regular for every* $k \in \mathbb{Z}$.

Proof $\operatorname{Supp} M$ is projective, hence contained in a translate of an abelian subvariety $A \subseteq T$; because Lemma 1.4 holds for polarizable complex Hodge modules, we may therefore assume without loss of generality that $T = A$ is an abelian variety.

As usual, let $\mathcal{M} = \mathcal{M}' \oplus \mathcal{M}'' = \ker(J - i \cdot \mathrm{id}) \oplus \ker(J + i \cdot \mathrm{id})$ be the decomposition into eigenspaces. The summand $\mathcal{M}'$ is a simple holonomic $\mathscr{D}$–module with positive Euler characteristic on an abelian variety, and so [34, Theorem 2.2 and Corollary 20.5] show that

$$\{\rho \in \mathrm{Char}(A) \mid H^i(A, \mathrm{DR}(\mathcal{M}') \otimes_{\mathbb{C}} \mathbb{C}_\rho) \neq 0\} \tag{15.2}$$

is equal to $\mathrm{Char}(A)$ when $i = 0$, and is equal to a finite union of translates of linear subvarieties of codimension $\geq 2i + 2$ when $i \geq 1$.

We have a one-to-one correspondence between $\mathrm{Pic}^0(A)$ and the subgroup of unitary characters in $\mathrm{Char}(A)$; it takes a unitary character $\rho \in \mathrm{Char}(A)$ to the holomorphic line bundle $L_\rho = \mathbb{C}_\rho \otimes_{\mathbb{C}} \mathscr{O}_A$. If $\rho \in \mathrm{Char}(A)$ is unitary, the twist $(M, J) \otimes_{\mathbb{C}} \mathbb{C}_\rho$ is still a polarizable complex Hodge module by Lemma 6.1, and so the complex computing its hypercohomology is strict. It follows that

$$H^i(A, \mathrm{gr}_k^F \mathrm{DR}(\mathcal{M}') \otimes_{\mathscr{O}_A} L_\rho) \quad \text{is a subquotient of} \quad H^i(A, \mathrm{DR}(\mathcal{M}') \otimes_{\mathbb{C}} \mathbb{C}_\rho).$$

If we identify $\mathrm{Pic}^0(A)$ with the subgroup of unitary characters, this means that

$$\{L \in \mathrm{Pic}^0(A) \mid H^i(A, \mathrm{gr}_k^F \mathrm{DR}(\mathcal{M}') \otimes_{\mathscr{O}_A} L) \neq 0\}$$

is contained in the intersection of (15.2) and the subgroup of unitary characters. When $i \geq 1$, this intersection is a finite union of translates of subtori of codimension $\geq i+1$; it follows that

$$\mathrm{codim}_{\mathrm{Pic}^0(A)}\{L \in \mathrm{Pic}^0(A) \mid H^i(A, \mathrm{gr}_k^F \mathrm{DR}(\mathcal{M}') \otimes_{\mathscr{O}_A} L) \neq 0\} \geq i+1.$$

Since the cotangent bundle of A is trivial, a simple induction on k as in the proof of [26, Lemma 1] gives

$$\mathrm{codim}_{\mathrm{Pic}^0(A)}\{L \in \mathrm{Pic}^0(A) \mid H^i(A, \mathrm{gr}_k^F \mathcal{M}' \otimes_{\mathscr{O}_A} L) \neq 0\} \geq i+1,$$

and so each $\mathrm{gr}_k^F \mathcal{M}'$ is indeed M–regular. □

Note In fact, the result still holds without the assumption that $\mathrm{Supp}\, M$ is projective; this is an easy consequence of the decomposition in (7.2).

16 Chen–Jiang decomposition and generic vanishing

Using the decomposition in Theorem 7.1 and the result of the previous section, we can now prove the most general version of the generic vanishing theorem, namely Theorem D in the introduction.

Proof of Theorem D We apply Theorem 7.1 to the complexification $(M \oplus M, J_M)$ in $\mathrm{HM}_{\mathbb{C}}(T, w)$. Passing to the associated graded in (7.2), we obtain a decomposition of the desired type with $\mathscr{F}_j = \mathrm{gr}_k^F \mathcal{N}_j'$ and $L_j = \mathbb{C}_{\rho_j} \otimes_{\mathbb{C}} \mathscr{O}_T$, where

$$\mathcal{N}_j = \mathcal{N}_j' \oplus \mathcal{N}_j'' = \ker(J_j - i \cdot \mathrm{id}) \oplus \ker(J_j + i \cdot \mathrm{id})$$

is as usual the decomposition into eigenspaces of $J_j \in \mathrm{End}(\mathcal{N}_j)$. Since $\mathrm{Supp}\, N_j$ is projective and $\chi(T_j, N_j, J_j) > 0$, we conclude from Lemma 15.1 that each coherent $\mathscr{O}_{T_j}$–module $\mathscr{F}_j$ is M–regular. □

Corollary 16.1 *If $M = (\mathcal{M}, F_\bullet\mathcal{M}, M_{\mathbb{R}}) \in \mathrm{HM}_{\mathbb{R}}(T, w)$, then for every $k \in \mathbb{Z}$ the coherent $\mathscr{O}_T$–module $\mathrm{gr}_k^F \mathcal{M}$ is a GV-sheaf.*

Proof This follows immediately from Theorem D and the fact that, if $p\colon T \to T_0$ is a surjective homomorphism of complex tori and $\mathscr{G}$ is a GV-sheaf on T_0, then $\mathscr{F} = f^*\mathscr{G}$ is a GV-sheaf on T. For this last statement and more refined facts (for instance when $\mathscr{G}$ is M–regular), see eg [9, Section 2], especially Proposition 2.6. The arguments in [9] are for abelian varieties, but given the remarks in Section 13, they work equally well on compact complex tori. □

By specializing to the direct image of the canonical Hodge module $\mathbb{R}_X[\dim X]$ along a morphism $f\colon X \to T$, we are finally able to conclude that each $R^j f_* \omega_X$ is a GV-sheaf. In fact, we have the more refined Theorem A; it was first proved for smooth projective varieties of maximal Albanese dimension by Chen and Jiang [9, Theorem 1.2], which was a source of inspiration for us.

Proof of Theorem A Denote by $\mathbb{R}_X[\dim X] \in \mathrm{HM}_{\mathbb{R}}(X, \dim X)$ the polarizable real Hodge module corresponding to the constant real variation of Hodge structure of rank one and weight zero on X. According to [29, Theorem 3.1], each $\mathcal{H}^j f_* \mathbb{R}_X[\dim X]$ is a polarizable real Hodge module of weight $\dim X + j$ on T; it also admits an integral structure [35, Section 1.2.2]. In the decomposition by strict support, let M be the summand with strict support $f(X)$; note that M still admits an integral structure by Lemma 5.2. Now $R^j f_* \omega_X$ is the first nontrivial piece of the Hodge filtration on the underlying regular holonomic $\mathscr{D}$–module [31], and so the result follows directly from Theorem D and Corollary 16.1. For the ampleness, see Corollary 20.1. □

Note Except for the assertion about finite order, Theorem A still holds for arbitrary coherent $\mathscr{O}_T$–modules of the form

$$R^j f_*(\omega_X \otimes L)$$

with $L \in \mathrm{Pic}^0(X)$. The point is that every such L is the holomorphic line bundle associated with a unitary character $\rho \in \mathrm{Char}(X)$; we can therefore apply the same argument as above to the polarizable complex Hodge module $\mathbb{C}_\rho[\dim X]$.

If the given morphism is generically finite over its image, we can say more:

Corollary 16.2 *If $f\colon X \to T$ is generically finite over its image, then $S^0(T, f_*\omega_X)$ is preserved by the involution $L \mapsto L^{-1}$ of $\mathrm{Pic}^0(T)$.*

Proof As before, we define $M = \mathcal{H}^0 f_* \mathbb{R}_X[\dim X] \in \mathrm{HM}_{\mathbb{R}}(T, \dim X)$. Recall from Corollary 7.3 that we have a decomposition

$$(M \oplus M, J_M) \simeq \bigoplus_{j=1}^{n} (q_j^{-1}(N_j, J_j) \otimes_{\mathbb{C}} \mathbb{C}_{\rho_j}).$$

Since f is generically finite over its image, there is a dense Zariski-open subset of $f(X)$ where M is a variation of Hodge structure of type $(0,0)$; the above decomposition shows that the same is true for N_j on $(q_j \circ f)(X)$. If we pass to the underlying regular holonomic $\mathscr{D}$–modules and remember Lemma 6.1, we see that

$$\mathcal{M} \oplus \mathcal{M} \simeq \bigoplus_{j=1}^{n} (q_j^* \mathcal{N}_j' \otimes_{\mathscr{O}_T} (L_j, \nabla_j)) \oplus \bigoplus_{j=1}^{n} (q_j^* \mathcal{N}_j'' \otimes_{\mathscr{O}_T} (L_j, \nabla_j)^{-1}),$$

where (L_j, ∇_j) is the flat bundle corresponding to the character ρ_j. By looking at the first nontrivial step in the Hodge filtration on $\mathcal{M}$, we then get

$$f_*\omega_X \oplus f_*\omega_X \simeq \bigoplus_{j=1}^{n} (q_j^* \mathscr{F}_j' \otimes_{\mathscr{O}_T} L_j) \oplus \bigoplus_{j=1}^{n} (q_j^* \mathscr{F}_j'' \otimes_{\mathscr{O}_T} L_j^{-1}),$$

where $\mathscr{F}_j' = F_{p(M)}\mathcal{N}_j'$ and $\mathscr{F}_j'' = F_{p(M)}\mathcal{N}_j''$, and $p(M)$ is the smallest integer with the property that $F_p\mathcal{M} \neq 0$. Both sheaves are torsion-free on $(q_j \circ f)(X)$, and can therefore be nonzero only when $\operatorname{Supp} N_j = (q_j \circ f)(X)$; after reindexing, we may assume that this holds exactly in the range $1 \leq j \leq m$.

Now we reach the crucial point of the argument: the fact that N_j is generically a polarizable real variation of Hodge structure of type $(0,0)$ implies that $\mathscr{F}_j'$ and $\mathscr{F}_j''$ have the same rank at the generic point of $(q_j \circ f)(X)$. Indeed, on a dense Zariski-open subset of $(q_j \circ f)(X)$, we have $\mathscr{F}_j' = \mathcal{N}_j'$ and $\mathscr{F}_j'' = \mathcal{N}_j''$, and complex conjugation with respect to the real structure on N_j interchanges the two factors.

Since $\mathscr{F}_j'$ and $\mathscr{F}_j''$ are M–regular by Lemma 15.1, we have (for $1 \leq j \leq m$)

$$S^0(T, q_j^* \mathscr{F}_j' \otimes_{\mathscr{O}_T} L_j) = L_j^{-1} \otimes S^0(T_j, \mathscr{F}_j') = L_j^{-1} \otimes \operatorname{Pic}^0(T_j),$$

and similarly for $q_j^* \mathscr{F}_j'' \otimes_{\mathscr{O}_T} L_j^{-1}$; to simplify the notation, we identify $\operatorname{Pic}^0(T_j)$ with its image in $\operatorname{Pic}^0(T)$. The decomposition from above now gives

$$S^0(T, f_*\omega_X) = \bigcup_{j=1}^{m} (L_j^{-1} \otimes \operatorname{Pic}^0(T_j)) \cup \bigcup_{j=1}^{m} (L_j \otimes \operatorname{Pic}^0(T_j)),$$

and the right-hand side is clearly preserved by the involution $L \mapsto L^{-1}$. □

17 Points of finite order on cohomology support loci

Let $f \colon X \to T$ be a holomorphic mapping from a compact Kähler manifold to a compact complex torus. Our goal in this section is to prove that the cohomology support loci of the coherent $\mathscr{O}_T$–modules $R^j f_*\omega_X$ are finite unions of translates of subtori by points of finite order. We consider the refined cohomology support loci

$$S_m^i(T, R^j f_*\omega_X) = \{L \in \operatorname{Pic}^0(T) \mid \dim H^i(T, R^j f_*\omega_X \otimes L) \geq m\} \subseteq \operatorname{Pic}^0(T).$$

The following result is well-known in the projective case:

Corollary 17.1 *Every irreducible component of $S_m^i(T, R^j f_*\omega_X)$ is a translate of a subtorus of $\operatorname{Pic}^0(T)$ by a point of finite order.*

Proof As in the proof of Theorem A (in Section 16), we let $M \in \mathrm{HM}_{\mathbb{R}}(T, \dim X + j)$ be the summand with strict support $f(X)$ in the decomposition by strict support of $\mathcal{H}^j f_* \mathbb{R}_X[\dim X]$; then M admits an integral structure, and

$$R^j f_* \omega_X \simeq F_{p(M)} \mathcal{M},$$

where $p(M)$ again means the smallest integer such that $F_p \mathcal{M} \neq 0$. Since M still admits an integral structure by Lemma 5.2, the result in Corollary 12.1 shows that the sets

$$S^i_m(T, M) = \{\rho \in \mathrm{Char}(T) \mid \dim H^i(T, M_{\mathbb{R}} \otimes_{\mathbb{R}} \mathbb{C}_\rho) \geq m\}$$

are finite unions of translates of linear subvarieties by points of finite order. As in the proof of Lemma 15.1, the strictness of the complex computing the hypercohomology of $(M \oplus M, J_M) \otimes_{\mathbb{C}} \mathbb{C}_\rho$ implies that

$$\dim H^i(T, M_{\mathbb{R}} \otimes_{\mathbb{R}} \mathbb{C}_\rho) = \sum_{p \in \mathbb{Z}} \dim H^i(T, \mathrm{gr}^F_p \mathrm{DR}(\mathcal{M}) \otimes_{\mathscr{O}_T} L_\rho)$$

for every unitary character $\rho \in \mathrm{Char}(T)$; here $L_\rho = \mathbb{C}_\rho \otimes_{\mathbb{C}} \mathscr{O}_T$. Note that $\mathrm{gr}^F_p \mathrm{DR}(\mathcal{M})$ is acyclic for $p \gg 0$, and so the sum on the right-hand side is actually finite. Intersecting $S^i_m(T, M)$ with the subgroup of unitary characters, we see that each set

$$\left\{ L \in \mathrm{Pic}^0(T) \,\middle|\, \sum_{p \in \mathbb{Z}} \dim H^i(T, \mathrm{gr}^F_p \mathrm{DR}(\mathcal{M}) \otimes_{\mathscr{O}_T} L) \geq m \right\}$$

is a finite union of translates of subtori by points of finite order. By a standard argument [1, page 312], it follows that the same is true for each of the summands; in other words, for each $p \in \mathbb{Z}$, the set

$$S^i_m(T, \mathrm{gr}^F_p \mathrm{DR}(\mathcal{M})) \subseteq \mathrm{Pic}^0(T)$$

is itself a finite union of translates of subtori by points of finite order. Since

$$\mathrm{gr}^F_{p(M)} \mathrm{DR}(\mathcal{M}) = \omega_T \otimes F_{p(M)} \mathcal{M} \simeq R^j f_* \omega_X,$$

we now obtain the assertion by specializing to $p = p(M)$. □

Note Alternatively, one can deduce Corollary 17.1 from Wang's theorem [41] about cohomology jump loci on compact Kähler manifolds, as follows. Wang shows that the sets $S^{p,q}_m(X) = \{L \in \mathrm{Pic}^0(X) \mid \dim H^q(X, \Omega^p_X \otimes L) \geq m\}$ are finite unions of translates of subtori by points of finite order; in particular, this is true for $\omega_X = \Omega^{\dim X}_X$. Takegoshi's results about higher direct images of ω_X in Theorem 14.1 imply the E_2–degeneration of the spectral sequence

$$E^{i,j}_2 = H^i(T, R^j f_* \omega_X \otimes L) \Rightarrow H^{i+j}(X, \omega_X \otimes f^* L)$$

for every $L \in \operatorname{Pic}^0(T)$, which means that

$$\dim H^q(X, \omega_X \otimes f^* L) = \sum_{k+j=q} \dim H^k(T, R^j f_* \omega_X \otimes L).$$

The assertion now follows from Wang's theorem by the same argument as above.

E Applications

18 Bimeromorphic characterization of tori

Our main application of generic vanishing for higher direct images of dualizing sheaves is an extension of the Chen–Hacon birational characterization of abelian varieties [6] to the Kähler case.

Theorem 18.1 *Let X be a compact Kähler manifold with $P_1(X) = P_2(X) = 1$ and $h^{1,0}(X) = \dim X$. Then X is bimeromorphic to a compact complex torus.*

Throughout this section, we take X to be a compact Kähler manifold, and denote by $f\colon X \to T$ its Albanese mapping; by assumption, we have

$$\dim T = h^{1,0}(X) = \dim X.$$

We use the following standard notation, analogous to that in Section 13:

$$S^i(X, \omega_X) = \{L \in \operatorname{Pic}^0(X) \mid H^i(X, \omega_X \otimes L) \neq 0\}$$

To simplify things, we shall identify $\operatorname{Pic}^0(X)$ and $\operatorname{Pic}^0(T)$ in what follows. We begin by recalling a few well-known results.

Lemma 18.2 *If $P_1(X) = P_2(X) = 1$, there cannot be any positive-dimensional analytic subvariety $Z \subseteq \operatorname{Pic}^0(X)$ such that both Z and Z^{-1} are contained in $S^0(X, \omega_X)$. In particular, the origin must be an isolated point in $S^0(X, \omega_X)$.*

Proof This result is due to Ein and Lazarsfeld [12, Proposition 2.1]; they state it only in the projective case, but their proof actually works without any changes on arbitrary compact Kähler manifolds. □

Lemma 18.3 *Assume that $S^0(X, \omega_X)$ contains isolated points. Then the Albanese map of X is surjective.*

Proof By Theorem A (for $j = 0$), $f_* \omega_X$ is a GV-sheaf. Proposition 13.6 shows that any isolated point in $S^0(T, f_* \omega_X) = S^0(X, \omega_X)$ also belongs to $S^{\dim T}(T, f_* \omega_X)$; but this is only possible if the support of $f_* \omega_X$ has dimension at least $\dim T$. □

To prove Theorem 18.1, we follow the general strategy introduced in [20, Section 4], which in turn is inspired by [12; 7]. The crucial new ingredient is of course Theorem A, which had only been known in the projective case. Even in the projective case however, the argument below is substantially cleaner than the existing proofs; this is due to Corollary 16.2.

Proof of Theorem 18.1 The Albanese map $f\colon X \to T$ is surjective by Lemmas 18.2 and 18.3; since $h^{1,0}(X) = \dim X$, this means that f is generically finite. To conclude the proof, we just have to argue that f has degree one; more precisely, we shall use Theorem A to show that $f_*\omega_X \simeq \mathscr{O}_T$.

As a first step in this direction, let us prove that $\dim S^0(T, f_*\omega_X) = 0$. If

$$S^0(T, f_*\omega_X) = S^0(X, \omega_X)$$

had an irreducible component Z of positive dimension, Corollary 16.2 would imply that Z^{-1} is contained in $S^0(X, \omega_X)$ as well. As this would contradict Lemma 18.2, we conclude that $S^0(T, f_*\omega_X)$ is zero-dimensional.

Now $f_*\omega_X$ is a GV-sheaf by Theorem A, and so Proposition 13.6 shows that

$$S^0(T, f_*\omega_X) = S^{\dim T}(T, f_*\omega_X).$$

Since f is generically finite, Theorem 14.1 implies that $R^j f_*\omega_X = 0$ for $j > 0$, which gives

$$S^{\dim T}(T, f_*\omega_X) = S^{\dim T}(X, \omega_X) = S^{\dim X}(X, \omega_X) = \{\mathscr{O}_T\}.$$

Putting everything together, we see that $S^0(T, f_*\omega_X) = \{\mathscr{O}_T\}$.

We can now use the Chen–Jiang decomposition for $f_*\omega_X$ to get more information. The decomposition in Theorem A (for $j = 0$) implies that

$$\{\mathscr{O}_T\} = S^0(T, f_*\omega_X) = \bigcup_{k=1}^{n} L_k^{-1} \otimes \operatorname{Pic}^0(T_k),$$

where we identify $\operatorname{Pic}^0(T_k)$ with its image in $\operatorname{Pic}^0(T)$. This equality forces $f_*\omega_X$ to be a trivial bundle of rank n; but then

$$n = \dim H^{\dim T}(T, f_*\omega_X) = \dim H^{\dim X}(X, \omega_X) = 1,$$

and so $f_*\omega_X \simeq \mathscr{O}_T$. The conclusion is that f is generically finite of degree one, and hence birational, as asserted by the theorem. □

19 Connectedness of the fibers of the Albanese map

As another application, one obtains the following analogue of an effective version of Kawamata's theorem on the connectedness of the fibers of the Albanese map, proved by Jiang [16, Theorem 3.1] in the projective setting. Note that the statement is more general than Theorem 18.1, but uses it in its proof.

Theorem 19.1 *Let X be a compact Kähler manifold with $P_1(X) = P_2(X) = 1$. Then the Albanese map of X is surjective, with connected fibers.*

Proof The proof goes entirely along the lines of [16]. We only indicate the necessary modifications in the Kähler case. We have already seen that the Albanese map $f\colon X \to T$ is surjective. Consider its Stein factorization:

$$\begin{tikzcd} X \arrow[d, "g"] \arrow[dr, "f"] & \\ Y \arrow[r, "h"] & T \end{tikzcd}$$

Up to passing to a resolution of singularities and allowing h to be generically finite, we can assume that Y is a compact complex manifold. Moreover, by [40, Théorème 3], after performing a further bimeromorphic modification, we can assume that Y is in fact compact Kähler. This does not change the hypothesis $P_1(X) = P_2(X) = 1$.

The goal is to show that Y is bimeromorphic to a torus, which is enough to conclude. If one could prove that $P_1(Y) = P_2(Y) = 1$, then Theorem 18.1 would do the job. In fact, one can show precisely as in [16, Theorem 3.1] that $H^0(X, \omega_{X/Y}) \neq 0$, and consequently that

$$P_m(Y) \le P_m(X) \quad \text{for all } m \ge 1.$$

The proof of this statement needs the degeneration of the Leray spectral sequence for $g_*\omega_X$, which follows from Theorem 14.1, and the fact that $f_*\omega_X$ is a GV-sheaf, which follows from Theorem A. Besides this, the proof is purely Hodge-theoretic, and hence works equally well in the Kähler case. □

20 Semipositivity of higher direct images

In the projective case, GV-sheaves automatically come with positivity properties; more precisely, on abelian varieties it was proved in [10, Corollary 3.2] that M–regular sheaves are ample, and in [24, Theorem 4.1] that GV-sheaves are nef. Due to Theorem D a stronger result in fact holds true for arbitrary graded quotients of Hodge modules on compact complex tori.

Recall that to a coherent sheaf $\mathscr{F}$ on a compact complex manifold one can associate the analytic space $\mathbb{P}(\mathscr{F}) = \mathbb{P}(\mathrm{Sym}^\bullet \mathscr{F})$, with a natural mapping to X and a line bundle $\mathscr{O}_{\mathbb{P}(\mathscr{F})}(1)$. If X is projective, the sheaf $\mathscr{F}$ is called *ample* if the line bundle $\mathscr{O}_{\mathbb{P}(\mathscr{F})}(1)$ is ample on $\mathbb{P}(\mathscr{F})$.

Corollary 20.1 *Let $M = (\mathcal{M}, F_\bullet\mathcal{M}, M_{\mathbb{R}})$ be a polarizable real Hodge module on a compact complex torus T. Then, for each $k \in \mathbb{Z}$, the coherent $\mathscr{O}_T$–module $\mathrm{gr}_k^F \mathcal{M}$ admits a decomposition*

$$\mathrm{gr}_k^F \mathcal{M} \simeq \bigoplus_{j=1}^{n} (q_j^* \mathscr{F}_j \otimes_{\mathscr{O}_T} L_j),$$

where $q_j\colon T \to T_j$ is a quotient torus, $\mathscr{F}_j$ is an ample coherent $\mathscr{O}_{T_j}$–module whose support $\operatorname{Supp} \mathscr{F}_j$ is projective, and $L_j \in \operatorname{Pic}^0(T)$.

Proof By Theorem D we have a decomposition as in the statement, where each $\mathscr{F}_j$ is an M–regular sheaf on the abelian variety generated by its support. But then [10, Corollary 3.2] implies that each $\mathscr{F}_j$ is ample. □

The ampleness part in Theorem A is then a consequence of the proof in Section 16 and the statement above. It implies that higher direct images of canonical bundles have a strong semipositivity property (corresponding to semiampleness in the projective setting). Even the following very special consequence seems to go beyond what can be said for arbitrary holomorphic mappings of compact Kähler manifolds (see eg [19]).

Corollary 20.2 *Let $f\colon X \to T$ be a surjective holomorphic mapping from a compact Kähler manifold to a complex torus. If f is a submersion outside of a simple normal crossings divisor on T, then each $R^i f_* \omega_X$ is locally free and admits a smooth hermitian metric with semipositive curvature (in the sense of Griffiths).*

Proof Note that if f is surjective, then Theorem 14.1 implies that $R^i f_* \omega_X$ are all torsion-free. If one assumes in addition that f is a submersion outside of a simple normal crossings divisor on T, then they are locally free; see [38, Theorem V]. Because of the decomposition in Theorem A, it is therefore enough to show that an M–regular locally free sheaf on an abelian variety always admits a smooth hermitian metric with semipositive curvature. But this is an immediate consequence of the fact that M–regular sheaves are continuously globally generated [22, Proposition 2.19]. □

The existence of a metric with semipositive curvature on a vector bundle E implies that the line bundle $\mathscr{O}_{\mathbb{P}(E)}(1)$ is nef, but is in general known to be a strictly stronger condition. Corollary 20.2 suggests the following question:

Problem Let T be a compact complex torus. Suppose that a locally free sheaf $\mathscr{E}$ on T admits a smooth hermitian metric with semipositive curvature (in the sense of Griffiths or Nakano). Does this imply the existence of a decomposition

$$\mathscr{E} \simeq \bigoplus_{k=1}^{n} (q_k^* \mathscr{E}_k \otimes L_k)$$

as in Theorem A, in which each locally free sheaf $\mathscr{E}_k$ has a smooth hermitian metric with strictly positive curvature?

21 Leray filtration

Let $f\colon X \to T$ be a holomorphic mapping from a compact Kähler manifold X to a compact complex torus T. We use Theorem A to describe the Leray filtration on the cohomology of ω_X, induced by the Leray spectral sequence associated to f. Recall that, for each k, the Leray filtration on $H^k(X, \omega_X)$ is a decreasing filtration $L^\bullet H^k(X, \omega_X)$ with the property that

$$\operatorname{gr}_L^i H^k(X, \omega_X) = H^i(T, R^{k-i} f_* \omega_X).$$

On the other hand, one can define a natural decreasing filtration $F^\bullet H^k(X, \omega_X)$ induced by the action of $H^1(T, \mathscr{O}_T)$, namely

$$F^i H^k(X, \omega_X) = \operatorname{Im}\bigl(\textstyle\bigwedge^i H^1(T, \mathscr{O}_T) \otimes H^{k-i}(X, \omega_X) \to H^k(X, \omega_X)\bigr).$$

It is obvious that the image of the cup product mapping

$$H^1(T, \mathscr{O}_T) \otimes L^i H^k(X, \omega_X) \to H^{k+1}(X, \omega_X) \tag{21.1}$$

is contained in the subspace $L^{i+1} H^{k+1}(X, \omega_X)$. This implies that

$$F^i H^k(X, \omega_X) \subseteq L^i H^k(X, \omega_X) \quad \text{for all } i \in \mathbb{Z}.$$

This inclusion is actually an equality, as shown by the following result:

Theorem 21.2 *The image of the mapping in* (21.1) *is equal to* $L^{i+1} H^{k+1}(X, \omega_X)$. *Consequently, the two filtrations* $L^\bullet H^k(X, \omega_X)$ *and* $F^\bullet H^k(X, \omega_X)$ *coincide.*

Proof By [18, Theorem A], the graded module

$$Q_X^j = \bigoplus_{i=0}^{\dim T} H^i(T, R^j f_* \omega_X)$$

over the exterior algebra on $H^1(T, \mathscr{O}_T)$ is 0–regular, hence generated in degree 0. (Since each $R^j f_* \omega_X$ is a GV-sheaf by Theorem A, the proof in [18] carries over to the case where X is a compact Kähler manifold.) This means that the cup product mappings

$$\textstyle\bigwedge^i H^1(T, \mathscr{O}_T) \otimes H^0(T, R^j f_* \omega_X) \to H^i(T, R^j f_* \omega_X)$$

are surjective for all i and j, which in turn implies that the mappings

$$H^1(T, \mathscr{O}_T) \otimes \operatorname{gr}_L^i H^k(X, \omega_X) \to \operatorname{gr}_L^{i+1} H^{k+1}(X, \omega_X)$$

are surjective for all i and k. This implies the assertion by ascending induction. □

If we represent cohomology classes by smooth forms, Hodge conjugation and Serre duality provide for each $k \geq 0$ a hermitian pairing

$$H^0(X, \Omega_X^{n-k}) \times H^k(X, \omega_X) \to \mathbb{C}, \quad (\alpha, \beta) \mapsto \int_X \alpha \wedge \overline{\beta},$$

where $n = \dim X$. The Leray filtration on $H^k(X, \omega_X)$ therefore induces a filtration on $H^0(X, \Omega_X^{n-k})$; concretely, with a numerical convention which again gives us a decreasing filtration with support in the range $0, \dots, k$, we have

$$L^i H^0(X, \Omega_X^{n-k}) = \{\alpha \in H^0(X, \Omega_X^{n-k}) \mid \alpha \perp L^{k+1-i} H^k(X, \omega_X)\}.$$

Using the description of the Leray filtration in Theorem 21.2, and the elementary fact that

$$\int_X \alpha \wedge \overline{\theta \wedge \beta} = \int_X \alpha \wedge \overline{\theta} \wedge \overline{\beta}$$

for all $\theta \in H^1(X, \mathscr{O}_X)$, we can easily deduce that $L^i H^0(X, \Omega_X^{n-k})$ consists of those holomorphic $(n-k)$–forms whose wedge product with

$$\textstyle\bigwedge^{k+1-i} H^0(X, \Omega_X^1)$$

vanishes. In other words, for all j we have:

Corollary 21.3 *The induced Leray filtration on $H^0(X, \Omega_X^j)$ is given by*

$$L^i H^0(X, \Omega_X^j) = \Big\{\alpha \in H^0(X, \Omega_X^j) \;\Big|\; \alpha \wedge \textstyle\bigwedge^{n+1-i-j} H^0(X, \Omega_X^1) = 0\Big\}.$$

Remark It is precisely the fact that we do not know how to obtain this basic description of the Leray filtration using standard Hodge theory that prevents us from giving a proof of Theorem A in the spirit of [13], and forces us to appeal to the theory of Hodge modules for the main results.

Acknowledgements We thank Christopher Hacon, who first asked us about the behavior of higher direct images in the Kähler setting some time ago, and with whom we have had numerous fruitful discussions about generic vanishing over the years. We also thank Claude Sabbah for advice about the definition of polarizable complex Hodge modules, and Jungkai Chen, János Kollár, Thomas Peternell and Valentino Tosatti for useful discussions.

During the preparation of this paper Schnell has been partially supported by the NSF grant DMS-1404947, Popa by the NSF grant DMS-1405516 and Pareschi by the MIUR PRIN project *Geometry of algebraic varieties*.

References

[1] **D Arapura**, *Higgs line bundles, Green–Lazarsfeld sets, and maps of Kähler manifolds to curves*, Bull. Amer. Math. Soc. 26 (1992) 310–314 MR

[2] **A A Beĭlinson**, **J Bernstein**, **P Deligne**, *Faisceaux pervers*, from "Analyse et topologie sur les espaces singuliers, I", Astérisque 100, Soc. Math. France, Paris (1982) 5–171 MR

[3] **O Ben-Bassat**, **J Block**, **T Pantev**, *Non-commutative tori and Fourier–Mukai duality*, Compos. Math. 143 (2007) 423–475 MR

[4] **C Birkenhake**, **H Lange**, *Complex tori*, Progress in Mathematics 177, Birkhäuser, Boston, MA (1999) MR

[5] **E Cattani**, **A Kaplan**, **W Schmid**, *Degeneration of Hodge structures*, Ann. of Math. 123 (1986) 457–535 MR

[6] **J A Chen**, **C D Hacon**, *Characterization of abelian varieties*, Invent. Math. 143 (2001) 435–447 MR

[7] **J A Chen**, **C D Hacon**, *On algebraic fiber spaces over varieties of maximal Albanese dimension*, Duke Math. J. 111 (2002) 159–175 MR

[8] **J A Chen**, **C D Hacon**, *Varieties with $P_3(X) = 4$ and $q(X) = \dim(X)$*, Ann. Sc. Norm. Super. Pisa Cl. Sci. 3 (2004) 399–425 MR

[9] **J A Chen**, **Z Jiang**, *Positivity in varieties of maximal Albanese dimension*, J. Reine Angew. Math. (online publication July 2015)

[10] **O Debarre**, *On coverings of simple abelian varieties*, Bull. Soc. Math. France 134 (2006) 253–260 MR

[11] **P Deligne**, *Un théorème de finitude pour la monodromie*, from "Discrete groups in geometry and analysis" (R Howe, editor), Progr. Math. 67, Birkhäuser, Boston, MA (1987) 1–19 MR

[12] **L Ein**, **R Lazarsfeld**, *Singularities of theta divisors and the birational geometry of irregular varieties*, J. Amer. Math. Soc. 10 (1997) 243–258 MR

[13] **M Green**, **R Lazarsfeld**, *Deformation theory, generic vanishing theorems, and some conjectures of Enriques, Catanese and Beauville*, Invent. Math. 90 (1987) 389–407 MR

[14] **M Green**, **R Lazarsfeld**, *Higher obstructions to deforming cohomology groups of line bundles*, J. Amer. Math. Soc. 4 (1991) 87–103 MR

[15] **C D Hacon**, *A derived category approach to generic vanishing*, J. Reine Angew. Math. 575 (2004) 173–187 MR

[16] **Z Jiang**, *An effective version of a theorem of Kawamata on the Albanese map*, Commun. Contemp. Math. 13 (2011) 509–532 MR

[17] **J Kollár**, *Higher direct images of dualizing sheaves, II*, Ann. of Math. 124 (1986) 171–202 MR

[18] **R Lazarsfeld**, **M Popa**, **Ch Schnell**, *Canonical cohomology as an exterior module*, Pure Appl. Math. Q. 7 (2011) 1529–1542 MR

[19] **C Mourougane**, **S Takayama**, *Extension of twisted Hodge metrics for Kähler morphisms*, J. Differential Geom. 83 (2009) 131–161 MR

[20] **G Pareschi**, *Basic results on irregular varieties via Fourier–Mukai methods*, from "Current developments in algebraic geometry" (L Caporaso, J McKernan, M Mustaţă, M Popa, editors), Math. Sci. Res. Inst. Publ. 59, Cambridge Univ. Press (2012) 379–403 MR

[21] **G Pareschi**, **M Popa**, *Regularity on abelian varieties, I*, J. Amer. Math. Soc. 16 (2003) 285–302 MR

[22] **G Pareschi**, **M Popa**, *Strong generic vanishing and a higher-dimensional Castelnuovo–de Franchis inequality*, Duke Math. J. 150 (2009) 269–285 MR

[23] **G Pareschi**, **M Popa**, *GV–sheaves, Fourier–Mukai transform, and generic vanishing*, Amer. J. Math. 133 (2011) 235–271 MR

[24] **G Pareschi**, **M Popa**, *Regularity on abelian varieties, III: Relationship with generic vanishing and applications*, from "Grassmannians, moduli spaces and vector bundles" (D A Ellwood, E Previato, editors), Clay Math. Proc. 14, Amer. Math. Soc., Providence, RI (2011) 141–167 MR

[25] **M Popa**, *Generic vanishing filtrations and perverse objects in derived categories of coherent sheaves*, from "Derived categories in algebraic geometry" (Y Kawamata, editor), Eur. Math. Soc. (2012) 251–278 MR

[26] **M Popa**, **Ch Schnell**, *Generic vanishing theory via mixed Hodge modules*, Forum Math. Sigma 1 (2013) art. id. e1 MR

[27] **C Sabbah**, **Ch Schnell**, *The MHM project*, book in progress (2016) Available at `http://www.cmls.polytechnique.fr/perso/sabbah.claude/MHMProject/mhm.html`

[28] **M Saito**, *Modules de Hodge polarisables*, Publ. Res. Inst. Math. Sci. 24 (1988) 849–995 MR

[29] **M Saito**, *Decomposition theorem for proper Kähler morphisms*, Tohoku Math. J. 42 (1990) 127–147 MR

[30] **M Saito**, *Mixed Hodge modules*, Publ. Res. Inst. Math. Sci. 26 (1990) 221–333 MR

[31] **M Saito**, *On Kollár's conjecture*, from "Several complex variables and complex geometry, II" (E Bedford, J P D'Angelo, R E Greene, S G Krantz, editors), Proc. Sympos. Pure Math. 52, Amer. Math. Soc., Providence, RI (1991) 509–517 MR

[32] **W Schmid**, *Variation of Hodge structure: the singularities of the period mapping*, Invent. Math. 22 (1973) 211–319 MR

[33] **W Schmid**, **K Vilonen**, *Hodge theory and unitary representations of reductive Lie groups*, from "Frontiers of mathematical sciences" (B Gu, S-T Yau, editors), Int. Press, Somerville, MA (2011) 397–420 MR

[34] **Ch Schnell**, *Holonomic D-modules on abelian varieties*, Publ. Math. Inst. Hautes Études Sci. 121 (2015) 1–55 MR

[35] **Ch Schnell**, *Torsion points on cohomology support loci: from $\mathcal{D}$–modules to Simpson's theorem*, from "Recent advances in algebraic geometry" (C D Hacon, M Mustaţă, M Popa, editors), London Math. Soc. Lecture Note Ser. 417, Cambridge Univ. Press (2015) 405–421 MR

[36] **Ch Schnell**, *On Saito's vanishing theorem*, Math. Res. Lett. 23 (2016) 499–527 MR

[37] **C Simpson**, *Subspaces of moduli spaces of rank one local systems*, Ann. Sci. École Norm. Sup. 26 (1993) 361–401 MR

[38] **K Takegoshi**, *Higher direct images of canonical sheaves tensorized with semi-positive vector bundles by proper Kähler morphisms*, Math. Ann. 303 (1995) 389–416 MR

[39] **K Ueno**, *Classification theory of algebraic varieties and compact complex spaces*, Lecture Notes in Mathematics 439, Springer (1975) MR

[40] **J Varouchas**, *Sur l'image d'une variété kählérienne compacte*, from "Fonctions de plusieurs variables complexes, V" (F Norguet, editor), Lecture Notes in Math. 1188, Springer (1986) 245–259 MR

[41] **B Wang**, *Torsion points on the cohomology jump loci of compact Kähler manifolds*, Math. Res. Lett. 23 (2016) 545–563 MR

[42] **F Zhang**, *Quaternions and matrices of quaternions*, Linear Algebra Appl. 251 (1997) 21–57 MR

[43] **S Zucker**, *Hodge theory with degenerating coefficients: L_2 cohomology in the Poincaré metric*, Ann. of Math. 109 (1979) 415–476 MR

Dipartimento di Matematica, Università di Roma "Tor Vergata"
Via Della Ricerca Scientifica, I-00133 Roma, Italy

Department of Mathematics, Northwestern University
2033 Sheridan Road, Evanston, IL 60208, United States

Department of Mathematics, Stony Brook University
Stony Brook, NY 11794, United States

pareschi@axp.mat.uniroma2.it, mpopa@math.northwestern.edu, cschnell@math.sunysb.edu

Proposed: Richard Thomas Received: 3 March 2016
Seconded: Jim Bryan, Dan Abramovich Accepted: 8 October 2016

Geometry & Topology 21 (2017) 2461–2473

Non-Kähler complex structures on $\mathbb{R}^4$

ANTONIO J DI SCALA
NAOHIKO KASUYA
DANIELE ZUDDAS

We construct the first examples of non-Kähler complex structures on $\mathbb{R}^4$. These complex surfaces have some analogies with the complex structures constructed in the early fifties by Calabi and Eckmann on the products of two odd-dimensional spheres. However, our construction is quite different from that of Calabi and Eckmann.

32Q15; 57R40, 57R42

1 Introduction

In the early fifties, Calabi and Eckmann [3] constructed an integrable complex structure on the Cartesian product of odd-dimensional spheres $M_{p,q} = S^{2p+1} \times S^{2q+1}$. These complex manifolds are nothing but complex tori for $p = q = 0$, while for $p \geq 1$ and $q = 0$ (or $p = 0$ and $q \geq 1$), they are Hopf manifolds; see Hopf [9]. It is remarkable that the manifolds $M_{p,q}$ (for all nonzero p and q) were the first examples of simply connected, compact complex manifolds which are not algebraic. Moreover, there is a holomorphic torus bundle $h_{p,q}\colon M_{p,q} \to \mathbb{CP}^p \times \mathbb{CP}^q$ given by the Hopf fibration on each factor.

By removing a point on each sphere and taking the product, we get an open subset $E_{p,q} \subset M_{p,q}$ which is diffeomorphic to $\mathbb{R}^{2p+2q+2}$. If $p, q \geq 1$, most fibers of the bundle $h_{p,q}$ are contained in $E_{p,q}$. Thus, $E_{p,q}$ contains embedded holomorphic tori. Therefore, it neither admits a Kähler metric nor can be covered by a single holomorphic coordinate chart. Calabi and Eckmann also proved that the only holomorphic functions on $E_{p,q}$ are the constants.

Definition 1.1 A complex manifold M is said to be of Calabi–Eckmann type if there exist a compact complex manifold X of positive dimension and a holomorphic immersion $k\colon X \to M$ which is null-homotopic as a continuous map.

