Geometry & Topology

Volume 28 (2024)

Issue 5 (pages 1995–2482)

msp

GEOMETRY & TOPOLOGY

msp.org/gt

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

The subscription price for 2024 is US $805/year for the electronic version, and $1135/year (+$70, if shipping outside the US) for print and electronic. Subscriptions, requests for back issues and changes of subscriber address should be sent to MSP. 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) is published 9 times per year and continuously online, by Mathematical Sciences Publishers, c/o Department of Mathematics, University of California, 798 Evans Hall #3840, Berkeley, CA 94720-3840. Periodical rate postage paid at Oakland, CA 94615-9651, and additional mailing offices. POSTMASTER: send address changes to Mathematical Sciences Publishers, c/o Department of Mathematics, University of California, 798 Evans Hall #3840, Berkeley, CA 94720-3840.

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

PUBLISHED BY

mathematical sciences publishers

nonprofit scientific publishing

http://msp.org/

msp

Geometry & Topology **28**:5 (2024) 1995–2124
DOI: 10.2140/gt.2024.28.1995
Published: 24 August 2024

Shear-shape cocycles for measured laminations and ergodic theory of the earthquake flow

AARON CALDERON
JAMES FARRE

We extend Mirzakhani's conjugacy between the earthquake and horocycle flows to a bijection, demonstrating conjugacies between these flows on all strata and exhibiting an abundance of new ergodic measures for the earthquake flow. The structure of our map indicates a natural extension of the earthquake flow to an action of the upper-triangular subgroup $P < \mathsf{SL}_2\,\mathbb{R}$ and we classify the ergodic measures for this action as pullbacks of affine measures on the bundle of quadratic differentials. Our main tool is a generalization of the shear coordinates of Bonahon and Thurston to arbitrary measured laminations.

30F30, 30F60, 32G15; 37D40

1 Main results

1.1 Conjugating earthquake and horocycle flow

This paper deals with two notions of unipotent flow over the moduli space $\mathcal{M}_g$ of Riemann surfaces. The first is the *Teichmüller horocycle flow*, defined on the bundle $\mathcal{Q}^1\mathcal{M}_g$ of unit-area quadratic differentials q by postcomposing the charts of the flat metric $|q|$ by the parabolic transformation $\left(\begin{smallmatrix} 1 & s \\ 0 & 1 \end{smallmatrix}\right)$. This flow is ergodic with respect to a finite measure induced by Lebesgue in local period coordinates [Masur 1982; Veech 1982] and is a fundamental object of study in Teichmüller dynamics.

The second is the *earthquake flow* on the bundle $\mathcal{P}^1\mathcal{M}_g$, whose fiber is the sphere of unit-length measured geodesic laminations on a hyperbolic surface. The earthquake flow is defined as a generalization of twisting about simple closed curves, or by postcomposing hyperbolic charts by certain piecewise-isometric transformations. While this flow is more mysterious, earthquakes are a familiar tool in Teichmüller theory, playing a central role in Kerckhoff's proof [1983] of the Nielsen realization conjecture, for example.

These two flows are both assembled from families of Hamiltonian flows (extremal length for horocycle [Papadopoulos 1986] and hyperbolic length for earthquake [Kerckhoff 1983; Wolpert 1983; Sözen and Bonahon 2001]) and exhibit similar nondivergence properties [Minsky and Weiss 2002], but the horocycle flow belongs properly to the flat-geometric viewpoint and the earthquake flow to the hyperbolic one. All the same, Mirzakhani [2008, Theorem 1.1] established a bridge between the two worlds, demonstrating a measurable conjugacy between the earthquake and horocycle flows. Consequently, the earthquake flow is ergodic with respect to the measure class of Lebesgue on $\mathcal{P}^1\mathcal{M}_g$.

In this article, we deepen this connection between flat and hyperbolic geometry, proving that the correspondence can be further upgraded to yield new results on both the ergodic theory of the earthquake flow and the structure of Teichmüller space.

Theorem A *Mirzakhani's conjugacy extends to a bijection*

$$\mathbb{O}\colon \mathcal{P}^1\mathcal{M}_g \leftrightarrow \mathcal{Q}^1\mathcal{M}_g$$

that conjugates earthquake flow to horocycle flow.

The moduli space of quadratic differentials is naturally partitioned into *strata* $\mathcal{Q}^1\mathcal{M}_g(\underline{\kappa})$, disjoint subsets parametrizing unit-area differentials with zeros of order $\underline{\kappa} = (\kappa_1, \dots, \kappa_n)$. Similarly, for any $\underline{\kappa}$ we may define the *regular locus* $\mathcal{P}^1\mathcal{M}_g^{\mathrm{reg}}(\underline{\kappa})$ to be the set of (X, λ) where λ cuts X into ideal polygons with $(\kappa_1 + 2, \dots, \kappa_n + 2)$ many sides, each with a cyclic symmetry of that order.

With this notation, Mirzakhani's conjugacy can more precisely be stated as the existence of a bijection

$$\mathcal{P}^1\mathcal{M}_g^{\mathrm{reg}}(1^{4g-4}) \leftrightarrow \mathcal{Q}^1\mathcal{M}_g^{\mathrm{nsc}}(1^{4g-4})$$

taking earthquake flow to horocycle flow, where the superscript nsc specifies the (full-measure) sublocus of the stratum consisting of those differentials with no horizontal saddle connections.

One of our main applications of Theorem A is to produce an analogue of Mirzakhani's conjugacy for components of strata (even those coming from global squares of abelian differentials), confirming a conjecture of Alex Wright [2022, Remark 5.6] (see also [loc. cit., Problems 12.5 and 12.6]).