It follows that a Calabi–Eckmann type complex manifold cannot be tamed by a symplectic form, and in particular it is not Kähler. As a consequence, Stein manifolds, complex

Published: 19 May 2017 DOI: 10.2140/gt.2017.21.2461

algebraic manifolds, and open subsets of $\mathbb{C}^n$ with the induced complex structure are not of Calabi–Eckmann type.

On the other hand, the manifolds $M_{p,q}$ and $E_{p,q}$ are of Calabi–Eckmann type for $p, q \geq 1$. Notice that $\mathbb{R}^4$ is not included in this list. Also notice that the Hopf manifolds $M_{p,0} = S^{2p+1} \times S^1$, $p \geq 1$, are not of Calabi–Eckmann type, since their universal cover is $\mathbb{C}^{p+1} - \{0\}$ (see Proposition 1.2).

The aim of this article is to construct Calabi–Eckmann type complex structures on $\mathbb{R}^4$. This represents a major improvement of our previous result [5], where we constructed a *nonintegrable* almost complex structure on $\mathbb{R}^4$ that contains embedded holomorphic tori and an immersed holomorphic sphere with one node. The methods used there were inspired by previous work of the second author [10; 11] (see also Di Scala and Vezzoni [6] and Di Scala and Zuddas [7] for related results).

The following proposition is an immediate consequence of the definition.

Proposition 1.2 *Let M and N be complex manifolds, with M of Calabi–Eckmann type. Then N is of Calabi–Eckmann type if either*

(1) *there is an immersion of M into N, or*
(2) *there is a covering map $p\colon N \to M$.*

Throughout this paper, we denote by P the open subset of the plane defined by

$$P = \{(\rho_1, \rho_2) \in \mathbb{R}^2 \mid 0 < \rho_1 < 1, 1 < \rho_2 < \rho_1^{-1}\},$$

and we always assume $(\rho_1, \rho_2) \in P$.

We are now ready to state our main theorem.

Theorem 1.3 *There is a family of Calabi–Eckmann-type complex structures $\{J(\rho_1, \rho_2)\}$ on $\mathbb{R}^4$, parametrized by $(\rho_1, \rho_2) \in P$, and a surjective map $f\colon \mathbb{R}^4 \to \mathbb{CP}^1$ with only one critical point, such that:*

(1) *f is holomorphic with respect to $J(\rho_1, \rho_2)$ and the complex hessian at the critical point of f is of maximal rank, for all $(\rho_1, \rho_2) \in P$;*
(2) *the only holomorphic functions on $(\mathbb{R}^4, J(\rho_1, \rho_2))$ are the constants;*
(3) *$J(\rho_1, \rho_2)$ depends smoothly on $(\rho_1, \rho_2) \in P$;*
(4) *$(\mathbb{R}^4, J(\rho_1, \rho_2))$ is not biholomorphic to $(\mathbb{R}^4, J(\rho'_1, \rho'_2))$ for any $(\rho_1, \rho_2) \neq (\rho'_1, \rho'_2)$;*
(5) *the fibers of f are either an immersed holomorphic sphere with one node, embedded holomorphic cylinders, or embedded holomorphic tori.*

We denote by $E(\rho_1, \rho_2)$ the complex manifold $(\mathbb{R}^4, J(\rho_1, \rho_2))$.

Remark Property (1) implies that f can be locally expressed by $f(z_1, z_2) = z_1^2 + z_2^2$ in a neighborhood of the critical point, with respect to suitable local holomorphic coordinates in the complex structure $J(\rho_1, \rho_2)$. In other words, f has a Lefschetz critical point. In fact, as we shall see below, f is the restriction of an achiral Lefschetz fibration on S^4.

As far as we know, $J(\rho_1, \rho_2)$ are the first examples of non-Kähler complex structures on $\mathbb{R}^4$. In [4] we prove some further properties of $E(\rho_1, \rho_2)$, as well as an existence result for Calabi–Eckmann-type complex structures on all smooth connected open oriented 4–manifolds.

Remark Ramanujam in [14] proved that a complex algebraic surface *homeomorphic* to $\mathbb{C}^2$ must be isomorphic to $\mathbb{C}^2$. Among many known constructions of nonstandard complex $\mathbb{R}^4$s, we mention Boc Thaler and Forstnerič [2] and Wold [15].

The following proposition gives a classification of the holomorphic curves of $E(\rho_1, \rho_2)$.

Proposition 1.4 *If S is a compact connected Riemann surface and $g\colon S \to E(\rho_1, \rho_2)$ is holomorphic, then either g is constant or $g(S)$ is a compact fiber of f. It follows that the only compact holomorphic curves of $E(\rho_1, \rho_2)$ are the compact fibers of f, namely embedded holomorphic tori or the immersed holomorphic sphere.*

The following is a corollary of Theorem 1.3.

Corollary 1.5 *The blowup $E(\rho_1, \rho_2) \# m\overline{\mathbb{CP}^2}$ is a Calabi–Eckmann-type complex manifold. In particular, $m\overline{\mathbb{CP}^2} - \{\mathrm{pt}\}$ admits uncountably many non-Kähler complex structures that are pairwise biholomorphically distinct. Moreover, the only holomorphic functions on the blowup $E(\rho_1, \rho_2) \# m\overline{\mathbb{CP}^2}$ are the constants.*

The paper is organized as follows. In Section 2, we recall the construction of the Matsumoto–Fukaya torus fibration on S^4, which is a genus-1 Lefschetz fibration over S^2. This fibration plays a central role in the proof of Theorem 1.3. As an application, we derive a certain decomposition of $\mathbb{R}^4$ in Proposition 2.1.

In Section 3, we construct the complex structure $J(\rho_1, \rho_2)$ and the holomorphic map f, and we prove Theorem 1.3 and Corollary 1.5.

2 The Matsumoto–Fukaya fibration on S^4

In the early eighties, Yukio Matsumoto [13] constructed a genus-1 achiral Lefschetz fibration $f\colon S^4 \to S^2$, having two critical points of opposite signs. As was remarked by

Matsumoto himself in the same article, Kenji Fukaya gave an important contribution in the understanding of this fibration. For this reason, in a private conversation Matsumoto suggested to us to call f the *Matsumoto–Fukaya fibration*, and we are glad to follow his suggestion.

Without going into details, f can be defined as follows. Start with the Hopf fibration $h\colon S^3 \to S^2$, and take the suspension $\Sigma h\colon \Sigma S^3 \to \Sigma S^2$. There is a canonical smoothing $\Sigma S^n \cong S^{n+1}$, which makes the suspension Σh into a smooth map $\Sigma h\colon S^4 \to S^3$; see also our paper [5] for an explicit computation.

The composition $f' = h \circ \Sigma h$ is a torus fibration with two Lefschetz singularities, but the two critical points belong to the same fiber. Indeed, the following formula can be easily obtained [5]:

$$f'(z_1, z_2, x) = \big(4z_1\bar{z}_2(|z_1|^2 - |z_2|^2 - ix\sqrt{2-x^2}), 8|z_1|^2|z_2|^2 - 1\big),$$

where S^{2n} is thought as the unit sphere in $\mathbb{C}^n \times \mathbb{R}$ defined by the equation

$$|z_1|^2 + \cdots + |z_n|^2 + x^2 = 1.$$

In order to get two distinct singular fibers, we slightly perturb f' and the result is the Matsumoto–Fukaya torus fibration $f\colon S^4 \to S^2$. A description of this fibration is given also in the book of Gompf and Stipsicz [8, Example 8.4.7] in terms of a Kirby diagram, which is depicted in Figure 1, where the framings are with respect to the blackboard framing.

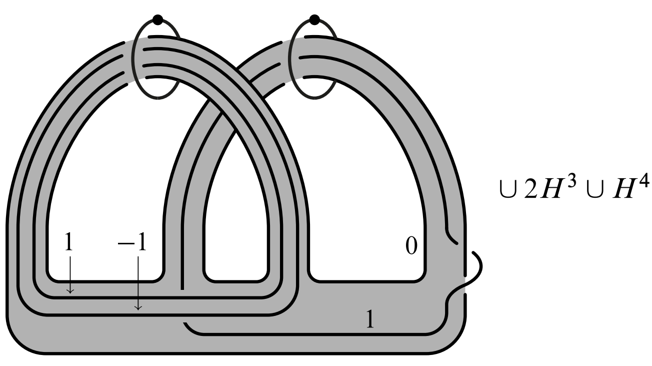


Figure 1: The Matsumoto–Fukaya fibration on S^4

We now explain this Kirby diagram and show how it can be derived. Let $a_1 \in S^2$ be the positive critical value of f, and let a_2 be the negative one. Decompose the base space S^2 as the union of two disks D_1 and D_2 such that $a_j \in \operatorname{Int} D_j$, and put $N_j = f^{-1}(D_j)$. Then N_j is a tubular neighborhood of $F_j = f^{-1}(a_j) \subset S^4$.

It follows that ∂N_1 and ∂N_2 are torus bundles over the circle that are identified by a fiber-reversing diffeomorphism. So ∂N_1 and ∂N_2 are essentially the same torus bundle, but with different orientations.

Let us consider the (achiral) Lefschetz fibration $f_j = f_{|N_j}\colon N_j \to D_j \cong B^2$, having only one critical point. It can be easily realized that the monodromy of f_1 is a right-handed Dehn twist δ about an essential simple curve $c \subset T^2$ that can be identified with a meridian of the torus, while the monodromy of f_2 is given by δ^{-1}. For generalities on Lefschetz fibrations and their monodromies, a good reference is [8]. A description of the monodromy and the induced handlebody decomposition is given also in Apostolakis, Piergallini and Zuddas [1].

Therefore, f_2 can be identified with f_1 by reversing the orientation of the base disk and keeping the same orientation on the fiber.

Since f is built on the Hopf fibration, the monodromy of the latter reflects on the gluing diffeomorphism ϕ between N_1 and N_2. Namely, we have

$$\partial N_1 = \frac{[0, 2\pi] \times T^2}{(0, \delta(x, y)) \sim (2\pi, x, y)},$$

where (x, y) are the angular coordinates in $T^2 = S^1 \times S^1$, and the attaching diffeomorphism $\phi\colon \partial N_1 \to \partial N_2 = -\partial N_1$ is given by $\phi(t, x, y) = (t, x, y + t)$, which passes to the quotient. In other words, while running over ∂D_1, the fiber rotates along the longitude of 2π radians.

In Figure 1 the two 2–handles attached along parallel curves correspond to the two Lefschetz critical points, giving the corresponding vanishing cycles that are parallel to the curve c. The 2–handle with framing 0 attached along the boundary of the punctured torus is needed to close the fiber.

At this point, the Kirby diagram of Figure 2 describes the fiber sum of f_1 and f_2 along an arc in ∂B^2.

In order to complete the fibration, we have to glue a trivial bundle $B^2 \times T^2$ by a fiber-preserving diffeomorphism given by 2π radians rotation in the longitudinal direction.

Considering B^2 as a 2–handle and taking the product with the standard handle decomposition of T^2, we get an extra 2–handle attached along a section, which follows the longitude. A simple computation shows that this 2–handle has framing 1. Also, we get two 3–handles and a 4–handle.

Removing a neighborhood X of the singular point of F_2 which is diffeomorphic to B^4, we obtain $\mathbb{R}^4$. We define the subset X of N_2 to be the standard model of the

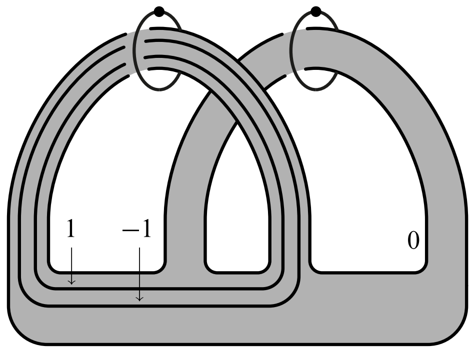


Figure 2: The fiber sum of f_1 and f_2

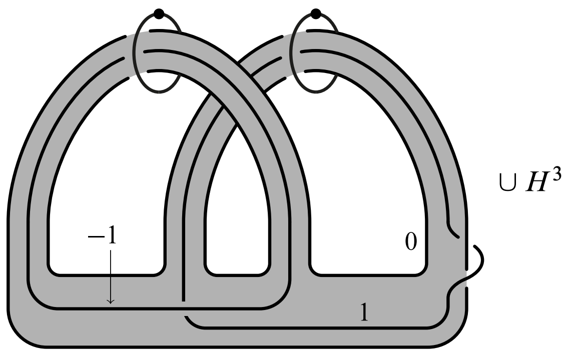


Figure 3: The map f on B^4

neighborhood of a negative Lefschetz singularity. That is, X is the total space of a singular annulus fibration over D_2 with one singular fiber and a left-handed Dehn twist as the monodromy. It is well-known that X is diffeomorphic to B^4, up to smoothing the corners. Then $N_2 - \operatorname{Int} X$ is the total space of a trivial annulus bundle over D_2.

In other words, from the Kirby diagram of Figure 1 we are removing a 4–handle, the 3–handle that comes from the longitude of the torus, and the 2–handle with framing 1 that comes from the negative Lefschetz critical point, thus obtaining the diagram of Figure 3. A simple computation shows that this represents B^4, taking into account that the 3–handle immediately cancels with the 0–framed 2–handle, as it results from its attaching map.

This Kirby diagram encodes the map $f_|\colon S^4 - \operatorname{Int} X \cong B^4 \to S^2$ as part of the Matsumoto–Fukaya fibration.

The above considerations can be summed up in the following proposition, where $A = S^1 \times [0,1]$.

Proposition 2.1 *If we glue $B^2 \times A$ to N_1 along $S^1 \times A$ so that for each $t \in \partial B^2 = -\partial D_1^2 \cong S^1$, the annulus $\{t\} \times A$ embeds in $f^{-1}(t) \cong T^2$ as a thickened meridian, and it rotates in the longitude direction once when $t \in S^1$ rotates once, then the resulting manifold is diffeomorphic to B^4, and so the interior is diffeomorphic to $\mathbb{R}^4$.*

3 Construction of the complex structures

In the following we make use of the notation $\Delta(r_0, r_1) = \{z \in \mathbb{C} \mid r_0 < |z| < r_1\}$, $\Delta(r_1) = \{z \in \mathbb{C} \mid |z| < r_1\}$, for $0 \le r_0 < r_1$. We put also $\mathbb{C}^* = \mathbb{C} - \{0\}$. For a given $w \in \mathbb{C}^*$ such that $|w| < 1$, we consider the smooth elliptic curve $T_w^2 = \mathbb{C}^*/\mathbb{Z} \cong T^2$, where the action is given by $n \cdot z = w^n z$. We call the curve $\mu = \{|z| = 1\} \subset T_w^2$ the meridian of T_w^2. For $0 < \arg w < 2\pi$, we also consider the curve $\lambda = \{w^t \mid t > 0\} \subset T_w^2$, which we call the longitude of the torus, where $w^t = |w|^t e^{it \arg w}$. These meridian and longitude can be identified with those of the previous section.

Proof of Theorem 1.3 We begin with the construction of the complex structure $J(\rho_1, \rho_2)$. Since N_1 is the total space of a positive genus-1 Lefschetz fibration over the 2–disk with one singular fiber, there exists a complex structure such that the fibration

$$f_{|N_1} \colon N_1 \to D_1$$

is holomorphic. Indeed, we consider a holomorphic elliptic fibration over S^2, and we take a tubular neighborhood of a singular fiber which is fiberwise diffeomorphic with N_1 (see also [8]). According to Kodaira [12], we can give a more explicit model of $\operatorname{Int} N_1$ by the Weierstrass curves. For $0 < \rho_1 < 1$, we consider the complex submanifold

$$S = \{([z_0 : z_1 : z_2], \tau) \mid z_1^2 z_2 - 4z_0^3 - z_0^2 z_2 + g_2(\tau) z_0 z_2^2 + g_3(\tau) z_2^3 = 0\} \subset \mathbb{CP}^2 \times \Delta(\rho_1),$$

where

$$g_2(\tau) = 20 \sum_{n=1}^{\infty} (1 - \tau^n)^{-1} n^3 \tau^n,$$

$$g_3(\tau) = \frac{1}{3} \sum_{n=1}^{\infty} (1 - \tau^n)^{-1} (7n^5 + 5n^3) \tau^n.$$

For each $\tau \in \Delta(\rho_1)$, the fiber $\{z_1^2 z_2 - 4z_0^3 - z_0^2 z_2 + g_2(\tau) z_0 z_2^2 + g_3(\tau) z_2^3 = 0\} \subset \mathbb{CP}^2$ is an elliptic curve and it is singular only for $\tau = 0$. Using the coordinates $(x, y) = (z_0/z_2, z_1/z_2)$, the singular elliptic fiber is defined by the equation

$$y^2 - 4x^3 - x^2 = 0$$

and it has an ordinary double point at $x = y = 0$. Hence, the canonical projection

$$\pi\colon \mathbb{CP}^2 \times \Delta(\rho_1) \to \Delta(\rho_1)$$

restricts to a holomorphic map $\pi_{|S}\colon S \to \Delta(\rho_1)$, which is a genus-1 holomorphic Lefschetz fibration over the 2–disk with one singular fiber. Thus, the complex manifold S is a complex model of $\operatorname{Int} N_1$.

Now, we consider the quotient $(\mathbb{C}^* \times \Delta(0, \rho_1))/\mathbb{Z}$, where for any $n \in \mathbb{Z}$, the action is given by

$$n \cdot (z, w) = (zw^n, w).$$

This elliptic fibration extends over $\Delta(\rho_1)$. Let us denote the completion by W. Kodaira gave an explicit biholomorphism between W and S (see [12, pages 597–599]). So, in the following we shall consider W as the model of $\operatorname{Int} N_1$, instead of the Weierstrass model S.

We fix a holomorphic atlas on the Riemann sphere given by two open disks $D_1' \supset D_1$ and $D_2' \supset D_2$. The disk D_1' is biholomorphic with $\Delta(\rho_1)$ and D_2' is biholomorphic with $\Delta(\rho_0^{-1})$, where $\rho_0 \in (0, \rho_1)$ is arbitrarily chosen, and the transition function $\psi\colon \Delta(\rho_0, \rho_1) \to \Delta(\rho_1^{-1}, \rho_0^{-1})$ is given by $\psi(z) = z^{-1}$.

We define the complex structure on the topologically trivial annulus bundle $N_2 - X$, considered over $\Delta(\rho_0^{-1}) \cong D_2'$, by the product structure $\Delta(1, \rho_2) \times \Delta(\rho_0^{-1})$.

Next, we want to glue $W \cong \operatorname{Int} N_1$ with $\Delta(1, \rho_2) \times \Delta(\rho_0^{-1}) \cong \operatorname{Int} N_2 - X$ analytically along $\Delta(1, \rho_2) \times \Delta(\rho_1^{-1}, \rho_0^{-1}) \subset \Delta(1, \rho_2) \times \Delta(\rho_0^{-1})$, so that the attaching map is isotopic to that of the Matsumoto–Fukaya fibration, implying that the resulting manifold is diffeomorphic to $\mathbb{R}^4$. In order to do this, we need to choose an attaching region in W which is biholomorphic to the product $\Delta(1, \rho_2) \times \Delta(\rho_1^{-1}, \rho_0^{-1})$. In the following argument, we show how to take such a region in W.

We begin with a multivalued holomorphic function $\phi\colon \Delta(\rho_0, \rho_1) \to \mathbb{C}^*$ such that multiplication by w^n for all $n \in \mathbb{Z}$ determines a transitive $\mathbb{Z}$–action on the set of the branches of ϕ. In other words, given any branch ϕ_0 of ϕ, all the other branches are of the form $w^n \phi_0(w)$ for an arbitrary $n \in \mathbb{Z}$.

For example, as can be easily verified, we can take

$$\phi(w) = \exp\bigl(\tfrac{1}{4\pi i}(\log w)^2 - \tfrac{1}{2}\log w\bigr),$$

where the two logarithms are taken simultaneously with all of their possible branches.

Next, consider the open subset $Y \subset \mathbb{C}^* \times \Delta(0, \rho_1)$ defined by

$$Y = \bigl\{(z, w) \in \mathbb{C}^* \times \Delta(\rho_0, \rho_1) \mid z\phi(w)^{-1} \in \Delta(1, \rho_2) \text{ for some value of } \phi(w)\bigr\}.$$

Then Y can be parametrized by the multivalued local holomorphic immersion

$$\Phi\colon \Delta(1,\rho_2)\times\Delta(\rho_0,\rho_1)\to\mathbb{C}^*\times\Delta(0,\rho_1)$$

defined by $\Phi(z,w)=(z\phi(w),w)$. Notice that Y is invariant under the action of $\mathbb{Z}$ on $\mathbb{C}^*\times\Delta(0,\rho_1)$.

It follows that the composition of Φ with the quotient map

$$\pi\colon \mathbb{C}^*\times\Delta(0,\rho_1)\to(\mathbb{C}^*\times\Delta(0,\rho_1))/\mathbb{Z}\subset W$$

is a single-valued holomorphic embedding, and we denote by V the image of Y in W.

Also notice that $f_{|V}\colon V\to\Delta(\rho_0,\rho_1)$ is a holomorphic annulus bundle, and Φ determines a trivialization of this bundle.

Let $j\colon \Delta(1,\rho_2)\times\Delta(\rho_0,\rho_1)\to\Delta(1,\rho_2)\times\Delta(\rho_1^{-1},\rho_0^{-1})$ be the biholomorphism defined by $j(z,w)=(z,w^{-1})$. We use $\pi\circ\Phi\circ j^{-1}\colon \Delta(1,\rho_2)\times\Delta(\rho_1^{-1},\rho_0^{-1})\to V$ as the attaching biholomorphism for making the union

$$E(\rho_1,\rho_2)=(\Delta(1,\rho_2)\times\Delta({\rho_0}^{-1}))\cup_V W,$$

which therefore is a complex manifold. We denote by $J(\rho_1,\rho_2)$ the complex structure of $E(\rho_1,\rho_2)$.

In order to identify the topology of $E(\rho_1,\rho_2)$, we consider how the annulus fiber of $V=\Delta(1,\rho_2)\times\Delta(\rho_0,\rho_1)$ looks inside the toric fiber of $W\cong\operatorname{Int}N_1$. Let $w\in\Delta(\rho_0,\rho_1)$ be a complex number of some fixed modulus, and of arbitrary argument $\arg w$. When $\arg w$ varies from 0 to 2π, a point z of the fiber $\Delta(1,\rho_2)$ moves accordingly with the $\mathbb{Z}$–action on the branches of ϕ, which encodes the attaching biholomorphism. Namely, z goes to wz, which is in the next fundamental domain of the $\mathbb{Z}$–action on $\mathbb{C}^*$ that gives the torus $T^2_w=\mathbb{C}^*/\mathbb{Z}$. This means that when w rotates once in the argument direction, the annulus $\Delta(1,\rho_2)$ rotates once in the longitude direction of the fiber torus T^2_w, and this coincides with the attaching map of the Matsumoto–Fukaya fibration restricted on S^4-X. This implies that $E(\rho_1,\rho_2)$ is diffeomorphic to $\mathbb{R}^4$.

Observe that the number ρ_0 is auxiliary. Indeed, its role is just to provide an open subset V of $\operatorname{Int}N_1$ which is necessary for the analytical gluing. Namely, ρ_0 only determines the size of the attaching region of the two complex manifolds. Hence, $E(\rho_1,\rho_2)$ does not depend on ρ_0 up to biholomorphisms.

Moreover, we can define a surjective holomorphic map

$$f_{\rho_1,\rho_2}\colon E(\rho_1,\rho_2)\to\mathbb{CP}^1,\quad f_{\rho_1,\rho_2}(z,w)=\begin{cases} w & \text{if } (z,w)\in W,\\ w^{-1} & \text{if } (z,w)\in\Delta(1,\rho_2)\times\Delta({\rho_0}^{-1}).\end{cases}$$

Notice that f_{ρ_1,ρ_2} is isotopic to the restriction of the Matsumoto–Fukaya fibration f (up to the diffeomorphism $E(\rho_1,\rho_2)\cong\mathbb{R}^4$). Actually, the isotopy can be chosen to be smooth with respect to the parameters (ρ_1,ρ_2), and up to this isotopy we can assume that $f_{|\mathbb{R}^4}$ itself is a holomorphic fibration with respect to all of the complex structures $J(\rho_1,\rho_2)$. Moreover, the complex hessian of f at the critical point, computed with respect to $J(\rho_1,\rho_2)$, is of maximal rank because this complex structure on W coincides with that of the Weierstrass curves model.

Since f has holomorphic tori as fibers, it follows that $E(\rho_1,\rho_2)$ is of Calabi–Eckmann type. Moreover, statements (1), (3) and (5) of the theorem are implicit in the construction.

The nonexistence of nonconstant holomorphic functions also follows easily, since there is an open subset of $E(\rho_1,\rho_2)$, namely W, which is foliated by compact holomorphic curves, hence a holomorphic function $g\colon E(\rho_1,\rho_2)\to\mathbb{C}$ must be constant on these fibers. Therefore, the differential of g is zero along those compact fibers. Since dg is a holomorphic 1-form, it follows that dg is zero even along the annulus fibers of f, and this implies that g is constant on the fibers of f. Thus, g factorizes by the fibration f. Namely, there is a holomorphic function $g'\colon\mathbb{CP}^1\to\mathbb{C}$ such that $g=g'\circ f$. Since g' is constant, g is also constant. This proves statement (2).

Now we give the proof of statement (4), which is based on Proposition 1.4, which in turn will be proved at the end of this section. By Proposition 1.4, the union of all compact holomorphic curves of $E(\rho_1,\rho_2)$ is the open subset W. Analogously, we denote by W' the union of the compact holomorphic curves of $E(\rho_1',\rho_2')$.

Suppose that there is a biholomorphism $g\colon E(\rho_1,\rho_2)\to E(\rho_1',\rho_2')$. We want to show that $(\rho_1,\rho_2)=(\rho_1',\rho_2')$.

The discussion above implies that g decomposes into two biholomorphisms $g_|\colon W\to W'$ and $g_|\colon \Delta(1,\rho_2)\times\Delta({\rho_1}^{-1})\to\Delta(1,\rho_2')\times\Delta({\rho_1'}^{-1})$. Moreover, g is fiber-preserving on W, and so $g_{|W}$ passes to the quotient, giving a biholomorphism $g'\colon\Delta(\rho_1)\to\Delta(\rho_1')$. By analyticity, g must be fiber-preserving also on $\Delta(1,\rho_2)\times\Delta({\rho_1}^{-1})$. This immediately gives $\rho_2=\rho_2'$, because of the well-known holomorphic classification of complex annuli.

The torus T^2_w that corresponds to the complex number $w\in\Delta(0,\rho_1)$ is isomorphic to a complex torus of the form $\mathbb{C}/(\mathbb{Z}\oplus\mathbb{Z}v)$, where

$$v=\tfrac{1}{2\pi i}\log w=\tfrac{1}{2\pi}\arg w-\tfrac{i}{2\pi}\log|w|,$$

and $\arg w\in[0,2\pi)$.

By the classification of complex nonsingular elliptic curves, g' must be the identity because T^2_w is isomorphic to $g(T^2_w) = T^2_{g'(w)}$ for all $w \in \Delta(0, \rho_1)$. Therefore, we obtain $\rho_1 = \rho'_1$. □

Proof of Proposition 1.4 It is sufficient to show that $f \circ g$ is constant. In fact, $f \circ g\colon S \to \mathbb{CP}^1$ is homotopic to a constant, since it factorizes through the contractible space $E(\rho_1, \rho_2)$. Therefore, $f \circ g$ is of degree zero. Since it is a holomorphic map between compact Riemann surfaces, it must be constant. □

Finally, we prove Corollary 1.5.

Proof of Corollary 1.5 Since the blowup affects $E(\rho_1, \rho_2)$ only at finitely many points, after blowing up, there are still embedded holomorphic tori which are homotopically trivial. Then, $E(\rho_1, \rho_2) \# m \overline{\mathbb{CP}^2}$ is of Calabi–Eckmann type.

Moreover, if $E(\rho_1, \rho_2) \# m \overline{\mathbb{CP}^2}$ and $E(\rho'_1, \rho'_2) \# m \overline{\mathbb{CP}^2}$ are biholomorphic, then it follows that $\rho_1 = \rho'_1$ and $\rho_2 = \rho'_2$ by the same argument as in the proof of Theorem 1.3(4). Since $E(\rho_1, \rho_2) \# m \overline{\mathbb{CP}^2}$ is diffeomorphic to $m \overline{\mathbb{CP}^2} - \{p\}$, there are uncountably many distinct non-Kähler complex structures on $m \overline{\mathbb{CP}^2} - \{p\}$.

Finally, if h is a holomorphic function on $E(\rho_1, \rho_2) \# m \overline{\mathbb{CP}^2}$, it must be constant on the exceptional spheres, and so it factorizes through the blowup map

$$\sigma\colon E(\rho_1, \rho_2) \# m \overline{\mathbb{CP}^2} \to E(\rho_1, \rho_2).$$

Hence, h is constant. This completes the proof. □

Remark Corollary 1.5 holds even for $m = \infty$, and the proof is essentially the same. By making the points for the blowups vary, we get even more pairwise inequivalent complex structures on $E(\rho_1, \rho_2) \# m \overline{\mathbb{CP}^2}$.

Acknowledgements

The authors wish to thank Yukio Matsumoto, Ryushi Goto, and Ichiro Enoki. Yukio Matsumoto gave us an important suggestion about Theorem 1.3. Ryushi Goto and Ichiro Enoki gave us helpful advice for the model of the holomorphic elliptic fibration in the proof of Theorem 1.3. Also thanks to Alberto Verjovsky for useful comments on related results. Part of this article was written when Kasuya and Zuddas were visiting the Department of Mathematical Sciences of Durham University, UK. They are grateful to Durham University for its hospitality. Zuddas also thanks the Grey College of Durham University for hospitality during his stay at Durham. We are also thankful to Wilhelm Klingenberg for helpful conversations and for having invited us to Durham University.

Antonio J Di Scala and Daniele Zuddas are members of GNSAGA of INdAM.

References

[1] **N Apostolakis**, **R Piergallini**, **D Zuddas**, *Lefschetz fibrations over the disc*, Proc. Lond. Math. Soc. 107 (2013) 340–390 MR

[2] **L Boc Thaler**, **F Forstnerič**, *A long* $\mathbb{C}^2$ *without holomorphic functions*, Anal. PDE 9 (2016) 2031–2050 MR

[3] **E Calabi**, **B Eckmann**, *A class of compact, complex manifolds which are not algebraic*, Ann. of Math. 58 (1953) 494–500 MR

[4] **A J Di Scala**, **N Kasuya**, **D Zuddas**, *Non-Kähler complex structures on* R^4*, II*, preprint (2015) arXiv

[5] **A J Di Scala**, **N Kasuya**, **D Zuddas**, *On embeddings of almost complex manifolds in almost complex Euclidean spaces*, J. Geom. Phys. 101 (2016) 19–26 MR

[6] **A J Di Scala**, **L Vezzoni**, *Complex submanifolds of almost complex Euclidean spaces*, Q. J. Math. 61 (2010) 401–405 MR

[7] **A J Di Scala**, **D Zuddas**, *Embedding almost-complex manifolds in almost-complex Euclidean spaces*, J. Geom. Phys. 61 (2011) 1928–1931 MR

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

[9] **H Hopf**, *Zur Topologie der komplexen Mannigfaltigkeiten*, from "Studies and Essays Presented to R. Courant on his 60th Birthday", Interscience, New York (1948) 167–185 MR

[10] **N Kasuya**, *On contact embeddings of contact manifolds in the odd dimensional Euclidean spaces*, Internat. J. Math. 26 (2015) art. id. 1550045, 10 pages MR

[11] **N Kasuya**, *An obstruction for codimension two contact embeddings in the odd dimensional Euclidean spaces*, J. Math. Soc. Japan 68 (2016) 737–743 MR

[12] **K Kodaira**, *On compact analytic surfaces, II*, Ann. of Math. 78 (1963) 563–626 MR

[13] **Y Matsumoto**, *On* 4*–manifolds fibered by tori*, Proc. Japan Acad. Ser. A Math. Sci. 58 (1982) 298–301 MR

[14] **C P Ramanujam**, *A topological characterisation of the affine plane as an algebraic variety*, Ann. of Math. 94 (1971) 69–88 MR

[15] **E F Wold**, *A long* $\mathbb{C}^2$ *which is not Stein*, Ark. Mat. 48 (2010) 207–210 MR

Dipartimento di Scienze Matematiche G L Lagrange, Politecnico di Torino
Corso Duca degli Abruzzi 24, 10129 Torino, Italy

School of Social Informatics, Aoyama Gakuin University
5-10-1 Fuchinobe, Chuo-ku, Sagamihara, Kanagawa 252-5258, Japan

School of Mathematics, Korea Institute for Advanced Study
85 Hoegiro, Dongdaemun-gu, Seoul 02455, South Korea

antonio.discala@polito.it, nkasuya@si.aoyama.ac.jp, zuddas@kias.re.kr

Proposed: Yasha Eliashberg Received: 23 March 2016
Seconded: Bruce Kleiner, András I. Stipsicz Revised: 28 April 2016

Geometry & Topology 21 (2017) 2475–2484

Infinite order corks

ROBERT E GOMPF

We construct a compact, contractible 4–manifold C, an infinite order self-diffeomorphism f of its boundary, and a smooth embedding of C into a closed, simply connected 4–manifold X, such that the manifolds obtained by cutting C out of X and regluing it by powers of f are all pairwise nondiffeomorphic. The manifold C can be chosen from among infinitely many homeomorphism types, all obtained by attaching a 2–handle to the meridian of a thickened knot complement.

57N13, 57R55

1 Introduction

The wild proliferation of exotic smoothings of 4–manifolds highlights the failure of high-dimensional topology to apply in dimension 4, notably through failure of the h–cobordism theorem. Attempts to understand this issue led to the notion of a *cork twist*. A cork, as originally envisioned, is a contractible, smooth submanifold C of a closed 4–manifold X, with an involution f of ∂C, such that cutting out C and regluing it by the twist f changes the diffeomorphism type of X (while necessarily preserving its homeomorphism type). We can think of C as a control knob with two settings, toggling between two smoothings of X. The first example of a cork was discovered by Akbulut [1]. Subsequently, various authors (Curtis, Freedman, Hsiang and Stong [6], Matveyev [15]; see Gompf and Stipsicz [14] for more history) showed that any two homeomorphic, simply connected (smooth) 4–manifolds are related by a cork twist. Since then, much work has been done (see eg Akbulut and Ruberman [3], Akbulut and Yasui [4]) to understand and apply cork twists. Various people, going back at least to Freedman in the 1990s, have asked whether higher-order corks may exist — that is, knobs with n settings for n different diffeomorphism types, or possibly even infinitely many settings all realizing distinct types. Recently, progress has been made by modifying known examples of corks: Tange [19] exhibited knobs with n settings for any finite n, displaying two diffeomorphism types on X. Independently, Auckly, Kim, Melvin and Ruberman [5] constructed the desired finite order corks. More generally, they constructed G–corks for any finite subgroup G of $\mathrm{SO}(4)$, where the control knob can be set to any element of G to yield $|G|$ diffeomorphism types. However, both of

Published: 19 May 2017 DOI: 10.2140/gt.2017.21.2475

these latter papers pose the infinite order case as a still unsolved problem, in spite of fruitless attacks by various mathematicians. The purpose of the present article is to exhibit a large family of infinite order corks, arising from a simple general construction.

There is variation in the literature about the definition of a cork. All approaches share the following:

Definition 1.1 A *cork* (C, f) is a smooth, compact, contractible 4–manifold C with a diffeomorphism $f\colon \partial C \to \partial C$. The cork will be called *nontrivial* if f does not extend to a self-diffeomorphism of C. If C is smoothly embedded in a 4–manifold X, cutting out C and regluing it by f to get $(X - \operatorname{int} C) \cup_f C$ will be called a *twist* by f.

Note that (C, f^k) is then a cork for any $k \in \mathbb{Z}$, so we also talk about twisting by powers f^k. By Freedman's topological h–cobordism theorem rel boundary [9; 10], f necessarily extends to a self-homeomorphism of C, so a cork twist does not change the homeomorphism type of a manifold. In some references, f is required to be an involution, or extend to a finite cyclic (or other finite) group action on ∂C. Since we are interested in $\mathbb{Z}$–actions, no additional hypothesis is needed. We can now state our main existence theorem, which is proved in Section 2.

Theorem 1.2 *There is a cork (C, f) and a smooth embedding of C into a closed, simply connected 4–manifold X, for which the manifolds X_k, $k \in \mathbb{Z}$, obtained by twisting by f^k are homeomorphic but pairwise nondiffeomorphic. Hence, the corks (C, f^k) are distinct (up to diffeomorphism commuting with the maps), and nontrivial unless $k = 0$.*

In the terminology of [5], the embedding $C \hookrightarrow X$ is $\mathbb{Z}$*–effective* and exhibits (C, f) as the first example of a $\mathbb{Z}$*–cork*.

Corollary 1.3 *The homology 3–sphere ∂C bounds infinitely many smooth, contractible manifolds that are all diffeomorphic, homeomorphic rel boundary and pairwise nondiffeomorphic rel boundary.*

Proof Identify ∂C as the boundary of C using each of the diffeomorphisms f^k. □

Corollary 1.4 *There is a compact, contractible 4–manifold admitting infinitely many nondiffeomorphic smooth structures.*

Proof This follows immediately from the previous corollary and Akbulut and Ruberman [3, Theorem 5.3]. □

We obtain infinitely many examples of corks (C, f) as in the theorem, distinguished by the homeomorphism types of their boundaries. However, our examples all have a simple form. For any knot $\kappa \subset S^3$, let P be its closed complement, and let $C(\kappa, m)$ be the oriented 4–manifold obtained from $I \times P$, where I denotes the interval $[-1, 1]$ throughout the paper, by attaching a 2–handle along the meridian to κ in $\{1\} \times P$ with framing m. Note that $I \times P$ can be identified with the obvious ribbon complement of $\kappa \# \bar{\kappa}$ in B^4, so this is a special case of removing a slice disk and regluing it with a twist. Either perspective reveals the identity $C(\bar{\kappa}, m) \approx C(\kappa, m)$, and these are orientation-reversingly diffeomorphic to $C(\kappa, -m)$. Clearly, $C(\kappa, m)$ is the 4–ball when $m = 0$ or κ is unknotted, but otherwise it is a contractible manifold whose boundary is irreducible and not S^3. In fact, $\partial C(\kappa, m)$ is obtained by $\left(-\frac{1}{m}\right)$–surgery on $\kappa \# \bar{\kappa}$, and contains two oppositely oriented copies of the complement P. When κ is prime, the JSJ decomposition of $\partial C(\kappa, m)$ begins by splitting out these complements. (This gives the entire decomposition unless κ is a satellite knot, in which case the splitting continues symmetrically.) Since the complements can then be recovered from $\partial C(\kappa, m)$, it follows that the manifolds $C(\kappa, m)$ (κ prime) are never orientation-preservingly homeomorphic unless the corresponding knots are the same up to orientation and the (nonzero) integers are equal. In our examples, κ is the double twist knot $\kappa(r, -s) = \kappa(-s, r)$ shown in Figure 1, where the boxes count full twists, right-handed when the integer is positive. The resulting oriented 4–manifolds $C(r, s; m) = C(\kappa(r, -s), m)$, for $r, s > 0$ and $m \neq 0$, are not orientation-preservingly homeomorphic to each other unless the integers m agree and the pairs (r, s) agree up to order. In general, the incompressible torus $\{0\} \times \partial P$ in $\partial C(\kappa, m)$ can be used to create self-diffeomorphisms of the latter: Let $f\colon \partial C(\kappa, m) \to \partial C(\kappa, m)$ be obtained by rotating the torus $\{t\} \times \partial P$ parallel to the canonical longitude of κ, through angle $(t + 1)\pi$, $t \in I = [-1, 1]$, as we pass through $I \times \partial P$, and extending as the identity. Our simplest cork, $(C(1, 1; -1), f)$ is made in this manner from the figure-eight knot $\kappa(1, -1)$. Its boundary is given by surgery with coefficient 1 on the connected sum of two figure-eight knots, with the obvious incompressible torus in the complement of this sum. More generally, we have:

Theorem 1.5 *The cork C appearing in Theorem 1.2 can be taken to be any of the infinitely many contractible manifolds $C(r, s; m)$ with $r, s > 0$ and $m \neq 0$, and f as specified above. The manifolds X_k can be assumed to be irreducible, except possibly if r, s or $|m|$ equals 2.*

Recall that a 4–manifold is *irreducible* if it cannot split as a smooth connected sum unless one summand is homeomorphic to S^4. Other explicit constructions of corks in the literature typically involve reducible (blown up) manifolds. It seems likely that the restriction avoiding 2 is unnecessary; see Remarks 2.1(a).

Our incompressible torus in $C(\kappa, m)$ can be also used to define other twists. Instead of twisting parallel to the longitude, we could twist parallel to the meridian, or more

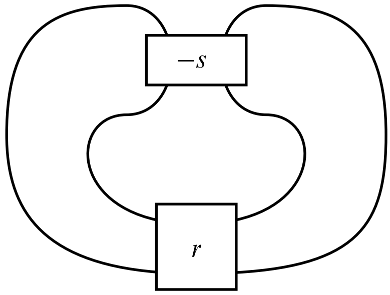


Figure 1: The double twist knot $\kappa(r,-s)$

generally, twist using any element of $H_1(T^2) \cong \mathbb{Z} \oplus \mathbb{Z}$. Thus, it is natural to ask both about other contractible manifolds and other twists:

Question 1.6 (a) Is every pair $(C(\kappa, m), f)$, for κ a nontrivial knot, $m \neq 0$ and f a longitudinal twist as given above, a $\mathbb{Z}$–cork?

(b) Does twisting by other elements of $H_1(T^2)$ ever extend these to $(\mathbb{Z} \oplus \mathbb{Z})$–corks?