Theorem B *For every $\underline{\kappa}$, the map $\mathbb{C}$ restricts to a bijection*

$$\mathcal{P}^1\mathcal{M}_g^{\mathrm{reg}}(\underline{\kappa}) \leftrightarrow \mathcal{Q}^1\mathcal{M}_g^{\mathrm{nsc}}(\underline{\kappa})$$

that takes earthquake to horocycle flow and (generalized) stretch rays to Teichmüller geodesics.

While strata of holomorphic quadratic differentials are generally not connected, for $g \neq 4$ their connected components are classified by whether or not they consist of squares of abelian differentials and the parity of the induced spin structure (both of which depend only on the horizontal foliation when there are no horizontal saddles), as well as hyperellipticity [Kontsevich and Zorich 2003; Lanneau 2008].[1] The bijection $\mathbb{C}$ respects both the horizontal direction and the $\mathrm{Mod}(S)$–action, so Theorem B can be refined to describe the preimages of these components.

As an immediate consequence of Theorem B, the earthquake flow is ergodic with respect to the pushforward by $\mathbb{C}^{-1}$ of the Masur–Veech measure on any component of any stratum of quadratic differentials.

1.2 Geodesic flows and P–invariant measures

Pulling back the Teichmüller geodesic flow via $\mathbb{C}$ allows us to specify a family of "dilation rays" which serve as a geodesic flow for the earthquake flow's parabolic action and in many cases project to geodesics for Thurston's Lipschitz asymmetric metric. Combining dilation rays and the earthquake flow therefore gives a action of the upper-triangular subgroup $P < \mathsf{SL}_2\,\mathbb{R}$ on $\mathcal{P}^1\mathcal{M}_g$ by "stretchquakes". See Section 15.3.

Due in part to the failure of $\mathbb{C}$ to be continuous, the stretchquake action on $\mathcal{P}^1\mathcal{M}_g$ is not by homeomorphisms but rather by measurable bijections. More precisely, it preserves the σ–algebra obtained by pulling back the Borel σ–algebra of $\mathcal{Q}^1\mathcal{M}(S)$ along $\mathbb{C}$. In a sequel [Calderon and Farre 2024a], we show that $\mathbb{C}$ is actually a measurable isomorphism with respect to the Borel σ–algebra on $\mathcal{P}^1\mathcal{M}_g$ and that the stretchquake action restricted to each $\mathcal{P}^1\mathcal{M}_g^{\mathrm{reg}}(\underline{\kappa})$ is by homeomorphisms; see also Remark 2.2.

Remark 1.1 In fact, Arana-Herrera and Wright [2024] have shown that there is *no* continuous map conjugating the earthquake flow to horocycle flow, at least when $\mathcal{P}^1\mathcal{M}_g$ and $\mathcal{Q}^1\mathcal{M}_g$ are equipped with their standard topologies.

In their foundational work on $\mathsf{SL}_2\,\mathbb{R}$–invariant ergodic measures on the moduli space of flat surfaces, Eskin and Mirzakhani [2018, Theorem 1.4] proved that the support of any P–invariant ergodic measure on $\mathcal{Q}^1\mathcal{M}_g$ is locally an affine manifold cut out by linear equations in period coordinates. Our conjugacy translates this classification into a classification of ergodic measures for the extension of the earthquake flow defined above:

[1] In genus 4, there are certain strata whose components have only been characterized via algebraic geometry [Chen and Möller 2014].

Theorem C *Every stretchquake-invariant ergodic measure is the pullback of an affine measure.*

Proof If ν is a stretchquake-invariant ergodic measure on $\mathcal{P}^1\mathcal{M}_g$, then $\mathbb{C}_*\nu$ is a P–invariant ergodic measure on $\mathcal{Q}^1\mathcal{M}_g$, which is affine by [Eskin and Mirzakhani 2018, Theorem 1.4]. □

Using this correspondence we obtain a geometric rigidity phenomenon for stretchquake-invariant ergodic measures on $\mathcal{P}^1\mathcal{M}_g$; the generic point is made out of a fixed collection of regular ideal polygons.

Corollary 1.2 *For any stretchquake-invariant ergodic probability measure ν on $\mathcal{P}^1\mathcal{M}_g$, there is some $\underline{\kappa}$ such that ν–almost every (X,λ) lies in $\mathcal{P}^1\mathcal{M}_g^{\mathrm{reg}}(\underline{\kappa})$.*

This in particular implies that the dynamics of the stretchquake action with respect to any ergodic probability measure are measurably the same as its restriction to a stratum, on which we can identify dilation rays as (directed, unit-speed) geodesics for the Lipschitz asymmetric metric on $\mathcal{T}(S)$ (see Proposition 15.12).

Remark 1.3 General ergodic measures for the stretchquake action can look quite different than the Lebesgue measure class on $\mathcal{P}^1\mathcal{M}_g$, even when pushed down to $\mathcal{M}_g$. For example, if ν gives full measure to $\mathcal{P}^1\mathcal{M}_g^{\mathrm{reg}}(4g-4)$, then a ν–generic point is obtained by gluing together a single regular ideal $(4g-2)$–gon; in particular, the injectivity radius at the center of the polygon can be arbitrarily large, allowing $g\to\infty$. This implies that ν gives zero mass to (the restriction of $\mathcal{P}^1\mathcal{M}_g$ to) sufficiently thin parts of the moduli space, as any (X,λ) where X has a very short pants decomposition has injectivity radius uniformly bounded above.

Remark 1.4 While an important result of [Eskin and Mirzakhani 2018] is that any P–invariant ergodic measure on $\mathcal{Q}^1\mathcal{M}_g$ is actually $\mathsf{SL}_2\,\mathbb{R}$–invariant, the circle action on $\mathcal{Q}^1\mathcal{M}_g$ (corresponding to rotating a quadratic differential) does not have an obvious geometric interpretation on $\mathcal{P}^1\mathcal{M}_g$. See also [Wright 2020, Problems 12.3 and 12.4].

1.3 Dual foliations from hyperbolic structures

A foundational result of Gardiner and Masur (Theorem 2.1 below) states that quadratic differentials are parametrized by their real and imaginary parts, or, equivalently, their vertical and horizontal foliations (or laminations). In particular, the real-analytic submanifold $\mathcal{F}^{uu}(\lambda)$ of all quadratic differentials with horizontal lamination λ can be identified with the space $\mathcal{MF}(\lambda)$ of foliations that bind together with λ. See Section 2 for a formal definition. As the horocycle flow preserves the horizontal foliation, it induces a flow on $\mathcal{MF}(\lambda)$.

Mirzakhani's conjugacy and our extension therefore both follow from the construction of flow-equivariant maps that assign to a hyperbolic surface X and a measured lamination λ a "dual" measured foliation.

For maximal laminations λ, this dual is the *horocyclic foliation* $F_\lambda(X)$ introduced by Thurston [1986], obtained by foliating the spikes of each triangle of $X \setminus \lambda$ by horocycles and extending across the leaves of λ. The measure of an arc transverse to $F_\lambda(X)$ is then the total distance along λ between horocycles meeting the arc at its endpoints. As $F_\lambda(X)$ necessarily binds S together with λ, this defines a map

$$F_\lambda : \mathcal{T}(S) \to \mathcal{MF}(\lambda).$$

We endow $\mathcal{MF}(\lambda)$ with the real-analytic structure coming from its identification with $\mathcal{F}^{uu}(\lambda)$. The main engine of Mirzakhani's conjugacy is the following theorem of [Bonahon 1996; Thurston 1986]; see also Section 2.1 for a discussion of her interpretation of this result.

Theorem 1.5 (Bonahon and Thurston) *For any maximal λ, the horocyclic foliation map F_λ is a real-analytic homeomorphism which takes the earthquake in λ to the horocycle flow restricted to $\mathcal{MF}(\lambda) \cong \mathcal{F}^{uu}(\lambda)$ in a time-preserving way. Moreover, the family $\{F_\lambda\}$ is equivariant with respect to the* $\mathrm{Mod}(S)$*–action. That is, $F_{g\lambda}(gX) = gF_\lambda(X)$ for all $g \in \mathrm{Mod}(S)$.*

When λ is not maximal, the horocyclic foliation is no longer defined. The first thing one might try is to simply choose a completion of λ, but this approach is too naive. Indeed, this would require choosing a completion of every lamination, which necessarily destroys $\mathrm{Mod}(S)$–equivariance because laminations (and differentials) can have symmetries.[2] Such a map will not descend to moduli space and is therefore unsuitable for our applications. Besides, for our purposes it is important that the geometry of the subsurfaces of $X \setminus \lambda$ predict the singularity structure of the corresponding differential.

If one restricts their attention to the case when λ is filling and cuts X into regular ideal polygons, then there is a canonical notion of horocyclic foliation. While this construction is equivalent on the regular locus to the more general procedure we describe just below, any attempt to prove Theorem B with this restricted viewpoint would necessarily rely on ($\mathrm{Mod}(S)$–equivariant) descriptions of the loci of surfaces built from regular polygons, as well as the intersection of $\mathcal{F}^{uu}(\lambda)$ with strata, results which (to the knowledge of the authors) were heretofore unknown. Compare Corollary 2.6 and Section 2.2.

We therefore place no restrictions on the topological type or the complementary geometry of λ. Following a suggestion of Yi Huang (communicated to us by Alex Wright), we prove that the correct analogue of the horocyclic foliation for nonmaximal λ is the *orthogeodesic foliation* $\mathcal{O}_\lambda(X)$, whose leaves are the fibers of the closest-point projection to λ and whose measure is given by length of the projection to λ. As in the maximal case, the orthogeodesic foliation binds together with λ, inducing a map

$$\mathcal{O}_\lambda : \mathcal{T}(S) \to \mathcal{MF}(\lambda).$$

See Section 5 for a more detailed discussion of this construction.

[2] For example, take γ to be a simple closed curve; completions of γ correspond to triangulations of $X \setminus \gamma$ where the boundaries are shrunk to cusps (up to a choice of spiraling about each side of γ). The space of such triangulations carries a rich $\mathrm{Stab}(\gamma)$–action, and a computation shows that the horocyclic foliations for two completions in the same $\mathrm{Stab}(\gamma)$ orbit need not be equal.

Theorem D *For any $\lambda \in \mathcal{ML}(S)$, the orthogeodesic foliation map $\mathbb{O}_\lambda$ is a homeomorphism which takes the earthquake in λ to the horocycle flow restricted to $\mathcal{MF}(\lambda) \cong \mathcal{F}^{uu}(\lambda)$ in a time-preserving way. Moreover, the family $\{\mathbb{O}_\lambda\}$ is equivariant with respect to the $\mathrm{Mod}(S)$–action. That is, $\mathbb{O}_{g\lambda}(gX) = g\mathbb{O}_\lambda(X)$ for all $g \in \mathrm{Mod}(S)$.*

Although $\mathcal{MF}(\lambda)$ does not have an obvious smooth structure, the map $\mathbb{O}_\lambda$ still exhibits a surprising amount of regularity; see Theorem E.