Akbulut, in a preliminary version of [2], previously studied the meridian twist for κ the trefoil and $m = -1$, trying to prove nontriviality. However, we show in [13] that the meridian twist extends over every $C(\kappa, \pm 1)$. Recently, Ray and Ruberman [17] answered (a) in the negative for torus knots κ when $|m| = 1$. It follows that every boundary diffeomorphism extends over $C(\kappa, \pm 1)$ for such knots [13]. See the latter paper for further discussion and the translation of the main proofs of this paper into the language of handle calculus. The question is still open for meridian twists when $|m| \geq 2$ and for longitudinal twists with nontorus knots κ outside our family $\{\kappa(r,-s) \mid r, s > 0\}$.

More recently, Tange has posted papers extending the methods of this article to exhibit n–fold boundary sums of our $\mathbb{Z}$–corks as $\mathbb{Z}^n$–corks [21] and providing constraints on families of manifolds that can be related by $\mathbb{Z}$–corks [20].

Acknowledgements

The author would like to thank Cameron Gordon, Danny Ruberman and András Stipsicz for helpful comments.

2 Constructing corks

The closed manifolds X_k in Theorem 1.2 are made from the elliptic surface $E(n)$, for a fixed $n \geq 1$, by the Fintushel–Stern knot construction [7]. Recall (see eg [14]) that $E(n)$

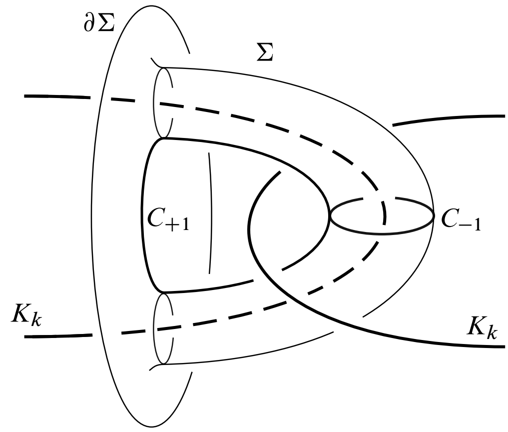


Figure 2: The punctured torus Σ in the knot complement $M = S^3 - \nu K_k$ $(k = 0)$

has a standard description in which it is built from $S^1 \times S^1 \times D^2$ (a neighborhood of a regular fiber $F = S^1 \times S^1$) by adding handles. Of most interest for present purposes, each of the two circle factors has $6n$ parallel copies (*vanishing cycles*) to which 2–handles are attached with framing -1 (relative to the product framing of the boundary 3–torus). We will use three of these 2–handles. Given a knot $K \subset S^3$, let M denote its closed complement. The knot construction consists of removing $S^1 \times S^1 \times D^2$ from $E(n)$ and replacing it by $M \times S^1$, gluing by a diffeomorphism of the boundary 3–torus that identifies the canonical longitude of K with $\{\text{point}\} \times \partial D^2$, and the meridian of K and circle $\{\text{point}\} \times S^1$ in $M \times S^1$ with copies of the two circle factors of F. As detailed in [7], Freedman's classification [9; 10] shows that the resulting manifold X_K retains the homeomorphism type of $E(n)$ (which is simply connected with $b_2 = 12n-2$ and signature $-8n$, and is even if and only if n is). However, when $n \geq 2$, varying the knot K results in diffeomorphism types that are distinguished by their Seiberg–Witten invariants if and only if the knots in question are distinguished by their Alexander polynomials. The structure of the Seiberg–Witten invariants then also shows that each X_K is irreducible. When $n = 1$ the discussion becomes more technical, but these statements remain true for the k–twist knots $K_k = \kappa(k, -1)$ with $k \in \mathbb{Z}$ [8], except that the unknot K_0 yields the reducible manifold $E(1)$, a sum of copies of $\pm\mathbb{CP}^2$. For fixed $n \geq 1$, let X_k be obtained as above from the twist knot K_k. Since these knots are distinguished by their Alexander polynomials, Theorem 1.2 follows once we locate a contractible $C \subset X = X_0 = E(n)$ with a twist f for which each power f^k gives the corresponding X_k.

Proof of Theorem 1.2 Let $\Sigma \subset M$ be the punctured torus depicted in Figure 2 (near the clasp of K_k, the $-s = -1$ twist box in Figure 1) with circles $C_{\pm 1}$ generating its homology. Set $k=0$ for this, but note that the corresponding picture for any k (with the correct longitude) is then obtained from the $k=0$ case by $\left(-\frac{1}{k}\right)$–surgery on the circle $\partial\Sigma$. To examine this surgery more closely, identify a tubular neighborhood of Σ in M with $I\times\Sigma$, where $I=[-1,1]$ and $\{1\}\times\Sigma$ contains the outer part of the boundary of $I\times\Sigma$ visible in the figure. Let A be a collar of $\partial\Sigma$ in Σ. Then we can perform the required surgery by cutting out and regluing the solid torus $I\times A$. Since the surgery coefficient has numerator 1, we can take the gluing diffeomorphism to be the identity everywhere except on the annulus $I\times\partial\Sigma$. That is, we get from the $k=0$ case to the case of arbitrary k by slitting M open along the annulus $I\times\partial\Sigma$ and regluing by g^k for a suitable Dehn twist g of the annulus. Hence, to transform X_0 to X_k, we slit X_0 open along the 3–manifold $N=I\times\partial\Sigma\times S^1\subset M\times S^1$ and reglue by $(g\times\mathrm{id}_{S^1})^k$. (This operation can be viewed as a torus surgery, also called a logarithmic transformation, and would be a Luttinger surgery if $\{0\}\times\partial\Sigma\times S^1$ could be made Lagrangian. The latter is ruled out, however, since X_k has no symplectic structure unless $|k|\le 1$.) Our goal is to find a contractible manifold $C\subset X_0$ whose boundary contains N. Extending $g\times\mathrm{id}_{S^1}$ as the identity over the rest of ∂C then gives the required diffeomorphism f completing the proof.

Our first approximation to C is the manifold $Y=I\times\Sigma\times S^1\subset M\times S^1\subset X_0$. Then ∂Y clearly contains N, but Y is far from being contractible. In fact, Y is homotopy equivalent to $(S^1\vee S^1)\times S^1$, so it has $b_1=3$ and $b_2=2$, but no higher-dimensional homology. Its fundamental group is generated by three circles C_i^*, $i=-1,0,1$ (suitably attached to the base point), where $C_i^*=\{i\}\times C_i\times\{\theta_i\}$ for $i=\pm1$ and distinct points $\theta_{\pm1}\in S^1$, and $C_0^*=\{1\}\times\{p\}\times S^1$ for some $p\in\operatorname{int}\Sigma-(C_{-1}\cup C_1)$. A basis for $H_2(Y)$ is given by the pair of tori $T_i=\{0\}\times C_i'\times S^1$, $i=\pm1$, where C_i' is parallel to C_i in $\Sigma-\{p\}$. To improve Y, observe in Figure 2 that the circles $C_{\pm1}$ in M are both meridians of the knot K_0. Thus, the knot construction matches all three circles $C_i^*\subset\partial Y$ with vanishing cycles of $E(n)$. We obtain a new manifold $Y'\subset X_0$ by ambiently attaching a (-1)–framed 2–handle h_i to Y along C_i^* for each $i=-1,0,1$. Then Y' is simply connected with T_i, $i=\pm1$, still giving a basis for $H_2(Y')$, and N still contained in $\partial Y'$. To eliminate the last homology, note that for $i=\pm1$, the core of the handle h_i fits together with the annulus $I\times C_i\times\{\theta_i\}$, forming a pair of disks D_i disjointly embedded rel boundary in Y' (with $\partial D_i=\{-i\}\times C_i\times\{\theta_i\}$). Since each $D_i\cap T_i$ is empty, and $D_i\cap T_{-i}$ is a single point of transverse intersection, deleting tubular neighborhoods of these disks from Y' gives a manifold C with no homology. To see that $\pi_1(C)$ vanishes, use the core of the 2–handle h_0 to surger the tori T_i to immersed spheres, without changing the intersections with the disks D_i. These spheres then provide nullhomotopies for the meridians of the disks. Thus, C is a contractible manifold whose boundary contains N, as required. $\square$

Proof of Theorem 1.5 To identify the cork C constructed in the proof of Theorem 1.2, first consider any framed sphere S in a manifold Q. If we add a handle h to $I\times Q$ along $1\times S$, and then delete a neighborhood of the core of h, extended down to $\{-1\}\times Q$ using the annulus $I\times S$, the result is easily seen to be $I\times P$, where P is made from Q by surgery on S. We apply this trick with $Q=\Sigma\times S^1$ from the previous proof. Attaching the handles $h_{\pm 1}$ to $Y=I\times Q$ and deleting their cores $D_{\pm 1}$ gives a manifold of the form $I\times P$ that will become C when h_0 is attached. The manifold P is obtained from Q by surgery on the disjoint curves $C_{\pm 1}\times\{\theta_{\pm 1}\}$, with the framings induced from their identification with vanishing cycles of $E(n)$. These framings are -1 relative to the oriented boundary of the fiber neighborhood in $E(n)$ on which we performed the Fintushel–Stern construction, and hence are -1 relative to ∂Y. However, the circles $C^*_{\pm 1}$ lie on opposite faces of Y (with I coordinate ± 1), which inherit opposite orientations from Q. Thus, the framing coefficients are ∓ 1 relative to Q. To construct a surgery diagram of Q, we cap off Σ to get an embedding $Q=\Sigma\times S^1\subset T^2\times S^1=T^3$, with the latter exhibited as 0–surgery on the Borromean rings B. To recover Q, we remove its complementary solid torus in T^3. This has the effect of undoing one Dehn filling, leaving one component of B unfilled. The curves $\partial\Sigma\times\{\theta\}$ correspond to canonical longitudes of this drilled-out link component, and $\{p\}\times S^1$ is a meridian of it. The surgery curves $C_{\pm 1}\times\{\theta_{\pm 1}\}$ are then ∓ 1–framed meridians of the other two components. Blowing down changes the unfilled curve into a figure-eight knot $\kappa(1,-1)$ in S^3, whose complement is P. Attaching h_0 to $I\times P$ along C^*_0 now gives $C=C(1,1;-1)$, and $\partial\Sigma\times S^1$ is identified with the incompressible torus boundary of the figure-eight complement inside ∂C, with f twisting longitudinally as required.

Now that we have realized $C(1,1;-1)$ as the cork C in Theorem 1.2, using 4–manifolds X_k generated from $E(n)$ (so irreducible except for X_0 when $n=1$), we can easily realize any $C(r,s;m)$ with $r,s>0>m$ by giving up irreducibility: Just blow up points on the cores of the handles h_i to suitably lower their framings (as measured in $E(n)$). This replaces the original manifolds X_k by their $(r+s+|m|-3)$–fold blowups, which remain pairwise nondiffeomorphic. To realize $m>0$, simply reverse the orientation on each X_k. Retaining irreducibility is no harder when the integers r, s and m are all odd. Simply choose n large enough that $E(n)$ contains $\frac{1}{2}(r+s+|m|-3)$ disjoint spheres of square -2 avoiding the submanifolds used in our construction. Tubing the 2–handle cores into these spheres has the same effect as blowing up, without changing X_k. When the integers r, s and $|m|$ are also allowed to be even but not 2, we need an additional trick. For $n\geq 3$ we locate an $E(2)$ fiber-summand in X_0 away from the construction site, then cut it out and reglue it by a cyclic permutation of the circle factors of the boundary 3–torus. This modifies the manifolds X_k so that they each contain three (and more) disjoint spheres of square -3, made from sections of $E(2)$ by capping off with vanishing cycles of $E(n-2)$. Using these along with our previous even spheres allows us to realize any positive values of r, s and $|m|$ except 2. The manifolds remain pairwise nondiffeomorphic by a useful result of Sunukjian [18]. (This shows that manifolds made by the Fintushel–Stern construction on a given manifold X_0 are distinguished by the associated Alexander polynomials, in spite of subtleties introduced by automorphisms of $\mathbb{Z}[H_2(X)]$.) Irreducibility follows by examining the Seiberg–Witten basic classes. These are all linear combinations of the fiber classes of the two elliptic summands (by Doug Park [16, Corollary 22] for X_0, extended to each X_k by the Fintushel–Stern formula [7]). Thus, all differences of basic classes have square 0. However, if any X_k were reducible, it would split off a negative definite summand carrying a homology class e with square -1. Any basic class c would have nonzero (odd) value on e. By the gluing formula of [11, Theorem 14.1.1], reversing the sign of $\langle c, e\rangle$ would give a new basic class c' with $(c-c')^2$ negative. □

Remarks 2.1 (a) This irreducibility argument misses the case with r, s or $|m|$ equal to 2, for the technical reason that a disjoint sphere of square -1 would mean our starting manifold was reducible. It seems reasonable to conjecture that irreducibility is still attainable by a different method in this case.

(b) Each of our $\mathbb{Z}$–corks $(C(r,s;m), f)$ (for $r,s>0>m$) generates many other similar families of closed manifolds. We can vary the starting manifold X_0 and distinguish the resulting manifolds X_k from each other by Sunukjian's result, then distinguish various families from each other by their Seiberg–Witten invariants. Alternatively, since our construction only uses a single clasp of the knots K_k, we can apply the construction to other families of knots (or links) related by the twisting of a clasp described by Figure 2.

(c) Our corks $C(r,s;m)$ all have Mazur type, built with a single handle of each index 0, 1 and 2. This is because they have the form $C(\kappa,m)$ for a 2–bridge knot κ (namely $\kappa(r,-s)$). The complement P of κ then has a handle decomposition with two 1–handles and a 2–handle, as does $I\times P$. The final 2–handle h_0 of $C(\kappa,m)$ cancels a 1–handle. (See Figure 3 of [13].)

(d) Each of these corks also embeds in the 4–sphere. In fact, the double of any $C(\kappa,m)$ is also obtained from the complement of the spin of κ by filling trivially to get S^4 (for even m) or by Gluck filling (which gives S^4 for all spun knots [12]). We are left with the following question, which can be restated as a problem about certain torus surgeries in S^4:

Question 2.2 Does iterated twisting on these corks in S^4 always give the standard S^4?

References

[1] **S Akbulut**, *An exotic* 4*–manifold*, J. Differential Geom. 33 (1991) 357–361 MR

[2] **S Akbulut**, *An infinite order cork*, preprint (2014) arXiv (withdrawn)

[3] **S Akbulut**, **D Ruberman**, *Absolutely exotic compact* 4*–manifolds*, Comment. Math. Helv. 91 (2016) 1–19 MR

[4] **S Akbulut**, **K Yasui**, *Cork twisting exotic Stein* 4*–manifolds*, J. Differential Geom. 93 (2013) 1–36 MR

[5] **D Auckly**, **H J Kim**, **P Melvin**, **D Ruberman**, *Equivariant corks*, preprint (2016) arXiv

[6] **C L Curtis**, **M H Freedman**, **W C Hsiang**, **R Stong**, *A decomposition theorem for h–cobordant smooth simply-connected compact* 4*–manifolds*, Invent. Math. 123 (1996) 343–348 MR

[7] **R Fintushel**, **R J Stern**, *Knots, links, and* 4*–manifolds*, Invent. Math. 134 (1998) 363–400 MR

[8] **R Fintushel**, **R J Stern**, *Double node neighborhoods and families of simply connected* 4*–manifolds with* $b^+ = 1$, J. Amer. Math. Soc. 19 (2006) 171–180 MR

[9] **M H Freedman**, *The topology of four-dimensional manifolds*, J. Differential Geom. 17 (1982) 357–453 MR

[10] **M H Freedman**, **F Quinn**, *Topology of* 4*–manifolds*, Princeton Mathematical Series 39, Princeton Univ. Press (1990) MR

[11] **K A Frøyshov**, *Compactness and gluing theory for monopoles*, Geom. Topol. Monogr. 15, Geom. Topol. Publ., Coventry (2008) MR

[12] **H Gluck**, *The embedding of two-spheres in the four-sphere*, Trans. Amer. Math. Soc. 104 (1962) 308–333 MR

[13] **R E Gompf**, *Infinite order corks via handle diagrams*, preprint (2016) arXiv To appear in Algebr. Geom. Topol.

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

[15] **R Matveyev**, *A decomposition of smooth simply-connected h–cobordant* 4*–manifolds*, J. Differential Geom. 44 (1996) 571–582 MR

[16] **B D Park**, *A gluing formula for the Seiberg–Witten invariant along* T^3, Michigan Math. J. 50 (2002) 593–611 MR

[17] **A Ray**, **D Ruberman**, 4*–dimensional analogues of Dehn's lemma*, preprint (2016) arXiv

[18] **N S Sunukjian**, *A note on knot surgery*, J. Knot Theory Ramifications 24 (2015) art. id. 1520003, 5 pages MR

[19] **M Tange**, *Finite order corks*, preprint (2016) arXiv

[20] **M Tange**, *Non-existence theorems on infinite order corks*, preprint (2016) arXiv

[21] **M Tange**, *Notes on Gompf's infinite order cork*, preprint (2016) arXiv

Department of Mathematics, The University of Texas at Austin
RLM 8.100, 2515 Speedway Stop C1200, Austin, TX 78712-1202, United States

gompf@math.utexas.edu

Proposed: Rob Kirby Received: 4 April 2016
Seconded: Ciprian Manolescu, András I. Stipsicz Accepted: 19 September 2016

msp

Geometry & Topology 21 (2017) 2485–2526

Existence of minimizing Willmore Klein bottles in Euclidean four-space

PATRICK BREUNING
JONAS HIRSCH
ELENA MÄDER-BAUMDICKER

Let $K = \mathbb{R}P^2 \sharp \mathbb{R}P^2$ be a Klein bottle. We show that the infimum of the Willmore energy among all immersed Klein bottles $f\colon K \to \mathbb{R}^n$ for $n \geq 4$ is attained by a smooth embedded Klein bottle. We know from work of M W Hirsch and W S Massey that there are three distinct regular homotopy classes of immersions $f\colon K \to \mathbb{R}^4$, each one containing an embedding. One is characterized by the property that it contains the minimizer just mentioned. For the other two regular homotopy classes we show $\mathcal{W}(f) \geq 8\pi$. We give a classification of the minimizers of these two regular homotopy classes. In particular, we prove the existence of infinitely many distinct embedded Klein bottles in $\mathbb{R}^4$ that have Euler normal number -4 or $+4$ and Willmore energy 8π. The surfaces are distinct even when we allow conformal transformations of $\mathbb{R}^4$. As they are all minimizers in their regular homotopy class, they are Willmore surfaces.

53C42; 53C28, 53A07

1 Introduction

For a two-dimensional manifold Σ immersed into $\mathbb{R}^n$ via $f\colon \Sigma \to \mathbb{R}^n$, the Willmore energy is defined as

$$\mathcal{W}(f) := \frac{1}{4} \int |H|^2 \, d\mu_g,$$

where H is the mean curvature vector of the immersed surface, ie the trace of the second fundamental form. Integration is due to the area measure with respect to the induced metric $g = f^* \delta_{\mathrm{eucl}}$.

In this paper, we consider closed nonorientable manifolds Σ of (nonorientable) genus $p = 1, 2$, ie our surfaces are of the type of $\mathbb{R}P^2$ or $K := \mathbb{R}P^2 \sharp \mathbb{R}P^2$ (a Klein bottle). We are interested in the existence and the properties of immersions $f\colon \Sigma \to \mathbb{R}^n$ for $n \geq 4$ that are regularly homotopic to an embedding and that have low Willmore energy.

Published: 19 May 2017 DOI: 10.2140/gt.2017.21.2485

Concerning a lower bound on the Willmore energy, a result of Li and Yau [19] is very useful for closed surfaces immersed into $\mathbb{R}^n$: Let $x \in \mathbb{R}^n$ be a point and $\theta(x) := |\{z \in \Sigma : f(z) = x\}|$ the (finite) number of distinct preimages of x. Then

$$\mathcal{W}(f) \geq 4\pi\theta(x).$$

As any immersed $\mathbb{R}P^2$ in $\mathbb{R}^3$ has at least one triple point (see Banchoff [3]) it follows that $\mathcal{W}(f) \geq 12\pi$ for any such immersion. Equality holds for example for Boy's surface; see Kusner [11]. Similarly, as an immersed Klein bottle in $\mathbb{R}^3$ must have double points we have that $\mathcal{W}(f) \geq 8\pi$ for such immersions. Kusner conjectured that Lawson's minimal Klein bottle in $\mathbb{S}^3$ is (after inverse stereographic projection) the minimizer of the Willmore energy for all Klein bottles immersed into $\mathbb{R}^3$; see Kusner [11] and Lawson [18]. This immersion has energy about 9.7π.

Since any m–dimensional manifold can be embedded into $\mathbb{R}^{2m}$ (Whitney [28]) it is natural to ask what is known about real projective planes and Klein bottles immersed into $\mathbb{R}^4$. Li and Yau [19] showed that $\mathcal{W}(f) \geq 6\pi$ for any immersed $\mathbb{R}P^2$ in $\mathbb{R}^4$, and equality holds if and only if the immersion is the Veronese. It turns out that the Veronese embedding and the reflected Veronese embedding are representatives of the only two distinct regular homotopy classes of immersions containing an embedding. The number of regular homotopy classes is due to Whitney and Massey (see Massey [21]) and Hirsch [9]; see Section 3.

As in the case of $\mathbb{R}P^2$, we can count the number of distinct regular homotopy classes of immersions of a Klein bottle containing an embedding. There are three of them. By a gluing construction of Bauer and Kuwert there is a Klein bottle embedded in $\mathbb{R}^4$ with Willmore energy strictly less than 8π; see Bauer and Kuwert [4, Theorem 1.3]. We repeat parts of this gluing construction in Section 4 and conclude that this gives a Klein bottle in the regular homotopy class characterized by Euler normal number zero. As we can add arbitrary dimensions this construction yields an embedded Klein bottle $f\colon K \to \mathbb{R}^n$ for $n \geq 4$ with $W(f) < 8\pi$. It follows that the infimum of the Willmore energy among all immersed Klein bottles is less than 8π. E Kuwert and Y Li proved in [14] a compactness theorem for so-called $W^{2,2}$–conformal immersions and a theorem about the removability of point singularities. With these methods we prove that the infimum among immersed Klein bottles is attained by an embedding. We know that the minimizer is smooth by the work of T Rivière [24; 25]. Note that Rivière proved independently a compactness result similar to the one of Kuwert and Li mentioned above; see [25, Theorem III.1].

The existence of the minimizer among immersed Klein bottles gives a partial answer to a question that was stated by F Marques and A Neves [20, Section 4]: they asked about the infimum of the Willmore energy in $\mathbb{R}^3$ or $\mathbb{R}^4$ among all nonorientable surfaces of a given genus or among all surfaces in a given regular homotopy class and they asked whether it is attained. Here is the first existence result:

Theorem 1.1 *Let S be the class of all immersions $f\colon \Sigma \to \mathbb{R}^n$ where Σ is a Klein bottle. Consider*

$$\beta_2^n := \inf\{\mathcal{W}(f) : f \in S\}.$$

Then we have that $\beta_2^n < 8\pi$ for $n \geq 4$. Furthermore, β_2^n is attained by a smooth embedded Klein bottle for $n \geq 4$.

We want to point out that the upper bound $\beta_2^n < 8\pi$ can be improved. Let $\tilde{\tau}_{3,1}$ be the bipolar surface of Lawson's $\tau_{3,1}$–torus [18]. It is an embedded minimal Klein bottle in $\mathbb{S}^4$. After stereographic projection, one obtains a Klein bottle $f\colon K \to \mathbb{R}^4$ with Willmore energy $\mathcal{W}(f) = 6\pi\,\mathrm{E}\big(\frac{2\sqrt{2}}{3}\big)$; see Lapointe [16]. Here, $\mathrm{E}(\cdot)$ is the complete elliptic integral of the second kind. We conclude that $\beta_2^n \leq 6\pi\,\mathrm{E}\big(\frac{2\sqrt{2}}{3}\big) \approx 6.682\pi < 8\pi$. There is some indication that $\tilde{\tau}_{3,1}$ is the actual minimizer among immersed Klein bottles in $\mathbb{R}^4$; compare the forthcoming paper of Hirsch and Mäder-Baumdicker [8].

We will show in Section 3 that immersions in one of the other two regular homotopy classes of immersed Klein bottles in $\mathbb{R}^4$ satisfy $\mathcal{W}(f) \geq 8\pi$. There are minimizing representative embeddings $f_i\colon K \to \mathbb{R}^4$ for $i = 1, 2$ with Euler normal number -4 for f_1 and $+4$ for f_2 (for the definition of the Euler normal number, see Section 3).

We prove the following:

Theorem 1.2 *There is a one-parameter family of smooth embedded Klein bottles $f_i^r\colon K \to \mathbb{R}^4$ for $i = 1, 2$ and $r \in \mathbb{R}^+$, with $W(f_i^r) = 8\pi$ for $i = 1, 2$. The embeddings f_1^r have Euler normal number $e(\nu) = -4$. The oriented double cover of the surfaces $\tilde{f}_1^r\colon M_r \to \mathbb{R}^4$ are conformal, where $M_r = \mathbb{C}/\Gamma_r$ is the torus generated by $(1, i\,r)$. Furthermore, the $\tilde{f}_1^r$ are twistor holomorphic. The second embeddings f_2^r are obtained by reflecting $f_1^r(K)$ in $\mathbb{R}^4$, and they have Euler normal number $+4$. Every embedding f_1^r and f_2^r is a minimizer of the Willmore energy in its regular homotopy class. Thus, all discovered surfaces are Willmore surfaces.*

For $r \neq r'$ the surfaces $f_1^r(K)$ and $f_1^{r'}(K)$ are different in the following sense: for all conformal transformations Φ of $\mathbb{R}^4$ we have $f_1^r(K) \neq \Phi \circ f_1^{r'}(K)$ for $r \neq r'$.

Furthermore, there is a classification (including a concrete formula) of immersed Klein bottles in $\mathbb{R}^4$ that satisfy $\mathcal{W}(f) = 8\pi$ and $|e(\nu)| = 4$.

Our techniques can also be used for any $\mathbb{R}P^2$ with $W(f) = 6\pi$. As such a surface must be a conformal transformation of the Veronese embedding (Li and Yau [19]) we get an explicit formula for this surface:

Proposition 1.3 *Define* $f\colon \mathbb{S}^2 \to \mathbb{C}^2 = \mathbb{R}^4$ *by*

$$f(z) = \left(\overline{z}\frac{|z|^4-1}{|z|^6+1}, z^2\frac{|z|^2+1}{|z|^6+1}\right).$$

Then $f(\mathbb{S}^2)$ *is the Veronese surface (up to conformal transformation of* $\mathbb{R}^4$ *).*

We give an overview of the structure of this paper. In Section 2 we prove that each torus carrying an antiholomorphic involution without fixpoints is biholomorphically equivalent to a torus T with a rectangular lattice generated by $(1,\tau)$. On T, the involution has the form $I(z) = \overline{z} + \frac{1}{2}$ up to Möbius transformations on T. Section 3 contains the proof in the nonorientable case of the so-called "Wintgen inequality", $\mathcal{W}(f) \geq 2\pi\big(\chi + |e(\nu)|\big)$; see Wintgen [29]. We then give an introduction to the theory of twistor holomorphic immersions into $\mathbb{R}^4$ (see Friedrich [7]) and construct the surfaces of Theorem 1.2 with this theory. The same methods yield the formula for the Veronese embedding. We explain in Section 4 that the gluing construction of Bauer and Kuwert [4] gives an embedded Klein bottle $f\colon K \to \mathbb{R}^n$ for $n \geq 4$ with Willmore energy strictly less that 8π (thus, with Euler normal number zero if $n = 4$). This embedding is not in one of the regular homotopy classes of the embeddings of Theorem 1.2. After this, we show that a sequence of Klein bottles $f_k\colon K \to \mathbb{R}^n$ where the oriented double covers diverge in moduli space satisfies

$$\liminf_{k\to\infty} W(f_k) \geq 8\pi.$$

We use this estimate together with techniques and results from Kuwert and Li [14] and Rivière [24; 25] to show Theorem 1.1.

Remark In $\mathbb{R}^3$, there is no immersed Klein bottle with Willmore energy 8π. If it existed then we could invert at one of the double points in $\mathbb{R}^3$. We would get a complete minimal immersion in $\mathbb{R}^3$ with two ends. But due to Kusner [11] this surface must be embedded, a contradiction.

Acknowledgment First of all we want to thank Ernst Kuwert for the initial idea. He posed the existence of Willmore-minimizing Klein bottles as an open problem and proposed the gluing of two Veronese surfaces as a first step to obtain a competitor below 8π. Furthermore, he gave us a sketch of the proof of Theorem 1.1. Secondly we want to thank Tobias Lamm for drawing our attention to this problem and for helpful discussions. Mäder-Baumdicker is funded by the DFG in the project "Willmore surfaces in Riemannian manifolds".

2 Antiholomorphic involutions on the torus

Let N be a nonorientable manifold of dimension two and $\widetilde{f}\colon N \to \mathbb{R}^n$ $(n \geq 3)$ an immersion. We equip N with the induced Riemannian metric $\widetilde{f}^*\delta_{\mathrm{eucl}}$. Consider $q\colon M \to N$, the conformal oriented two-sheeted cover of N, and define $f := \widetilde{f} \circ q$. As every $2-$dimensional oriented manifold can be locally conformally reparametrized, M is a Riemann surface that is conformal to $(M, f^*\delta_{\mathrm{eucl}})$. Let $I\colon M \to M$ be the antiholomorphic order-two deck transformation for q. The map I is an antiholomorphic involution without fixpoints such that $f \circ I = f$.

Now consider the situation where N is the Klein bottle, ie N is compact, without boundary and has nonorientable genus two. In this case, the oriented two-sheeted cover $q\colon T^2 \to N$ lives on the two-dimensional torus T^2. It is the aim of this section to classify all antiholomorphic involutions without fixpoints on a torus T^2 up to Möbius transformation. A Möbius transformation is a biholomorphic map $\varphi\colon T^2 \to T^2$. We use the fact that every torus is a quotient space $\mathbb{C}/\Gamma$, where Γ is a lattice in $\mathbb{C}$, ie

$$\Gamma = \{m\omega + n\omega' : m, n \in \mathbb{Z}\},$$

where ω, $\omega' \in \mathbb{C} = \mathbb{R}^2$ are vectors that are linearly independent over $\mathbb{R}$. We call (ω, ω') a *generating pair* of Γ.

Theorem 2.1 *Consider a lattice Γ in $\mathbb{C}$ generated by a pair $(1, \tau)$ where $\Im(\tau) > 0$, $-\frac{1}{2} < \Re(\tau) \leq \frac{1}{2}$ and $|\tau| \geq 1$. Let $I\colon \mathbb{C}/\Gamma \to \mathbb{C}/\Gamma$ be an antiholomorphic involution without fixpoints. Then Γ must be a rectangular lattice, ie $\tau \in i\mathbb{R}^+$, and, up to Möbius transformation, the induced doubly periodic map $\widehat{I}\colon \mathbb{C} \to \mathbb{C}$ is of the form*

$$\widehat{I}(z) = \overline{z} + \tfrac{1}{2} \quad \text{or} \quad \widehat{I}(z) = -\overline{z} + \tfrac{1}{2}\tau.$$

Remarks (i) A similar result can be found in [13, Appendix F]. For the sake of completeness we give a full proof of Theorem 2.1 in the following. The case that Γ is a hexagonal lattice, ie generated by $(1, e^{i\pi/3})$, and $\alpha\Gamma = \overline{\Gamma}$ with $\alpha = e^{li\pi/3}$ for $l = 1, 2, 4, 5$ is not considered in the proof of [13].

(ii) The expression "up to Möbius transformation" means that there is a Möbius transformation $\varphi\colon \mathbb{C}/\Gamma \to \mathbb{C}/\Gamma$ such that $\varphi^{-1} \circ I \circ \varphi$ is of the claimed form. If I is an antiholomorphic involution without fixpoints on a torus $\mathbb{C}/\Gamma$ then $\varphi^{-1} \circ I \circ \varphi$ is also an antiholomorphic involution without fixpoints on that torus. Therefore, it only makes sense to classify such involutions up to Möbius transformation.

(iii) Every map $I\colon \mathbb{C}/\Gamma \to \mathbb{C}/\Gamma$ induces a map $\widehat{I}\colon \mathbb{C} \to \mathbb{C}$ that is doubly periodic with respect to Γ. From now on we denote $\widehat{I}$ simply by I.

We prove this theorem in several steps. But first we explain how we come to the case of a general lattice.

Proposition 2.2 *Let Γ be a lattice in $\mathbb{C}$. Then there exists a generating pair (ω, ω') such that $\tau := \omega'/\omega$ satisfies $\Im(\tau) > 0$, $-\frac{1}{2} < \Re(\tau) \le \frac{1}{2}$, $|\tau| \ge 1$ and, if $|\tau| = 1$, then $\Re(\tau) \ge 0$. Let $\widetilde{\Gamma}$ be the lattice generated by $(1, \tau)$. Then there exists a biholomorphic map $\varphi\colon \mathbb{C}/\Gamma \to \mathbb{C}/\widetilde{\Gamma}$.*

Proof The pair (ω, ω') is sometimes called the "canonical basis". The proof of the existence of this basis can be found in [1, Chapter 7, Theorem 2]. For the biholomorphic map we define $\widetilde{\varphi}(z) := z/\omega$ for $z \in \mathbb{C}$. Then $\varphi([z]) := \widetilde{\varphi}(z)$ for $[z] \in \mathbb{C}/\Gamma$ defines a biholomorphic map $\varphi\colon \mathbb{C}/\Gamma \to \mathbb{C}/\widetilde{\Gamma}$ □

Lemma 2.3 *Let Γ be a lattice in $\mathbb{C}$ and $I\colon \mathbb{C}/\Gamma \to \mathbb{C}/\Gamma$ an antiholomorphic involution. Then I is of the form $I(z) = a\overline{z} + b$, where $a, b \in \mathbb{C}$ with $a\overline{\Gamma} = \Gamma$, $|a| = 1$ and $a\overline{b} + b \in \Gamma$. Here, $\overline{\Gamma}$ is the complex conjugation of Γ.*

Proof Define $\psi(z) := I(\overline{z})$. Notice $\psi\colon \mathbb{C}/\overline{\Gamma} \to \mathbb{C}/\Gamma$ is holomorphic. Let Γ be generated by (τ_1, τ_2). The derivative $\psi'\colon \mathbb{C} \to \mathbb{C}$ is holomorphic and bounded on the compact fundamental domain $F := \{t_1\tau_1 + t_2\tau_2 : 0 \le t_1, t_2 \le 1\}$. The periodicity of ψ' implies that it is bounded in all of $\mathbb{C}$. By Liouville's theorem we get that $\psi' = a$ for some $a \in \mathbb{C}$. Therefore, we have that $\psi(z) = az + b$ for some $b \in \mathbb{C}$. Since $I\colon \mathbb{C}/\Gamma \to \mathbb{C}/\Gamma$, we have that

$$\Gamma \ni I(z+\omega) - I(z) = \psi(\overline{z} + \overline{\omega}) - \psi(\overline{z}) = a\overline{\omega} \quad \text{for all } \omega \in \Gamma,$$

which implies $a\overline{\Gamma} \subset \Gamma$. For the other implication we use that ψ is one-to-one (if restricted to the fundamental domain F). The map $\Phi := \overline{I}$ is an inverse of ψ because $\overline{I} \circ \psi(z) = \overline{I} \circ I(\overline{z}) = z \bmod \overline{\Gamma}$ and $\psi \circ \overline{I}(z) = I(I(z)) = z \bmod \Gamma$. The same argument as above implies that there are complex numbers $c, d \in \mathbb{C}$ such that $\Phi(z) = cz + d$. So we have that

$$\overline{\Gamma} \ni \overline{I}(z+\omega) - \overline{I}(z) = \Phi(z+\omega) - \Phi(z) = c\omega \quad \text{for all } \omega \in \Gamma,$$

which implies $c\Gamma \subset \overline{\Gamma}$. We get that

$$\mathrm{id}|_{\mathbb{C}/\Gamma}(z) = \psi \circ \Phi(z) = acz + ad + b,$$

which implies $ac = 1$ and $\frac{1}{a}\Gamma \subset \overline{\Gamma}$. □

Lemma 2.4 *Let Γ be a lattice in $\mathbb{C}$ generated by $(1, \tau)$ with $\Im(\tau) > 0$. Then all Möbius transformations $\varphi\colon \mathbb{C}/\Gamma \to \mathbb{C}/\Gamma$ are of the form $\varphi(z) = \alpha z + \delta$ with $\delta \in \mathbb{C}$ and:*

(i) *If* $\tau = i$ *(quadratic lattice) then* $\alpha \in \{1, -1, i, -i\}$.

(ii) *If* $\tau = e^{i\pi/3}$ *or* $\tau = e^{2i\pi/3}$ *(hexagonal lattice) then* $\alpha \in \{e^{li\pi/3} : l = 1, \dots, 6\}$.

(iii) *If* Γ *is neither the quadratic lattice nor the hexagonal lattice then* $\alpha \in \{1, -1\}$.

Proof First we note that a translation $\varphi(z) = z + \delta$ for a $\delta \in \mathbb{C}$ is always a Möbius transformation. Therefore, we assume that $\varphi(0) = 0$ (by composing with a translation). The rest of the proof can be found in [23, Chapter III, Proposition 1.12.]. $\square$

Lemma 2.5 *Let* Γ *be a lattice in* $\mathbb{C}$ *generated by* $(1, \tau)$ *with* $\Im(\tau) > 0$ *and* $|\tau| = 1$. *Let* I *be an antiholomorphic involution on* $\mathbb{C}/\Gamma$ *of the form* $I(z) = a\overline{z} + b$ *with* $a \in \{+\tau, -\tau\}$. *Then* I *has a fixpoint.*

Proof Let $\varphi(z) = z + \delta$ be a translation on $\mathbb{C}/\Gamma$. We have that

$$\varphi^{-1} \circ I \circ \varphi(z) = a(\overline{z} + \overline{\delta}) + b - \delta = a\overline{z} + b + a\overline{\delta} - \delta.$$

Consider now a translation with $\delta \in \mathbb{R}$. Then we have that

$$\widetilde{I}(z) := \varphi^{-1} \circ I \circ \varphi(z) = \pm\tau\overline{z} + (\pm\tau - 1)\delta + b.$$

By $\Im(\pm\tau - 1) \neq 0$ we can choose $\delta \in \mathbb{R}$ such that $(\pm\tau - 1)\delta + b \in \mathbb{R}$. Hence by passing from I to $\widetilde{I}$ we can assume that the involution is of the form $I(z) = \pm\tau\overline{z} + b$ with $b \in \mathbb{R}$. Lemma 2.3 implies that $\pm\tau b + b \in \Gamma$. Since $(1, \tau)$ is the generating pair of Γ we get that $b \in \mathbb{Z}$. But $I(z) = \pm\tau\overline{z} + n$ with $n \in \mathbb{Z}$ has the fixpoint 0. Then the original involution also had a fixpoint. $\square$

Lemma 2.6 *Let* Γ *be a lattice in* $\mathbb{C}$ *generated by* $(1, \tau)$ *with* $\Im(\tau) > 0$ *and* $|\tau| = 1$. *Let* I *be an antiholomorphic involution on* $\mathbb{C}/\Gamma$ *of the form* $I(z) = a\overline{z} + b$ *with* $a \notin \{+1, -1\}$. *Then* I *has a fixpoint.*

Proof Since a satisfies $a\overline{\Gamma} = \Gamma$ and $|a| = 1$ (see Lemma 2.3) we want to know how many lattice points lie on the unit circle $\mathbb{S}^1$. There are two cases.

Case 1 ($\tau - 1 \notin \mathbb{S}^1$) But $|\tau - 1|^2 \neq 1$ is here equivalent to $\Re(\tau) \neq \frac{1}{2}$ since $|\tau - 1|^2 = 2 - 2\Re(\tau)$. Therefore we know that Γ cannot be the hexagonal lattice and there are exactly four lattice points on $\mathbb{S}^1$, namely 1, -1, τ and $-\tau$. Since $a\overline{\Gamma} = \Gamma$ and $1 \in \overline{\Gamma}$ we have that $a \in \Gamma \cap \mathbb{S}^1$, which implies $a \in \{1, -1, \tau, -\tau\}$. But we assumed $a \notin \{+1, -1\}$, and $a \in \{\tau, -\tau\}$ implies that I has a fixpoint by the previous lemma.

Case 2 ($\tau - 1 \in \mathbb{S}^1$) This corresponds to the hexagonal lattice, $\tau = e^{i\pi/3}$. There are six lattice points lying on $\mathbb{S}^1$, namely $e^{li\pi/3}$ for $l = 1, \dots, 6$. Again, as in the first case, we have that $a \in \Gamma \cap \mathbb{S}^1$. The cases $l = 3$ and $l = 6$ are not possible by assumption, therefore we get that $a \in \{\tau^l : l = 1, 2, 4, 5\}$. Now consider a Möbius transformation of the hexagonal lattice $\varphi(z) = \alpha z$ with $\alpha \neq 0$. Lemma 2.4 yields $\overline{\alpha} \in \{\tau^k : k = 1, \dots, 6\}$. We compose

$$\widetilde{I}(z) := \varphi^{-1} \circ I \circ \varphi(z) = \frac{\overline{\alpha}}{\alpha} a\overline{z} + \frac{b}{\alpha} = \tau^{2k+l}\overline{z} + \overline{\alpha} b.$$

If l is even, then we choose k such that $2k + l = 6$. Thus, we are in the case $a = 1$. If $l = 5$ then we compose with the Möbius transformation $\varphi(z) = \alpha z$, where $\alpha = \tau^4$ (which is equivalent to $k = -2$). We have then reduced it to the case $a = \tau$, which is Lemma 2.5. □

Lemma 2.7 *Consider a lattice Γ in $\mathbb{C}$ generated by a pair $(1, \tau)$ with $\Im(\tau) > 0$, $-\frac{1}{2} < \Re(\tau) \leq \frac{1}{2}$ and $|\tau| > 1$. Let $I\colon \mathbb{C}/\Gamma \to \mathbb{C}/\Gamma$ be an antiholomorphic involution. Then we have that $I(z) = a\overline{z} + b$ with $a \in \{-1, 1\}$.*

Proof By Lemma 2.3 we know that $a\overline{\Gamma} = \Gamma$ and $|a| = 1$. Hence $a \in \mathbb{S}^1 \cap \Gamma$. We claim that $\mathbb{S}^1 \cap \Gamma = \{-1, 1\}$. Since $|\tau| > 1$, we know that $\pm\tau \notin \mathbb{S}^1 \cap \Gamma$. But then we only have to consider the case that $z \in \mathbb{S}^1 \cap \Gamma$ is of the form $z = -1 + l\tau$ for an $l \in \mathbb{Z} \setminus \{0\}$. We use the assumptions on τ and get

$$|-1 + l\tau|^2 = 1 + l^2|\tau|^2 - 2l\Re(\tau) > 1 + l^2 - l \geq 1.$$

This strict inequality shows the lemma. □

Definition 2.8 A lattice Γ in $\mathbb{C}$ is called a *real lattice* if it is stable under complex conjugation, ie $\overline{\Gamma} = \Gamma$.