The proof of Theorem D requires generalizing Bonahon's machinery of transverse cocycles to new combinatorial objects, called "shear-shape cocycles", which capture the essential structure of the orthogeodesic foliation; see Section 2.1. The space of shear-shape cocycles forms a common coordinatization of both $\mathcal{T}(S)$ and $\mathcal{MF}(\lambda)$ that is compatible with the map $\mathbb{O}_\lambda$ and reveals an abundance of structure encoded in the orthogeodesic foliation:

- When λ cuts X into regular ideal polygons, the orthogeodesic and horocyclic foliations agree.
- The locus of points of X which are closest to at least two leaves of λ forms a piecewise-geodesic spine for $X \setminus \lambda$ which captures the geometry and topology of the complementary subsurfaces (see Theorem 6.4). Moreover, this spine is exactly the diagram of horizontal separatrices for the quadratic differential with horizontal foliation λ and vertical foliation $\mathbb{O}_\lambda(X)$.
- For every measure μ on λ, the intersection of μ and $\mathbb{O}_\lambda(X)$ is the hyperbolic length of μ on X.
- The pullbacks of Teichmüller geodesics with no horizontal saddle connections are geodesics with respect to Thurston's Lipschitz (asymmetric) metric (Proposition 15.12).

The orthogeodesic foliation map can also be thought of as relating the hyperbolic and extremal length functions $\ell_\lambda(\cdot)$ and $\mathrm{Ext}_\lambda(\cdot)$ for any fixed λ. Indeed, a seminal theorem of [Hubbard and Masur 1979] states that the natural projection

$$\pi : \mathcal{F}^{uu}(\lambda) \to \mathcal{T}(S)$$

that records only the complex structure underlying a differential is a homeomorphism. Combining this with the fact that the extremal length of λ on Y is exactly the area of the differential $\pi^{-1}(Y)$, we deduce that:

Corollary 1.6 *For every $\lambda \in \mathcal{ML}(S)$, the map $\pi \circ \mathbb{O}_\lambda$ is a $\mathrm{Stab}(\lambda)$–equivariant self-homeomorphism of $\mathcal{T}(S)$ that takes the hyperbolic length function $\ell_\lambda(\cdot)$ to the extremal length function $\mathrm{Ext}_\lambda(\cdot)$.*

Acknowledgments

The authors would firstly like to thank Alex Wright for providing the germ of this project and useful suggestions, as well as for helpful comments on a preliminary draft of this paper. We are grateful to Francis Bonahon, Feng Luo, Howie Masur and Jing Tao for lending their expertise and for enlightening discussions.

The authors would also like to thank those who contributed helpful comments or with whom we had clarifying conversations, including Daniele Alessandrini, Francisco Arana-Herrera, Mladen Bestvina, Jon Chaika, Vincent Delecroix, Valentina Disarlo, Spencer Dowdall, Ben Dozier, Aaron Fenyes, Ser-Wei Fu, Curt McMullen, Mareike Pfeil, Beatrice Pozzetti, John Smillie and Sam Taylor. We also want to thank the referee for their careful reading and useful comments.

Finally, the authors are indebted to Yair Minsky for his dedicated guidance and mentorship throughout all stages of this project, as well as his generous listening and insightful comments.

Calderon gratefully acknowledges support from NSF grants DGE-1122492, DMS-161087 and DMS-2005328, and travel support from NSF grants DMS-1107452, DMS-1107263 and DMS-1107367 *RNMS*: *Geometric structures and representation varieties* (the GEAR Network). Farre gratefully acknowledges support from NSF grants DMS-1246989, DMS-1509171, DMS-1902896, DMS-161087 and DMS-2005328.

Portions of this work were accomplished while the authors were visiting MSRI for the Fall 2019 program *Holomorphic differentials in mathematics and physics*, and the authors would like to thank the venue for its hospitality and excellent working environment. Part of this material is based upon work supported by the NSF under grant DMS-1928930 while Farre participated in the MSRI Fall 2020 program *Random and arithmetic structures in topology*. Farre would also like to thank the both the Department of Mathematics at the University of Utah and the Mathematics Institute at Universität Heidelberg for their hospitality and rich working environments.

2 About the proof

Given Theorem D, which associates to (X, λ) a dual foliation $\mathcal{C}_\lambda(X)$ describing the geometry of the pair, it is not difficult to prove Theorems A and B. First we recall the relationship between differentials, foliations and laminations in a little more detail.

The space of measured foliations (up to equivalence) on a closed surface S of genus $g \geq 2$ is denoted by $\mathcal{MF}(S)$. There is a canonical identification [Levitt 1983] between $\mathcal{MF}(S)$ and $\mathcal{ML}(S)$, the space of measured laminations on S; throughout this paper we will implicitly pass between the two notions at will, depending on our situation. By $\mathcal{Q}\mathcal{T}_g$ and $\mathcal{Q}^1\mathcal{T}_g$ we mean the bundle of holomorphic quadratic differentials over the Teichmüller space and the locus of unit-area quadratic differentials, respectively. We similarly let $\mathcal{P}\mathcal{T}_g = \mathcal{T}(S) \times \mathcal{ML}(S)$ and $\mathcal{P}^1\mathcal{T}_g$ be the locus of pairs (X, λ) where λ has unit length on X.

To every $q \in \mathcal{Q}\mathcal{T}_g$ one may associate the real measured foliation $|\mathrm{Re}(q)|$ which measures the total variation of the real part of the holonomy of an arc; the imaginary foliation $|\mathrm{Im}(q)|$ is defined similarly. These

foliations have vertical and horizontal trajectories, respectively, and so we will also refer to them as the vertical and horizontal foliations (or laminations) of q and write

$$q = q\big(|\mathrm{Re}(q)|, |\mathrm{Im}(q)|\big).$$

A foundational theorem of Gardiner and Masur implies that the real and imaginary foliations completely determine q, and that, given any two foliations which "fill up" the surface, one can integrate against their measures to recover a quadratic differential.

A pair of measured foliations/laminations (η, λ) is said to *bind* S if, for every $\gamma \in \mathcal{ML}(S)$,

$$i(\gamma, \eta) + i(\gamma, \lambda) > 0,$$

where $i(\cdot, \cdot)$ is the geometric intersection pairing. In the literature, such pairs are sometimes called *filling*, though we choose to distinguish the topological notion of filling from the measure-theoretic notion of binding.

Theorem 2.1 [Gardiner and Masur 1991, Thereom 3.1] *There is a* $\mathrm{Mod}(S)$*–equivariant homeomorphism*

$$\mathcal{QT}(S) \cong \mathcal{MF}(S) \times \mathcal{MF}(S) \setminus \Delta,$$

where Δ *is the set of all nonbinding pairs* (η, λ)*. In particular, the set* $\mathcal{F}^{uu}(\lambda)$ *of all quadratic differentials with* $|\mathrm{Im}(q)| = \lambda$ *may be identified with* $\mathcal{MF}(\lambda)$*, the set of foliations which together bind with* λ*.*

Proof of Theorems A and B By definition, there is a $\mathrm{Mod}(S)$–equivariant projection $\mathcal{PT}_g \to \mathcal{ML}(S)$ with fiber $\mathcal{T}(S)$. Theorem 2.1 implies there is a $\mathrm{Mod}(S)$–equivariant projection $\mathcal{QT}_g \to \mathcal{ML}(S)$ whose fiber over λ may be identified with $\mathcal{MF}(\lambda)$. Applying Theorem D on the fibers therefore yields an equivariant bijection

$$\mathbb{C}\colon \mathcal{PT}_g \leftrightarrow \mathcal{QT}_g$$

which takes unit-length laminations to unit-area differentials (Corollary 13.14), and quotienting by the $\mathrm{Mod}(S)$–action proves Theorem A.

Furthermore, we observe that the spine of the orthogeodesic foliation of a regular ideal $(k+2)$–gon is just a star with $k+2$ edges, which corresponds to the separatrix diagram of a zero of order k when there are no horizontal saddle connections. Thus $\mathbb{C}$ restricts to the promised conjugacy on strata (Theorem B). □

Remark 2.2 Mirzakhani's conjugacy is defined on the Borel subset $\mathcal{PT}_g^{\mathrm{reg}}(1^{4g-4}) \subset \mathcal{PT}_g$ of full Lebesgue measure and is moreover Borel measurable on its domain of definition. The latter assertion is a consequence of a stronger result, namely that $\mathcal{PT}_g^{\mathrm{reg}}(1^{4g-4}) \to \mathcal{QT}_g$ is continuous (with respect to the subspace topology on $\mathcal{PT}_g^{\mathrm{reg}}(1^{4g-4})$).

While convergence of measured laminations (in measure) does not typically imply Hausdorff convergence of the supports, whenever a sequence $\{\lambda_n\}$ of maximal measured laminations converges to a maximal

measured lamination λ, λ_n is eventually carried (snugly) on a maximal train track also carrying λ. From here, it is not difficult to deduce that $\lambda_n \to \lambda$ in the Hausdorff topology [Zhu and Bonahon 2004] and thus the horocyclic foliations $F_{\lambda_n}(X)$ converge to $F_{\lambda(X)}$. Intuitively, the leaves of λ_n intersect the leaves of λ with small angle (depending on the specific surface on which they are realized), so the orthogonal directions become more parallel.

In [Calderon and Farre 2024a], we extend these ideas and prove that $\mathbb{O}$ is (everywhere) Borel measurable with Borel measurable inverse by identifying a countable partition of $\mathcal{PT}_g$ and $\mathcal{QT}_g$ into Borel subsets on which $\mathbb{O}$ is homeomorphic. See also Section 16.

In general, the compact edges of the spine of a pair (X, λ) correspond exactly to horizontal saddle connections in the differential $\mathbb{O}(X, \lambda)$. This observation allows us to prove that the generic point for a P–invariant ergodic probability measure on $\mathcal{P}^1\mathcal{M}_g$ consists of pairs (X, λ) where λ cuts X into a fixed set of regular ideal polygons.

Proof of Corollary 1.2 Using our conjugacy, the desired statement is equivalent to the fact that any P–invariant ergodic probability measure on $\mathcal{Q}^1\mathcal{M}_g$ is

(a) supported in a single stratum, and

(b) gives 0 measure to the set of differentials with horizontal saddle connections.

The first statement is implied by ergodicity, while the second follows from the fact that the measure is actually $\mathsf{SL}_2\,\mathbb{R}$–invariant [Eskin et al. 2015]. Indeed, for any quadratic differential q, the Lebesgue measure of the set of directions θ such that $e^{i\theta}q$ has a saddle connection is 0, so Fubini's theorem implies (b). □

Refining the proof by considering connected components of strata, we can also conclude that ν–almost every pair has the same orientability, spin and hyperellipticity properties.

2.1 Shear-shape coordinates

Our strategy to prove Theorem D follows Mirzakhani's interpretation of Theorem 1.5, in which she clarifies the relationship between Thurston's geometric perspective on the horocyclic foliation and Bonahon's powerful analytic approach in terms of transverse cocycles. Namely, she shows that the horocyclic foliation map F_λ is compatible with shearing coordinates for both hyperbolic structures and measured foliations. To motivate our construction, we give a brief outline of Mirzakhani's proof below.