Lemma 2.9 *Let Γ be a real lattice generated by $(1, \tau)$ with $-\frac{1}{2} < \Re(\tau) \leq \frac{1}{2}$. Then we have that $\Re(\tau) \in \{0, \frac{1}{2}\}$.*

Proof Let $\tau = x + iy$. Then there are $m, n \in \mathbb{Z}$ such that

$$\begin{aligned} \overline{\tau} = x - iy = n + m(x + iy) \quad &\iff \quad x - n - mx - i(my + y) = 0 \\ &\iff \quad m = -1 \quad \text{and} \quad x(1 - m) = n. \end{aligned}$$

This implies that $\Re(\tau) = x \in \left(-\frac{1}{2}, \frac{1}{2}\right] \cap \left\{\frac{1}{2}n : n \in \mathbb{Z}\right\} = \left\{0, \frac{1}{2}\right\}$. □

Lemma 2.10 *Let Γ be a lattice generated by $(1, \tau)$ with $-\frac{1}{2} < \Re(\tau) \leq \frac{1}{2}$ and let $I(z) = a\overline{z} + b$ be an antiholomorphic involution with $a = -1$. Then the lattice is real and $\Re(\tau) \in \{0, \frac{1}{2}\}$. If $\Re(\tau) = \frac{1}{2}$ then I has a fixpoint. If $\Re(\tau) = 0$ then $I(z) = -\overline{z} + \frac{1}{2}\tau$ (up to Möbius transformation) and I has no fixpoints.*

Proof As every lattice satisfies $-\Gamma = \Gamma$ we have by Lemma 2.3 that $\overline{\Gamma} = -\overline{\Gamma} = \Gamma$, ie the lattice is real. The previous lemma yields $\Re(\tau) \in \{0, \frac{1}{2}\}$.

Case 1 $\left(\Re(\tau) = \frac{1}{2}\right)$ We note that

$$i\mathbb{R} \cap \Gamma = \{2mi\Im(\tau) : m \in \mathbb{Z}\}. \tag{1}$$

By composing I with a translation we can assume that $b \in i\mathbb{R}$: Consider the translation $\varphi(z) = z + \delta$; then

$$\widetilde{I}(z) := \varphi^{-1} \circ I \circ \varphi(z) = -\overline{z} + b - \overline{\delta} - \delta = -\overline{z} + b - 2\Re(\delta).$$

Thus, we can subtract the real part of b and consider $\widetilde{I}$ instead of I.

But from $a\overline{b} + b \in \Gamma$ (see Lemma 2.3) and (1) we have that $2b = -\overline{b} + b \in \Gamma$ and $2b = 2mi\Im(\tau)$ for an $m \in \mathbb{Z}$. Composing the involution with another translation yields that $b = m\tau$ for an $m \in \mathbb{Z}$. Hence $I(z) = -\overline{z} + m\tau$, which has the fixpoint 0.

Case 2 $(\Re(\tau) = 0)$ Here, Γ is a rectangular lattice. By translation, as in the first case, we assume $b \in i\,\mathbb{R}$. Therefore, we get that $-\overline{b} + b = 2b \in \Gamma \cap i\,\mathbb{R} = \{m\tau : m \in \mathbb{Z}\}$, hence $b = \frac{1}{2}m\tau$ for an $m \in \mathbb{Z}$. Observe that m cannot be even because otherwise I would have a fixpoint. As the formula for I is only defined modulo Γ, we have that $I(z) = -\overline{z} + \frac{1}{2}\tau$. We only have to show that this I has no fixpoint; an equality like

$$I(z) - z = -\overline{z} - z + \tfrac{1}{2}\tau = -2\Re(z) + \tfrac{1}{2}\tau = n + m\tau$$

cannot hold for numbers $m, n \in \mathbb{Z}$ because τ is purely imaginary. □

Lemma 2.11 *Let Γ be a lattice generated by $(1, \tau)$ with $-\frac{1}{2} < \Re(\tau) \le \frac{1}{2}$ and let $I(z) = \overline{z} + b$ be an antiholomorphic involution. Then, up to Möbius transformation, I is of the form $I(z) = \overline{z} + \frac{1}{2}$ and the lattice satisfies $\Re(\tau) \in \{0, \frac{1}{2}\}$. If $\Re(\tau) = \frac{1}{2}$ then I has fixpoints*

Proof By composing with a translation $\varphi(z) = z + \delta$ we get

$$\widetilde{I}(z) := \varphi^{-1} \circ I \circ \varphi(z) = \overline{z} + b + \overline{\delta} - \delta = -\overline{z} + b - 2i\Im(\delta).$$

Thus, we can assume that $b \in \mathbb{R}$. Now we have that $2b = \overline{b} + b \in \Gamma \cap \mathbb{R} = \mathbb{Z}$ and therefore $b = \frac{1}{2}m$ for an $m \in \mathbb{Z}$. If m was even, then I would have the fixpoint 0, and since the formula is only defined modulo Γ we have that $b = \frac{1}{2}$. As $a\overline{\Gamma} = \Gamma$ with $a = 1$ we know that the lattice is real and hence satisfies $\Re(\tau) \in \{0, \frac{1}{2}\}$ (see Lemma 2.9). It remains to check in which cases I has fixpoints: Let $m, n \in \mathbb{Z}$. If

$$I(z) - z = \overline{z} - z + \tfrac{1}{2} = -2i\Im(z) + \tfrac{1}{2} = n + m\Re(\tau) + mi\Im(\tau)$$

then $\Im(z) = -\frac{1}{2}m\Im(\tau)$ and $\Re(\tau) = (1-2n)/2m$. Hence, if the real part of τ is an odd number divided by an even number, then I has a whole line of fixpoints, otherwise it has no fixpoints. □

We are now able to prove Theorem 2.1:

Proof Any involution is of the form $I(z) = a\overline{z} + b$ by Lemma 2.3. If $|\tau| > 1$ then Lemma 2.7 implies that $a \in \{1, -1\}$. The case $a = -1$ is Lemma 2.10 and the case $a = 1$ is Lemma 2.11. If $|\tau| = 1$ then we have that $a \in \{-1, 1\}$ by Lemma 2.6. Lemmas 2.10 and 2.11 apply also for this case. □

3 Willmore surfaces of Klein bottle type in $\mathbb{R}^4$ with energy 8π

Let M be a closed manifold of dimension two (orientable or nonorientable) immersed into an oriented 4–dimensional Riemannian manifold (X^4, h). The immersion induces a metric g on M, a connection ∇ on tangential bundle TM and a connection $\nabla^\perp$ on the normal bundle NM. Since TM and NM are both two-dimensional, their curvature operator is determined by scalars. Let $\{E_1, E_2, N_1, N_2\}$ be an orthonormal oriented frame of X^4 in a neighborhood U of $x_0 \in M$ such that E_1, E_2 is a basis for T_xM and N_1, N_2 a basis for N_xM for all $x \in U$. The scalars of interest are the Gauss curvature, given on U by

$$K(x) = R(E_1, E_2, E_2, E_1) = \langle \nabla_{E_1}\nabla_{E_2}E_2 - \nabla_{E_2}\nabla_{E_1}E_2 - \nabla_{[E_1,E_2]}E_2, E_1\rangle,$$

and the trace of the curvature tensor of the normal connection, given on U by

$$K^\perp(x) := \langle R^\perp(E_1, E_2)N_2, N_1\rangle = \langle \nabla^\perp_{E_1}\nabla^\perp_{E_2}N_2 - \nabla^\perp_{E_2}\nabla^\perp_{E_1}N_2 - \nabla^\perp_{[E_1,E_2]}N_2, N_1\rangle.$$

We introduce the connection 1–forms $\{w_i^j\}_{i,j=1,2,3,4}$ given by

$$D_v e_i := w_i^j(v)e_j \quad \text{for } v \in T_xM, \tag{2}$$

where $\{E_1, E_2, N_1, N_2\} = \{e_1, e_2, e_3, e_4\}$ and D is the Levi-Civita connection of X. Classical calculations show that

$$R(X,Y)E_2 = dw_2^1(X,Y)E_1 \quad \text{and} \quad R^\perp(X,Y)N_2 = dw_4^3(X,Y)N_1,$$

hence the definition of K and $K^\perp$ is independent of the orientation of E_1, E_2.

The Weingarten equation relates D to the connection ∇ and the second fundamental form $A(v,w) = (D_v w)^\perp$ for vector fields v and w on M by $D_v w = \nabla_v w + A(v,w)$.

We can express R and $R^\perp$ in terms of the second fundamental form and the curvature operator R^X of the ambient manifold X^4 using A_{ij} for $A(E_i, E_j)$:

$$\begin{aligned} K(x) &= R^X(E_1,E_2,E_2,E_1) + \langle (D_{E_1}E_1)^\perp, (D_{E_2}E_2)^\perp\rangle \\ &\qquad\qquad\qquad\qquad - \langle (D_{E_1}E_2)^\perp, (D_{E_1}E_2)^\perp\rangle \\ &= R^X(E_1,E_2,E_2,E_1) + \langle A_{11}, A_{22}\rangle - \langle A_{12}, A_{12}\rangle. \end{aligned} \tag{3}$$

Similarly, one gets for the normal curvature

$$\begin{aligned} K^\perp(x) &= R^X(E_1,E_2,N_2,N_1) + \langle (D_{E_1}N_1)^\top, (D_{E_2}N_2)^\top\rangle - \langle (D_{E_1}N_2)^\top, (D_{E_2}N_1)^\top\rangle \\ &= R^X(E_1,E_2,N_2,N_1) + \Big(\sum_{j=1,2} \langle A_{1j}, N_1\rangle\langle A_{2j}, N_2\rangle - \langle A_{1j}, N_2\rangle\langle A_{2j}, N_1\rangle\Big). \end{aligned}$$

Observe that the second part can be expressed as $\big\langle \sum_{j=1,2} A_{1j} \wedge A_{2j}, N_1 \wedge N_2\big\rangle$. Introducing the tracefree part $A^\circ_{ij} = A_{ij} - \frac{1}{2}Hg_{ij}$ using $A^\circ_{11} + A^\circ_{22} = 0$ and $A_{12} = A^\circ_{12}$, the equation for $K^\perp$ simplifies to

$$K^\perp(x) = R^X(E_1,E_2,N_2,N_1) + 2\langle A^\circ_{11} \wedge A^\circ_{12}, N_1 \wedge N_2\rangle. \tag{4}$$

Recall that the Euler number of the normal bundle can be expressed similarly to the Gauss–Bonnet formula [22] as

$$e(\nu) = \frac{1}{2\pi}\int_M K^\perp. \tag{5}$$

As a corollary of these calculations we obtain a classical inequality by Wintgen. This inequality was known to be true for oriented surfaces. We extend the result to nonorientable surfaces.

Theorem 3.1 (Wintgen [29]) *Let M be a closed manifold of dimension two* (orientable or nonorientable) *and Euler characteristic χ. Let $f\colon M \to \mathbb{R}^4$ be an immersion and denote by $e(\nu)$ the Euler normal number of f. Then we have that*

$$\mathcal{W}(f) \geq 2\pi\big(\chi + |e(\nu)|\big), \tag{6}$$

and equality holds if and only if

$$|A^\circ_{11}|^2 = |A^\circ_{12}|^2, \langle A^\circ_{11}, A^\circ_{12}\rangle = 0 \quad \textit{and} \quad K^\perp \textit{ does not change sign.} \tag{7}$$

Proof The proof for the orientable case can be found in [29]. Note that in this case $e(\nu) = 2I$ (see [17]), where I is the self-intersection number due to Whitney; see [28]. And we have the equality $\chi = 2 - 2p$, where p is the genus of M.

The general case follows from (3) and (4) and the flatness of $\mathbb{R}^4$. Equality (3) becomes $K = \langle A_{11}, A_{22}\rangle - \langle A_{12}, A_{12}\rangle$ and so $|H|^2 = |A|^2 + 2K$. Together with $|A^\circ|^2 = |A|^2 - \frac{1}{2}|H|^2$, we have

$$\tfrac{1}{2}|H|^2 = 2K + |A^\circ|^2.$$

Equation (4) becomes $K^\perp = 2\langle A^\circ_{11} \wedge A^\circ_{12}, N_1 \wedge N_2\rangle$ and we can estimate

$$\begin{aligned}|K^\perp| &= 2|A^\circ_{11} \wedge A^\circ_{12}||N_1 \wedge N_2| = 2(|A^\circ_{11}|^2|A^\circ_{12}|^2 - \langle A^\circ_{11}, A^\circ_{12}\rangle^2)^{1/2} \\ &\le 2|A^\circ_{11}||A^\circ_{12}| \le |A^\circ_{11}|^2 + |A^\circ_{12}|^2 = \tfrac{1}{2}|A^\circ|^2,\end{aligned}$$

with equality if and only if the first part of (7) holds. Combining both gives

$$\tfrac{1}{2}|H|^2 = 2K + |A^\circ|^2 \ge 2K + 2|K^\perp|.$$

Multiplying by $\frac{1}{2}$ and integrating over M gives

$$\mathcal{W}(f) \ge \int_M K + \int_M |K^\perp| \ge \int_M K + \left|\int_M K^\perp\right|,$$

with equality if and only if $K^\perp$ does not change sign. □

Remark As we are interested in the case $p = 2$, ie $N = \mathbb{R}P^2 \sharp \mathbb{R}P^2$ is a Klein bottle, the inequality above does not give us any information about the Willmore energy in the case $e(\nu) = 0$. But we get information about the energy if the immersion is an embedding, due to the following theorem.

Theorem 3.2 (Whitney, Massey [21]) *Let N be a closed, connected, nonorientable manifold of dimension two with Euler characteristic χ. Consider an embedding $f\colon N \to \mathbb{R}^4$ with Euler normal number $e(\nu)$. Then $e(\nu)$ can take the following values:*

$$-4 + 2\chi,\ \ 2\chi,\ \ 2\chi + 4,\ \ 2\chi + 8,\ \ \ldots,\ \ 4 - 2\chi.$$

Furthermore, any of these possible values is attained by an embedding of N into $\mathbb{R}^4$.

Corollary 3.3 *Let $N = \mathbb{R}P^2 \sharp \mathbb{R}P^2$ be a Klein bottle. Consider an immersion $f\colon N \to \mathbb{R}^4$ that is regularly homotopic to an embedding, and denote by $e(\nu)$ the Euler normal number of f. If $e(\nu) \ne 0$ then $\mathcal{W}(f) \ge 8\pi$.*

Proof By [9, Theorem 8.2], two immersions $f, g\colon N \to \mathbb{R}^4$ are regularly homotopic if and only if they have the same normal class. By assumption, the given immersion f is regularly homotopic to an embedding $g\colon N \to \mathbb{R}^4$. Theorem 3.2 and $\chi(N) = 2 - p = 0$ implies that $e(\nu_f) = e(\nu_g) \in \{-4, 0, 4\}$. As $e(\nu) \ne 0$ we use Theorem 3.1 to see that $\mathcal{W}(f) \ge 8\pi$. □

Remark In the case of genus one, we get from Theorem 3.2 that the Euler normal number of the Veronese embedding $f\colon \mathbb{R}P^2 \to \mathbb{R}^4$ must be $e(\nu) \in \{-2, +2\}$. By the work of Hirsch [9] we get that there are exactly two regular homotopy classes of surfaces of $\mathbb{R}P^2$ type containing an embedding. Each regular homotopy class is represented by a Veronese embedding, one is the reflected surface of the other.

For the construction of immersed Klein bottles with $\mathcal{W}(f) = 8\pi$ and $e(\nu) \in \{-4, +4\}$ we need the theory of twistor holomorphic immersions. They were studied in [7], and we follow that paper.

Definition 3.4 Let (X^4, h) be an oriented, 4–dimensional Riemannian manifold. Consider a point $x \in X^4$ and let P_x be the set of all linear maps $J\colon T_x X^4 \to T_x X^4$ satisfying the following conditions:

(i) $J^2 = -\mathrm{id}$.

(ii) J is compatible with the metric h, ie J is an isometry.

(iii) J preserves the orientation.

(iv) If we define the 2–form $\Omega(t_1, t_2) := h(Jt_1, t_2)$ for t_1, $t_2 \in T_x X^4$, then $-\Omega \wedge \Omega$ equals the given orientation of X^4.

The set $P := \bigcup_{x \in X^4} P_x$ is a $\mathbb{C}P^1$–fiber bundle over X^4 (note $\mathrm{SO}(4)/U(2) \cong \mathbb{C}P^1$). We call P the *twistor space of* X^4 and denote by $\pi\colon P \to X^4$ the projection of the bundle.

Definition 3.5 (the lift of an immersion into the twistor space) Let M be an oriented manifold of dimension two and $f\colon M \to X^4$ an immersion. We decompose the tangent space $T_{f(x)} X^4$ of the ambient manifold into the sum of the tangent space $T_x M$ and the normal space $N_x M$. Let E_1, E_2 be a positively oriented orthonormal basis of $T_x M$ and N_1, N_2 an orthonormal basis of $N_x M$ such that $\{E_1, E_2, N_1, N_2\}$ is a positively oriented basis of $T_{f(x)} X^4$. We define the *lift of the immersion* f by

$$
\begin{aligned}
&F(x)\colon T_{f(x)} X^4 \to T_{f(x)} X^4, \\
&F(x)E_1 = E_2, \qquad F(x)E_2 = -E_1, \\
&F(x)N_1 = -N_2, \quad F(x)N_2 = N_1,
\end{aligned} \tag{8}
$$

ie $F(x)$ is the rotation around the angle $\frac{\pi}{2}$ in the positive (negative) direction on $T_x M$ (on $N_x M$). In this way,[1] $F\colon M \to P$ is a lift of f.

[1] The frame $\{E_1, E_2, N_1, N_2\}$ gives a local bundle chart of the pullback bundle f^*P around x. The defined linear map $F(x)\colon T_{f(x)} X^4 \to T_{f(x)} X^4$ is an element of the fiber $P_{f(x)}$. Hence we can either consider F to be a map into the pullback bundle f^*P that is the identity on M or as a map into P by $\pi \circ F(x) := f(x)$. We follow the classical line and think of F as a map into P.

Definition 3.6 (twistor holomorphic) There exists an almost complex structure Y on P coinciding with the canonical complex structure on the fibers $\mathrm{SO}(4)/U(2) \cong \mathbb{C}P^1$. For a point $J \in P$ the horizontal part $T_J^H P$ is determined by the Levi-Civita connection on X^4 and the complex structure on it is $d\pi^{-1} J\, d\pi$ [7, Section 1]. The pair (P, Y) is a complex manifold if and only if the manifold X^4 is self-dual; see [2]. Let M be an oriented two-dimensional manifold and $f\colon M \to X^4$ an immersion. Denote by $I\colon T_x M \to T_x M$ the complex structure of M with respect to the induced metric f^*h. The immersion f is called *twistor holomorphic* if the lift $F\colon (M, I) \to (P, Y)$ is holomorphic, ie $dF(I(t)) = Y(dF(t))$.

Remark The pair (M, I) from the definition above is a Riemann surface and I only depends on the conformal class of f^*h. The map F has the property that for any conformal coordinates $\varphi\colon U \subset \mathbb{C} \to M$ the map $F \circ \varphi$ is holomorphic. Furthermore, the metric $(f \circ \varphi)^*h$ is conformal to the standard metric on $\mathbb{C}$.

On the other hand, if a the lift $F\colon M \to P$ of a map $f\colon M \to X^4$ from a Riemann surface M has the property that $F \circ \varphi\colon U \to P$ is holomorphic for any conformal coordinates $\varphi\colon U \to M$, it is not hard to check that F is twistor holomorphic, as defined above. In fact, this is the picture we will use in the following.

Remark As we only want to use the construction of twistor spaces for $X^4 = \mathbb{R}^4$ we have more information about the structure of P: using an isomorphism $\mathrm{SO}(4)/U(2) \cong \mathbb{C}P^1 \cong \mathbb{S}^2$, the twistor space P of $\mathbb{R}^4$ is (as a set) the trivial $\mathbb{S}^2-$fiber bundle over $\mathbb{R}^4$, ie $P = \mathbb{R}^4 \times \mathbb{S}^2$ (see [2, Section 4]). On the other hand P carries a holomorphic structure which is not the standard holomorphic structure on $\mathbb{C}^2 \times \mathbb{S}^2$ but a twisted one: If H is the standard positive line bundle over $\mathbb{C}P^1$, then P is isomorphic to $H \oplus H$ (the Whitney sum of H with itself); see [2, Section 4]. This is a bundle over $\mathbb{S}^2$ with projection $p\colon H \oplus H \to \mathbb{S}^2$. Thus, we are in the following situation:

$$\begin{tikzcd} & M \arrow[dl, "p\circ F"'] \arrow[d, "F"] \arrow[dr, "f"] & \\ \mathbb{S}^2 & H \oplus H \cong P \arrow[l, "p"] \arrow[r, "\pi"'] & \mathbb{R}^4 \end{tikzcd}$$

Proposition 3.7 [7] *Let $f\colon M \to X^4$ be an immersion of an oriented two-dimensional manifold M; then the following conditions are equivalent:*

(i) *f is twistor holomorphic.*

(ii) *The connection forms defined above satisfy on every neighborhood U*

$$w_2^4 + w_1^3 - \star w_2^3 + \star w_1^4 = 0,$$

where $\star$ is the Hodge star operator with respect to the induced metric f^*h.

(iii) *For all* $x \in U$,

$$F(x)A^\circ_{11}(x) = A^\circ_{12}(x).$$

Proof Although this proposition corresponds to [7, Proposition 2] we give here a direct proof. Fix a point x and choose an orthonormal frame $\{E_1, E_2, N_1, N_2\}$ in a neighborhood U as in Definition 3.5. As described in Definition 3.5 the lift corresponds to a matrix $F(y) \in \mathrm{SO}(4)$ for all $y \in U$. By definition of $F(y)$, being twistor holomorphic is a condition on the vertical part of $T_{f(y)}P$, ie

$$F(x)D_v F(x) = D_{I(v)} F(x) \quad \text{for all } v \in T_x M. \tag{9}$$

Observe that conditions (8) imply $F(y)^2 = -\mathbf{1}$ and $F(y)^t = -F(y)$ (where A^t denotes the transpose of A) and so

$$DF(y)F(y) = -F(y)DF(y), \quad DF(y)^t = -DF(y).$$

Therefore $DF(y)$ maps the tangent space $T_y M$ into the normal space $N_y M$. This can be seen as follows:

$$\begin{aligned}\langle E_1, DF(y)E_2\rangle &= \langle E_1, DF(y)F(y)E_1\rangle \\ &= -\langle E_1, F(y)DF(y)E_1\rangle = \langle E_2, DF(y)E_1\rangle.\end{aligned}$$

But the antisymmetry of $DF(y)$ implies

$$\langle E_1, DF(y)E_2\rangle = -\langle E_2, DF(y)E_1\rangle,$$

so $\langle E_1, DF(y)E_2\rangle = 0$. Similarly one shows that $\langle N_1, DF(y)N_2\rangle = 0$. Furthermore, we have

$$\begin{aligned}DF(x)E_2 = DF(x)F(x)E_1 &= -F(x)DF(x)E_1, \\ \langle N_i, DF(x)E_j\rangle &= -\langle E_j, DF(x)N_i\rangle \quad \text{for } i, j = 1, 2.\end{aligned}$$

and conclude that (9) is satisfied if and only if

$$F(x)D_v F(x)E_1 = D_{I(v)} F(x)E_1 \quad \text{for all } v \in T_x M.$$

To calculate $DF(x)E_1$ we differentiate $0 = \langle N_i, F(y)E_1\rangle$ along $v \in T_x M$ and obtain

$$\begin{aligned}0 &= \langle N_i, DF(x)E_1\rangle + \langle D_v N_i, F(x)E_1\rangle - \langle F(x)N_i, D_v E_1\rangle \\ &= \langle N_i, DF(x)E_1\rangle - \langle N_i, D_v E_2\rangle - \sum_{j=1,2} \langle F(x)N_i, N_j\rangle\langle N_j, D_v E_1\rangle \\ &= \langle N_i, DF(x)E_1\rangle - \Big(w_2^{i+2}(v) + \sum_{j=1,2} \langle F(x)N_i, N_j\rangle w_1^{j+2}(v)\Big)\end{aligned}$$

using the 1–forms introduced in (2). We calculate

$$D_{(\cdot)}F(x)E_1 = (w_2^3 - w_1^4)N_1 + (w_2^4 + w_1^3)N_2,$$

showing the equivalence of (i) and (ii), since

$$\begin{aligned} &F(x)D_{(\cdot)}F(x)E_1 - D_{I(\cdot)}F(x)E_1 \\ &\qquad = (w_2^4 + w_1^3 - \star w_2^3 + \star w_1^4)N_1 + (-w_2^3 + w_1^4 + \star w_2^4 + \star w_1^3)N_2. \end{aligned}$$

It remains to check that (ii) is equivalent to (iii). Evaluating (ii) in E_1 and E_2, recalling $w_i^{k+2}(E_j) = \langle N_k, D_{E_j}E_i\rangle = \langle N_k, A_{ij}\rangle$ we have

$$\begin{aligned} \langle N_2, A_{12}\rangle + \langle N_1, A_{11}\rangle - \langle N_1, A_{22}\rangle + \langle N_2, A_{12}\rangle &= 2(\langle N_2, A^\circ_{12}\rangle + \langle N_1, A^\circ_{11}\rangle), \\ \langle N_2, A_{22}\rangle + \langle N_1, A_{12}\rangle + \langle N_1, A_{12}\rangle - \langle N_2, A_{11}\rangle &= 2(-\langle N_2, A^\circ_{11}\rangle + \langle N_1, A^\circ_{12}\rangle). \end{aligned}$$

This shows that (ii) holds if and only if $F(x)A^\circ_{11} = A^\circ_{12}$. □

Corollary 3.8 *Let M be an oriented two dimensional manifold and $f\colon M \to \mathbb{R}^4$ an immersion into $\mathbb{R}^4$; then the following are equivalent:*

(i) *f is twistor holomorphic.*

(ii) $\mathcal{W}(f) = 2\pi(\chi - e(\nu)) = 2\pi\big(\chi + |e(\nu)|\big)$, *ie equality holds in* (6) *and* $e(\nu) \le 0$.

Proof The equivalence follows from the fact that (7) is equivalent to condition (iii) in the previous proposition, ie $F(x)A^\circ_{11} = A^\circ_{12}$, because in this case

$$\begin{aligned} K^\perp &= 2\langle A^\circ_{11} \wedge A^\circ_{12}, N_1 \wedge N_2\rangle = 2(\langle A^\circ_{11}, N_1\rangle\langle A^\circ_{12}, N_2\rangle - \langle A^\circ_{11}, N_2\rangle\langle A^\circ_{12}, N_1\rangle) \\ &= -2|A^\circ_{11}|^2 = -2|A^\circ_{11}||A^\circ_{12}|. \end{aligned}$$

If (7) holds then either $F(x)A^\circ_{11} = A^\circ_{12}$ or $-F(x)A^\circ_{11} = A^\circ_{12}$, but, since $K^\perp$ must be nonpositive so that equality holds, the second is excluded. □

Remark As the Veronese surface satisfies $\mathcal{W}(f) = 6\pi$ and $e(\nu) = -2$ (when the orientation of $\mathbb{R}^4$ is chosen appropriately), we get that the oriented double cover $\tilde{f}\colon \mathbb{S}^2 \to \mathbb{R}^4$ is twistor holomorphic.

Friedrich [7] considered twistor holomorphic immersions into $\mathbb{R}^4$ in detail. He used the special structure of P to prove a kind of "Weierstrass representation" for such immersions.

Theorem 3.9 (Friedrich [7]) *Let M be an oriented two-dimensional manifold. Let $P \cong H \oplus H$ be the twistor space of $\mathbb{R}^4$* (see the remark before Proposition 3.7). *A holomorphic map $F\colon M \to P$ corresponds to a triple (g, s_1, s_2), where*

(i) $g\colon M \to \mathbb{S}^2$ *is a meromorphic function;*

(ii) s_1 *and* s_2 *are holomorphic sections of the bundle* $g^*(H)$ *over* M.

Furthermore, there are holomorphic maps φ^i, $\psi^i\colon M_i \to \mathbb{C}$, *where* $i = 1, 2$ *and* $M_1 := \{g \neq \infty\}$ *and* $M_2 := \{g \neq 0\}$, *such that*

$$s_1 = (\varphi^1, \varphi^2), \quad s_2 = (\psi^1, \psi^2)$$

with

$$\varphi^2 = \frac{1}{g}\varphi^1 \quad \text{and} \quad \psi^2 = \frac{1}{g}\psi^1 \quad \text{on } M_1 \cap M_2.$$

A holomorphic map $F\colon M \to P$ *defines a twistor holomorphic immersion* $f\colon M \to \mathbb{R}^4$ *via* $f = \pi \circ F$ *if and only if*

$$|ds_1| + |ds_2| > 0. \tag{10}$$

If (10) *is satisfied, then* f *is given by the formula*

$$f = \left(\frac{\varphi^1\overline{g} - \overline{\psi}^1}{1 + |g|^2}, \frac{\overline{\psi}^1 g + \varphi^1}{1 + |g|^2}\right). \tag{11}$$

Conversely, if f *is given by* (11) *with* s_1 *and* s_2 *satisfying* (10) *then* f *is a twistor holomorphic immersion. Any such twistor holomorphic immersion satisfies the formula*

$$\mathcal{W}(f) = 4\pi \deg(g). \tag{12}$$

Proof The proof is done in [7, Section 1, Remark 2 and Section 4, Example 4]. We remark that the meromorphic function g is defined by $g = p \circ F$. □

Corollary 3.10 *Let* M *be an oriented two-dimensional manifold and* $f\colon M \to \mathbb{R}^4$ *a twistor holomorphic immersion. Let* (g, s_1, s_2) *be the triple corresponding to the lift* $F\colon M \to P$ (see Theorem 3.9). *Then the maps* φ^i *and* ψ^i *for* $i = 1, 2$ *can be extended to meromorphic functions* φ^i, $\psi^i\colon M \to \mathbb{S}^2$ *for* $i = 1, 2$ *with the following properties: Denote by* $S_P(h) := \{x \in M : h(x) = \text{north pole} = \infty\}$ *the poles and by* $S_N(h) := \{x \in M : h(x) = \text{south pole} = 0\}$ *the zeros of a meromorphic function* $h\colon M \to \mathbb{S}^2$, *and let* $\mathrm{ord}_h(b)$ *for* $b \in S_P(h)$ *or* $b \in S_N(h)$ *be the order of the poles or zeros of* h. *Then we have:*

(i) $S_P(\varphi^1) \subset S_P(g)$ *and* $\mathrm{ord}_{\varphi^1}(b) \le \mathrm{ord}_g(b)$ *for all* $b \in S_P(\varphi^1)$.

(ii) $S_P(\varphi^2) \subset S_N(g)$ *and* $\mathrm{ord}_{\varphi^2}(a) \le \mathrm{ord}_g(a)$ *for all* $a \in S_P(\varphi^2)$.

(iii) $S_P(\psi^1) \subset S_P(g)$ *and* $\mathrm{ord}_{\psi^1}(b) \le \mathrm{ord}_g(b)$ *for all* $b \in S_P(\psi^1)$.

(iv) $S_P(\psi^2) \subset S_N(g)$ *and* $\mathrm{ord}_{\psi^2}(a) \le \mathrm{ord}_g(a)$ *for all* $a \in S_P(\psi^2)$.

Proof On $\mathbb{C} \setminus (S_P(g) \cup S_N(g))$, we have $\varphi^2 g = \varphi^1$, and $\varphi^1 \colon \mathbb{C} \setminus S_P(g) \to \mathbb{C}$ and $\varphi^2 \colon \mathbb{C} \setminus S_N(g) \to \mathbb{C}$ are holomorphic. Thus, either $\lim_{z \to b} \varphi^1(z) = \infty$ for $b \in S_P(g)$ or $\varphi^1(b)$ is a zero of order greater than or equal to $-\operatorname{ord}_g(b)$. In the latter case, φ^1 has a removable singularity in b and can be extended smoothly. In the first case, φ^1 has a pole in b. There are no other poles of φ^1, and the order of a pole of φ^1 cannot be bigger than that of g. Therefore, (i) holds. The other three claims follow in the same way. □

Proposition 3.11 *The twistor space $P = \mathbb{R}^4 \times \mathrm{SO}(4)/U(2)$ of $\mathbb{R}^4$ naturally carries an antiholomorphic involution J defined as being the identity on $\mathbb{R}^4$ and the multiplication by -1 on $\mathrm{SO}(4)/U(2)$. The composition of this involution with a lift of an immersed surface into the twistor space gives the lift of the same surface with reversed orientation. Furthermore, the involution is fiber-preserving and induces the antiholomorphic involution $z \mapsto -1/\bar{z}$ on $\mathbb{C}P^1$.*

Proof As already mentioned, the twistor space P of $\mathbb{R}^4$ is isomorphic to $H \oplus H$ in the sense that the following diagram commutes:

$$\begin{tikzcd} P \arrow[r, "\psi"] \arrow[dr, "\pi"'] & H \oplus H \arrow[d, "\tilde{\pi}"] \\ & \mathbb{R}^4 \end{tikzcd}$$

The projection $\tilde{\pi}$ is given by (11); compare [7, Section 4, Example 4]. One can understand $\tilde{\pi}$ as follows: Define the local sections around a point $z \in \mathbb{C}P^1 \setminus \{[(0,1)], [(1,0)]\}$ with representative $(u_1, u_2) \in \mathbb{C}^2$ as

$$\alpha(z) = \left(\frac{u_1}{u_2}, 1\right), \quad \beta(z) = \left(1, \frac{u_2}{u_1}\right).$$

A holomorphic section in $H \oplus H$ can be parametrized by the real 4–parameter family

$$\zeta = (A\alpha(z) + B\beta(z), \bar{B}\alpha(z) - \bar{A}\beta(z)). \tag{13}$$

The projection $\tilde{\pi}(\zeta)$ is then $(A, B) \in \mathbb{C}^2 = \mathbb{R}^4$.

The space $H \oplus H$ is holomorphically embedded in $\mathbb{C}^4$ by inclusion. We define the antiholomorphic involution

$$\tilde{I} \colon \mathbb{C}^4 \to \mathbb{C}^4, \quad u = (u_1, u_2, u_3, u_4) \mapsto (\bar{u}_4, -\bar{u}_3, -\bar{u}_2, \bar{u}_1).$$

Let $z \in \mathbb{C}P^1 \setminus \{[(0,1)]\}$ and $u \in H \oplus H$ with $p(u) = z$, ie $z = u_1/u_2 = u_3/u_4$; then $p \circ \tilde{I}(u) = -\bar{u}_2/\bar{u}_1$, hence $\tilde{I}$ defines an antiholomorphic involution on $\mathbb{C}P^1$ by

$z \mapsto -1/\bar{z}$. Using the parametrization (13) one readily checks that

$$\widetilde{I}(\zeta) = \Big(A\alpha\Big(-\frac{1}{\bar{z}}\Big) + B\beta\Big(-\frac{1}{\bar{z}}\Big), \bar{B}\alpha\Big(-\frac{1}{\bar{z}}\Big) - \bar{A}\beta\Big(-\frac{1}{\bar{z}}\Big)\Big). \tag{14}$$

Due to the isomorphism $\psi\colon H \oplus H \to P$ we obtain an antiholomorphic involution on P by $J := \psi \circ \widetilde{I} \circ \psi^{-1}$. Equation (14) implies that J is the identity on $\mathbb{R}^4$. It remains to show that J corresponds to the multiplication by -1 on $\mathrm{SO}(4)/U(2)$. This can be seen as follows: Let $f\colon M \to \mathbb{R}^4$ be a twistor holomorphic immersion with holomorphic lift $F\colon M \to P$; compare Definition 3.5. We denote by $\widetilde{F} := \psi \circ F$ the associated holomorphic map into $H \oplus H$ and let $\sigma\colon M \to M$ be an antiholomorphic involution on the Riemann surface M reversing the orientation. Using $\widetilde{I}$ we obtain a new holomorphic map $\widetilde{F}_2\colon M \to H \oplus H$ by $\widetilde{F}_2 := \widetilde{I} \circ \widetilde{F} \circ \sigma$. Furthermore, we have $\tilde{\pi} \circ \widetilde{F}_2(p) = f(\sigma(p)) = f(p)$ for all $p \in M$, due to (14). Hence, $F_2 := \psi^{-1} \circ \widetilde{F}_2\colon M \to P$ has to be the lift corresponding to the immersion $f \circ \sigma\colon M \to \mathbb{R}^4$ and therefore J must be the multiplication by -1 on $\mathrm{SO}(4)/U(2)$. □

Corollary 3.12 *Let M be a two-dimensional oriented manifold and $f\colon M \to \mathbb{R}^4$ a twistor holomorphic immersion. Assume that M is equipped with an antiholomorphic involution $I\colon M \to M$ without fixpoints such that $f \circ I = f$. Let $F\colon M \to P$ be the lift into the twistor space and (g, s_1, s_2) the corresponding triple* (see Theorem 3.9). *Then we have that*

$$g \circ I = -\frac{1}{\bar{g}} \tag{15}$$

and

$$\varphi^1 \circ I = \frac{\bar{\psi}^1}{\bar{g}} = \bar{\psi}^2 \quad \text{and} \quad \psi^1 \circ I = -\frac{\bar{\varphi}^1}{\bar{g}} = -\bar{\varphi}^2. \tag{16}$$

The immersion f is given by the formula

$$f = (f_1, f_2) = \left(\frac{\varphi^1 - \varphi^1 \circ I}{g - g \circ I}, \frac{\bar{\psi}^1 - \bar{\psi}^1 \circ I}{\bar{g} - \bar{g} \circ I}\right) = \left(\frac{\varphi^1 - \bar{\psi}^2}{g + \frac{1}{\bar{g}}}, \frac{\bar{\psi}^1 + \varphi^2}{\bar{g} + \frac{1}{g}}\right). \tag{17}$$

Proof Let $F\colon M \to P$ be the holomorphic lift of f as in the statement above. Then $g = p \circ F$, where $p\colon P \to \mathbb{C}P^1$ is the projection in $H \oplus H \cong P$. Consider the holomorphic map $\widetilde{F} := J \circ F \circ I\colon M \to P$, where $J\colon P \to P$ is the antiholomorphic involution from Proposition 3.11. By assumption we have $f \circ I = f$, which implies (together with the properties of J) that $\widetilde{F}$ is the lift of f, ie $\widetilde{F} = F$. As J induces the antipodal map on $\mathbb{C}P^1$ we have that

$$g \circ I = p \circ F \circ I = p \circ J \circ F = -\frac{1}{\bar{g}},$$

which is (15). For (16), we use the same argument but now on $H \oplus H$: Denote by $\psi\colon H \oplus H \to P$ the isomorphism as in the proof of Proposition 3.11. The antiholomorphic involution $\widetilde{I}$ on $H \oplus H$ from the same proposition has the property $\widetilde{I} \circ \widetilde{F} \circ I = \widetilde{F}$, where $\widetilde{F} := \psi \circ F$. By definition of $\widetilde{I}$ we get $\overline{\psi}^2 \circ I = \varphi^1$ and $-\overline{\psi}^1 \circ I = \varphi^2$, which implies (16). Formula (17) is a consequence of (11) and (16). $\square$

Lemma 3.13 *Let $M = \mathbb{C}/\Gamma$ be a torus that carries an antiholomorphic involution $I\colon M \to M$ without fixpoints. Then M is biholomorphically equivalent to a torus with a rectangular lattice. Moreover, there is a set of admissible parameters $\Lambda_0 \neq \varnothing$ and a family of meromorphic functions $g_\lambda\colon M \to \mathbb{S}^2$ for $\lambda \in \Lambda_0$ with $\deg(g_\lambda) = 4$ and $g_\lambda \circ I = -1/\overline{g}_\lambda$.*

Proof By Proposition 2.2 we get a biholomorphic map $\varphi\colon M \to \widetilde{M}$, where $\widetilde{M}$ is generated by a "canonical basis" $(1, \tau)$. As $\widetilde{I} := \varphi \circ I \circ \varphi^{-1}$ is an antiholomorphic involution without fixpoints on $\widetilde{M}$ we know that $\widetilde{M}$ is generated by a rectangular lattice and $\widetilde{I}$ must be $\widetilde{I}(z) = \overline{z} + \frac{1}{2}$ or $\widetilde{I}(z) = -\overline{z} + \frac{1}{2}\tau$; see Theorem 2.1. If $\widetilde{I}(z) = -\overline{z} + \frac{1}{2}\tau$, then we go again to another lattice by a biholomorphic map $\psi\colon \widetilde{M} \to \mathbb{C}/\Gamma_1$, $\psi(z) = z/\tau$. Then Γ_1 has generator $(\tau_1, 1)$, where $\tau_1 = -\tau/|\tau|^2$ and

$$\psi \circ I \circ \psi^{-1}(z) = \psi\Big(-\overline{\tau z} + \frac{\tau}{2}\Big) = -\frac{\overline{\tau}}{\tau}\overline{z} + \frac{1}{2} = \overline{z} + \frac{1}{2}$$

because τ is purely imaginary. Thus, we can assume that $I(z) = \overline{z} + \frac{1}{2}$ and we have a rectangular lattice generated by $(1, \tau)$.