A (real-valued) *transverse cocycle* for λ is a finitely additive signed measure on arcs transverse to λ that is invariant under isotopy transverse to λ; observe that transverse measures are themselves transverse cocycles. These objects equivalently manifest as transverse Hölder distributions, cohomology classes, or weight systems on snug train tracks [Bonahon 1997a; 1996; 1997b]. The space $\mathcal{H}(\lambda)$ of transverse

cocycles forms a finite-dimensional vector space which carries a natural homological intersection pairing which is nondegenerate when λ is maximal. The intersection pairing then identifies a "positive locus" $\mathscr{H}^+(\lambda) \subset \mathscr{H}(\lambda)$ cut out by finitely many geometrically meaningful linear inequalities. See also Section 7.1.

Bonahon [1996, Theorem A] proved that, for any maximal geodesic lamination λ, there is a real-analytic homeomorphism $\sigma_\lambda \colon \mathcal{T}(S) \to \mathscr{H}^+(\lambda)$ that takes a hyperbolic metric to its "shearing cocycle", which essentially records the signed distance along λ between the centers of ideal triangles in the complement of λ. Mirzakhani [2008, Sections 5.2 and 6.2] then constructed a homeomorphism I_λ (essentially by a well-chosen system of period coordinates) that coordinatizes $\mathcal{MF}(\lambda)$ by $\mathscr{H}^+(\lambda)$ and for which the following diagram commutes:

$$\begin{array}{ccc} \mathcal{T}(S) & \xrightarrow{\quad F_\lambda \quad} & \mathcal{MF}(\lambda) \\ & {\scriptstyle \sigma_\lambda} \searrow \quad \swarrow {\scriptstyle I_\lambda} & \\ & \mathscr{H}^+(\lambda) & \end{array} \tag{1}$$

Since $F_\lambda = I_\lambda^{-1} \circ \sigma_\lambda$ is a composition of homeomorphisms, it is itself a homeomorphism. As the construction of the horocyclic foliation requires no choices, the family $\{F_\lambda\}$ is necessarily $\mathrm{Mod}(S)$–equivariant. Finally, a direct computation shows that σ_λ transports the earthquake in λ to translation in $\mathscr{H}^+(\lambda)$ by λ, and I_λ similarly takes horocycle to translation, demonstrating Theorem 1.5.

Shear-shape cocycles When λ is not maximal, the space of transverse cocycles is no longer suitable to coordinatize hyperbolic structures (or transverse foliations). Indeed, in this case the vector space $\mathscr{H}(\lambda)$ has dimension less than $6g-6$ and its intersection form may be degenerate; this is a consequence of the fact that the Teichmüller space of $S \setminus \lambda$ now has a rich analytic structure that transverse cocycles cannot see.

In order to imitate (1) and its concomitant arguments for arbitrary $\lambda \in \mathcal{ML}(S)$, we therefore introduce the notion of *shear-shape cocycles* on λ. Roughly, a shear-shape cocycle consists of finitely additive signed data on certain arcs transverse to λ together with a weighted arc system that cuts $S \setminus \lambda$ into cells; this pair is also required to satisfy a certain compatibility condition mimicking features of the orthogeodesic foliation (Definition 7.11). Generalizing results of [Luo 2007, Theorem 1.2 and Corollary 1.4], we show that such an arc system is equivalent to a hyperbolic structure on $S \setminus \lambda$ (Theorem 6.4), so shear-shape cocycles may equivalently be thought of as transverse data together with a compatible hyperbolic structure on the complementary subsurface(s). Like transverse cocycles, shear-shape cocycles also admit realizations as cohomology classes or weight systems on certain train tracks (Definition 7.5 and Proposition 9.5).

Remark 2.3 Only certain classes of arcs admit consistent weights when measured by a shear-shape cocycle, whereas transverse cocycles provide a measure to any arc transverse to λ. While this subtlety is exactly what allows us to understand how to relate shear-shape cocycles with the geometry of complementary subsurfaces, it also presents a number of technical challenges throughout the paper.

Unlike transverse cocycles, the space $\mathcal{SH}(\lambda)$ of shear-shape cocycles is not a vector space, instead forming a principal $\mathcal{H}(\lambda)$–bundle over a contractible analytic subvariety of $\mathcal{T}(S \setminus \lambda)$ (Theorem 8.1). All the same, the cohomological realization of shear-shape cocycles equips $\mathcal{SH}(\lambda)$ with an intersection form

$$\omega_{\mathcal{SH}} \colon \mathcal{SH}(\lambda) \times \mathcal{H}(\lambda) \to \mathbb{R}$$

that identifies a "positive locus" $\mathcal{SH}^+(\lambda)$ and equips both $\mathcal{SH}(\lambda)$ and $\mathcal{SH}^+(\lambda)$ with piecewise-integral-linear structures. The positive locus forms an $\mathcal{H}^+(\lambda)$–cone bundle over the same subvariety of $\mathcal{T}(S \setminus \lambda)$ (Proposition 8.5) and fits into the familiar-looking commutative diagram

(2)
$$\begin{array}{ccc} \mathcal{T}(S) & \xrightarrow{\quad \mathbb{O}_\lambda \quad} & \mathcal{MF}(\lambda) \\ & {\scriptstyle \sigma_\lambda} \searrow \quad \swarrow {\scriptstyle I_\lambda} & \\ & \mathcal{SH}^+(\lambda) & \end{array}$$

where σ_λ and I_λ record shearing data along λ as well as shape data in the complementary subsurfaces. These maps can be thought of as a common generalization of Bonahon and Mirzakhani's shear coordinates as well as Fenchel–Nielsen and Dehn–Thurston coordinates adapted to a pants decomposition (see Section 2.2). In the case when λ is orientable, the map I_λ can also be viewed as an extension of Minsky and Weiss's description [2014, Theorem 1.2] of the set of abelian differentials with given horizontal foliation.[3]

The conjugacy of Theorem D is then a consequence of the following structural theorem, which is an amalgam of the main technical results of the paper (compare Theorems 10.15, 12.1 and 13.13).

Theorem E *For any measured lamination λ, diagram* (2) *commutes and all arrows are* $\mathrm{Stab}(\lambda)$*–equivariant homeomorphisms. Moreover*:

- *σ_λ is (stratified) real-analytic and transports the earthquake flow to translation by λ and the hyperbolic length of λ to $\omega_{\mathcal{SH}}(\cdot, \lambda)$.*
- *The weighted arc system underlying $\sigma_\lambda(X)$ records the hyperbolic structure $X \setminus \lambda$ under the correspondence of Theorem 6.4.*
- *I_λ is piecewise-integral-linear and transports horocycle flow to translation by λ and intersection with λ to $\omega_{\mathcal{SH}}(\cdot, \lambda)$.*
- *The weighted arc system underlying $I_\lambda(\eta)$ records the compact horizontal separatrices of $q(\eta, \lambda)$.*

In the course of our proof, we also describe new "shape-shifting deformations" of hyperbolic surfaces which generalize Bonahon and Thurston's cataclysms by shearing along a lamination while also varying the hyperbolic structures on complementary pieces. See Section 15.1.

[3]Technically, Minsky and Weiss [2014] investigate the family of abelian differentials with a fixed horizontal foliation and fixed topological type of horizontal separatrix diagram, whereas our map applies to quadratic differentials (whether or not they are globally the squares of abelian differentials) and packages together all possible types of separatrix diagrams.

One particularly interesting family of deformations is obtained by dilation. The space $\mathcal{SH}^+(\lambda)$ admits a natural scaling action by $\mathbb{R}_{>0}$, and, since both earthquake and horocycle flow are carried to translation in coordinates, this scaling action indicates extensions of each to P–actions. A quick computation (Lemma 11.1) shows that the pullback of a dilation ray by I_λ is (a variant of) the Teichmüller geodesic flow, so the P–action on the flat side is just the standard P–action on $\mathcal{QT}_g$.

On the hyperbolic side, these dilation rays define our extension of the earthquake flow, and correspond to families of hyperbolic metrics on which the length of λ is scaled by a uniform factor. They are therefore natural candidates for (directed, unit-speed) geodesics for the Lipschitz asymmetric metric on $\mathcal{T}(S)$, and in some cases we can identify them as such (see Propositions 15.12 and 15.18, as well as Remarks 15.19 and 15.14).

Remark 2.4 Over the course of the paper we formalize the notion that shear-shape coordinates for hyperbolic structures are essentially the "real part" of period coordinates for $\mathcal{PT}_g$. Interpreting $\sigma_\lambda(X)+i\lambda$ as a complex weight system on a train track, Theorem C implies that the support of every stretchquake-invariant ergodic measure on $\mathcal{P}^1\mathcal{M}_g$ is locally an affine measure in train track charts. See Lemma 10.10.

Coordinatizing horospheres Since the Thurston intersection form $\omega_{\mathcal{SH}}$ captures both the hyperbolic length of and geometric intersection with λ, the coordinate systems of Theorem E also allow us to give global descriptions of the level sets of these functions. In particular, we can recover Gardiner and Masur's description [1991, page 236] of extremal-length horospheres as well as Bonahon's description of the hyperbolic-length ones (which is implicit in the structure of shear coordinates for maximal completions).

Corollary 2.5 *Suppose that λ supports k ergodic transverse measures $\lambda_1,\dots,\lambda_k$. Then, for all $L_1,\dots,L_k\in\mathbb{R}_{>0}$, the level sets*

$$\{X\in\mathcal{T}(S)\mid \ell_X(\lambda_i)=L_i \text{ for all } i\}\quad\text{and}\quad\{\eta\in\mathcal{MF}(\lambda)\mid i(\eta,\lambda_i)=L_i \text{ for all } i\}$$

are both homeomorphic to $\mathbb{R}^{6g-6-k}$.

Analyzing this coordinatization more closely, in fact both level sets can be described as affine bundles of dimension $\dim_{\mathbb{R}}\mathcal{H}^+(\lambda)-k$ over the same subvariety of $\mathcal{T}(S\setminus\lambda)$ as underlies $\mathcal{SH}(\lambda)$.

From this refinement, we are able to describe the intersection of the leaf $\mathcal{F}^{uu}(\lambda)$ with strata. The decomposition of period coordinates into real and imaginary parts shows that this intersection (when not empty) is locally homeomorphic to $\mathbb{R}^d$, where d is the complex dimension of the stratum; our work shows that these local homeomorphisms patch together to a global one. Compare [Minsky and Weiss 2014, Theorem 1.2].

Corollary 2.6 *Suppose that λ is a filling measured lamination that cuts a surface into polygons with $\kappa_1+2,\dots,\kappa_n+2$ many sides, and let $\varepsilon=+1$ if λ is orientable and -1 otherwise. Let $\mathcal{QT}_g(\underline{\kappa};\varepsilon)$ denote the union of the components of the stratum $\mathcal{QT}_g(\underline{\kappa})\subset\mathcal{QT}_g$ that either are $(\varepsilon=+1)$ or are not $(\varepsilon=-1)$*

global squares of abelian differentials. Then

$$\{q \in \mathbb{Q}\mathcal{T}_g(\underline{\kappa};\varepsilon) : |\mathrm{Im}(q)| = \lambda\} \cong \mathscr{H}^+(\lambda) \cong \mathbb{R}^d,$$

where d is the complex dimension of $\mathbb{Q}\mathcal{T}_g(\underline{\kappa};\varepsilon)$.