The second step is the proof of the existence of g. As we are looking for an elliptic function of degree 4, g must have four poles b_k for $k = 1, \dots, 4$, and four zeros a_k for $k = 1, \dots, 4$ (counting with multiplicities). For the theory of elliptic functions, see for example [10]. Such an elliptic function exists if and only if $\sum_{k=1}^4 b_k - \sum_{k=1}^4 a_k \in \Gamma$; see [10, Section 1.6]. Consider the function

$$h(z) := g\big(\overline{z} + \tfrac{1}{2}\big)\overline{g}(z).$$

We show that we can choose the poles and zeros of g so that $h \equiv -1$, which is equivalent to (15).

As g only has poles in b_k, we require $a_k = I(b_k) = \overline{b}_k + \frac{1}{2}$. Then

$$\sum_{i=1}^4 b_k - \sum_{k=1}^4 \big(\overline{b}_k + \tfrac{1}{2}\big) = 2i \sum_{k=1}^4 \Im(b_k) - 2 \in \Gamma$$

is a necessary condition for the existence of such g. Thus, there must be an $m \in \mathbb{Z}$ with $\sum_{k=1}^4 \Im(b_k) = \frac{1}{2} m \Im(\tau)$. As I is an involution we have that $I(a_k) = b_k$, thus if $g \circ I$ has a pole in a point, then $\overline{g}$ has a zero of the same order at that point (and vice versa). It follows that h has no poles. As $\overline{h}$ is elliptic without poles it is constant, and h is constant as well. We have to find out if this constant can be -1. Define $\omega_0 := \sum_{i=1}^4 b_k - \sum_{k=1}^4 \bigl(\overline{b}_k + \frac{1}{2}\bigr) = m\tau - 2 \in \Gamma$. Then, up to a complex constant factor c, the function g is of the form

$$g(z) = e^{-\eta(\omega_0)z} \frac{\prod_{k=1}^4 \sigma\bigl(z - \overline{b}_k - \frac{1}{2}\bigr)}{\prod_{k=1}^4 \sigma(z - b_k)},$$

where $\sigma\colon \mathbb{C} \to \mathbb{C}$ denotes the Weierstrass sigma function and $\eta\colon \Gamma \to \mathbb{C}$ is the group homomorphism that satisfies the Legendre relation, ie

$$\eta(\omega_2)\,\omega_1 - \eta(\omega_1)\,\omega_2 = 2\pi i \quad \text{if } \Im\left(\frac{\omega_1}{\omega_2}\right) > 0. \tag{18}$$

We collect some facts about σ and η; see [10, Section 1.6]: This function σ is an entire function that has in all lattice points zeros of order one and no other zeros. As it is nonconstant and has no poles it cannot be doubly periodic. But it has the property

$$\sigma(z + \omega) = -e^{\eta(\omega)(z+\omega/2)}\sigma(z),$$

when $\frac{1}{2}\omega \notin \Gamma$. If the lattice is real, then $\overline{\sigma}(z) = \sigma(\overline{z})$. This can be seen in the representation formula

$$\sigma(z) = z \prod_{0 \neq \omega \in \Gamma} \Bigl(1 - \frac{z}{\omega}\Bigr) e^{z/\omega + (z/\omega)^2/2}.$$

For a rectangular lattice, η has the property that

$$\begin{aligned} &\eta(\omega) \in i\mathbb{R} && \text{for } \omega \in \Gamma \cap i\mathbb{R}, \\ &\eta(\widetilde{\omega}) \in \mathbb{R} && \text{for } \widetilde{\omega} \in \Gamma \cap \mathbb{R}. \end{aligned} \tag{19}$$

We use these properties to get

$$\begin{aligned} h(z) &= \exp\bigl(-\eta(\omega_0)\bigl(\overline{z} + \tfrac{1}{2}\bigr) - \overline{\eta}(\omega_0)\overline{z}\bigr) \frac{\prod_{k=1}^4 \sigma(\overline{z} - \overline{b}_k)}{\prod_{k=1}^4 \overline{\sigma}(z - b_k)} \frac{\prod_{k=1}^4 \overline{\sigma}\bigl(z - \overline{b}_k - \frac{1}{2}\bigr)}{\prod_{k=1}^4 \sigma\bigl(\overline{z} - b_k + \frac{1}{2}\bigr)} \\ &= \exp\bigl(-2\Re(\eta(\omega_0))\overline{z} - \tfrac{1}{2}\eta(\omega_0)\bigr) \frac{\prod_{k=1}^4 \overline{\sigma}\bigl(z - \overline{b}_k - \frac{1}{2}\bigr)}{\prod_{k=1}^4 \sigma\bigl(\overline{z} - b_k + \frac{1}{2}\bigr)} \end{aligned} \tag{20}$$

$$= \exp\big(-2\Re(\eta(\omega_0))\overline{z} - \tfrac{1}{2}\eta(\omega_0)\big)\frac{\prod_{k=1}^4 \overline{\sigma}\big(z-\overline{b}_k-\frac{1}{2}\big)}{\prod_{k=1}^4 \sigma\big(\overline{z}-b_k-\frac{1}{2}\big)} \times (-1)^4 \exp\Big(-\eta(1)\sum_{k=1}^4(\overline{z}-b_k)\Big)$$

$$= \exp\Big(-2\Re(\eta(\omega_0))\overline{z} - \tfrac{1}{2}\eta(\omega_0) - 4\eta(1)\overline{z} + \eta(1)\sum_{k=1}^4 b_k\Big).$$

As $\eta(\omega+\widetilde{\omega}) = \eta(\omega)+\eta(\widetilde{\omega})$ for all $\omega, \widetilde{\omega} \in \Gamma$ (η is a group homomorphism) and $\eta(0)=0$ we get that

$$\eta(\omega_0) = m\eta(\tau) - 2\eta(1) \quad \text{and} \quad \eta(\tau) \in i\mathbb{R} \quad \text{and} \quad \eta(1) \in \mathbb{R}.$$

Thus, (20) yields

$$h(z) = \exp\Big(+4\eta(1)\overline{z} + \eta(1) - \frac{m}{2}\eta(\tau) - 4\eta(1)\overline{z} + \eta(1)\sum_{k=1}^4 \Re(b_k) + \eta(1)\frac{m}{2}\tau\Big)$$

$$= \exp\Big(\eta(1)\Big(1+\sum_{k=1}^4 \Re(b_k)\Big)\Big)\exp(m\pi i),$$

where we used the Legendre relation (18) and property (19) in the last step. Thus, for every combination of poles b_k for $k=1,\dots,4$ that satisfies $i\sum_{k=1}^4 \Im(b_k) = \frac{1}{2}(2l+1)\tau$ for an $l \in \mathbb{Z}$, we define $R := \sum_{k=1}^4 \Re(b_k)$ and choose $c := e^{-\eta(1)(1+R)/2}$. Then we have, with $\widetilde{g}(z) := cg(z)$ (g as above), that

$$h(z) = |c|^2 e^{\eta(1)(1+R)} \cdot (-1) = -1,$$

which is equivalent to (15). As we can assume that $b_k \in [0,1]\times[0,\Im(\tau)]$, it suffices to consider poles such that $i\sum_{k=1}^4 \Im(b_k) = \frac{1}{2}(2l+1)\tau$ with $l \in \{0,1,2,3\}$. We collect all such possible b_k in Λ_0. The set Λ_0 is obviously not empty. □

Proposition 3.14 *Let $M = \mathbb{C}/\Gamma$ be a torus that carries an antiholomorphic involution $I\colon M \to M$ without fixpoints. Let $g\colon M \to \mathbb{S}^2$ be meromorphic with $g \circ I = -1/\overline{g}$* (coming from Lemma 3.13). *If there is a meromorphic function $\varphi^1\colon M \to \mathbb{S}^2$ with $2 \le \deg(\varphi^1) \le 4$, $S_P(\varphi^1) \subset S_P(g)$ and $\mathrm{ord}_{\varphi^1}(b) \le \mathrm{ord}_g(b)$ for all $b \in S_P(\varphi^1)$ and $\varphi^1 \neq cg + \widetilde{c}$ for all $c, \widetilde{c} \in \mathbb{C}$ then there are unique meromorphic functions $\psi^1, \psi^2, \varphi^2\colon M \to \mathbb{S}^2$ such that the triple (g, s_1, s_2) with $s_1 = (\varphi^1, \varphi^2)$ and $s_2 = (\psi^1, \psi^2)$ corresponds to a twistor holomorphic immersion $f\colon M \to \mathbb{R}^4$. In particular, the properties of Theorem 3.9 and Corollary 3.10 are satisfied.*

Proof As the existence of φ^1 is assumed, we define $\varphi^2 = \varphi^1/g$, $\psi^2 = \overline{\varphi}^1 \circ I$ and $\psi^1 = -\overline{\varphi}^2 \circ I$. This defines ψ^1, ψ^2, φ^2: $M \to \mathbb{S}^2$ uniquely and we have all the properties of Corollary 3.10. Then we define

$$f := \left(\frac{\varphi^1 \overline{g} - \overline{\psi}^1}{1+|g|^2}, \frac{\overline{\psi}^1 g + \varphi^1}{1+|g|^2} \right) = \left(\frac{\varphi^1 - \varphi^1 \circ I}{g - g \circ I}, \frac{\overline{\psi}^1 - \overline{\psi}^1 \circ I}{\overline{g} - \overline{g} \circ I} \right),$$

which is (11) and (17). In this way, we also know $f \circ I = f$. We claim that f is not constant. If $f_1 = c$ for a constant $c \in \mathbb{C}$ then $\varphi^1 - \varphi^1 \circ I = c(g - g \circ I)$. This is equivalent to $\varphi^1 - cg = (\varphi^1 - cg) \circ I$, which implies that $\varphi^1 - cg$ is holomorphic (as a map into $\mathbb{S}^2$) and antiholomorphic. Thus, it must be a constant. But this contradicts $\varphi^1 \neq cg + \widetilde{c}$ for all $c, \widetilde{c} \in C$.

We do not know yet if $|d\varphi^1| + |d\varphi^2| + |d\psi^1| + |d\psi^2| > 0$. This is necessary for f to be an immersion; see Theorem 3.9. Define

$$B := \bigl\{ z \in M : |d\varphi^1|(z) + |d\varphi^2|(z) + |d\psi^1|(z) + |d\psi^2|(z) = 0 \bigr\}.$$

We assume $B \neq \varnothing$. Considering φ^i and ψ^i as elliptic functions with finite degree we know that $|B| < \infty$. As I has no fixpoints, $|B|$ is an even number. By Friedrich's construction, f: $M \to \mathbb{R}^4$ is a branched conformal immersion with branch points in B. The Riemann–Hurwitz formula for covering maps with ramification points yields the formula

$$\mathcal{W}(f) = 4\pi \deg(g) = 16\pi,$$

as shown by Friedrich; see [7, Section 4, Example 4]. We combine this with the Gauss–Bonnet formula for conformal branched immersions [6, Theorem 4],

$$\int_M K = 2\pi \biggl(\chi(M) + \sum_{p \in B} m(p) \biggr),$$

where $m(p)$ is the branching order in p, to get

$$\frac{1}{4} \int_M |A|^2 = \mathcal{W}(f) - \frac{1}{2} \int_M K < \infty.$$

Since M is compact we have that $\mathrm{Vol}(f(M)) < \infty$. Thus, f: $M \to \mathbb{R}^4$ is a $W^{2,2}$–conformal branched immersion and we can apply [14]. For that, fix any $x_0 = f(p)$ for some $p \in B$. Then $\sum_{p \in f^{-1}(x_0)} (m(p)+1) \geq 4$. Define $\widehat{f} := S \circ J \circ f$, where $J(x) := x_0 + (x - x_0)/|x - x_0|^2$ and S: $\mathbb{R}^4 \to \mathbb{R}^4$ is any reflection. Then $\widehat{f}$: $M \setminus B \to \mathbb{R}^4$ is twistor holomorphic because A° does not change by a conformal transformation and by Proposition 3.7(iii) (note that the reflection makes sure that $S \circ J$

is orientation-preserving). We apply [14, (3.1)] and get

$$\mathcal{W}(\hat{f}) = \mathcal{W}(f) - 4\pi \sum_{p \in f^{-1}(x_0)} (m(p)+1) \leq 0. \tag{21}$$

Hence, $\hat{f}\colon M \setminus B \to \mathbb{R}^4$ is superminimal (ie twistor holomorphic and minimal). By a classical result of Eisenhart [5], $\hat{f}$ is locally given by two (anti)holomorphic functions $\hat{f} = (h_1, h_2)$. But this yields a contradiction because $\hat{f} \circ I = \hat{f}$ implies that the components of $\hat{f}$ are holomorphic and antiholomorphic, and hence constant. □

We now restate and prove our main theorem:

Theorem 3.15 *On each torus M_r with rectangular lattice generated by $(1, ir)$ for $r \in \mathbb{R}^+$, there is a set of admissible parameters $\Lambda \neq \varnothing$ such that there are smooth conformal immersions $\hat{f}^r_\lambda\colon M_r \to \mathbb{R}^4$ that are twistor holomorphic and double covers of Klein bottles. The corresponding immersed Klein bottles $f^r_\lambda\colon K \to \mathbb{R}^4$ for $\lambda \in \Lambda$ satisfy $\mathcal{W}(f^r_\lambda) = 8\pi$ and $e(\nu^r_\lambda) = -4$. By reversing the orientation of $\mathbb{R}^4$ we get a family of immersions $\tilde{f}^r_\lambda$ with $W(\tilde{f}^r_\lambda) = 8\pi$ and $e(\tilde{\nu}^r_\lambda) = +4$.*

Every immersion $f\colon K \to \mathbb{R}^4$ with $\mathcal{W}(f) = 8\pi$ and $e(\nu) \in \{-4, +4\}$ is an embedding and is either an element of $\{f^r_\lambda : \lambda \in \Lambda\}$ or of $\{\tilde{f}^r_\lambda : \lambda \in \Lambda\}$. Furthermore, every such immersion is a minimizer of the Willmore energy in its regular homotopy class, thus it is a Willmore surface.

Proof Every Klein bottle N is the quotient of its oriented double cover $q\colon M \to N$ and the group $\{\mathrm{id}, I\} = \langle I \rangle$, where $I\colon M \to M$ is the antiholomorphic order-two deck transformation on a torus M, ie $N = M/\langle I \rangle$ and $q\colon M \to M/\langle I \rangle$.

We consider the rectangular lattice generated by $(1, \tau)$ with $\Im(\tau) > 0$ and the involution $I(z) = \bar{z} + \frac{1}{2}$. As the imaginary part of τ is not fixed we get the parameter $r := \Im(\tau)$ and a family of tori $\{M_r\}_{r \in \mathbb{R}^+}$ with the involution I. From now on we keep r fixed and denote $N := M_r/\langle I \rangle$. We choose $b_k \in [0,1] \times [0, \Im(\tau)] =: F$ for $k = 1, \dots, 4$, with $i \sum_{k=1,\dots,4} \Im(b_k) = \frac{1}{2} m\tau$ with m odd. As in the proof of Lemma 3.13, each such combination yields a meromorphic functions g with (15). Consider any $p_1 \in F \setminus \{b_1, b_2\}$ and define $p_2 := b_1 + b_2 - p_1$. Then

$$b_1 + b_2 - p_1 - p_2 = 0 \in \Gamma. \tag{22}$$

If $p_2 \in \{b_1 + \Gamma, b_2 + \Gamma\}$ then we go to $(\tilde{p}_1, \tilde{p}_2) = (p_1 + \epsilon, p_2 - \epsilon)$ such that (22) is still satisfied and $\{\tilde{p}_1, \tilde{p}_2\} \cap (\{b_1 + \Gamma\} \cup \{b_2 + \Gamma\}) = \varnothing$. By the existence theorem for elliptic functions there exists a meromorphic $\varphi^1\colon M \to \mathbb{S}^2$ with poles in b_1 and b_2 and zeros in $\tilde{p}_1$ and $\tilde{p}_2$. By construction and with $\deg(\varphi^1) = 2 \neq \deg(g)$ we have proven the existence of a φ^1 that we need for Proposition 3.14.[2] Of course, this φ^1 is only one choice. In Λ we include each possible g and φ^1. Then,

$$\tilde{f} := \left(\frac{\varphi^1 \bar{g} - \bar{\psi}^1}{1+|g|^2}, \frac{\bar{\psi}^1 g + \varphi^1}{1+|g|^2} \right)$$

is a twistor holomorphic immersion with $\mathcal{W}(\tilde{f}) = 16\pi$ and

$$e(\nu) = \chi(M) - \frac{1}{2\pi}\mathcal{W}(\tilde{f}) = -8,$$

due to Corollary 3.8(ii). We define f by $\tilde{f} = f \circ q$ and get immersions $f \colon N \to \mathbb{R}^4$ with $\mathcal{W}(f) = 8\pi$ and $e(\nu_f) = -4$ (the equality $e(\nu_{\tilde{f}}) = 2e(\nu_f)$ can be seen for example in (5)). By reversing the orientation of $\mathbb{R}^4$ and repeating the construction of $\tilde{f}$ and f we get immersions $\hat{f} \colon N \to \mathbb{R}^4$ with $\mathcal{W}(\hat{f}) = 8\pi$ and $e(\nu_{\hat{f}}) = +4$. Note that in this case $\tilde{\hat{f}} \colon M \to \mathbb{R}^2$ is not twistor holomorphic (Corollary 3.8).

On the other hand, every immersion $f \colon N \to \mathbb{R}^4$ with $\mathcal{W}(f) = 8\pi$ and $e(\nu) \in \{+4, -4\}$ has all the properties shown in Theorem 3.9 and Corollaries 3.10 and 3.12. The proof of Lemma 3.13 shows that every g from the triple (g, s_1, s_2) must be one of the g_λ that we found for our surfaces. Also φ^1 must be one of ours. Thus, by the "Weierstrass representation" of Friedrich, f must be in $\{f_\lambda : \lambda \in \Lambda\}$ or $\{\hat{f}_\lambda : \lambda \in \Lambda\}$.

It remains to check every immersion $f \colon N \to \mathbb{R}^4$ with $\mathcal{W}(f) = 8\pi$ and $e(\nu) \in \{+4, -4\}$ is an embedding. We repeat an argument from the proof of Proposition 3.14. If $e(\nu) = +4$, then we reverse the orientation of $\mathbb{R}^4$ and get an immersion with $e(\nu) = -4$. We go to the oriented double cover and get an immersion $\tilde{f} \colon M \to \mathbb{R}^4$ with $\mathcal{W}(\tilde{f}) = 16\pi$ and $e(\nu) = -8$. As equality is satisfied in the Wintgen inequality, $\tilde{f}$ is twistor holomorphic (Corollary 3.8). If f has a double point, then $\tilde{f}$ has a quadruple point $x_0 \in \mathbb{R}^4$. Inverting at $\partial B_1(x_0)$ and reflecting in $\mathbb{R}^4$ yields $\mathcal{W}(\hat{f}) = 0$ as in (21), where $\hat{f} := S \circ J \circ \tilde{f}$ (J is the inversion, S the reflection). As $S \circ J$ is conformal and orientation-preserving, $\hat{f}$ is still twistor holomorphic (Proposition 3.7(iii)). But every superminimal immersion into $\mathbb{R}^4$ is locally given by two (anti)holomorphic functions. As $\hat{f} \circ I = \hat{f}$ for I antiholomorphic, $\hat{f}_1$ and $\hat{f}_2$ must be constant. Thus, it cannot be an immersion, a contradiction.

Corollary 3.3 shows that every immersion with the properties above is a minimizer in its regular homotopy class. As Willmore surfaces are defined as critical points of the Willmore energy under compactly supported variations, the discovered immersions are Willmore surfaces. □

[2] The existence of φ^1 seems to be clear, but there are indeed cases where we have to be careful. If $b_1 = \cdots = b_4 =: b$ are the poles of g then we cannot construct φ with a double pole in b and a double zero in $I(b)$, because then we have that $i\Im(b) = \frac{1}{8}m\tau$ with m odd and $i\Im(b) = \frac{1}{4}k\tau$ with $k \in \mathbb{Z}$, a contradiction.

Corollary 3.16 *Let K be a Klein bottle and $f\colon K \to \mathbb{R}^4$ be an immersion with $\mathcal{W}(f) = 8\pi$ and $|e(\nu)| = 4$. Let $q\colon M \to K$ be the oriented double cover on the corresponding torus M. Then M is biholomorphically equivalent to a torus M_r generated by $(1, ir)$ for $r \in \mathbb{R}^+$, via a map $\varphi\colon M \to M_r$. Moreover, there are f_λ^r, $\widetilde{f}_\lambda^r$ and q_r coming from Theorem 3.15 such that $f \circ q = f_\lambda^r \circ q_r \circ \varphi$ for $e(\nu) = -4$ or $f \circ q = \widetilde{f}_\lambda^r \circ q_r \circ \varphi$ for $e(\nu) = +4$. Thus, the surface $f(K)$ is one of the surfaces obtained in Theorem 3.15.*

Proof By possibly changing the orientation of $\mathbb{R}^4$ we can assume $e(\nu) = -4$. By Corollary 3.8 we know that f is twistor holomorphic. The work of Friedrich [7] yields the existence of a triple (g, s_1, s_2) as in Theorem 3.9. By Corollary 3.12 we have $g \circ I = -1/\overline{g}$, where I comes from the order-two deck transformation of the oriented cover. Lemma 3.13 shows that there is a biholomorphic map $\varphi\colon M \to M_r$, where M_r is generated by $(1, ir)$ for an $r \in \mathbb{R}^+$. From Theorem 3.15 we get that $f \circ q$ must be $f_\lambda^r \circ q_r \circ \varphi$ for a $\lambda \in \Lambda$. $\square$

Proposition 3.17 *The Lie group $\mathrm{SO}(4)$ acts naturally and fiber-preserving on the twistor space P. It induces a fiber-preserving action on $H \oplus H$. The induced action on $\mathbb{C}P^1$ is the action of the 3–dimensional Lie subgroup G of the Möbius group on $\mathbb{C}P^1$ that commutes with the antipodal map $z \mapsto -1/\overline{z}$:*

$$G := \left\{ m(z) = \frac{az+b}{\overline{b}z - \overline{a}} : a, b \in \mathbb{C},\ |a|^2 + |b|^2 = 1 \right\}.$$

Proof We proceed via several claims:

Claim 1 *The $\mathrm{SO}(4)$ action $\varphi\colon \mathrm{SO}(4) \times P \to P$ defined as*

$$O \cdot (y, j) := (Oy, OjO^t)$$

is natural and fiber-preserving.

Proof of Claim 1 The action preserves by definition the fibers. It is natural in the sense that if $f\colon M \to \mathbb{R}^4$ is a given immersion with corresponding lift $F\colon M \to \mathbb{R}^4$ then for any $O \in \mathrm{SO}(4)$ the map $O \cdot F$ is the lift of the immersion Of. Fix a point $x \in M$ and an orthonormal frame $\{E_1, E_2, N_1, N_2\}$ in a neighborhood U of x as in Definition 3.5 with related matrix $F(y) \in \mathrm{SO}(4)$ for $y \in U$. As F satisfies conditions (i)–(iv), so does $OF(y)O^t$. Furthermore $\{OE_1, OE_2, ON_1, ON_2\}$ is an orthonormal frame of Of and $OF(y)O^t$ obviously satisfies the conditions (8) with respect to the new orthonormal frame. $\square$

Claim 2 *The action is holomorphic.*

Proof of Claim 2 Given a point $(p, j) \in P$ the complex structure $\mathcal{J}_{(p,j)}$ on

$$T_{(p,j)}P = T^H_{(p,j)}P \oplus T^V_{(p,j)}P = T_p\mathbb{R}^4 \oplus T_j\,\mathrm{SO}(4)/U(2)$$

is by definition the multiplication by j on $T_p\mathbb{R}^4$ and by j on $T_j\,\mathrm{SO}(4)/U(2)$. Hence, given $X + Y \in T_p\mathbb{R}^4 \oplus T_j\,\mathrm{SO}(4)/U(2)$ and $O \in \mathrm{SO}(4)$ we have

$$\begin{aligned} dO\mathcal{J}_{(p,j)}(X+Y) &= dO(jX + jY) \\ &= OjO^t\,OX + OjO^t\,OYO^t = \mathcal{J}_{O\cdot(p,j)}dO(X+Y). \end{aligned}$$

Thus, the action is holomorphic. □

Claim 3 $\mathrm{SO}(4)$ *acts naturally and holomorphically and is fiber-preserving on* $H \oplus H$. *Hence it induces a group homomorphism* $h\colon \mathrm{SO}(4) \mapsto \mathrm{Aut}(\widehat{\mathbb{C}})$, *where* $\mathrm{Aut}(\widehat{\mathbb{C}})$ *is the Möbius group of the Riemann sphere.*

Proof of Claim 3 The isomorphism $\psi\colon P = \mathbb{R}^4 \times \mathrm{SO}(4)/U(2) \to H \oplus H$ induces a natural, holomorphic action of $\mathrm{SO}(4)$ on $H \oplus H$ by composition:

$$O \cdot (u, v) = \psi \circ O \cdot \circ \psi^{-1}(u, v).$$

Recall that parallel transport (translation in $\mathbb{R}^4$) defines a fibration of P over one of its fibers; compare [7, Remark 2]. This fibration defines the isomorphism ψ. Hence we have a commutative diagram (compare the remark below):

$$\begin{array}{ccc} \mathrm{SO}(4)/U(2) & \xrightarrow[\phi]{} & \mathbb{C}P^1 \\ \uparrow & & \uparrow p \\ P & \xrightarrow[\psi]{} & H \oplus H \end{array}$$

The action of $\mathrm{SO}(4)$ on the $\mathrm{SO}(4)/U(2)$–factor of P is independent of the basepoint in $\mathbb{R}^4$ and therefore the induced action on $H \oplus H$ is fiber-preserving. Therefore, the $\mathrm{SO}(4)$ action on P induces an action of $\mathrm{SO}(4)$ on $\mathrm{SO}(4)/U(2)$ and via the isomorphism $\phi\colon \mathrm{SO}(4)/U(2) \to \mathbb{C}P^1$ it induces also an action on $\mathbb{C}P^1$. The action is holomorphic, as proven in Claim 2. Therefore, $\mathrm{SO}(4)$ acts on $\mathbb{C}P^1$ as biholomorphic maps. The holomorphic automorphism group of the Riemann sphere $\mathbb{C}P^1 \cong \widehat{\mathbb{C}}$ is the Möbius group, ie all rational functions of the form

$$m(z) = \frac{az+b}{cz+d} \quad \text{with } a, b, c, d \in \mathbb{C},\ ad - bc \neq 0. \qquad \square$$

Claim 4 *Let* G *be the image of the group homomorphism* $h\colon \mathrm{SO}(4) \to \mathrm{Aut}(\widehat{\mathbb{C}})$; *then*

$$G = \left\{ m \in \mathrm{Aut}(\widehat{\mathbb{C}}) : \frac{-1}{\overline{m(z)}} = m\left(\frac{-1}{\bar{z}}\right) \right\},$$

ie G *is a* 3*–dimensional Lie subgroup of* $\mathrm{Aut}(\widehat{\mathbb{C}})$.

Proof of Claim 4 The map h is induced by the group homomorphism ϕ. Therefore, its kernel corresponds to the normal subgroup

$$N := \{O \in \mathrm{SO}(4) : O \cdot j = O j O^t = j \text{ for all } j \in \mathrm{SO}(4)/U(2)\},$$

which can be determined explicitly, using the isomorphism $\mathrm{Sp}(1) \otimes \mathrm{Sp}(1) \to \mathrm{SO}(4)$ defined by

$$(a, b) \cdot q = a q \overline{b} \quad \text{for } q \in \mathbb{H} \cong \mathbb{R}^4 \text{ and } (a, b) \in \mathrm{Sp}(1) \otimes \mathrm{Sp}(1),$$

where $\mathrm{Sp}(1)$ is the group of unit quaternions; compare [26, Proposition 1.1]. Conditions (i)–(iv) in Definition 3.4 determine

$$\mathrm{SO}(4)/U(2) \cong \{(1, c) : \overline{c} = -c \in \mathrm{Sp}(1)\}.$$

Hence $(a, b) \in N$ if and only if $(a, b)(1, c)(a, b)^t = (1, c)$ for all $c \in \mathbb{H}$ with $|c| = 1$ and $\overline{c} = -c$. This simplifies to

$$\begin{aligned} (a, b)(1, c)(a, b)^t = (a, b)(1, c)(\overline{a}, \overline{b}) = (1, bc\overline{b}) = (1, c) \quad &\text{for all } c \in \mathrm{Sp}(1),\ \overline{c} = -c \\ \text{or, equivalently,} \quad cb = bc \quad &\text{for all } c \in \mathrm{Sp}(1),\ \overline{c} = -c. \end{aligned}$$

The last line implies that b has to be real, and since $|b| = 1$ we conclude $b \in \{-1, +1\}$. Furthermore, we have

$$N = \{(a, 1) : a \in \mathrm{Sp}(1)\},$$

which is a 3–dimensional Lie subgroup of $\mathrm{SO}(4)$. The Lie group $\mathrm{SO}(4)/N$ is also 3–dimensional, so G is isomorphic to $\mathrm{SO}(4)/N$ by the first isomorphism theorem. □

Recall the antiholomorphic involution J on the twistor space P and the corresponding involution $\widetilde{I}$ on $H \oplus H$ introduced in Proposition 3.11. Applying J corresponds to reversing the orientation of an immersed surface $f\colon M \to \mathbb{R}^4$. Since reversing the orientation of the manifold M commutes with the $\mathrm{SO}(4)$ action on $\mathbb{R}^4$, the natural associated maps on the whole space and the base have to commute as well, ie

$$\begin{aligned} O \cdot \widetilde{I}(u, v) &= \widetilde{I}(O \cdot (u, v)) \quad &&\text{for all } (u, v) \in H \oplus H,\ O \in \mathrm{SO}(4), \\ \frac{-1}{\overline{m(z)}} &= m\Big(\frac{-1}{\overline{z}}\Big) &&\text{for all } z \in \mathbb{C},\ m \in G, \end{aligned} \tag{23}$$

by the properties of J. The subgroup H of $\mathrm{Aut}(\widehat{\mathbb{C}})$ that commutes with the antipodal map $z \mapsto -1/\overline{z}$ is readily calculated to be

$$H = \Big\{m \in \mathrm{Aut}(\widehat{\mathbb{C}}) : m(z) = \frac{az + b}{\overline{b}z - \overline{a}} \text{ with } a, b \in \mathbb{C},\ |a|^2 + |b|^2 = 1\Big\}.$$

We observe that H is also a 3–dimensional connected Lie subgroup. Property (23) implies that $G \subset H$. Since $h\colon \mathrm{SO}(4)/N \to G$ is a Lie group isomorphism, G is open and closed. As observed before, G and H are both connected and of dimension 3. Hence we finally conclude that $H = G$. □

Remark The translation-invariance of the isomorphism $\psi\colon P \to H \oplus H$ can also be seen in the formulas of Friedrich as follows: Let $\mathbb{C}P^1 \to H \oplus H$, $z \mapsto (u, v)$, be holomorphic sections with $u(z_0) = A\alpha(z_0) + B\beta(z_0)$ and $v(z_0) = \bar{B}\alpha(z_0) - \bar{A}\beta(z_0)$. Consider the holomorphic sections $\tilde{u}(z) = u(z) - (A\alpha(z) + B\beta(z))$ and $\tilde{v}(z) = v(z) - (\bar{B}\alpha(z) - \bar{A}\beta(z))$. These sections still satisfy $p(u) = z = p(\tilde{u})$ and $p(v) = z = p(\tilde{v})$, but $\tilde{\pi}(u, v) = \tilde{\pi}(\tilde{u}, \tilde{v}) - (A, B)$.

The isomorphism $\varphi\colon \mathbb{C}P^1 \to \mathrm{SO}(4)/U(2)$ can explicitly be stated, identifying $\mathbb{C}^2$ with the quaternions $\mathbb{H}$. Fix $g \in \mathbb{C} \cup \{\infty\}$ and let $\gamma \in \mathbb{H}$ be the unit quaternion $\gamma = \bigl(-1/\sqrt{1+|g|^2}, g/\sqrt{1+|g|^2}\bigr)$. The map φ can now be stated, using the quaternionic multiplication, to be

$$\mathbb{C}P^1 \to \mathrm{SO}(4)/U(2), \quad g \mapsto A_g, \quad \text{where } A_g q = q(-\bar{\gamma} i \gamma) \text{ for all } q \in \mathbb{H} \cong \mathbb{R}^4,$$

which is equivalent to

$$g \mapsto \frac{1}{|g|^2+1} \begin{pmatrix} 0 & -|g|^2+1 & 2g_2 & -2g_1 \\ |g|^2-1 & 0 & 2g_1 & 2g_2 \\ -2g_2 & -2g_1 & 0 & |g|^2-1 \\ 2g_1 & -2g_2 & -|g|^2+1 & 0 \end{pmatrix}.$$

Corollary 3.18 *Consider $f_i^{r_i}\colon K \to \mathbb{R}^4$ for $i = 1, 2$, a pair of Klein bottles with $\mathcal{W}(f_i^{r_i}) = 8\pi$ and $|e(\nu_i^{r_i})| = 4$. Let $\Phi\colon \mathbb{R}^4 \to \mathbb{R}^4$ be a conformal diffeomorphism such that $f_1^{r_1}(K) = \Phi \circ f_2^{r_2}(K)$; then $r_1 = r_2$.*

Proof We use the notation $f_i := f_i^{r_i}$ and keep in mind that the double covers possibly live on different lattices. After changing the orientation of $\mathbb{R}^4$ we may assume without loss of generality that $e(\nu_1) = -4$. The Willmore energy and the Euler normal number are conformally invariant, hence $f_2' := \Phi \circ f_2\colon K \to \mathbb{R}^4$ is a Klein bottle with $\mathcal{W}(f_2') = 8\pi$ and $|e(\nu_2')| = 4$. The Euler normal number of an immersion only depends on the image and not on the particular chosen immersion. Since $f_1(K) = f_2'(K)$ we deduce $e(\nu_1) = e(\nu_2') = -4$. Hence it is sufficient to prove the statement under the assumption that Φ is the identity and $e(\nu_i) = -4$ for $i = 1, 2$.

Let $q_i\colon M_{r_i} = \mathbb{C}/\Gamma_i \to K$ be the oriented double cover of the related tori. By Theorem 3.15, $f_i \circ q_i\colon M_{r_i} \to \mathbb{R}^4$ are twistor holomorphic with related holomorphic lifts $F_i\colon M_{r_i} \to P \cong H \oplus H$. By assumption we have $F_1(M_{r_1}) = F_2(M_{r_2})$. We also have that $f_i \circ q_i \circ I(z) = f_i \circ q_i(z)$ for all $z \in \mathbb{C}$, where $I(z) = \bar{z} + \frac{1}{2}$. The maps F_i restricted to one fundamental domain are homeomorphisms onto their images. Hence $G = F_1^{-1} \circ F_2\colon M_{r_2} \to M_{r_1}$ is a homeomorphism between tori. As F_1 and F_2 are conformal and orientation-preserving, G is conformal and orientation-preserving and therefore holomorphic. As a biholomorphic map between tori, G is of the form $G(z) = az + b$ with inverse $G^{-1}(z) = \frac{z}{a} - \frac{b}{a}$ and $a\Gamma_2 = \Gamma_1$. Furthermore, $\widetilde{I} = G^{-1} \circ I \circ G$ is an antiholomorphic fixpoint-free involution on M_{r_2}. Arguing as in Corollary 3.12, using the natural involution $J\colon P \to P$, we deduce

$$J \circ F_2 \circ \widetilde{I}(z) = J \circ F_1 \circ I \circ G(z) = F_1 \circ G(z) = F_2(z) = J \circ F_2 \circ I(z).$$

As J is an involution and F_2 is (restricted to a fundamental domain) a homeomorphism onto its image, we hence conclude $\widetilde{I} = I$. By direct computation following

$$G^{-1} \circ I \circ G(z) = \frac{\overline{a}}{a}\overline{z} + \frac{1}{2a} + \frac{\overline{b} - b}{a} = \overline{z} + \frac{1}{2},$$

we deduce $a = 1$ and $\Im(b) = 0$. This implies $r_1 = r_2$ and $F_1(z + b) = F_2(z)$ for some $b \in [0, 1)$. □

Corollary 3.19 *Let $f_i^{r_i}\colon K \to \mathbb{R}^4$ for $i = 1, 2$ be a pair of Klein bottles with $\mathcal{W}(f_i^{r_i}) = 8\pi$, $e(\nu_i^{r_i}) = -4$ and $f_1^{r_1}(K) = \lambda R \circ f_2^{r_2}(K)$ for a rigid motion $R(x) = O(x + v)$ for $x \in \mathbb{R}^4$, ie some $O \in \mathrm{SO}(4)$, $v \in \mathbb{R}^4$ and a scaling factor $\lambda \in \mathbb{R}^+$. Then we have that $r_1 = r_2$. Furthermore, if $g_i\colon M_{r_i} \to \mathbb{C}P^1$ are the related projections of the holomorphic lifts, there is a Möbius transform $m_O \in G$ of Proposition 3.17 and $b \in [0, 1)$ such that $g_1(z) = m_O \circ g_2(z + b)$.*

Proof Firstly observe that scaling and rigid motions on $\mathbb{R}^4$ are conformal transformations. Corollary 3.18 hence implies that $r_1 = r_2$. We use the notation $f_i := f_i^{r_i}$.

Secondly we can assume that $\lambda = 1$ and $v = 0$, ie $f_1(K) = O \cdot f_2(K)$ for some $O \in \mathrm{SO}(4)$. Otherwise we may consider additionally the immersion $f_2' := \lambda(f_2 + v)$ and observe that

$$p \circ F_2(z) = p \circ F_2'(z) \quad \text{for all } z \in M_{r_2},$$

where $F_2, F_2'\colon M_{r_2} \to H \oplus H$ are the related lifts and $p\colon H \oplus H \to \mathbb{C}P^1$ is the projection. This is easily seen because scaling and translation in $\mathbb{R}^4$ does not affect the tangent space and so the lift in the twistor space is unaffected.

Let $F_i\colon M_{r_i} \to H \oplus H$ be the lift of $f_i\colon K \to \mathbb{R}^4$ and $O \in \mathrm{SO}(4)$ such that $f_1(K) = O \cdot f_2(K)$. The group SO(4) acts naturally on P — compare Proposition 3.17 — hence $O \cdot F_2$ is the lift of $O \cdot f_2$. Using the proof of Corollary 3.18 we conclude the existence of $b \in [0, 1)$ such that

$$F_1(z) = O \cdot F_2(z + b) \quad \text{for all } z \in M_{r_1} = M_{r_2}.$$

Furthermore, Proposition 3.17 implies the existence of a Möbius transform $m_O \in G$ such that $p \circ O \cdot F_2 = m_O \circ p \circ F_2$. With $g_i = p \circ F_i$ we get

$$g_1(z) = m_O \circ g_2(z+b) \quad \text{for all } z \in M_{r_1} = M_{r_2}. \qquad \square$$

Remark Concerning the question how many surfaces we have found in Theorem 3.15 we can say the following: As shown in the proof of Lemma 3.13 the parameter set (of g_λ and therefore of f_λ^r) is at least of the size of $[0,1]^7$. In Corollary 3.19 we studied how rigid motions and scaling in $\mathbb{R}^4$ and admissible reparametrizations of the tori affect our surfaces, in particular the g_λ. Counting dimensions we still have a parameter set $[0,1]^3$ for every torus M_r.

We finish this section with deducing the explicit formula for the double cover of the Veronese embedding:

Proof of Proposition 1.3 Consider the triple (g, s_1, s_2) with $g = z^3$, $s_1 = (\varphi^1, \varphi^2) = (z^2, 1/z)$ and $s_2 = (\psi^1, \psi^2) = (z, 1/z^2)$. Then g satisfies $g \circ I = -1/\overline{g}$ for the antiholomorphic involution $z \mapsto -1/\overline{z}$ without fixpoints on $\mathbb{S}^2$. Furthermore, our choice of s_1 and s_2 yields (16). The immersion f is defined so that

$$f(z) = \left(\frac{\varphi^1 \overline{g} - \overline{\psi}^1}{1+|g|^2}, \frac{\overline{\psi}^1 g + \varphi^1}{1+|g|^2} \right) = \left(\frac{\varphi^1 - \varphi^1 \circ I}{g - g \circ I}, \frac{\overline{\psi}^1 - \overline{\psi}^1 \circ I}{\overline{g} - \overline{g} \circ I} \right),$$

which is (11) and (17). It follows that $f \circ I = f$. As $|ds_1| + |ds_2| > 0$ on $\mathbb{S}^2$, we defined a twistor holomorphic immersion with $W(f) = 12\pi$; see Theorem 3.9 or [7]. As S^2 carries the involution I, we consider $q\colon \mathbb{S}^2 \to \mathbb{S}^2/\langle I \rangle$ and $\widetilde{f} := f \circ q^{-1}\colon \mathbb{S}^2/\langle I \rangle \to \mathbb{R}^4$. We get an $\mathbb{R}P^2$ with $W(\widetilde{f}) = 6\pi$. By the work of Li and Yau [19], $\widetilde{f}$ must be a conformal transformation of the Veronese embedding. $\square$

4 A Klein bottle in $\mathbb{R}^4$ with Willmore energy less than 8π

In this final section we would like to consider the case of immersions $f\colon K \to \mathbb{R}^4$ with Euler normal number $e(\nu) = 0$. Our goal is to show the existence of the minimizer of immersed Klein bottles in $\mathbb{R}^n$ for $n \geq 4$. For this, we need the following theorem:

Theorem 4.1 *Let $K = \mathbb{R}P^2 \sharp \mathbb{R}P^2$ be a Klein bottle. Then there exists an embedding $f\colon K \to \mathbb{R}^4$ with $e(\nu) = 0$ and $\mathcal{W}(f) < 8\pi$.*

For the proof of the preceding theorem we use a construction by Bauer and Kuwert:

Theorem 4.2 (M Bauer and Kuwert [4]) *Let $f_i\colon \Sigma_i \to \mathbb{R}^n$ for $i = 1, 2$ be two smoothly immersed, closed surfaces. If neither f_1 nor f_2 is a round sphere (ie totally umbilical), then there is an immersed surface $f\colon \Sigma \to \mathbb{R}^n$ with topological type of the connected sum $\Sigma_1 \sharp \Sigma_2$ such that*

$$\mathcal{W}(f) < \mathcal{W}(f_1) + \mathcal{W}(f_2) - 4\pi. \tag{24}$$

Remark Notice that the strategy for the gluing construction implemented by Bauer and Kuwert was proposed by R Kusner [12].

A rough sketch of the construction of the connected sum in Theorem 4.2 is as follows: The surface f_1 is inverted at an appropriate sphere in order to obtain a surface $\widehat{f_1}$ with a planar end and energy $\mathcal{W}(\widehat{f_1}) = \mathcal{W}(f_1) - 4\pi$. Then a small disk is deleted from f_2 and a suitably scaled copy of $\widehat{f_1}$ is implanted. An interpolation yields the strict inequality (24). For the details we refer to [4].

For the Veronese embedding $V\colon \mathbb{R}P^2 \to \mathbb{R}^4$ we have $\mathcal{W}(V) = 6\pi$. Hence we can connect two Veronese surfaces and obtain a new surface f with $\mathcal{W}(f) < 8\pi$. However, by the previous sections we know that there is no Klein bottle in $\mathbb{R}^4$ with Euler normal number 4 or -4 and Willmore energy less than 8π. In order to obtain a better understanding of this situation we have to take a closer look on the construction of Bauer and Kuwert:

For that, let $f_i\colon \Sigma_i \to \mathbb{R}^n = \mathbb{R}^{2+k}$ for $i = 1, 2$ be two immersions that are not totally umbilical (ie no round spheres). Let A and B denote the second fundamental forms of f_1 and f_2, respectively. Moreover let $p_i \in \Sigma_i$ be two points such that $A^\circ(p_1)$ and $B^\circ(p_2)$ are both nonzero. After a translation and a rotation we may assume

$$f_i(p_i) = 0, \quad \operatorname{im} Df_i(p_i) = \mathbb{R}^2 \times \{0\} \quad \text{for } i = 1, 2.$$

Then $A^\circ(p_1), B^\circ(p_2)\colon \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}^k$ are symmetric, tracefree, nonzero bilinear forms.

In [4, (4.34), page 574], it is shown that Theorem 4.2 is true provided

$$\langle A^\circ(p_1), B^\circ(p_2)\rangle > 0. \tag{25}$$

In order to achieve inequality (25), one exploits the freedom to rotate the surface f_1 by an orthogonal transformation $R \simeq (S, T) \in \mathbb{O}(2) \times \mathbb{O}(k) \subset \mathbb{O}(n)$ before performing the connected sum construction.

The second fundamental form $A_{S,T}$ of the rotated surface Rf_1 at the origin is given by

$$A_{S,T}(\zeta, \zeta) = TA(S^{-1}\zeta, S^{-1}\zeta) \quad \text{for all } \zeta \in \mathbb{R}^2.$$

For the tracefree part we obtain

$$A^\circ_{S,T}(\zeta,\zeta) = TA^\circ(S^{-1}\zeta, S^{-1}\zeta) \quad \text{for all } \zeta \in \mathbb{R}^2.$$

We need the following linear algebra fact, which will be applied to A° and B°:

Lemma 4.3 *Let $P, Q\colon \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}^k$ be bilinear forms that are symmetric, tracefree and both nonzero.*

(a) *There exist orthogonal transformations $S \in \mathrm{SO}(2)$ and $T \in \mathbb{O}(k)$ such that the form $P_{S,T}(\zeta,\zeta) = TP(S^{-1}\zeta, S^{-1}\zeta)$ satisfies $\langle P_{S,T}, Q\rangle > 0$.*

(b) *We can choose $S \in \mathrm{SO}(2)$ and $T \in \mathrm{SO}(k)$ such that $\langle P_{S,T}, Q\rangle > 0$, except for the case that all of the following properties are satisfied:*
- $k = 2$,
- $|P(e_1,e_1)|^2 = |P(e_1,e_2)|^2$ *and* $|Q(e_1,e_1)|^2 = |Q(e_1,e_2)|^2$,
- $\langle P(e_1,e_1), P(e_1,e_2)\rangle = 0$ *and* $\langle Q(e_1,e_1), Q(e_1,e_2)\rangle = 0$,
- $\{P(e_1,e_1), P(e_1,e_2)\}$ *and* $\{Q(e_1,e_1), Q(e_1,e_2)\}$ *determine opposite orientations of* $\mathbb{R}^2$.

In this case, if $S \in \mathrm{SO}(2)$ and $T \in \mathbb{O}(2)$ with $\langle P_{S,T}, Q\rangle > 0$, then we have $T \in \mathbb{O}(2) \setminus \mathrm{SO}(2)$.

In (b)*, the ordered set $\{e_1, e_2\}$ is any positively oriented orthonormal basis of $\mathbb{R}^2$. If all four properties in* (b) *are satisfied for one such basis, they are also satisfied for any other positively oriented orthonormal basis of $\mathbb{R}^2$.*

Proof See [4, Lemma 4.5, page 574]. The exceptional case in (b) is the case in [4] in which $k = 2$, $|b| = a$, $|d| = c$, and b and d have opposite signs. □

Proof of Theorem 4.1 Let $V\colon \mathbb{R}P^2 \to \mathbb{R}^4$ be the Veronese embedding. Consider two copies $f_i\colon \Sigma_i \to \mathbb{R}^4$ of V, ie $\Sigma_i = \mathbb{R}P^2$ and $f_i = V$ for $i = 1, 2$. Let A and B denote the second fundamental forms of f_1 and f_2, respectively. Of course, $A = B$ and $A^\circ = B^\circ$. Choose $p \in \mathbb{R}P^2$ such that $A^\circ(p)$ is nonzero (in fact, this is satisfied for any $p \in \mathbb{R}P^2$) and set $P := A^\circ(p)$ and $Q := B^\circ(p)$. Then P and Q are two bilinear forms as in Lemma 4.3 with $P = Q$. Surely, the last condition in the exceptional case of Lemma 4.3(b) fails to be true. Hence, by Lemma 4.3 we can rotate f_1 by rotations $S, T \in \mathrm{SO}(2)$ such that (25) is satisfied. Now we are able to perform the connected sum construction: Inverting f_1 and connecting f_1 and f_2 as described in [4] yields a surface $f\colon K \to \mathbb{R}^4$ with $e(\nu) = 0$ and $\mathcal{W}(f) < 8\pi$. As any closed surface with Willmore energy less than 8π is injective, f is an embedding. □

Let us finally explain why it is not possible to construct an immersion $f\colon K \to \mathbb{R}^4$ with $|e(\nu)| = 4$ with the method above: A direct calculation shows that the Veronese embedding V satisfies $|A^\circ_{11}|^2 = |A^\circ_{12}|^2 = 1$ and $\langle A^\circ_{11}, A^\circ_{12}\rangle = 0$ in any point of $\mathbb{R}P^2$. Let P and Q be defined as in the preceding paragraph. Then P and Q satisfy the second and the third condition of the exceptional case in Lemma 4.3. In order to obtain a surface with $|e(\nu)| = 4$ we have to reflect one of the Veronese surfaces before rotating f_1 and performing the gluing construction. But then also the last condition of the exceptional case in Lemma 4.3(b) is satisfied. Hence we cannot choose $T \in \mathrm{SO}(2)$, ie f_1 has to be reflected another time. But then, after inverting f_1 and connecting the surfaces, $e(\nu) = 0$ for the new surface. Hence, in this very special case, the construction above fails.

Remarks 4.4 (i) We can also argue the other way round: Theorem 3.1 implies that we cannot choose $S \in \mathrm{SO}(2)$ in Lemma 4.3. This implies that $|A^\circ_{11}|^2 = |A^\circ_{12}|^2$ and $\langle A^\circ_{11}, A^\circ_{12}\rangle = 0$ for the Veronese embedding. Moreover, the surface $f\colon K \to \mathbb{R}^4$ that we obtain from Theorem 4.2 must have Euler normal number 0 as $\mathcal{W}(f) < 8\pi$.

(ii) As we can add arbitrary dimensions to $\mathbb{R}^4$ we get by Theorem 4.1 that every Klein bottle can be embedded into $\mathbb{R}^n$, $n \geq 4$, with $\mathcal{W}(f) < 8\pi$.

The existence of a smooth embedding $\widetilde{f_0}\colon K \to \mathbb{R}^n$ for $n \geq 4$ minimizing the Willmore energy in the class of all immersions $\widetilde{f}\colon K \to \mathbb{R}^n$ can be deduced by a compactness theorem of Kuwert and Li [14, Proposition 4.1, Theorem 4.1] and the regularity results of Kuwert and R Schätzle [15] or Rivière [24; 25] if one can rule out diverging in moduli space. We note that Rivière showed independently a compactness theorem similar to the one of Kuwert and Li; see [25]. The nondegenerating property is shown by combining the subsequent Theorem 4.5 and Theorem 4.1. We get the following theorem:

Theorem 1.1 *Let S be the class of all immersions $f\colon \Sigma \to \mathbb{R}^n$ where Σ is a Klein bottle. Consider*

$$\beta_2^n := \inf\{\mathcal{W}(f) : f \in S\}.$$

Then we have that $\beta_2^n < 8\pi$ for $n \geq 4$. Furthermore, β_2^n is attained by a smooth embedded Klein bottle for $n \geq 4$.

Before we prove that a sequence of degenerating Klein bottles always has 8π Willmore energy we explain how we apply certain techniques from [14] to nonorientable closed surfaces.

We repeat our general set-up from the beginning of Section 2: Let N be a nonorientable closed manifold of dimension two and $\widetilde{f}\colon N \to \mathbb{R}^n$ $(n \geq 3)$ an immersion. Consider $q\colon M \to N$, the conformal oriented two-sheeted cover of N, and define $f := \widetilde{f} \circ q$. As every 2–dimensional oriented manifold can be locally conformally reparametrized, M is a Riemann surface that is conformal to $(M, f^*\delta_{\mathrm{eucl}})$. Let $I\colon M \to M$ be the antiholomorphic order-two deck transformation for q. The map I is an antiholomorphic involution without fixpoints such that $f \circ I = f$. From now on we will work with the immersion f on the Riemann surface M equipped with an antiholomorphic involution I. We are not arguing on the quotient space $N = M/\langle I\rangle$.

For the Willmore energy of the immersion f we have

$$\mathcal{W}(f) = 2\mathcal{W}(\widetilde{f}).$$

If $p \in f^{-1}(y)$ then $I(p) \in f^{-1}(y)$, ie the number of preimages of f is always even. We describe this in other words: Consider M as a varifold and consider the push-forward of M via f, ie $f_\sharp M$. Then $f_\sharp M$ is a compactly supported rectifiable varifold with at least multiplicity 2 at every point.

We now consider the case that f is a proper branched conformal immersion (compare [14, page 323]), ie there exists $\Sigma \subset M$ discrete such that $f \in W^{2,2}_{\mathrm{conf,loc}}(M \setminus \Sigma, \mathbb{R}^n)$ and

$$\int_U |A|^2 \, d\mu_{f^*\delta_{\mathrm{eucl}}} < \infty \quad \text{and} \quad \mu_{f^*\delta_{\mathrm{eucl}}}(U) < \infty \quad \text{for all } U \Subset M.$$

We note once again that we have $I(\Sigma) = \Sigma$ since $f \circ I = f$. If $\varphi\colon B_\sigma \to M$ is a local conformal parametrization around $\varphi(0) \in \Sigma$ such that $\varphi(B_\sigma) \cap \Sigma = \varphi(0)$, we may apply the classification of isolated singularities result of Kuwert and Li [14, Theorem 3.1] to $f \circ \varphi$ and conclude that

$$\theta^2(f \circ \varphi_\sharp [\![B_\sigma]\!], f \circ \varphi(0)) = m + 1 \quad \text{for some } m \geq 0.$$

Here, we considered $[\![B_\sigma]\!]$ as a varifold itself. Furthermore $I \circ \varphi\colon B_\sigma \to M$ is an antiholomorphic parametrization around the point $I \circ \varphi(0)$. Applying once more [14, Theorem 3.1] ($I \circ \varphi(B_\sigma) \cap \varphi(B_\sigma) = \varnothing$ by the choice of σ)

$$\theta^2(f \circ I \circ \varphi_\sharp [\![B_\sigma]\!], f \circ I \circ \varphi(0)) = m' + 1 \quad \text{for some } m' \geq 0.$$

We have $m = m'$ since $f \circ I = f$. Combining both local estimates with the monotonicity formula of Simon (which extends to branched conformal immersions) we obtain, for $q = f \circ \varphi(0) = f \circ I \circ \varphi(0)$,

$$\begin{aligned}\mathcal{W}(f) &\geq \theta^2(f_\sharp M, q) \\ &\geq \theta^2(f \circ \varphi_\sharp [\![B_\sigma]\!], q) + \theta^2(f \circ I \circ \varphi_\sharp [\![B_\sigma]\!], q) \\ &\geq 2(m+1)4\pi.\end{aligned} \tag{26}$$

We remark that in general we could have started working on $N = M/\langle I \rangle$ with the associated varifold $\widetilde{f}_\sharp N$ which has density 1 at most points. But we decided to stick to the oriented double cover M since all theorems in the literature are proven on orientable Riemann surfaces.

The following theorem can be considered as the analog of [14, Theorem 5.2] for the nonorientable situation. Our argumentation is inspired by the arguments of Kuwert and Li.

Theorem 4.5 *Let K_m be a sequence of Klein bottles diverging in moduli space. Then for any sequence of conformal immersions $\widetilde{f}_m \in W^{2,2}_{\mathrm{conf}}(K_m, \mathbb{R}^n)$ we have*

$$\liminf_{m\to\infty} \mathcal{W}(\widetilde{f}_m) \geq 8\pi.$$

Proof Let $q_m\colon T_m^2 \to K_m$ be the two sheeted oriented double cover and $I_m\colon T_m^2 \to T_m^2$ the associated antiholomorphic order-two deck transformation. By Theorem 2.1 we may assume that $T_m^2 = \mathbb{C}/\Gamma_m$, where Γ_m is a lattice generated by $(1, ib_m)$ with $b_m \geq 1$ and I_m is given by

$$I_m(z) = \overline{z} + \tfrac{1}{2} \quad \text{or} \quad I_m(z) = -\overline{z} + \tfrac{1}{2} i b_m. \tag{27}$$

Diverging in moduli space implies $\lim_{m\to\infty} b_m = \infty$. We lift the maps $\widetilde{f}_m$ to the double cover T_m^2 and then to Γ_m–periodic maps from $\mathbb{C}$ into $\mathbb{R}^n$, and denote the lifted maps by f_m, ie $f_m \circ I_m = f_m$. By Gauss–Bonnet we may also assume that the maps $f_m\colon \mathbb{C} \to \mathbb{R}^n$ satisfy

$$\limsup_{m\to\infty} \frac{1}{4} \int_{T_m^2} |A_{f_m}|^2 \, d\mu_{g_m} = \limsup_{m\to\infty} \mathcal{W}(f_m) \leq W_0 < \infty.$$

The theorem is proven if we show that

$$\liminf_{m\to\infty} \mathcal{W}(f_m) \geq 16\pi. \tag{28}$$

We have to distinguish two cases. They are determined by the form of the involution. After passing to a subsequence the involution is either of the second kind in (27) for all m (Case 1) or it is the involution $I(z) = \overline{z} + \frac{1}{2}$ for all m (Case 2).

Following the notation of [14] we will denote by I_p the inversion at $\partial B_1(p)$ in $\mathbb{R}^n$, ie $I_p(x) = p + (x-p)/|x-p|^2$ for $x \in \mathbb{R}^n$. Furthermore, for $\delta \in \mathbb{R}$ we define the translations $\eta_\delta(z) := z + i\delta$ for $z \in \mathbb{C}$.

Case 1 *$I_m(z) = -\overline{z} + \frac{1}{2} i b_m$ for all m.*

Proof of (28) in Case 1 The local L^∞–bound of the conformal factor [14, Corollary 2.2] implies that f_m is not constant on any circle $C_v = [0,1] \times \{v\}$. As $f_m \circ I_m = f_m$ we have that $f_m\big([0,1] \times \big[0, \frac{1}{2} b_m\big]\big) = f_m\big([0,1] \times \big[\frac{1}{2} b_m, b_m\big]\big)$. Thus, there exists $v_m \in \big[0, \frac{1}{2} b_m\big)$ such that

$$\lambda_m := \operatorname{diam}(f_m(C_{v_m})) \le \operatorname{diam}(f_m(C_v)) \quad \text{for all } v \in \mathbb{R}.$$

As already mentioned in Lemma 2.5 the involution is not affected by these translations because $\eta_\delta^{-1} \circ I_m \circ \eta_\delta = I_m - 2\Re(i\delta) = I_m$. Consider the two sequences

$$\begin{aligned} h_m(z) &= \lambda_m^{-1}(f_m \circ \eta_{v_m}(z) - f_m \circ \eta_{v_m}(0)), \\ k_m(z) &= \lambda_m^{-1}(f_m \circ \eta_{b_m/2+v_m}(z) - f_m \circ \eta_{b_m/2+v_m}(0)). \end{aligned}$$

We have that $1 = \operatorname{diam}(h_m(C_0)) = \operatorname{diam}(k_m(C_0))$, $0 = h_m(0) = k_m(0)$ for all m and $h_m(z) = k_m(-\overline{z})$. The immersions h_m and k_m are immersed tori diverging in moduli space. We can therefore repeat the proof of Kuwert and Li from [14, Theorem 5.2]. We find a suitable inversion I_{x_0} at a sphere $\partial B_1(x_0)$ and deduce that $\widehat{h}_m := I_{x_0} \circ h_m$, $\widehat{k}_m := I_{x_0} \circ k_m$ converge locally uniformly to branched conformal immersions $\widehat{h}$ and $\widehat{k}$ satisfying $\mathcal{W}(\widehat{h}) \ge 8\pi$ and $\mathcal{W}(\widehat{k}) \ge 8\pi$. Observe that

$$\begin{aligned} \mathcal{W}(f_m) &= \mathcal{W}(I_{x_0} \circ f_m) \\ &= \frac{1}{4} \int_{[0,1]\times[v_m - b_m/4, v_m + b_m/4]} |H_{I_{x_0}\circ f_m}|^2 \, d\mu_{\widehat{g}_m} \\ &\qquad\qquad + \frac{1}{4} \int_{[0,1]\times[v_m + b_m/4, v_m + 3b_m/4]} |H_{I_{x_0}\circ f_m}|^2 \, d\mu_{\widehat{g}_m} \\ &= \mathcal{W}(\widehat{h}_m|_{[0,1]\times[-b_m/4, b_m/4]}) + \mathcal{W}(\widehat{k}_m|_{[0,1]\times[-b_m/4, b_m/4]}). \end{aligned}$$

We pass to the limit and get

$$\liminf_{m\to\infty} \mathcal{W}(f_m) \ge \liminf_{m\to\infty} \mathcal{W}(\widehat{h}_m) + \liminf_{m\to\infty} \mathcal{W}(\widehat{k}_m) \ge \mathcal{W}(\widehat{h}) + \mathcal{W}(\widehat{k}) \ge 16\pi.$$

Note that $\widehat{h}$ and $\widehat{k}$ parametrize the same sphere because of $\widehat{h}(z) = \widehat{k}(-\overline{z})$. This sphere has a double point, as shown in the proof of [14, Theorem 5.2].

Case 2 *$I_m(z) = \overline{z} + \frac{1}{2}$ for all m.*

Proof of (28) in Case 2 Observe that we cannot translate into "imaginary direction" without changing the involution because $\eta_\delta^{-1} \circ I \circ \eta_\delta(z) = I(z) - 2i\delta$. Another delicate point is that the form of the involution does not help to find a "second" torus.

We fix a large integer $M \in \mathbb{N}$ such that $4\pi M \geq W_0$.

For each $m \in \mathbb{N}$ pick $u_m \in \{-M, \dots, M\}$ such that

$$\lambda_m = \operatorname{diam}(f_m(C_{u_m})) = \min_{u \in \{-M, \dots, M\}} \operatorname{diam}(f_m(C_u)). \tag{29}$$

By passing to a subsequence we may assume that $u_m = u_0$ for all m. Furthermore, arguing as in [14, Propositon 4.1] we obtain $B_1(x_1) \subset \mathbb{R}^n$ such that $f_m(T_m^2) \cap B_1(x_1) = \varnothing$ for all m. We consider the sequence

$$h_m(z) := I_{x_1}\big(\lambda_m^{-1}(f_m(z) - f_m \circ \eta_{u_0}(0))\big). \tag{30}$$

Repeat the procedure and fix $v_m \in \{-M, \dots, M\}$ such that

$$\mu_m = \operatorname{diam}(h_m(C_{b_m/2+v_m})) = \min_{v \in \{-M, \dots, M\}} \operatorname{diam}(h_m(C_{b_m/2+v})). \tag{31}$$

By passing to a subsequence we may assume $v_m = v_0$ for all m and define

$$k_m(z) := \mu_m^{-1}(h_m \circ \eta_{b_m/2}(z) - h_m \circ \eta_{b_m/2+v_0}(0)).$$

The translations were chosen so that we still have $h_m \circ I = h_m$ and $k_m \circ I = k_m$ for all m. As before we find $x_2 \in \mathbb{R}^n$ with $k_m(T_m^2) \cap B_1(x_2) = \varnothing$ and consider $\widehat{k}_m = I_{x_2}(k_m)$. We have achieved that $h_m(T_m^2) \subset \overline{B_1(x_1)}$ and $\widehat{k}_m(T_m^2) \subset \overline{B_1(x_2)}$. Lemma 1.1 of [27] implies area bounds $\mu_{g_m}(T_m^2) \leq C$ for both sequences. Up to a subsequence, we have $|A_{h_m}|^2 \, d\mu_{g_m} \to \alpha_1$ and $|A_{\widehat{k}_m}|^2 \, d\mu_{g_m} \to \alpha_2$ as Radon measures on the cylinder $C = [0,1] \times \mathbb{R}$. The sets $\Sigma_i := \{z \in \mathbb{C} : \alpha_i(\{z\}) \geq 4\pi\}$ for $i = 1, 2$ are discrete. Theorem 5.1 in [14] yields that h_m and $\widehat{k}_m$ converge locally uniformly on $C \setminus \Sigma_1$ and $C \setminus \Sigma_2$, respectively. The limits either are conformal immersions $h\colon C \setminus \Sigma_1 \to \mathbb{R}^n$, $\widehat{k}\colon C \setminus \Sigma_2 \to \mathbb{R}^n$ or points p_1 and p_2. Note that by construction

$$\begin{aligned} h_m(C_{u_0}) &\subset I_{x_1}(\overline{B_1(0)}) \subset \mathbb{R}^n \setminus B_{\theta_1}(x_1) \quad \text{with } \theta_1 = \frac{1}{1+|x_1|}, \\ \widehat{k}_m(C_{v_0}) &\subset I_{x_2}(\overline{B_1(0)}) \subset \mathbb{R}^n \setminus B_{\theta_2}(x_2) \quad \text{with } \theta_2 = \frac{1}{1+|x_2|}. \end{aligned} \tag{32}$$

Assume the second alternative holds for h_m, ie $h_m \to p_1$ locally uniformly. Observe that $C_u \cap \Sigma_1 = \varnothing$ for at least one $u_* \in \{-M, \dots, M\}$. Otherwise there would be points $z_u \in C_u$ with $\alpha_1(B_{1/4}(z_u)) > 4\pi$ for each $u \in \{-M, \dots, M\}$ contradicting $\sum_{u=-M}^{M} \alpha_1(B_{1/4}(z_u)) \leq \alpha_1(C) \leq W_0$.

Due to (32) we have $|p_1 - x_1| \geq \theta_1 > 0$ and hence $I_{x_1}(h_m(C_{u_*})) \to I_{x_1}(p_1)$ uniformly. But $\operatorname{diam}(I_{x_1}(h_m(C_{u_*}))) \geq 1$ by (29) and (30), a contradiction. In the same way we exclude $\widehat{k}_m \to p_2$.

By uniform local convergence we get

$$\begin{aligned} h &= \lim_{m\to\infty} h_m = \lim_{m\to\infty} h_m \circ I = h \circ I, \\ \widehat{k} &= \lim_{m\to\infty} \widehat{k}_m = \lim_{m\to\infty} \widehat{k}_m \circ I = \widehat{k} \circ I, \end{aligned}$$

and we are in the situation of branched $W^{2,2}$–conformal immersions that are invariant under I. We now investigate the behavior of h and $\widehat{k}$ at the ends $\{\pm\infty\}$ of the cylinder C. We present the argument for h; the argument for $\widehat{k}$ works analogously. We note that $\varphi_+(z) := \frac{-i}{2\pi}\ln(z)$ is a holomorphic chart around $+\infty$ and $I \circ \varphi_+(z) = \frac{i}{2\pi}\ln(\overline{z}) + \frac{1}{2}$ is an antiholomorphic chart around $-\infty$. Since $\int_C |A_h|^2\, d\mu_g \leq \alpha_1(C) < \infty$, the map $h_+(z) := h \circ \varphi_+(z)$ is a $W^{2,2}_{\mathrm{loc}}(B_\sigma \backslash \{0\}, \mathbb{R}^n)$–conformal immersion with $\Sigma_1 \cap \varphi_+(B_\sigma \setminus \{0\}) = \varnothing$ for $\sigma > 0$ sufficiently small. We follow the explanations presented in front of Theorem 4.5 (page 2519) to conclude that the varifold $h_{+\sharp}[\![B_\sigma]\!]$ extends continuously to 0. This implies that $h(C_v) \to q_1$ for $\nu \to \pm\infty$ using the fact that $I(C_v) = C_{-v}$. Furthermore, applying the Li–Yau inequality (a version for branched immersion can be found in [14, Formula (3.1)]) yields the following lower bound of the Willmore energy of h. A detailed explanation how we apply this inequality to the oriented double covers was done on page 2519; see (26). We have that

$$\begin{aligned} \mathcal{W}(h) &\geq \theta^2(h_\sharp C, q_1) \\ &\geq \theta^2(h_{+\sharp}[\![B_\sigma]\!], q) + \theta^2(h \circ I \circ \varphi_{+\sharp}[\![B_\sigma]\!], q) \\ &\geq 2(m(+\infty) + 1)\, 4\pi. \end{aligned}$$

Thus, if $m(+\infty) \geq 1$ we have that $\mathcal{W}(h) \geq 4 \cdot 4\pi$. The very same argument applies to $\widehat{k}$.

Similarly we exclude branch points for the maps h and $\widehat{k}$ in the interior of the cylinder C as follows. Suppose that the application of the classification theorem of isolated singularities [14, Theorem 3.1] to a point $z \in \Sigma_1$ reveals a point with branching order $m(z) \geq 1$; then, by (26), we conclude $\mathcal{W}(h) \geq 4 \cdot 4\pi = 16\pi$. In the same way we can assume that all points $z \in \Sigma_2$ are removable singularities, ie $m(z) = 0$ and $\widehat{k}$ has removable singularities in $\pm\infty$.

It remains to handle the situation where h and $\widehat{k}$ are unbranched. Since $h \circ I = h$ and $\widehat{k} \circ I = \widehat{k}$, and h and $\widehat{k}$ extend smoothly to $\pm\infty$, they are double covers of immersions of an $\mathbb{R}P^2$ into $\mathbb{R}^n$. By the work of Li and Yau [19] we get $\min\{\mathcal{W}(h), \mathcal{W}(\widehat{k})\} \geq$

$2 \cdot 6\pi = 12\pi$ as h and $\widehat{k}$ are the oriented double covers of the unoriented surfaces. Recall once more that this also implies that $\#h^{-1}(\{x\})$ and $\#\widehat{k}^{-1}(\{x\})$ are even for all $x \in \mathbb{R}^n$.

If $\#\widehat{k}^{-1}(\{x_1\}) > 2$, the Li–Yau inequality implies $\mathcal{W}(\widehat{k}) \geq 4\pi \cdot \#\widehat{k}^{-1}(\{x_1\}) \geq 16\pi$. Otherwise let $C_m := [0,1] \times \left[-\frac{1}{4}b_m, \frac{1}{4}b_m\right]$. We observe that $k_m = I_{x_2} \circ \widehat{k}_m \to I_{x_2} \circ \widehat{k}$ and

$$\mathcal{W}(h_m) = \mathcal{W}(h_m|_{C_m}) + \mathcal{W}(h_m|_{\eta_{b_m/2}(C_m)}) = \mathcal{W}(h_m|_{C_m}) + \mathcal{W}(k_m|_{C_m}).$$

With $\#\widehat{k}(x_1) \leq 2$, we conclude by [14, Formula (3.1)] that

$$\begin{aligned}\liminf_{m\to\infty} W(h_m) &\geq \liminf_{m\to\infty} \mathcal{W}(h_m|_{C_m}) + \liminf_{m\to\infty} \mathcal{W}(k_m|_{C_m}) \\ &\geq \mathcal{W}(h) + \mathcal{W}(I_{x_1}(\widehat{k})) \geq 12\pi + (12\pi - 8\pi) = 16\pi. \qquad \square\end{aligned}$$

References

[1] **L V Ahlfors**, *Complex analysis: an introduction of the theory of analytic functions of one complex variable*, 2nd edition, McGraw-Hill, New York (1966) MR

[2] **M F Atiyah**, **N J Hitchin**, **I M Singer**, *Self-duality in four-dimensional Riemannian geometry*, Proc. Roy. Soc. London Ser. A 362 (1978) 425–461 MR

[3] **T F Banchoff**, *Triple points and surgery of immersed surfaces*, Proc. Amer. Math. Soc. 46 (1974) 407–413 MR

[4] **M Bauer**, **E Kuwert**, *Existence of minimizing Willmore surfaces of prescribed genus*, Int. Math. Res. Not. 2003 (2003) 553–576 MR

[5] **L P Eisenhart**, *Minimal surfaces in Euclidean four-space*, Amer. J. Math. 34 (1912) 215–236 MR

[6] **J Eschenburg**, **R Tribuzy**, *Branch points of conformal mappings of surfaces*, Math. Ann. 279 (1988) 621–633 MR

[7] **T Friedrich**, *On surfaces in four-spaces*, Ann. Global Anal. Geom. 2 (1984) 257–287 MR

[8] **J Hirsch**, **E Mäder-Baumdicker**, *A note on Willmore minimizing Klein bottles in Euclidean space*, preprint (2016) arXiv

[9] **M W Hirsch**, *Immersions of manifolds*, Trans. Amer. Math. Soc. 93 (1959) 242–276 MR

[10] **M Koecher**, **A Krieg**, *Elliptische Funktionen und Modulformen*, revised edition, Springer (2007) MR

[11] **R Kusner**, *Conformal geometry and complete minimal surfaces*, Bull. Amer. Math. Soc. 17 (1987) 291–295 MR

[12] **R Kusner**, *Estimates for the biharmonic energy on unbounded planar domains, and the existence of surfaces of every genus that minimize the squared-mean-curvature integral*, from "Elliptic and parabolic methods in geometry" (B Chow, R Gulliver, S Levy, J Sullivan, editors), A K Peters, Wellesley, MA (1996) 67–72 MR

[13] **R Kusner**, **N Schmitt**, *The spinor representations of surfaces in space*, preprint (1996) arXiv

[14] **E Kuwert**, **Y Li**, *$W^{2,2}$–conformal immersions of a closed Riemann surface into $\mathbb{R}^n$*, Comm. Anal. Geom. 20 (2012) 313–340 MR

[15] **E Kuwert**, **R Schätzle**, *Minimizers of the Willmore functional under fixed conformal class*, J. Differential Geom. 93 (2013) 471–530 MR

[16] **H Lapointe**, *Spectral properties of bipolar minimal surfaces in $\mathbb{S}^4$*, Differential Geom. Appl. 26 (2008) 9–22 MR

[17] **R Lashof**, **S Smale**, *On the immersion of manifolds in euclidean space*, Ann. of Math. 68 (1958) 562–583 MR

[18] **H B Lawson, Jr**, *Complete minimal surfaces in S^3*, Ann. of Math. 92 (1970) 335–374 MR

[19] **P Li**, **S T Yau**, *A new conformal invariant and its applications to the Willmore conjecture and the first eigenvalue of compact surfaces*, Invent. Math. 69 (1982) 269–291 MR

[20] **F C Marques**, **A Neves**, *The Willmore conjecture*, Jahresber. Dtsch. Math.-Ver. 116 (2014) 201–222 MR

[21] **W S Massey**, *Proof of a conjecture of Whitney*, Pacific J. Math. 31 (1969) 143–156 MR

[22] **J W Milnor**, **J D Stasheff**, *Characteristic classes*, Annals of Mathematics Studies 76, Princeton Univ. Press (1974) MR

[23] **R Miranda**, *Algebraic curves and Riemann surfaces*, Graduate Studies in Mathematics 5, Amer. Math. Soc., Providence, RI (1995) MR

[24] **T Rivière**, *Analysis aspects of Willmore surfaces*, Invent. Math. 174 (2008) 1–45 MR

[25] **T Rivière**, *Variational principles for immersed surfaces with L^2–bounded second fundamental form*, J. Reine Angew. Math. 695 (2014) 41–98 MR

[26] **S Salamon**, *Topics in four-dimensional Riemannian geometry*, from "Geometry seminar 'Luigi Bianchi'" (E Vesentini, editor), Lecture Notes in Math. 1022, Springer (1983) 33–124 MR

[27] **L Simon**, *Existence of surfaces minimizing the Willmore functional*, Comm. Anal. Geom. 1 (1993) 281–326 MR

[28] **H Whitney**, *The self-intersections of a smooth n–manifold in $2n$–space*, Ann. of Math. 45 (1944) 220–246 MR

[29] **P Wintgen**, *Sur l'inégalité de Chen–Willmore*, C. R. Acad. Sci. Paris Sér. A–B 288 (1979) A993–A995 MR

Pastor-Felke-Str. 1, D-76131 Karlsruhe, Germany

Scuola Internazionale Superiore di Studi Avanzati
Via Bonomea, 265, 34136 Trieste, Italy

Institute for Analysis, Karlsruhe Institute of Technology
Englerstr. 2, D-76131 Karlsruhe, Germany

pbreuning@yahoo.de, jhirsch@sissa.it, elena.maeder-baumdicker@kit.edu

Proposed: Tobias H. Colding — Received: 6 April 2016
Seconded: Ian Agol, Leonid Polterovich — Revised: 8 August 2016

msp

Geometry & Topology 21 (2017) 2527–2555

A spectral sequence for stratified spaces and configuration spaces of points

DAN PETERSEN

We construct a spectral sequence associated to a stratified space, which computes the compactly supported cohomology groups of an open stratum in terms of the compactly supported cohomology groups of closed strata and the reduced cohomology groups of the poset of strata. Several familiar spectral sequences arise as special cases. The construction is sheaf-theoretic and works both for topological spaces and for the étale cohomology of algebraic varieties. As an application we prove a very general representation stability theorem for configuration spaces of points.

55R80; 55T05, 32S60, 14F25, 55N30

1 Introduction

Let $X = \bigcup_{\alpha \in P} S_\alpha$ be a stratified space. By this we mean that the topological space X is the union of disjoint locally closed subspaces S_α called the *strata*, and that the closure of each stratum is itself a union of strata. The set P of strata becomes partially ordered by declaring that $\alpha \le \beta$ if $\overline{S}_\alpha \supseteq S_\beta$.

Let $\chi_c(-)$ denote the compactly supported Euler characteristic of a space. Since this invariant is additive over stratifications, one has an equality

$$\chi_c(\overline{S}_\alpha) = \sum_{\alpha \le \beta} \chi_c(S_\beta) \tag{1}$$

for all $\alpha \in P$. By the Möbius inversion formula for the poset P, it therefore holds that

$$\chi_c(S_\alpha) = \sum_{\alpha \le \beta} \mu_P(\alpha, \beta) \cdot \chi_c(\overline{S}_\beta), \tag{2}$$

where μ_P is the Möbius function of the poset. This expresses the simple combinatorial fact that if one knows all the integers $\chi_c(\overline{S}_\alpha)$, then one can also determine the integers $\chi_c(S_\alpha)$ by inclusion-exclusion.

Equation (1) can be upgraded (or "categorified") to a relationship between actual cohomology groups. Suppose $\sigma \colon P \to \mathbb{Z}$ is a function such that $\sigma(\alpha) < \sigma(\beta)$ if $\alpha < \beta$. Such a function defines a filtration of $\overline{S}_\alpha$ by closed subspaces, and the corresponding

Published: 19 May 2017 DOI: 10.2140/gt.2017.21.2527

spectral sequence in compactly supported cohomology reads

$$E_1^{pq} = \bigoplus_{\substack{\alpha \leq \beta \\ \sigma(\beta) = -p}} H_c^{p+q}(S_\beta, \mathbb{Z}) \implies H_c^{p+q}(\bar{S}_\alpha, \mathbb{Z}). \tag{3}$$

By equating the Euler characteristics of the E_1 and E_∞ pages of this spectral sequence one recovers (1).

It is then natural to ask whether also the dual equation (2) admits a similar interpretation. The quantity $\mu_P(\alpha, \beta)$ is *also* an Euler characteristic, by Philip Hall's theorem: the Möbius function $\mu_P(\alpha, \beta)$ equals the reduced Euler characteristic of $\mathrm{N}(\alpha, \beta)$, by which we mean the nerve of the poset (α, β), where (α, β) denotes an open interval in P. (The preceding is valid only if $\alpha < \beta$: in the degenerate case $\alpha = \beta$ it is natural to define the reduced cohomology of $\mathrm{N}(\alpha, \beta)$ to be $\mathbb{Z}$ in degree -2, as we explain in Section 2.) In any case, one can expect such a categorification to also involve the reduced cohomology groups of the poset.

In this paper, we construct a spectral sequence accomplishing this goal:

Theorem 1.1 *There exists a spectral sequence*

$$E_1^{pq} = \bigoplus_{\substack{\alpha \leq \beta \\ \sigma(\beta) = p}} \bigoplus_{i+j+2=p+q} H_c^j(\bar{S}_\beta, \widetilde{H}^i(\mathrm{N}(\alpha, \beta), \mathbb{Z})) \implies H_c^{p+q}(S_\alpha, \mathbb{Z}).$$

Taking Euler characteristics of both sides, we recover (2). This would seem a very natural question — given the cohomology of the closed strata, how does one compute the cohomology of open strata? — and it is close in spirit to the work of Vassiliev [30; 31]. Yet to my knowledge the result is new.

The proof is elementary and completely sheaf-theoretic, and the theorem we prove in the body of the paper is a more general statement that is valid with coefficients given by any sheaf or complex of sheaves $\mathcal{F}$ on X. It also works in the setting of ℓ–adic sheaves, if X is an algebraic variety: in this case, the spectral sequence is a spectral sequence of ℓ–adic Galois representations.

As an application of our result we prove a very general representation stability theorem for configuration spaces of points. In particular, a novel feature is that if one is willing to work with Borel–Moore homology (or, dually, compact support cohomology), then one can prove homological stability results for an arbitrary topological space M satisfying rather mild hypotheses; to my knowledge, all existing results in the literature prove homological stability for configuration spaces of points on *manifolds*. In this introduction we focus on the case when M is a (possibly singular) algebraic variety, in which case the result is easier to state.

Let M be a space, and let $\mathcal{A}$ be a finite collection of closed subspaces $A_i \subset M^{n_i}$. We define a configuration space $F_{\mathcal{A}}(M,n)$, parametrizing n ordered points on M "avoiding all $\mathcal{A}$–configurations". For instance, if $\mathcal{A}$ consists only of the diagonal inside M^2, then $F_{\mathcal{A}}(M,n)$ is the usual configuration space of distinct ordered points on M.

Theorem 1.2 *Let M be a geometrically irreducible d–dimensional algebraic variety over a field κ, and $\mathcal{A}$ an arbitrary finite collection of closed subvarieties $A_i \subset M^{n_i}$.*

(1) *For $\kappa = \mathbb{C}$, the (singular) Borel–Moore homology groups*

$$H^{\mathrm{BM}}_{i+2dn}(F_{\mathcal{A}}(M,n)(\mathbb{C}),\mathbb{Z})$$

form a finitely generated FI-module for all $i \in \mathbb{Z}$.

(2) *The (étale) Borel–Moore homology groups*

$$H^{\mathrm{BM}}_{i+2dn}(F_{\mathcal{A}}(M,n)_{\overline{\kappa}},\mathbb{Z}_\ell(-dn))$$

form a finitely generated FI-module in ℓ–adic $\mathrm{Gal}(\overline{\kappa}/\kappa)$–representations, for all $i \in \mathbb{Z}$, whenever ℓ is a prime different from $\mathrm{char}(\kappa)$.

In particular, the homology groups $H^{\mathrm{BM}}_{i+2dn}(F_{\mathcal{A}}(M,n),\mathbb{Q})$ form a representation stable sequence of $\mathbb{S}_n$–representations, and the $\mathbb{S}_n$–invariants $H^{\mathrm{BM}}_{i+2dn}(F_{\mathcal{A}}(M,n)/\mathbb{S}_n,\mathbb{Q})$ satisfy homological stability as $n \to \infty$.

If M is smooth, or at least a homology manifold, we may conclude instead that the cohomology groups $H^i(F_{\mathcal{A}}(M,n),\mathbb{Z})$ form a finitely generated FI-module, by Poincaré duality. See Remark 4.17.

The fact that we obtain a finitely generated FI-module over $\mathbb{Z}$ gives a homological stability result with rational coefficients, but it also has interesting consequences for the mod p homology of the unordered configuration spaces: by results of Nagpal [24], our theorem implies that the groups $H^{\mathrm{BM}}_{i+2dn}(F_{\mathcal{A}}(M,n)/\mathbb{S}_n,\mathbb{F}_p)$ become eventually periodic.

Vakil and Wood [29] introduced certain configuration spaces $\overline{w}^c_\lambda(M)$ depending on a partition λ. For a suitable choice of $\mathcal{A}$, one has $F_{\mathcal{A}}(M,n)/\mathbb{S}_n = \overline{w}^c_\lambda(M)$, so Theorem 1.2 implies in particular a homological stability theorem for the spaces $\overline{w}^c_\lambda(M)$ as $n \to \infty$, which gives a proof of [29, Conjecture F]. This conjecture has previously been proven by Kupers, Miller and Tran [21]. Compared to their proof, our proof gives the stronger assertion of representation stability, and makes no smoothness assumptions about M (they assume M is a smooth manifold). On the other hand, their proof gives in many cases integral stability for the unordered configuration space, and they give an explicit stability range. The latter should be possible in our setting, too, but we have not done so.

Vakil and Wood formulated their conjecture after making point counts of varieties over finite fields, and using the Grothendieck–Lefschetz trace formula to guess what the cohomology should look like. Since the Grothendieck–Lefschetz trace formula concerns *compact support* cohomology, it is in a sense natural that we obtain stronger results when working with compact support cohomology/Borel–Moore homology from the start.

As mentioned, one can prove also a version of Theorem 1.2 for an arbitrary topological space, but the assumptions on M and $\mathcal{A}$ become more cumbersome to state. However, if we let $\mathcal{A}$ be the arrangement leading to the configuration spaces considered by Vakil and Wood, the hypotheses are quite simple: if M is any locally compact topological space with finitely generated Borel–Moore homology groups, and such that there exists an integer $d \geq 2$ for which $H_d^{\mathrm{BM}}(M,\mathbb{Z}) \cong \mathbb{Z}$ and $H_i^{\mathrm{BM}}(M,\mathbb{Z}) = 0$ for $i > d$, then $H_{i+dn}^{\mathrm{BM}}(F_{\mathcal{A}}(M,n),\mathbb{Z})$ is a finitely generated FI-module if d is even; if d is odd, one needs to twist by the sign representation, and $H_{i+dn}^{\mathrm{BM}}(F_{\mathcal{A}}(M,n),\mathbb{Z}) \otimes \mathrm{sgn}_n$ is a finitely generated FI-module.

2 Generalities on posets

Let P be a poset, always assumed to be finite. We define its *nerve* $\mathrm{N}P$ to be the simplicial complex with vertices the elements of P, and a subset $S \subseteq P$ forms a face if and only if all elements of S are pairwise comparable. The corresponding simplicial set is exactly the usual nerve of P, when P is thought of as a category.

We use $\widetilde{C}_\bullet(\Delta)$ to denote the *reduced cellular chains* of a simplicial complex Δ. The group $\widetilde{C}_i(\Delta)$ is free abelian on the set of i–dimensional faces; we include the empty set as a (-1)–dimensional face. The homology of this chain complex is $\widetilde{H}_\bullet(\Delta,\mathbb{Z})$. However, we will prefer to work with cohomology. The usual definition would be to set

$$\widetilde{C}^\bullet(\Delta) = \mathrm{Hom}(\widetilde{C}_\bullet(\Delta),\mathbb{Z})$$

but we will find it more convenient to use the distinguished basis of $\widetilde{C}_i(\Delta)$ to consider $\widetilde{C}^i(\Delta)$ as *also* being free abelian on the set of i–dimensional faces; then the differential becomes an alternating sum over ways of *adding* an element to a face.

If $x \leq y$ in P, we denote by $\widetilde{C}^\bullet(x,y)$ the chain complex which in degree d is the free abelian group spanned by the increasing sequences

$$x = z_{-1} < z_0 < z_1 < \cdots < z_d < z_{d+1} = y,$$

and whose differential $\partial\colon \widetilde{C}_d \to \widetilde{C}_{d+1}$ is an alternating sum over ways of adding an element to the sequence. For $x < y$, $\widetilde{C}^\bullet(x,y)$ is equal to $\widetilde{C}^\bullet(\mathrm{N}(x,y))$, where (x,y)

denotes an open interval in P; for $x = y$, it consists of $\mathbb{Z}$ placed in degree -2. We denote by $\widetilde{H}^\bullet(x,y)$ the cohomology of this cochain complex.

Let $\sigma\colon P \to \mathbb{Z}$ be a strictly increasing function, eg a grading or a linear extension.

Proposition 2.1 *For $x < y$ in P, there exists a spectral sequence*

$$E_1^{pq} = \bigoplus_{\substack{x \le z \le y \\ \sigma(z) = p}} \widetilde{H}^{p+q-1}(x,z)$$

converging to zero.

Proof Consider the chain complex $\widetilde{C}^\bullet = \widetilde{C}^\bullet(\mathrm{N}(x,y])$, where $(x,y]$ denotes a half-open interval in P. Since $(x,y]$ has a unique maximal element its nerve is contractible, so the complex $\widetilde{C}^\bullet$ is acyclic. We identify $\widetilde{C}^d$ with the set of increasing sequences

$$x = z_{-1} < z_0 < z_1 < \cdots < z_d \le y.$$

Define a decreasing filtration on this complex by taking $F^p\widetilde{C}^d$ to be the span of all sequences with $\sigma(z_d) \ge p$. This makes $\widetilde{C}^\bullet$ a filtered complex. Consider the quotient

$$F^p\widetilde{C}^\bullet / F^{p+1}\widetilde{C}^\bullet$$

and its induced differential. Then $F^p\widetilde{C}_d / F^{p+1}\widetilde{C}_d$ has a basis consisting of sequences such that $\sigma(z_d)$ is *exactly* equal to p, and the differential is a sum over all ways of adding an element to the sequence coming *before* z_d. It therefore follows that the quotient is isomorphic to the direct sum

$$\bigoplus_{\substack{x \le z \le y \\ \sigma(z) = p}} \widetilde{C}^{\bullet-1}(x,z),$$

by an isomorphism taking the sequence

$$x = z_{-1} < z_0 < z_1 < \cdots < z_d \le y \in F^p\widetilde{C}^d / F^{p+1}\widetilde{C}^d$$

to the sequence

$$x = z_{-1} < z_0 < z_1 < \cdots z_{d-1} < z_d = z_d \in \widetilde{C}^{d-1}(x, z_d).$$

Thus the spectral sequence associated to this filtration has the required form. □

Let $\mathrm{Int}(P)$ denote the set of pairs $(x,y) \in P \times P$ with $x \le y$. We define the *Möbius function*

$$\mu\colon \mathrm{Int}(P) \to \mathbb{Z}$$

by $\mu(x,y) = \sum_i (-1)^i \operatorname{rank} \widetilde{H}^i(x,y)$.

Proposition 2.2 *If $x < y$, then $\sum_{x \le z \le y} \mu(x,z) = 0$.*

Proof The left-hand side is the Euler characteristic of the E_1 page of the spectral sequence constructed in Proposition 2.1, and the right-hand side is the Euler characteristic of the E_∞ page. □

A consequence of this is a simple recursive procedure for calculating the Möbius function: the Möbius function could equivalently have been defined as

$$\mu(x,y) = \begin{cases} 1 & \text{if } x = y, \\ -\sum_{x \le z < y} \mu(x,z) & \text{if } x < y. \end{cases}$$

In most treatments this is taken as the definition of the Möbius function. The fact that $\mu(x,y)$ for $x < y$ equals the reduced Euler characteristic of the nerve of the interval (x,y) is then called Philip Hall's theorem. We may think of Proposition 2.1 as a categorification of the usual recursion for the Möbius function.

3 The construction and several examples

Let $X = \bigcup_{\alpha \in P} S_\alpha$ be a stratified space. By this we mean that the space X is the union of disjoint locally closed subspaces S_α called the *strata*, and that the closure of each S_α is itself a union of strata. By a "space" we mean *either*:

(1) X is a locally compact Hausdorff topological space,

(2) X is an algebraic variety over some field.

In the former case, "sheaf" will just mean "sheaf of abelian groups"; in the latter case, "sheaf" will mean "constructible ℓ–adic sheaf, for ℓ different from the characteristic".

The set P of strata becomes partially ordered by declaring that $\alpha \le \beta$ if $\overline{S}_\alpha \supseteq S_\beta$. We assume for simplicity (and without loss of generality) that P has a unique minimal element 0, ie a unique open dense stratum S_0.

For $\alpha \in P$ we denote by j_α the locally closed inclusion $S_\alpha \hookrightarrow X$, and by i_α the inclusion $\overline{S}_\alpha \hookrightarrow X$ of the closure of a stratum.

For $d \ge 0$, define a sheaf

$$L^d(\mathcal{F}) = \bigoplus_{0=\alpha_0<\alpha_1<\cdots<\alpha_d \in P} (i_{\alpha_d})_*(i_{\alpha_d})^*\mathcal{F}$$

on X. In particular, $L^0(\mathcal{F}) = \mathcal{F}$. We may define a differential

$$L^d(\mathcal{F}) \to L^{d+1}(\mathcal{F})$$

as an alternating sum over ways of adding an element to the sequence $\alpha_0 < \alpha_1 < \cdots < \alpha_d$, just as in our definition of $\widetilde{C}^\bullet(\Delta)$ for a simplicial complex Δ; when the element we add appears at the end of the sequence, ie as $\alpha_{d+1} > \alpha_d$, then the differential uses the map

$$(i_{\alpha_d})_*(i_{\alpha_d})^*\mathcal{F} \to (i_{\alpha_{d+1}})_*(i_{\alpha_{d+1}})^*\mathcal{F}$$

obtained from the fact that $i_{\alpha_{d+1}}$ factors through i_{α_d}. This makes $L^\bullet(\mathcal{F})$ into a complex, for the same reason that $\widetilde{C}^\bullet(\Delta)$ is.

Proposition 3.1 *The complex $L^\bullet(\mathcal{F})$ is quasi-isomorphic to $j_{0!}j_0^{-1}\mathcal{F}$, where j_0 is the inclusion of the open stratum.*

Proof We show that

$$j_{0!}j_0^{-1}\mathcal{F} \to L^0(\mathcal{F}) \to L^1(\mathcal{F}) \to L^2(\mathcal{F}) \to \cdots$$

is an acyclic complex of sheaves. It suffices to check this on stalks. For $x \in S_0$ the induced sequence on stalks reads

$$\mathcal{F}_x \to \mathcal{F}_x \to 0 \to 0 \to \cdots,$$

with the map $\mathcal{F}_x \to \mathcal{F}_x$ the identity. Thus we may restrict attention to x in some stratum S_β, $\beta \neq 0$. In this case $(j_{0!}j_0^{-1}\mathcal{F})_x$ will of course vanish, and we will have

$$L^d(\mathcal{F})_x \cong \bigoplus_{\substack{0=\alpha_0<\alpha_1<\cdots<\alpha_d \in P\\ \alpha_d \leq \beta}} \mathcal{F}_x,$$

with the differential on $L^\bullet(\mathcal{F})_x$ given by adding an element to the sequence of α_i's. But this means that $L^\bullet(\mathcal{F})_x$ is (up to a degree shift) the tensor product of $\mathcal{F}_x$ with the complex $\widetilde{C}^\bullet(\mathrm{N}(0,\beta])$, which is acyclic because the poset $(0,\beta]$ has a unique maximal element. □

Remark 3.2 Another way to think about the complex $L^\bullet(\mathcal{F})$ is that $j_{0!}j_0^{-1}\mathcal{F}[-1]$ can be calculated as the cone of $\mathcal{F} \to i_*i^*\mathcal{F}$, where $i\colon (X \setminus S_0) \to X$ is the inclusion. Now $i_*i^*\mathcal{F}$ may be calculated as the homotopy limit of the various $(i_\alpha)_*(i_\alpha)^*\mathcal{F}$ (for $\alpha \neq 0$), and $L^{\geq 1}(\mathcal{F})$ is the bar resolution computing this homotopy limit.

Suppose we are given $\sigma\colon P \to \mathbb{Z}$ an increasing function. We may now define a decreasing filtration of $L^\bullet(\mathcal{F})$ by taking

$$F^pL^d(\mathcal{F}) = \bigoplus_{\substack{0=\alpha_0<\alpha_1<\cdots<\alpha_d \in P\\ \sigma(\alpha_d) \geq p}} (i_{\alpha_d})_*(i_{\alpha_d})^*\mathcal{F}.$$

The compactly supported hypercohomology spectral sequence associated to this filtration reads

$$E_1^{pq} = \mathbb{H}_c^{p+q}(X, \mathrm{Gr}_F^p L^\bullet(\mathcal{F})) \implies \mathbb{H}_c^{p+q}(X, L^\bullet(\mathcal{F})) = H_c^{p+q}(S_0, j_0^{-1}\mathcal{F}),$$

where the second equality is the preceding proposition. Thus we should understand the associated graded $\mathrm{Gr}_F^p L^\bullet(\mathcal{F})$. By arguments just like those in the proof of Proposition 2.1, the associated graded can be written as

$$\mathrm{Gr}_F^p L^d(\mathcal{F}) = \bigoplus_{\substack{\beta\in P\\ \sigma(\beta)=p}} \bigoplus_{0=\alpha_0<\alpha_1<\cdots<\alpha_d=\beta} (i_\beta)_*(i_\beta)^*\mathcal{F},$$

and hence

$$\mathrm{Gr}_F^p L^\bullet(\mathcal{F}) = \bigoplus_{\substack{\beta\in P\\ \sigma(\beta)=p}} \widetilde{C}^{\bullet+2}(0,\beta)\otimes(i_\beta)_*(i_\beta)^*\mathcal{F} = \bigoplus_{\substack{\beta\in P\\ \sigma(\beta)=p}} \widetilde{H}^{\bullet+2}(0,\beta)\otimes^L (i_\beta)_*(i_\beta)^*\mathcal{F}.$$

The last equality uses that $\widetilde{C}^{\bullet+2}(0,\beta)$ is a complex of free modules, so it calculates the derived tensor product, and that any complex of abelian groups is quasi-isomorphic to its cohomology.

In full generality, this cannot be simplified further. However, in most cases occurring in practice it can:

(1) If $\mathcal{F} = R$ is a constant sheaf associated to the ring R, then $\widetilde{H}^{\bullet+2}(0,\beta)\otimes^L (i_\beta)_*(i_\beta)^*\mathcal{F}$ is the constant sheaf $\widetilde{H}^{\bullet+2}(0,\beta;R)$.

(2) If the cohomology groups $\widetilde{H}^i(0,\beta)$ are torsion-free, or if $\mathcal{F}$ is a sheaf of k–vector spaces for some field k, then we may replace the derived tensor product with the usual tensor product.

Let us state our main result only in these two simpler situations.

Theorem 3.3 *Let $X = \bigcup_{\beta\in P} S_\beta$ be a stratified space, where the set P is partially ordered by reverse inclusion of the closures of strata. Choose a function $\sigma\colon P\to\mathbb{Z}$ such that $\sigma(x)<\sigma(y)$ if $x<y$.*

(i) *For any ring R, there is a spectral sequence*

$$E_1^{pq} = \bigoplus_{\substack{\beta\in P\\ \sigma(\beta)=p}} \bigoplus_{i+j=p+q-2} H_c^j(\overline{S}_\beta, \widetilde{H}^i(0,\beta;R)) \implies H_c^{p+q}(S_0,R).$$

(ii) *If $\mathcal{F}$ is a sheaf on X, and we assume either that $\mathcal{F}$ is a sheaf of k–vector spaces or that each interval $(0,\beta)$ in P has torsion-free cohomology, then there is a spectral sequence*

$$E_1^{pq} = \bigoplus_{\substack{\beta \in P \\ \sigma(\beta)=p}} \bigoplus_{i+j=p+q-2} \widetilde{H}^i(0,\beta) \otimes H_c^j(\bar{S}_\beta, i_\beta^* \mathcal{F}) \implies H_c^{p+q}(S_0, j_0^{-1}\mathcal{F}).$$

If in (ii) *X is an algebraic variety and $\mathcal{F}$ is an ℓ–adic sheaf, then this spectral sequence is a spectral sequence of Galois representations, if the cohomology groups $\widetilde{H}^i(0,\beta)$ are given the trivial Galois action.*

3.1 Examples and applications

Example 3.4 If the stratification consists only of a closed subspace $i\colon Z \hookrightarrow X$, then the complex $L^\bullet(\mathcal{F})$ reduces to the two-term complex $\mathcal{F} \to i_* i^* \mathcal{F}$, and the spectral sequence reduces to the long exact sequence

$$\cdots \to H_c^k(X,\mathcal{F}) \to H_c^k(Z,\mathcal{F}) \to H_c^{k+1}(X \setminus Z,\mathcal{F}) \to H_c^{k+1}(X,\mathcal{F}) \to \cdots . \quad \triangleleft$$

Example 3.5 Let X be a complex manifold, and $D = D_1 \cup \cdots \cup D_k$ a strict normal crossing divisor. Consider the stratification of X by the various intersections of the components of D. For $I \subset \{1,\ldots,k\}$, let $D_I = \bigcap_{i\in I} D_i$, including $D_\varnothing = X$. Each interval in the poset of strata is a boolean lattice, so its reduced cohomology vanishes below the top degree, where it is one-dimensional. The spectral sequence therefore reduces to

$$E_1^{pq} = \bigoplus_{|I|=p} H_c^q(D_I,\mathbb{Z}) \implies H_c^{p+q}(X \setminus D,\mathbb{Z}).$$

This is the Poincaré dual of the spectral sequence used by Deligne to construct the mixed Hodge structure on a smooth noncompact complex algebraic variety [11]. In the algebraic case, it is a spectral sequence of mixed Hodge structures/Galois representations. $\triangleleft$

Example 3.6 Suppose $X = \mathbb{A}^n$ is affine space over a field, and the stratification consists of all the intersections in some subspace arrangement. Let $\sigma(\alpha) = -\dim(S_\alpha)$. Let $\mathcal{F} = \mathbb{Q}_\ell$. In this case the spectral sequence simplifies to

$$E_1^{pq} = \bigoplus_{\substack{\beta \in P \\ \sigma(\beta)=p}} \widetilde{H}^{q+3p-2}(0,\beta) \otimes \mathbb{Q}_\ell(p) \implies H_c^{p+q}(S_0,\mathbb{Q}_\ell).$$

Since all columns have different weight there can be no differentials in the spectral sequence. It follows that (up to semisimplification of the Galois representation)

$$H_c^n(S_0, \mathbb{Q}_\ell) = \bigoplus_{j=0}^{n} \mathbb{Q}_\ell(-j) \otimes \Bigg(\bigoplus_{\beta : \dim(S_\beta) = j} \widetilde{H}^{n-2j-2}(0, \beta) \Bigg),$$

which re-proves a result of Björner and Ekedahl [4]. ◁

Example 3.7 The aforementioned result of Björner and Ekedahl is the algebro-geometric analogue of a theorem of Goresky and MacPherson [17] about real subspace arrangements; the latter result, too, can be given an easy proof using our spectral sequence. Goresky and MacPherson originally proved it as an application of their stratified Morse theory; many different authors have subsequently given alternative proofs and/or strengthenings. Their result, in turn, is a refinement of the work of Orlik and Solomon on complex hyperplane arrangements [25]. In any case, suppose that X is a real vector space, stratified according to intersections in a real subspace arrangement. Let $\sigma(\alpha) = -\dim_{\mathbb{R}} S_\alpha$. The result of Goresky and MacPherson is equivalent to our spectral sequence degenerating at E_1. The weight argument used in the case of a complex subspace arrangement is of course not valid in this setting. We can instead argue as follows:

Choose for each α an open ball U_α inside S_α. Then $C_c^\bullet(U_\alpha, \mathbb{Z})$ (compactly supported cochains) is a subcomplex of $C_c^\bullet(\bar{S}_\alpha, \mathbb{Z})$ for all α. The inclusion of each of these subcomplexes is a quasi-isomorphism, and the restriction maps between these subcomplexes can be chosen to be identically zero (since U_α may be taken to be disjoint from all S_β with $\beta > \alpha$). By additionally choosing an arbitrary quasi-isomorphism between $C_c^\bullet(U_\alpha, \mathbb{Z})$ and $H_c^\bullet(\bar{S}_\alpha, \mathbb{Z})$ we thus get a quasi-isomorphism between the two functors $P \to \mathrm{Ch}_k$ (where the poset P is thought of as a category) given by $\alpha \mapsto S_c^\bullet(\bar{S}_\alpha, \mathbb{Z})$ and $\alpha \mapsto H_c^\bullet(\bar{S}_\alpha, \mathbb{Z})$.

We can compute $R\Gamma_c(X, L^\bullet(\mathbb{Z}))$ by means of a double complex, with each vertical row a direct sum of complexes $S_c^\bullet(\bar{S}_\alpha, \mathbb{Z})$, and the differentials in the horizontal row given by the differentials in the complex $L^\bullet(\mathbb{Z})$; equivalently, given by the functor $\alpha \mapsto S_c^\bullet(\bar{S}_\alpha, \mathbb{Z})$. If we apply the quasi-isomorphism of functors constructed in the previous paragraph we can replace this double complex with one in which the vertical rows have zero differential, and the horizontal rows are direct sums of complexes of the form $\widetilde{C}^{\bullet+2}(0, \beta)$.

Our spectral sequence arises from a filtration of this double complex. In this case, the filtration clearly splits, and the spectral sequence degenerates immediately. ◁

Example 3.8 Let us give two variations of our spectral sequence.

(1°) Let $\mathbb{D}$ denote Verdier's duality functor. The filtration on $L^\bullet(\mathcal{F})$ induces a filtration on $\mathbb{D}L^\bullet(\mathcal{F})$, satisfying $\mathrm{Gr}_F\mathbb{D}L^\bullet(\mathcal{F}) \simeq \mathbb{D}\mathrm{Gr}_F L^\bullet(\mathcal{F})$ (see eg [18, (2.2.8.1)]). Thus the associated graded pieces of $\mathbb{D}L^\bullet(\mathcal{F})$ are quasi-isomorphic to

$$\mathbb{D}\Bigl(\bigoplus_{\sigma(\beta)=p} \widetilde{H}^{\bullet+2}(0,\beta) \otimes^L (i_\beta)_*(i_\beta)^*\mathcal{F}\Bigr) = \bigoplus_{\sigma(\beta)=p} \widetilde{H}_{-\bullet-2}(0,\beta) \otimes^L (i_\beta)_!(i_\beta)^!\mathbb{D}\mathcal{F}.$$

Since the cohomology of $\mathbb{D}R$ in negative degrees equals Borel–Moore homology with coefficients in R, our filtration of $\mathbb{D}L^\bullet(\mathcal{F})$ gives rise to a Borel–Moore homology spectral sequence

$$E^1_{pq} = \bigoplus_{\substack{\beta\in P\\ \sigma(\beta)=p}} \bigoplus_{i+j=p+q-2} H^{\mathrm{BM}}_j(\bar{S}_\beta, \widetilde{H}_i(0,\beta;R)) \implies H^{\mathrm{BM}}_{p+q}(S_0,R).$$

(2°) Instead of taking the compact support cohomology of $L^\bullet(\mathcal{F}) \simeq j_{0!}j_0^{-1}\mathcal{F}$, we may take the usual cohomology. Since $j_{0!}j_0^{-1}\mathcal{F}[1]$ is the cone of $\mathcal{F} \to i_*i^*\mathcal{F}$, where i is the inclusion $(X \setminus S_0) \hookrightarrow X$, this gives instead a spectral sequence

$$E_1^{pq} = \bigoplus_{\substack{\beta\in P\\ \sigma(\beta)=p}} \bigoplus_{i+j=p+q-2} H^j(\bar{S}_\beta, \widetilde{H}^i(0,\beta;R)) \implies H^{p+q}(X, X\setminus S_0; R). \quad \triangleleft$$

Example 3.9 Let $X=\{X(n)\}$ be a topological operad. Suppose that $X(n)$ is stratified in such a way that the strata correspond to trees with n legs, the closed stratum corresponding to a tree T is $\prod_{v\in\mathrm{Vert}(T)} X(n_v)$, and the composition maps in the operad X are given tautologically by grafting of trees. Let $Y(n)$ be the open stratum in $X(n)$ corresponding to the unique tree with a single vertex. Examples of such operads abound: the Stasheff associahedra, the Fulton–MacPherson model of the $\boldsymbol{e}_n$–operads, the Deligne–Mumford spaces $\overline{\mathcal{M}}_{0,n}$, the Boardman–Vogt W–construction applied to an arbitrary topological operad, Devadoss's mosaic operad, the cactus operad, Brown's dihedral moduli spaces $\mathcal{M}^\delta_{0,n}$, the brick operad $\mathcal{B}(n)$ of Dotsenko, Shadrin and Vallette, etc.

Clearly, the compact support cohomology $H^\bullet_c(X(n),\mathbb{Q})$ will form a cooperad. Moreover, the degree-shifted cohomologies $H^{\bullet-1}_c(Y(n),\mathbb{Q})$ will form an operad: $Y(n)\times Y(m)$ will be a stratum adjacent to $Y(n+m-1)$ inside $X(n+m-1)$, and there is a connecting homomorphism $H^\bullet_c(Y(n)\times Y(m),\mathbb{Q}) \to H^{\bullet+1}_c(Y(n+m-1),\mathbb{Q})$ coming from the long exact sequence of a pair in compact support cohomology.

If σ is the function taking a stratum to the number of vertices in the corresponding tree, then we get a filtration of $X(n)$. The corresponding spectral sequence in compact support cohomology (equation (3) from the introduction) is a cooperad in the category of spectral sequences. Its E_∞ page is the associated graded for a filtration on $H_c^\bullet(X(n),\mathbb{Q})$, and its E_1 page is exactly the *bar construction* on the operad $H_c^{\bullet-1}(Y(n),\mathbb{Q})$. This construction seems to have first been considered in [16, Section 3.3], where it was used to prove Koszul self-duality of the $\boldsymbol{e}_n$–operads (equivalently, collapse of the spectral sequence), using the Fulton–MacPherson compactification.

Our Theorem 1.1 then gives a dual spectral sequence. All intervals in the poset of trees are boolean lattices, and the spectral sequence of Theorem 1.1 takes the simple form

$$E_1^{pq} = \bigoplus_{\#\mathrm{Vert}(T)=p} H_c^q\Big(\prod_{v\in\mathrm{Vert}(T)} X(n_v),\mathbb{Q}\Big) \implies H_c^{p+q-1}(Y(n),\mathbb{Q}).$$

This is now an *operad* in the category of spectral sequences, whose E_∞ page is the associated graded for a filtration on $H_c^{\bullet-1}(Y(n),\mathbb{Q})$, and whose E_1 page is exactly the *cobar construction* on the operad $H_c^\bullet(X(n),\mathbb{Q})$.

Thus we see that working with compact support cohomology gives a quite general setting for proving bar/cobar-duality results for such pairs of operads X, Y. $\lhd$

Example 3.10 Suppose that the poset P is Cohen–Macaulay, or more generally that P is graded with rank function ρ and that $\widetilde{H}_i(0,\beta)=0$ for $i<\rho(\beta)-2=\dim \mathrm{N}(0,\beta)$. Then if we apply the spectral sequence for the function ρ, the spectral sequence simplifies to

$$E_1^{pq} = \bigoplus_{\substack{\beta\in P\\ \rho(\beta)=p}} \widetilde{H}^{p-2}(0,\beta)\otimes H_c^q(\overline{S}_\beta, i_\beta^*\mathcal{F}) \implies H_c^{p+q}(S_0, j_0^{-1}\mathcal{F}).$$

In fact, something stronger is true: the chain complex $L^\bullet(\mathcal{F})$ is filtered quasi-isomorphic to a complex of sheaves $K^\bullet(\mathcal{F})$, with

$$K^d(\mathcal{F}) = \bigoplus_{\substack{\beta\in P\\ \rho(\beta)=d}} \widetilde{H}^{d-2}(0,\beta)\otimes (i_\beta)_*(i_\beta)^*\mathcal{F},$$

and which is filtered by the "stupid filtration". The Cohen–Macaulay condition is extremely well studied and is known for large classes of posets. See eg [32, Section 4]. $\lhd$

Example 3.11 Suppose that X is a complex manifold, and that we are given an "arrangement-like" divisor D on X, ie D can locally be defined by a product of linear forms. Then the poset of strata is a geometric lattice and therefore Cohen–Macaulay. The complex of sheaves $K^\bullet(\mathcal{F})$ is the Verdier dual of the one constructed in [22, Section 2]. As part of his construction, he needs to inductively choose a certain free

$\mathbb{Z}$–module E_S for each stratum S — the fact that such a choice is possible is not obvious, and requires the Cohen–Macaulay condition! $\triangleleft$

Example 3.12 Suppose that P is Cohen–Macaulay and the two contravariant functors $P \to \mathrm{Ch}_k$ given by $\alpha \mapsto R\Gamma_c(\overline{S}_\alpha, i_\alpha^* \mathcal{F})$ and $\alpha \mapsto H_c^\bullet(\overline{S}_\alpha, i_\alpha^* \mathcal{F})$ are quasi-isomorphic. Then the spectral sequence degenerates at E_2.

Indeed, we can realize $R\Gamma_c(X, K^\bullet(\mathcal{F}))$ as a double complex, with each column a direct sum of complexes $R\Gamma_c(\overline{S}_\alpha, i_\alpha^* \mathcal{F})$, and the differentials in each row given by the differentials in the complex $K^\bullet(\mathcal{F})$. If we have such a quasi-isomorphism we can therefore replace this double complex with one in which all vertical differentials vanish. Our spectral sequence is the spectral sequence given by filtering this double complex column-wise, since $K^\bullet(\mathcal{F})$ has the stupid filtration. Thus it will indeed be the case that the spectral sequence has nontrivial differential only on E_1.

Suppose that each closed stratum $\overline{S}_\alpha$ is a compact complex manifold on which the dd^c–lemma holds, eg a Kähler or Moishezon manifold, and that the sheaf $\mathcal{F}$ is the constant sheaf $\mathbb{R}$. Then the criterion stated in the first sentence of this example is satisfied. Indeed, we may take as our model for $R\Gamma_c(\overline{S}_\alpha, \mathbb{R}) = R\Gamma(\overline{S}_\alpha, \mathbb{R})$ the real de Rham complex of $\overline{S}_\alpha$, and then the validity of the above criterion is a particular case of [12, Section 6, Main Theorem (ii)]. $\triangleleft$

Example 3.13 Suppose that P is Cohen–Macaulay, and that each closed stratum is an algebraic variety whose compact support cohomology is of pure weight in each degree (eg a smooth projective variety). Then the spectral sequence also degenerates at E_2, using instead a weight argument. $\triangleleft$

Example 3.14 For a space M, let $F(M,n)$ denote the configuration space of n distinct ordered points on M. If M is an oriented manifold, a spectral sequence calculating the cohomology of $F(M,n)$ was constructed by Cohen and Taylor [10]. Their construction was later simplified by Totaro, who noticed that the spectral sequence is just the Leray spectral sequence for the inclusion $j\colon F(M,n) \hookrightarrow M^n$ [28]. Getzler [15] then realized that the spectral sequence exists for a more or less arbitrary topological space, if one works with *compactly supported* cohomology instead: more precisely, Getzler constructed a complex of sheaves quasi-isomorphic to $j_! j^{-1}\mathcal{F}$, whose compactly supported hypercohomology spectral sequence was Poincaré dual to the spectral sequence of Cohen and Taylor in the case of an oriented manifold.

So suppose that $X = M^n$ for some space M, and let us stratify X according to points coinciding. Then the poset of strata is the partition lattice Π_n, which is Cohen–Macaulay. Our complex $K^\bullet(\mathcal{F})$ is exactly the one considered by Getzler, and the resulting spectral sequence

$$E_1^{pq} = \bigoplus_{\substack{\beta\in\Pi_n \\ \rho(\beta)=p}} \widetilde{H}^{p-2}(0,\beta)\otimes H_c^q(M^{n-p},\mathbb{Z}) \implies H_c^{p+q}(F(M,n),\mathbb{Z})$$

is the Poincaré dual of Cohen and Taylor's if M is an oriented manifold. To see the identification of $K^\bullet(\mathcal{F})$ with Getzler's resolution we need to know the cohomology of the partition lattice.

Note first of all that each lower interval $[0,\beta]$ in the partition lattice is itself a product of partition lattices: eg if β corresponds to the partition $(136|27|45)$, then $[0,\beta]\cong \Pi_3\times\Pi_2\times\Pi_2$. Thus by the Künneth theorem we only need to know the top cohomology group $\widetilde{H}^{n-3}(\Pi_n,\mathbb{Z})$. This calculation is hard to attribute correctly — it follows by combining the results of [20] and [27]; see also [19, Section 4]. The result is in any case that $\widetilde{H}^{n-3}(\Pi_n,\mathbb{Z})$ has rank $(n-1)!$ and that as a representation of the symmetric group $\mathbb{S}_n$, it is isomorphic to $\mathsf{Lie}(n)\otimes\mathrm{sgn}_n$, where $\mathsf{Lie}(n)$ is the arity-n component of the Lie operad. But the same is also true for the cohomology group $H^{n-1}(F(\mathbb{C},n),\mathbb{Z})$, by the results of Cohen [9]; specifically, since the homology of the little disk operad is the Gerstenhaber operad, and the Gerstenhaber operad in top degree is just a suspension of the Lie operad, we get the above identification. This explains why the cohomology groups of $F(\mathbb{C},n)$ appear in Getzler's construction of the resolution: the decomposition of $H^\bullet(F(\mathbb{C},n))$ into summands corresponding to different partitions of $\{1,\dots,n\}$ used by Getzler corresponds to

$$H^k(F(\mathbb{C},n),\mathbb{Z})\cong \bigoplus_{\substack{\beta\in\Pi_n \\ \rho(\beta)=k}} \widetilde{H}^{k-2}(0,\beta).$$

If we instead consider the spectral sequence described in Example 3.8(2°), applied to the stratification of $X=M^n$ according to points coinciding, then we recover the spectral sequence of Bendersky and Gitler [3]. ◁

Example 3.15 Consider the example $X=M^n$ of a configuration space. Let $A_c^\bullet(M)$ be a cdga model for the compactly supported cochains on M with rational coefficients. Then the criterion described in Example 3.12 is equivalent to $A_c^\bullet(M)$ being formal, ie that $A_c^\bullet(M)$ and $H_c^\bullet(M)$ are quasi-isomorphic. Hence if $A_c^\bullet(M)$ is formal then the Cohen–Taylor–Totaro spectral sequence degenerates after the first differential.

If in the same situation we consider the second variant of Example 3.8 (ie the Bendersky–Gitler spectral sequence), then we see by the same argument that the spectral sequence degenerates after the first differential whenever M is a formal space, a result which is also proven in Bendersky and Gitler's original paper. ◁

3.2 Compatibility with Hodge theory

We have already mentioned several times that in the algebraic case, we obtain a spectral sequence of ℓ–adic Galois representations. It is natural to ask whether in the complex algebraic setting we get a spectral sequence of mixed Hodge structures.

The answer is yes, and it follows from Saito's theory of mixed Hodge modules [26]. Yet some care must be taken here. Saito proves the existence of a six functors formalism on the level of the derived categories $D^b(\mathsf{MHM}(X))$, where X is a complex algebraic variety. We defined the complex $L^\bullet(\mathcal{F})$ in such a way that $L^d(\mathcal{F})$ is a sum of objects of the form $(i_{\alpha_d})_*(i_{\alpha_d})^*\mathcal{F}$. Now if $\mathcal{F}$ is a mixed Hodge module then $(i_{\alpha_d})_*(i_{\alpha_d})^*\mathcal{F}$ is in general only going to be an object of $D^b(\mathsf{MHM}(X))$, and this is not good enough: a "chain complex" of objects in a triangulated category T can in general have several nonisomorphic totalizations to an object of T, or none at all.

Thus the construction can only be carried out if i_*i^* is a t–exact functor, for i a closed immersion. This is of course true for the usual t–structure of constructible sheaves, but it is *false* for the perverse t–structure on $D^b_c(X)$: i_* is still t–exact, but i^* clearly is not. Since a mixed Hodge module does not have an underlying constructible sheaf but instead an underlying perverse sheaf, i^* is not t–exact for mixed Hodge modules.

However, one can choose instead to give $D^b(\mathsf{MHM}(X))$ a constructible (ie nonperverse) t–structure, uniquely characterized by the functor rat: $D^b(\mathsf{MHM}(X)) \to D^b_c(X)$ being t–exact for the constructible t–structure on $D^b_c(X)$ [26, Remark 4.6]. In other words, an object $\mathcal{F} \in D^b(\mathsf{MHM}(X))$ is in the heart of the constructible t–structure if and only rat($\mathcal{F}$) is quasi-isomorphic to a constructible sheaf. In particular, i^* will be t–exact for this t–structure, and i_* will be t–exact whenever i is a closed immersion.

Let $\mathsf{H}(X)$ be the heart of the constructible t–structure of $D^b(\mathsf{MHM}(X))$, and let $\mathcal{F}$ be an object of $\mathsf{H}(X)$. Then for $d \geq 0$ we obtain an object

$$L^d(\mathcal{F}) = \bigoplus_{0=\alpha_0<\alpha_1<\cdots<\alpha_d \in P} (i_{\alpha_d})_*(i_{\alpha_d})^*\mathcal{F}$$

of $\mathsf{H}(X)$, and we can define a differential $L^d(\mathcal{F}) \to L^{d+1}(\mathcal{F})$ just as before. Thus we get an object $L^\bullet(\mathcal{F})$ of $D^b(\mathsf{H}(X))$, quasi-isomorphic to $j_{0!}j_0^{-1}\mathcal{F} \in \mathsf{H}(X)$. Moreover, we obtain a filtration of $L^\bullet(\mathcal{F})$ by the same procedure as before. This filtration allows us to write down a *Postnikov system* in the triangulated category $D^b(\mathsf{H}(X))$, with totalization $L^\bullet(\mathcal{F})$ [14, page 262]:

$$\begin{tikzcd}[column sep=small]
 & \mathrm{Gr}_F^{p-1} L^\bullet(\mathcal{F})[-1] \arrow[dl] & & \mathrm{Gr}_F^{p-2} L^\bullet(\mathcal{F})[-1] \arrow[dl] & & \mathrm{Gr}_F^{p-3} L^\bullet(\mathcal{F})[-1] \arrow[dl] & \\
\cdots\ F^p L^\bullet(\mathcal{F}) \arrow[rr] & & F^{p-1} L^\bullet(\mathcal{F}) \arrow[rr] \arrow[ul] & & F^{p-2} L^\bullet(\mathcal{F}) \arrow[rr] \arrow[ul] & & F^{p-3} L^\bullet(\mathcal{F}) \cdots \arrow[ul]
\end{tikzcd}$$

Let T be a triangulated category with a t–structure, and $\mathsf{T}^{\heartsuit}$ its heart. A *realization functor* is an exact functor $D^b(\mathsf{T}^{\heartsuit}) \to \mathsf{T}$ whose restriction to the full subcategory $\mathsf{T}^{\heartsuit}$ is the inclusion into T. If T is itself the derived category of an abelian category (not necessarily with its standard t–structure), then a realization functor always exists [2, Section 3]; more generally, a realization functor always exists if T is the homotopy category of a stable ∞–category. In particular, we obtain a realization functor

$$\mathsf{real}\colon D^b(\mathsf{H}(X)) \to D^b(\mathsf{MHM}(X)).$$

Exact functors preserve Postnikov systems, and we get a Postnikov system in $D^b(\mathsf{MHM}(X))$ whose terms are of the form $\mathsf{real}(\mathrm{Gr}_F^p L^\bullet(\mathcal{F}))$, up to a degree shift. Now we note that the functor real commutes with tensoring with a bounded complex of free abelian groups. Hence, the terms in the Postnikov system in $D^b(\mathsf{MHM}(X))$ are given by

$$\begin{aligned}
\mathsf{real}(\mathrm{Gr}_F^p L^\bullet(\mathcal{F})) &\cong \mathsf{real} \bigoplus_{\substack{\beta\in P\\ \sigma(\beta)=p}} \widetilde{C}^{\bullet+2}(0,\beta)\otimes (i_\beta)_*(i_\beta)^*\mathcal{F} \\
&\cong \bigoplus_{\substack{\beta\in P\\ \sigma(\beta)=p}} \widetilde{C}^{\bullet+2}(0,\beta)\otimes \mathsf{real}(i_\beta)_*(i_\beta)^*\mathcal{F} \\
&\cong \bigoplus_{\substack{\beta\in P\\ \sigma(\beta)=p}} \widetilde{C}^{\bullet+2}(0,\beta)\otimes (i_\beta)_*(i_\beta)^*\mathcal{F},
\end{aligned}$$

using in the last step that $(i_\beta)_*(i_\beta)^*\mathcal{F}$ is in $\mathsf{H}(X)$. Applying $Rf_!$ to this Postnikov system, where $f\colon X \to \mathrm{Spec}(\mathbb{C})$ is the projection to a point, gives a Postnikov system in the derived category of mixed Hodge structures (the category of mixed Hodge modules over a point). The associated spectral sequence [14, page 263] is the one of Theorem 3.3(ii), now equipped with the canonical mixed Hodge structure coming from the fact that $\mathcal{F}$ is a mixed Hodge module.

Remark 3.16 It seems likely that $D^b(\mathsf{H}(X)) \to D^b(\mathsf{MHM}(X))$ is an equivalence of categories, which would be the analogue for mixed Hodge modules of Beĭlinson's theorem that the realization functor from the derived category of perverse sheaves to the derived category of constructible sheaves is an equivalence [1], but I do not know if this is known and I have not attempted to prove it.

4 Representation stability

The notion of *representation stability* was introduced by Church and Farb [8] as an extension of homological stability to situations where the Betti numbers do not actually stabilize. Roughly, a sequence $\{V(n)\}$ of representations of $\mathbb{S}_n$ over $\mathbb{Q}$ is said to be *representation stable* if, for $n \gg 0$, the representation $V(n+1)$ is obtained from the representation $V(n)$ by adding a single box to the top row of the Young diagram of each irreducible representation occurring in $V(n)$. Thus $V(n+1)$ is completely determined from $V(n)$ for sufficiently large n. Note in particular that the $\mathbb{S}_n$–invariants satisfy $V(n+1)^{\mathbb{S}_{n+1}} \cong V(n)^{\mathbb{S}_n}$ for $n \gg 0$; if $\{V(n)\}$ were a sequence of homology groups, the $\mathbb{S}_n$–invariants would satisfy homological stability in the usual sense.

The theory was clarified by the introduction of *FI-modules* [6]. The key point is that the underlying sequence of $\mathbb{S}_n$–representations of an FI-module in $\mathbb{Q}$–vector spaces is representation stable if and only if the FI-module is *finitely generated.* Most examples of representation stability arise from an FI-module in this way.

One of the main examples of representation stability is given by the following theorem of Church [5]: if M is an oriented manifold, then $H^i(F(M,n),\mathbb{Q})$ is a representation stable sequence of $\mathbb{S}_n$–representations for any i. In this example, it was known since [23] that the cohomology of the unordered configuration space $F(M,n)/\mathbb{S}_n$ satisfies homological stability for *integer* coefficients if M is an *open* manifold, but also that integral homological stability is false in general. Church's result shows in particular that, with $\mathbb{Q}$–coefficients, the unordered configuration space always satisfies homological stability.

This result of Church fits well into the general framework of FI-modules. The assignment $S \mapsto F(M,S)$ is a contravariant functor from FI to spaces: if $S \subset T$, then $F(M,T) \to F(M,S)$ is the map that forgets all the points indexed by elements of $T \setminus S$. On applying $H^i(-,\mathbb{Q})$ one gets an FI-module, which turns out to be finitely generated; in fact, finite generation holds already with integral coefficients [7].

For the remainder of this paper, we will prove a theorem extending Church's result in several ways:

(1) Our proof works in a uniform way for a much larger class of configuration-like spaces, such as "k–equals" configuration spaces, the spaces $\overline{w}^c_\lambda(M)$ considered by Vakil and Wood, etc.

(2) We give a proof valid also in the algebrogeometric setting, so we get eg representation stability in the category of ℓ–adic Galois representations. (This was previously proven for the spaces $F(M,n)$ in [13] under more restrictive assumptions on M.)

(3) We allow M to have singularities. In the paper we focus on the case when M is an algebraic variety (with arbitrary singularities); we comment towards the end on the differences in the topological setting.

In order to have representation stability in the more general setting of a singular space, one needs to work with Borel–Moore homology/compactly supported cohomology. Note that compactly supported cohomology is only contravariant for *proper* maps, and the map $F(M,T) \to F(M,S)$ is (almost) never proper, so $H_c^\bullet(F(M,S),\mathbb{Q})$ is not directly an FI-module. This should in any case not be surprising: if we want to recover Church's theorem by Poincaré duality when M is an oriented manifold, then we had better prove that the cohomology $H_c^\bullet(F(M,S),\mathbb{Q})$ satisfies representation stability up to a degree shift by the dimension of $F(M,S)$.

4.1 Twisted commutative algebras and FI-modules

Our proof of representation stability uses the formalism of FI-modules and twisted commutative algebras. We briefly recall the definitions for the reader's convenience. Let $\mathcal{C}$ be a symmetric monoidal category (the reader is encouraged to take $\mathcal{C}$ to be the category of dg vector spaces over a field of characteristic zero). By a *species in* $\mathcal{C}$ we mean a functor $\boldsymbol{B} \to \mathcal{C}$, where $\boldsymbol{B}$ the category of finite sets and bijections. The category of species is equivalent to the category of sequences of representations of the symmetric groups $\mathbb{S}_n$ in $\mathcal{C}$. We write a species as $S \mapsto A(S)$ or $n \mapsto A(n)$, depending on whether we wish to consider it as a functor of finite sets or as a sequence of representations. We call $A(n)$ the *arity-n* component of the species A.

Let us consider $\boldsymbol{B}$ as a symmetric monoidal category, with monoidal structure given by disjoint union. A *twisted commutative algebra* (tca) in $\mathcal{C}$ is a lax symmetric monoidal functor $\boldsymbol{B} \to \mathcal{C}$. Thus a twisted commutative algebra in $\mathcal{C}$ consists of a sequence $\{A(n)\}$ of $\mathbb{S}_n$–representations in $\mathcal{C}$, equipped with multiplication maps $A(n)\otimes A(m) \to A(n+m)$ which are $\mathbb{S}_n \times \mathbb{S}_m$–equivariant and satisfy suitable commutativity and associativity axioms. An equivalent definition is that a tca is a left module over the commutative operad Com in $\mathcal{C}$. A third equivalent definition is that a tca is an algebra over the operad Com in the category of species in $\mathcal{C}$, where the tensor product on the category of species is given by Day convolution:

$$(A \otimes B)(S) = \bigoplus_{S = T_1 \sqcup T_2} A(T_1) \otimes B(T_2).$$

Suppose that $\mathcal{C}$ is the category of dg R–modules. Let A be a species in $\mathcal{C}$. We define the *suspension* $\mathsf{S}A$ by

$$\mathsf{S}A(n) = A(n)[-n] \otimes \mathrm{sgn}_n.$$

The suspension is a symmetric monoidal endofunctor on the symmetric monoidal category of species. In particular, if A is a tca, then so is $\mathsf{S}A$.

Let FI denote the category of finite sets and injections. By an *FI-module* in the category $\mathcal{C}$ we mean a functor $\mathrm{FI} \to \mathcal{C}$.

Let Com be the twisted commutative algebra which has the monoidal unit with trivial $\mathbb{S}_n$–action in each arity, and for which all multiplication maps $\mathsf{Com}(S) \otimes \mathsf{Com}(T) \to \mathsf{Com}(S \sqcup T)$ are given by the canonical isomorphism $\mathbf{1} \otimes \mathbf{1} \cong \mathbf{1}$. In other words, we are considering the commutative operad as a left module over itself. There is a general notion of a module over an algebra over any operad, which in this case specializes to an evident notion of a module over a tca.

Lemma 4.1 *Every module over the tca* Com *is in a canonical way an FI-module, and vice versa.*

Sketch of proof Let M be a module over the tca Com, and let $S \subset T$. Then we have a map

$$M(S) = M(S) \otimes \mathbf{1} = M(S) \otimes \mathsf{Com}(T \setminus S) \to M(S \sqcup (T \setminus S)) = M(T),$$

where $\mathbf{1}$ denotes the monoidal unit in $\mathcal{C}$ and $=$ denotes a canonical isomorphism. This makes M into an FI-module. The converse construction is similar. □

If A is a tca in $\mathcal{C}$, then the choice of a morphism $a\colon \mathbf{1} \to A(1)$ is the same as the choice of a tca morphism $\mathsf{Com} \to A$. Thus the choice of such an a defines the structure of FI-module on the underlying species of the tca A.

In particular, let $\mathcal{C}$ be the category of graded R–modules. For a tca A in $\mathcal{C}$, let us write $A^i(n)$ for the degree-i component of $A(n)$. Then any $a \in A^0(1)$ defines a structure of FI-module on the collection $A^i(n)$, for all $i \in \mathbb{Z}$. We write $|a|$ for the degree of a homogeneous element in a graded vector space.

Lemma 4.2 *Suppose A is a tca in graded R–modules, and that $\{a_0, a_1, a_2, \dots\}$ is a set of generators. Suppose that $a_0 \in A^0(1)$, $|a_i| < 0$ for $i > 0$, and $\lim_{i\to\infty} |a_i| = -\infty$. Then the FI-module $n \mapsto A^i(n)$ defined by multiplication with a_0 is finitely generated for all $i \in \mathbb{Z}$.*

Proof The hypotheses imply that there are only finitely many monomials in $\{a_1, a_2, \dots\}$ (ie all generators except a_0) of given degree. Those monomials of degree i generate the FI-module $n \mapsto A^i(n)$. □

Applying the previous lemma to the d–fold suspension $\mathsf{S}^d A$, we get the following:

Lemma 4.3 *Suppose A is a tca in graded R–modules, and that $\{a_0, a_1, a_2, \dots\}$ is a set of generators. Suppose that if $a_k \in A^i(n)$ then $i \leq nd$, with equality only for a single element $a_0 \in A^d(1)$, and that for each $p \in \mathbb{Z}$, there are only finitely many k for which $a_k \in A^i(n)$ and $i \geq nd - p$. If d is even, then $n \mapsto A^{i+dn}(n)$ becomes an FI-module by multiplication with a_0, and if d is odd, then $n \mapsto A^{i+dn}(n) \otimes \mathrm{sgn}_n$ becomes an FI-module in this way. This FI-module is finitely generated for all $i \in \mathbb{Z}$.*

4.2 The "$\mathcal{A}$–avoiding" configuration spaces

For each finite set S, let M^S be the cartesian product of $|S|$ copies of some space M. The functor $S \mapsto M^S$ is a co-FI-space. If $S \hookrightarrow T$, then we denote by $\pi^T_S \colon M^T \to M^S$ the projection.

Let $\mathcal{A}$ be a finite collection of closed subspaces $\{A_i \subset M^{S_i}\}_{i=1}^{\ell}$, where each S_i is some finite set. For every finite set T, consider the stratification of M^T given by all subspaces

$$(\pi^T_{S_i})^{-1}(A_i) \subset M^T, \quad i = 1, \dots, \ell,$$

ranging over all inclusions $S_i \hookrightarrow T$, and all intersections of those subspaces. Let $P_{\mathcal{A}}(T)$ be the poset of strata in this stratification, and let $F_{\mathcal{A}}(M, T)$ denote the open stratum which is the complement of all of the $(\pi^T_{S_i})^{-1}(A_i)$.

Example 4.4 If $\mathcal{A}$ is a singleton with $A_1 = \Delta \subset M^2$, then $F_{\mathcal{A}}(M, n)$ is the classical configuration space of n points on M. If $\mathcal{A}$ instead consists only of the small diagonal in M^k, then $F_{\mathcal{A}}(M, n)$ is the "k–equals" configuration space of points on M. ◁

Example 4.5 If a finite group G acts on M, then we can let $\mathcal{A}$ consist of all subspaces $\{(x, g\cdot x) : x \in M\}$ inside M^2, in which case $F_{\mathcal{A}}(M, n)$ parametrizes n distinct ordered points all of which are in distinct G–orbits. An example of this is the complement of hyperplanes in the Coxeter arrangement associated to the wreath product $(\boldsymbol{\mu}_r)^n \rtimes \mathbb{S}_n$ acting on $\mathbb{C}^n$. ◁

Example 4.6 Suppose that $M = Y^2$ for some other space Y, and that $\mathcal{A}$ consists of the collection $\{\Delta_{13}, \Delta_{14}, \Delta_{23}, \Delta_{24}\}$ of diagonals inside $M^2 = Y^4$. Then $F_{\mathcal{A}}(M, n)$ parametrizes points $x_1, \dots, x_n$ and $y_1, \dots, y_n$ in Y such that the x_i may collide amongst each other, and so may the y_i, but $x_i \neq y_j$ for all i, j. ◁

Example 4.7 Let λ be a partition of n. As our notation for partitions we use both $\lambda = (\lambda_1, \lambda_2, \lambda_3, \dots)$ and $\lambda = (1^{n_1}\, 2^{n_2}\, 3^{n_3} \dots)$, so that $n = \sum_{j \geq 1} \lambda_j = \sum_{i \geq 1} i \cdot n_i$. Vakil and Wood [29] defined an open subspace $\overline{w}^c_\lambda(M) \subset M^n/\mathbb{S}_n$, and they studied the behavior of these spaces under the operation of "padding λ with ones", ie letting n_1 approach ∞ while keeping all n_i, $i \geq 2$, fixed.

We can understand their construction in our terms as follows: if λ is a partition of n with $n_1 = 0$, let $\Delta_\lambda \subset M^n$ be the locus where the first λ_1 points coincide, the subsequent λ_2 points coincide, etc., and put $\mathcal{A} = \{\Delta_\lambda\}$. If λ' is the partition of $N \geq n$ obtained by padding λ with ones, then $\overline{w}^c_{\lambda'}(M) = F_{\mathcal{A}}(M, N)/\mathbb{S}_N$. In particular, rational homological stability for the spaces $\overline{w}^c_\lambda(M)$ under the operation of padding λ with ones follows from representation stability for the spaces $F_{\mathcal{A}}(M, n)$ as $n \to \infty$. $\triangleleft$

Example 4.8 The configuration space of n points in $\mathbb{P}^2$ such that no three of them lie on a line and no six lie on a conic is of the form $F_{\mathcal{A}}(M, n)$, where $M = \mathbb{P}^2$ and $\mathcal{A}$ has two elements which are closed subvarieties of M^3 and M^6, respectively. $\triangleleft$

We are going to prove a homological stability result for the spaces $F_{\mathcal{A}}(M, n)$. To avoid dealing with trivial cases we will assume that $|S_i| \geq 2$ for all i, and that no subspace A_i can be written as

$$A_i = (\pi^{S_i}_{S'_i})^{-1}(A'_i)$$

where S'_i is a proper subset of S.

4.3 The setup

Observe that there is an open embedding

$$F_{\mathcal{A}}(M, S \sqcup T) \hookrightarrow F_{\mathcal{A}}(M, S) \times F_{\mathcal{A}}(M, T).$$

This makes $S \mapsto F_{\mathcal{A}}(M, S)$ a twisted cocommutative coalgebra of spaces. Since Borel–Moore homology is contravariant for open embeddings and admits cross products

$$H^{\mathrm{BM}}_\bullet(X, R) \otimes H^{\mathrm{BM}}_\bullet(Y, R) \to H^{\mathrm{BM}}_\bullet(X \times Y, R)$$

(which are isomorphisms if R is a field), we get a twisted commutative algebra in graded R–modules:

$$S \mapsto H^{\mathrm{BM}}_\bullet(F_{\mathcal{A}}(M, S), R),$$

for any choice of coefficients R.

The functor $S \mapsto P_{\mathcal{A}}(S)$ forms a twisted commutative algebra in the category of posets: the product of a stratum in M^S and a stratum in M^T is a stratum in $M^{S \sqcup T}$, which identifies $P_{\mathcal{A}}(S) \times P_{\mathcal{A}}(T)$ with an order ideal in $P_{\mathcal{A}}(S \sqcup T)$. If $\beta \in P_{\mathcal{A}}(S)$ and $\gamma \in P_{\mathcal{A}}(T)$, then we write $\beta \times \gamma$ for their product in $P_{\mathcal{A}}(S \sqcup T)$.

Let $L^\bullet(S)$ denote the complex of sheaves $L^\bullet(R)$ on M^S constructed in the previous part, associated to the stratification $P_{\mathcal{A}}(S)$. The previous paragraph implies that $L^\bullet(S) \boxtimes L^\bullet(T)$ is a subcomplex of $L^\bullet(S \sqcup T)$, and $\mathbb{D}L^\bullet(S) \boxtimes \mathbb{D}L^\bullet(T)$ is a quotient of $\mathbb{D}L^\bullet(S \sqcup T)$. Applying $R\Gamma(-)$, we see that the functor

$$S \mapsto R\Gamma^{-\bullet}(M^S, \mathbb{D}L^\bullet(S))$$

is a twisted commutative algebra of chain complexes, whose homology is the tca $S \mapsto H_\bullet^{\mathrm{BM}}(F_{\mathcal{A}}(M,S),R)$.

4.4 The hypotheses

Let us now describe the hypotheses on M and $\mathcal{A}$ that will lead to a proof of representation stability. Fix M and $\mathcal{A}$ as above, and a coefficient ring R.

Hypothesis 4.9 *We assume that $H_d^{\mathrm{BM}}(M,R) \cong R$, and that homology vanishes above this degree. We assume that (possibly after refining the stratifications) all strata in all spaces M^n have finitely generated homology groups, and there exists an increasing function $\sigma\colon P_{\mathcal{A}}(n) \to \mathbb{Z}$ for all n such that:*

(1) *If $\beta \in P_{\mathcal{A}}(S)$ satisfies $\sigma(\beta) = p$ and $\gamma \in P_{\mathcal{A}}(T)$ satisfies $\sigma(\gamma) = q$, then $\sigma(\beta \times \gamma) = p + q$.*

(2) *If $\beta \in P_{\mathcal{A}}(n)$ satisfies $\sigma(\beta) = p$, then $H_i^{\mathrm{BM}}(\overline{S}_\beta, R)$ vanishes above degrees $dn - 2p$, and $H_{dn-2p}^{\mathrm{BM}}(\overline{S}_\beta, R)$ is a projective R–module.*

Example 4.10 Suppose that M is a geometrically irreducible algebraic variety of dimension $\frac{d}{2}$, and that $\mathcal{A}$ consists of closed subvarieties. Then it will indeed be the case that $H_d^{\mathrm{BM}}(M,\mathbb{Z}) \cong \mathbb{Z}$, that $H_i^{\mathrm{BM}}(M,\mathbb{Z}) = 0$ for $i > d$, and that all strata have finitely generated homology. Let $\sigma(\beta)$ be the codimension of S_β. After refining the stratifications we may assume all strata irreducible, in which case σ becomes a strictly increasing function, and conditions (1) and (2) are clearly satisfied. $\triangleleft$

Example 4.11 Suppose that all the subspaces in $\mathcal{A}$ are given by diagonals, so all closed strata are products of the same space M. This covers eg all the configuration spaces considered by Vakil and Wood. In this case, for $\overline{S}_\beta \cong M^k \subset M^n$, we can take $\sigma(\beta) = (n-k)$. If we suppose that M has finitely generated homology and finite dimension $d > 1$, and $H_d^{\mathrm{BM}}(M,R) \cong R$, then Hypothesis 4.9 is satisfied. To verify the second condition, note that if $\beta \in P_{\mathcal{A}}(n)$ satisfies $\sigma(\beta) = p$, then $\overline{S}_\beta \cong M^{n-p}$, whose highest nonzero Borel–Moore homology group is $H_{d(n-p)}^{\mathrm{BM}}(M^{n-p},R) \cong R$. Since we assumed $d > 1$, we get in particular vanishing above degree $dn - 2p$ and that the homology group in degree $dn - 2p$ is projective. $\triangleleft$

From now on we shall assume that Hypothesis 4.9 is satisfied.

4.5 Proof with coefficients in a field

In this subsection, we fix a field k of coefficients, and all homology groups will be taken with coefficients in k. In the algebraic case we take $k = \mathbb{Q}_\ell$, where ℓ is not equal to the characteristic. We will later see that the proof can be modified to work also for integral coefficients, but the added complications arising from the lack of a Künneth isomorphism obscure the ideas somewhat.

Lemma 4.12 *There exists a twisted commutative algebra of spectral sequences which satisfies*

$$E^1_{pq}(S) = \bigoplus_{\substack{\beta \in P_{\mathcal{A}}(S) \\ \sigma(\beta)=p}} \bigoplus_{i+j=p+q-2} \widetilde{H}_i(0,\beta) \otimes H_j^{\mathrm{BM}}(\bar{S}_\beta, k),$$

and which converges to the twisted commutative algebra $S \mapsto H_\bullet^{\mathrm{BM}}(F_{\mathcal{A}}(M,S),k)$.

Proof By condition (2) in Hypothesis 4.9, the filtration on $L^\bullet(S) \boxtimes L^\bullet(T)$ induced by σ agrees with the one on $L^\bullet(S \sqcup T)$, when we consider $L^\bullet(S) \boxtimes L^\bullet(T)$ as a subcomplex of $L^\bullet(S \sqcup T)$. This makes the twisted commutative algebra $S \mapsto R\Gamma^{-\bullet}(M^S, \mathbb{D}L^\bullet(S))$ a tca in filtered chain complexes, and the associated spectral sequence is given as above. □

We say that an element $\beta \in P_{\mathcal{A}}(T)$ is *indecomposable* if, whenever $T = S \sqcup S'$ and β is in the image of the multiplication map $P_{\mathcal{A}}(S) \times P_{\mathcal{A}}(S') \to P_{\mathcal{A}}(T)$, then S or S' is empty.

Lemma 4.13 *There exists a constant C such that if $\beta \in P_{\mathcal{A}}(T)$ is indecomposable, then $\sigma(\beta) \geq C \cdot |T|$.*

Proof The stratum S_β is (an open stratum inside) an intersection of subspaces of the form $(\pi^T_{S_i})^{-1}(A_i)$, for some collections of inclusions $S_i \hookrightarrow T$. We may assume this collection of subspaces to be irredundant. In order that β be indecomposable, it is certainly necessary that the images of the $S_i \hookrightarrow T$ cover T, which means that the number of subspaces that one needs to intersect to obtain S_β grows linearly in $|T|$. Moreover, since we assumed the collection irredundant and σ increasing, the value of σ must go up by at least 1 for each subspace we intersect, which proves the result. □

Let E be the twisted commutative algebra in graded vector spaces given by

$$S \mapsto E_\bullet(S); \quad E_i(S) = \bigoplus_{p+q=i} E^1_{pq}(S).$$

Lemma 4.14 *The tca E satisfies the hypotheses of Lemma 4.3, so that $\mathsf{S}^d E$ is a finitely generated FI-module.*

Proof The tca E is generated by classes $\widetilde{H}_i(0,\beta) \otimes H_j^{\mathrm{BM}}(\bar{S}_\beta, k)$ where β ranges over indecomposable elements of $P_{\mathcal{A}}(n)$.

If $\beta \in P_{\mathcal{A}}(n)$ satisfies $\sigma(\beta) = p$, then $\widetilde{H}_i(0,\beta)$ vanishes above degree $p-2$, and $H_j^{\mathrm{BM}}(\bar{S}_\beta, k)$ vanishes above degree $dn-2p$. Thus the corresponding generators in $E_i(n)$ satisfy $i \le dn - p$. In particular, we get a generator in degree $i = dn$ only for $p = 0$. But the open stratum is indecomposable only when $n = 1$, in which case we get a single generator in this degree from the one-dimensional space $H_d^{\mathrm{BM}}(M,k)$.

Moreover, by Lemma 4.13, for each $p \ge 0$ there are only finitely many strata in $P_{\mathcal{A}}(n)$ (summed over all n) with $\sigma(\beta) \le p$, which means that only finitely many of these generators satisfy $i \ge dn - p$. Thus the hypotheses of Lemma 4.3 are satisfied. $\square$

Theorem 4.15 *The FI-module given by $n \mapsto H_{i+dn}^{\mathrm{BM}}(F_{\mathcal{A}}(M,n),k) \otimes \mathrm{sgn}_n^{\otimes d}$ is finitely generated for all $i \in \mathbb{Z}$.*

Proof By the previous lemma,

$$n \mapsto (\mathsf{S}^d E)(n) = \bigoplus_{p+q=i+dn} E_{pq}^1(n) \otimes \mathrm{sgn}_n^{\otimes d}$$

is a finitely generated FI-module. Then so is $n \mapsto \bigoplus_{p+q=i+dn} E_{pq}^\infty(n) \otimes \mathrm{sgn}_n^{\otimes d}$, being a subquotient of a finitely generated FI-module [6, Theorem 1.3]. But the latter is just the associated graded of the FI-module $n \mapsto H_{i+dn}^{\mathrm{BM}}(F_{\mathcal{A}}(M,n),k) \otimes \mathrm{sgn}_n^{\otimes d}$ for the Leray filtration. $\square$

Remark 4.16 The sign representation which appears for odd d arises from Lemma 4.3, and does not play a role in the algebraic case since an algebraic variety has even (real) dimension. That the sign representation should appear is clear, if we want to recover representation stability for the usual cohomology $H^i(F(M,n),k)$ from Poincaré duality when M is an oriented manifold. Indeed, the Poincaré duality isomorphism for $F(M,n)$ involves capping with the fundamental class, which generates the 1–dimensional vector space $H_{dn}^{\mathrm{BM}}(F(M,n),k) \cong H_{dn}^{\mathrm{BM}}(M^n,k)$. When d is even this vector space carries the trivial representation of $\mathbb{S}_n$, but when d is odd it has the sign representation.

Remark 4.17 A space M is called an R–homology manifold of dimension d if for all $x \in M$ one has

$$H_i(M, M \setminus \{x\}; R) \cong \begin{cases} R & \text{if } i = d, \\ 0 & \text{if } i \neq d. \end{cases}$$

Equivalently, $\mathbb{D}R \cong R[d]$. A trivial example is an oriented d–manifold for any R; a more typical example is that a complex algebraic variety with finite quotient singularities is a $\mathbb{Q}$–homology manifold (of dimension twice its dimension over $\mathbb{C}$). An R–homology manifold of dimension d satisfies Poincaré duality in the form $H_i^{\mathrm{BM}}(M,R) \cong H^{d-i}(M,R)$. If we add to Hypothesis 4.9 the condition that M is an R–homology manifold of dimension d, then the conclusion of Theorem 4.15 (if R is a field, or the results of the next subsection if R is a PID) become equivalent to the claim that $n \mapsto H^i(F_{\mathcal{A}}(M,n),R)$ is a finitely generated FI-module. For instance, if M is a connected oriented manifold of dimension $d>1$, then $H^i(F(M,n),\mathbb{Z})$ is a finitely generated FI-module for all i; this is how the results of [5; 7] can be obtained as specializations of those in this paper.

4.6 Proof for integral coefficients

In the proof of Theorem 4.15 in the preceding subsection, we used that any subquotient of a finitely generated FI-module is finitely generated. This was proven for field coefficients in [6], but the result was then extended to any noetherian ring in [7]. We will now use the latter result to give a proof also for integral coefficients.

However, the real reason we used field coefficients in the preceding subsection was to have a robust Künneth isomorphism: without it, it would not be true that generators for the twisted commutative algebra we considered arise from indecomposable strata. Namely, if β is decomposable — say $\bar{S}_\beta = \bar{S}_\gamma \times \bar{S}_{\gamma'}$ — then the cross product map

$$H_\bullet^{\mathrm{BM}}(\bar{S}_\gamma,\mathbb{Z}) \otimes H_\bullet^{\mathrm{BM}}(\bar{S}_{\gamma'},\mathbb{Z}) \to H_\bullet^{\mathrm{BM}}(\bar{S}_\beta,\mathbb{Z})$$

is not necessarily surjective. To remedy this, we will need to work on the chain level, analogous to [7, Lemma 4.1].

For the remainder of this section we fix a coefficient ring R which we assume to be a PID, eg $R=\mathbb{Z}$ or $R=\mathbb{Z}_\ell$. In order for the proof to work, we shall need to verify a refinement of Hypothesis 4.9.

Lemma 4.18 *Assume Hypothesis 4.9. For each stratum $\bar{S}_\beta$ in any of the spaces M^n, we can choose a quasi-isomorphism*

$$C_\bullet(\bar{S}_\beta) \simeq R\Gamma^{-\bullet}(\bar{S}_\beta,\mathbb{D}R)$$

where $C_\bullet(\bar{S}_\beta)$ is a bounded complex of finitely generated free modules, and such that

- *for any decomposable stratum $\bar{S}_\alpha \times \bar{S}_\beta$, we have an equality*

$$C_\bullet(\bar{S}_\alpha) \otimes C_\bullet(\bar{S}_\beta) = C_\bullet(\bar{S}_\alpha \times \bar{S}_\beta)$$

compatible with the quasi-isomorphism

$$R\Gamma^{-\bullet}(\bar{S}_\alpha, \mathbb{D}R) \otimes R\Gamma^{-\bullet}(\bar{S}_\beta, \mathbb{D}R) \simeq R\Gamma^{-\bullet}(\bar{S}_\beta \times \bar{S}_\alpha, \mathbb{D}R);$$

- *for $\bar{S}_\alpha \subset \bar{S}_\beta$, there is a map*

$$C_\bullet(\bar{S}_\alpha) \to C_\bullet(\bar{S}_\beta)$$

compatible with the map

$$R\Gamma^{-\bullet}(\bar{S}_\alpha, \mathbb{D}R) \to R\Gamma^{-\bullet}(\bar{S}_\beta, \mathbb{D}R);$$

- *if $\bar{S}_\beta \subset M^n$ has $\sigma(\beta) = p$ then $C_\bullet^{\mathrm{BM}}(\bar{S}_\beta, R)$ vanishes above degree $dn - 2p$, and $C_d(M) \cong R$.*

It is immediate from Hypothesis 4.9 that such a complex $C_\bullet(\bar{S}_\beta)$ can be constructed for each individual stratum, but we need the choices to satisfy various compatibilities.

Taking the lemma for granted for the moment, the idea will be to run nearly the same proof, but instead of starting at the E^1 page of the spectral sequence, we start at E^0. Equivalently, we work directly on the level of the double complex computing $R\Gamma^{-\bullet}(M^n, \mathbb{D}L^\bullet(R))$, associated to our filtration of $\mathbb{D}L^\bullet(R)$. The filtration on $\mathbb{D}L^\bullet(R)$ has its associated graded pieces quasi-isomorphic to sums of complexes of the form $\widetilde{C}_{-\bullet-2}(0,\beta) \otimes (i_\beta)_*\mathbb{D}R$, where i_β is the inclusion of a closed stratum, so the columns of this double complex are of the form $R\Gamma^{-\bullet}(M^n, \widetilde{C}_{-\bullet-2}(0,\beta) \otimes (i_\beta)_*\mathbb{D}R)$. Under our Lemma 4.18 we may (for all β) replace this with the totalization of the double complex $\widetilde{C}_{\bullet+2}(0,\beta) \otimes C_\bullet(\bar{S}_\beta)$. Then this collection of double complexes becomes a twisted commutative algebra, and that we get a tca of spectral sequences:

Lemma 4.19 *There exists a twisted commutative algebra of spectral sequences which satisfies*

$$E^0_{pq}(S) = \bigoplus_{\substack{\beta \in P_{\mathcal{A}}(S) \\ \sigma(\beta)=p}} \ \bigoplus_{i+j=p+q-2} \widetilde{C}_i(0,\beta) \otimes C_j(\bar{S}_\beta),$$

and which converges to the twisted commutative algebra $S \mapsto H_\bullet^{\mathrm{BM}}(F_{\mathcal{A}}(M,S), R)$.

If we consider the twisted commutative algebra in graded abelian groups given by

$$S \mapsto E_\bullet(S); \quad E_i(S) = \bigoplus_{p+q=i} E^0_{pq}(S),$$

then this tca *will* be generated by classes $\widetilde{C}_i(0,\beta) \otimes C_j^{\mathrm{BM}}(\bar{S}_\beta, R)$ where β ranges over indecomposable elements of $P_{\mathcal{A}}(n)$. In particular, it satisfies the hypotheses of

Lemma 4.3, for the same reason as the tca E considered in the previous subsection, and the rest of the proof carries over without any changes.

Let us now argue that Lemma 4.18 is satisfied.

Proof of Lemma 4.18 First off, we replace each $R\Gamma^{-\bullet}(\bar{S}_\beta, \mathbb{D}R)$ by a functorial free resolution, and then apply the "wise" truncation functor $\tau^{\leq(dn-2p)}$. Let us call the resulting complexes $C_\bullet^{\mathrm{BM}}(\bar{S}_\beta, R)$. Unfortunately, the truncation functor has the wrong functoriality: there is for any chain complex $C_\bullet$ a map $C_\bullet \to \tau^{\leq n} C_\bullet$, but we need a map in the opposite direction. However, the assumption that $H^{\mathrm{BM}}_{dn-2p}(\bar{S}_\beta, R)$ is projective and that R is a PID implies that $C^{\mathrm{BM}}_{dn-2p}(\bar{S}_\beta, R)$ is itself free. In particular, $C_\bullet^{\mathrm{BM}}(\bar{S}_\beta, R)$ is itself a free resolution, and the truncation map has a section. One checks that any choice of section gives rise to a well-defined chain map $C_\bullet^{\mathrm{BM}}(\bar{S}_\beta, R) \to C_\bullet^{\mathrm{BM}}(\bar{S}_\alpha, R)$ for $\bar{S}_\beta \subset \bar{S}_\alpha$, making the assignment $\beta \mapsto C_\bullet^{\mathrm{BM}}(\bar{S}_\beta, R)$ functorial.

To replace these complexes with ones that are finitely generated in each degree, we work inductively, starting with M itself. If we have chosen complexes $C_\bullet(\bar{S}_\beta) \xrightarrow{\sim} C_\bullet^{\mathrm{BM}}(\bar{S}_\beta, R)$ for all $\beta \in P_{\mathbb{A}}(n)$, $n < N$, then the condition that we have an "on-the-nose" Künneth isomorphism determines our choice of $C_\bullet(\bar{S}_\beta)$ for all decomposable strata $\beta \in P_{\mathbb{A}}(N)$. Now I claim that if we have compatible choices of $C_\bullet(\bar{S}_\beta)$ for all β in some order ideal $I \subset P_{\mathbb{A}}(N)$, and α is any minimal element of $P_{\mathbb{A}}(N) \setminus I$, then we can also choose $C_\bullet(\bar{S}_\alpha)$ compatibly (and thus the inductive procedure can be continued). Indeed, consider the composition

$$\operatorname*{colim}_{\beta<\alpha}(C_\bullet(\bar{S}_\beta)) \to \operatorname*{colim}_{\beta<\alpha}(C_\bullet^{\mathrm{BM}}(\bar{S}_\beta, R)) \to C_\bullet^{\mathrm{BM}}(\bar{S}_\alpha, R).$$

We note that the colimit over a finite poset can be defined as the totalization of a functorially defined chain complex, which is of finite rank in each degree if each chain complex in the colimit is. By induction, the image of this composition is then finitely generated in each degree, so we may choose a quasi-isomorphism $C_\bullet(\bar{S}_\alpha) \to C_\bullet^{\mathrm{BM}}(\bar{S}_\alpha, R)$ where $C_\bullet(\bar{S}_\alpha)$ is again finitely generated in each degree, such that the image contains the image of $\operatorname{colim}_{\beta<\alpha}(C_\bullet(\bar{S}_\beta))$. Then there is a factorization

$$\operatorname*{colim}_{\beta<\alpha}(C_\bullet(\bar{S}_\beta)) \to C_\bullet(\bar{S}_\alpha) \xrightarrow{\sim} C_\bullet^{\mathrm{BM}}(\bar{S}_\alpha, R),$$

which proves the claim. □

References

[1] **A A Beĭlinson**, *On the derived category of perverse sheaves*, from "K–theory, arithmetic and geometry" (Y I Manin, editor), Lecture Notes in Math. 1289, Springer, Berlin (1987) 27–41 MR

[2] **A A Beĭlinson**, **J Bernstein**, **P Deligne**, *Faisceaux pervers*, from "Analysis and topology on singular spaces, I", Astérisque 100, Soc. Math. France, Paris (1982) 5–171 MR

[3] **M Bendersky**, **S Gitler**, *The cohomology of certain function spaces*, Trans. Amer. Math. Soc. 326 (1991) 423–440 MR

[4] **A Björner**, **T Ekedahl**, *Subspace arrangements over finite fields: cohomological and enumerative aspects*, Adv. Math. 129 (1997) 159–187 MR

[5] **T Church**, *Homological stability for configuration spaces of manifolds*, Invent. Math. 188 (2012) 465–504 MR

[6] **T Church**, **J S Ellenberg**, **B Farb**, *FI-modules and stability for representations of symmetric groups*, Duke Math. J. 164 (2015) 1833–1910 MR

[7] **T Church**, **J S Ellenberg**, **B Farb**, **R Nagpal**, *FI-modules over Noetherian rings*, Geom. Topol. 18 (2014) 2951–2984 MR

[8] **T Church**, **B Farb**, *Representation theory and homological stability*, Adv. Math. 245 (2013) 250–314 MR

[9] **F R Cohen**, **T J Lada**, **J P May**, *The homology of iterated loop spaces*, Lecture Notes in Math. 533, Springer, Berlin (1976) MR

[10] **F R Cohen**, **L R Taylor**, *Computations of Gel'fand–Fuks cohomology, the cohomology of function spaces, and the cohomology of configuration spaces*, from "Geometric applications of homotopy theory, I" (M Barratt, M Mahowald, editors), Lecture Notes in Math. 657, Springer, Berlin (1978) 106–143 MR

[11] **P Deligne**, *Théorie de Hodge, II*, Inst. Hautes Études Sci. Publ. Math. 40 (1971) 5–57 MR

[12] **P Deligne**, **P Griffiths**, **J Morgan**, **D Sullivan**, *Real homotopy theory of Kähler manifolds*, Invent. Math. 29 (1975) 245–274 MR

[13] **B Farb**, **J Wolfson**, *Étale homological stability and arithmetic statistics*, preprint (2015) arXiv

[14] **S I Gelfand**, **Y I Manin**, *Methods of homological algebra*, 2nd edition, Springer, Berlin (2003) MR

[15] **E Getzler**, *Resolving mixed Hodge modules on configuration spaces*, Duke Math. J. 96 (1999) 175–203 MR

[16] **E Getzler**, **J D S Jones**, *Operads, homotopy algebra and iterated integrals for double loop spaces*, unpublished manuscript (1994) arXiv

[17] **M Goresky**, **R MacPherson**, *Stratified Morse theory*, Ergeb. Math. Grenzgeb. 14, Springer, Berlin (1988) MR

[18] **L Illusie**, *Complexe cotangent et déformations, I*, Lecture Notes in Mathematics 239, Springer, Berlin (1971) MR

[19] **A Joyal**, *Foncteurs analytiques et espèces de structures*, from "Combinatoire énumérative" (G Labelle, P Leroux, editors), Lecture Notes in Math. 1234, Springer, Berlin (1986) 126–159 MR

[20] **A A Klyachko**, *Lie elements in a tensor algebra*, Sibirsk. Mat. Ž. 15 (1974) 1296–1304 MR In Russian; translated in Siberian Math. J. 15 (1974) 914–920

[21] **A Kupers**, **J Miller**, **T Tran**, *Homological stability for symmetric complements*, Trans. Amer. Math. Soc. 368 (2016) 7745–7762 MR

[22] **E Looijenga**, *Cohomology of $\mathcal{M}_3$ and $\mathcal{M}_3^1$*, from "Mapping class groups and moduli spaces of Riemann surfaces" (C-F Bödigheimer, R M Hain, editors), Contemp. Math. 150, Amer. Math. Soc., Providence, RI (1993) 205–228 MR

[23] **D McDuff**, *Configuration spaces of positive and negative particles*, Topology 14 (1975) 91–107 MR

[24] **R Nagpal**, *FI-modules and the cohomology of modular representations of symmetric groups*, preprint (2015) arXiv

[25] **P Orlik**, **L Solomon**, *Combinatorics and topology of complements of hyperplanes*, Invent. Math. 56 (1980) 167–189 MR

[26] **M Saito**, *Mixed Hodge modules*, Publ. Res. Inst. Math. Sci. 26 (1990) 221–333 MR

[27] **R P Stanley**, *Some aspects of groups acting on finite posets*, J. Combin. Theory Ser. A 32 (1982) 132–161 MR

[28] **B Totaro**, *Configuration spaces of algebraic varieties*, Topology 35 (1996) 1057–1067 MR

[29] **R Vakil**, **M M Wood**, *Discriminants in the Grothendieck ring*, Duke Math. J. 164 (2015) 1139–1185 MR

[30] **V A Vassiliev**, *Complements of discriminants of smooth maps: topology and applications*, Transl. Math. Monogr. 98, Amer. Math. Soc., Providence, RI (1992) MR

[31] **V A Vassiliev**, *Topology of discriminants and their complements*, from "Proceedings of the International Congress of Mathematicians" (S D Chatterji, editor), volume 1, Birkhäuser, Basel (1995) 209–226 MR

[32] **M L Wachs**, *Poset topology: tools and applications*, from "Geometric combinatorics" (E Miller, V Reiner, B Sturmfels, editors), IAS/Park City Math. Ser. 13, Amer. Math. Soc., Providence, RI (2007) 497–615 MR

Department of Mathematical Sciences, University of Copenhagen, Universitetsparken 5, DK-2100 Copenhagen, Denmark

danpete@math.ku.dk

Proposed: Benson Farb | Received: 17 May 2016
Seconded: Ralph Cohen, Dan Abramovich | Revised: 14 July 2016

Guidelines for Authors

Submitting a paper to Geometry & Topology

Papers must be submitted using the upload page at the GT website. You will need to choose a suitable editor from the list of editors' interests and to supply MSC codes.

Preparing your article for Geometry & Topology

The normal language used by the journal is English. Articles written in other languages are acceptable, provided your chosen editor is comfortable with the language and you supply an additional English version of the abstract.

At the time of submission you need only supply a PDF file. Once accepted for publication, the paper must be supplied in LaTeX, preferably using the journal's class file. More information on preparing articles in LaTeX for publication in GT is available on the GT website.

`arXiv` papers

GT papers are published simultaneously on the `arXiv`. If your paper has previously been deposited there, we shall need the `arXiv` password at acceptance time, to submit the published version to the `arXiv`.

References

Bibliographical references should be listed alphabetically at the end of the paper. All references in the bibliography should be cited at least once in the text. Use of BibTeX is preferred but not required. Any bibliographical citation style may be used, but will be converted to the house style (see a current issue for examples).

Figures

Figures, whether prepared electronically or hand-drawn, must be of publication quality. Figures prepared electronically should be submitted in Encapsulated PostScript (EPS) or in a form that can be converted to EPS, such as GnuPlot, Maple, Mathematica or XFig. Many drawing tools such as Adobe Illustrator and Aldus FreeHand also produce EPS output. Figures containing bitmaps should be generated at the highest possible resolution. If there is doubt whether a particular figure is in an acceptable format, check with production by sending an email to gt@msp.warwick.ac.uk.

Proofs

Page proofs will be made available to authors (or to the designated corresponding author) in PDF format. Failure to acknowledge the receipt of proofs or to return corrections within the requested deadline may cause publication to be postponed.

Geometry & Topology

Volume 21 Issue 4 (pages 1931–2555) 2017

1465-3060(2017)21:4;1-S