Proof Theorem E indicates that the metric graph of compact horizontal separatrices of $q(\eta,\lambda)$ is encoded by the weighted arc system underlying $I_\lambda(\eta)$. These weighted arc systems are organized in a piecewise-linear subvariety $\mathscr{B}(S\setminus\lambda)$ of a product of *weighted filling arc complexes* that encode the combinatorics of how a zero of order κ_i can split up into lower-order zeros joined by horizontal saddle connections (see Sections 6, 7.3 and 10.1 and Figure 5). For differentials in the indicated set, there are no compact horizontal separatrices, and so the underlying arc system is always the *empty* (filling) arc system $\varnothing \in \mathscr{B}(S\setminus\lambda)$. In other words, the image of $\{q \in \mathbb{Q}\mathcal{T}_g(\underline{\kappa};\varepsilon) : |\mathrm{Im}(q)| = \lambda\}$ in coordinates is just the fiber over $\varnothing$, where Proposition 8.5 identifies $\mathcal{SH}^+(\lambda)$ as an $\mathscr{H}^+(\lambda)$–bundle over $\mathscr{B}(S\setminus\lambda)$.

The second isomorphism $\mathscr{H}^+(\lambda) \cong \mathbb{R}^d$ is just a dimension count (see Lemmas 4.6 and 7.3 in particular). □

In general, $\mathcal{F}^{uu}(\lambda) \cap \mathbb{Q}\mathcal{T}_g(\underline{\kappa};\varepsilon)$ forms a $\mathscr{H}^+(\lambda)$–bundle over a union of faces of an arc complex of $S\setminus\lambda$. As a consequence, the only obstruction to completeness of any such leaf comes from zeros colliding along a horizontal saddle connection (see also [Minsky and Weiss 2014, Theorem 11.2]). This global description of $\mathcal{F}^{uu}(\lambda) \cap \mathbb{Q}\mathcal{T}_g(\underline{\kappa};\varepsilon)$ also allows the importation of arguments from homogeneous dynamics to investigate equidistribution in both $\mathbb{Q}^1\mathcal{M}_g$ and $\mathbb{P}^1\mathcal{M}_g$ and their strata; see the discussion in Section 16.

2.2 Generalized Fenchel–Nielsen coordinates

Our shear-shape coordinates for hyperbolic structures can be thought of as interpolating between the classical Fenchel–Nielsen coordinates adapted to a pants decomposition and Bonahon and Thurston's shear coordinates. In both cases, one remembers the shapes of the complementary subsurfaces (pairs of pants and ideal triangles, respectively) and the space of all hyperbolic structures with given complementary shape is parametrized by gluing data (twist/shear parameters).

For general λ, there is a map

$$\mathsf{cut}_\lambda : \mathcal{T}(S) \to \mathcal{T}(S\setminus\lambda)$$

that remembers the induced hyperbolic structure on each complementary subsurface. Theorem 12.1 then implies that the image of cut_λ is a real-analytic subvariety $\mathscr{B}(S\setminus\lambda)$ of $\mathcal{T}(S\setminus\lambda)$ consisting of those structures satisfying a "metric residue condition" (see Lemma 13.1). In the case where each component of λ is either nonorientable or a simple closed curve, $\mathscr{B}(S\setminus\lambda)$ is just the space of hyperbolic structures for which the two boundary components of the cut surface corresponding to a simple curve component of λ have equal length. Theorem 12.1 together with the structure of $\mathcal{SH}^+(\lambda)$ also allows us to identify the fiber $\mathsf{cut}_\lambda^{-1}(Y)$ over any $Y \in \mathscr{B}(S\setminus\lambda)$ with the gluing data $\mathscr{H}^+(\lambda)$ (though not in a canonical way).[4]

[4]See the discussion around (18) in regards to the positivity condition for disconnected λ; in essence, $\mathscr{H}^+(\lambda)$ is the product of $\mathscr{H}^+(\lambda_i)$ for each nonclosed minimal component together with the twisting data around simple closed curves.

We summarize this discussion in the following triptych:

(3)
$$\begin{array}{ccc} \text{Fenchel–Nielsen} & \text{shear-shape} & \text{shear} \\ \mathbb{R}^{3g-3} \longrightarrow \mathcal{T}(S) & \mathscr{H}^+(\lambda) \longrightarrow \mathcal{T}(S) & \mathscr{H}^+(\lambda) \longrightarrow \mathcal{T}(S) \\ \downarrow & \downarrow & \downarrow \\ \mathbb{R}^{3g-3}_{>0} & \mathscr{B}(S \setminus \lambda) & \{\text{pt}\} \\ \lambda \text{ a pants decomposition} & \lambda \text{ arbitrary} & \lambda \text{ maximal} \end{array}$$

In each coordinate system, $\mathcal{T}(S)$ is the total space of a fiber bundle over a base space of allowable shape data on the subsurface complementary to λ, while the fiber consists of gluing data.

A completely analogous picture also holds for foliations transverse to λ, demonstrating I_λ as a common generalization of both Dehn–Thurston and Mirzakhani's shear coordinates.

2.3 Fenchel–Nielsen and Dehn–Thurston via shears and shapes

In order to give the reader a concrete example of shear-shape coordinates, we include here a discussion of our construction for $\lambda = P$ a pants decomposition. In this case, shear-shape coordinates are just a (mild) reformulation of the classical Fenchel–Nielsen and Dehn–Thurston ones.

First we consider a hyperbolic structure X. A pair of pants in $X \setminus P$ is typically parametrized by its boundary lengths (a, b, c) or, equivalently, by the alternating side lengths of either of the right-angled hexagons coming from cutting along seams. The orthogeodesic foliation on a pair of pants picks out either a pair or a triple of seams (those which are realized as leaves of $\mathcal{O}_P(X)$), each weighted by the length of a boundary arc consisting of endpoints of leaves of $\mathcal{O}_P(X)$ isotopic to the seam. See Figure 1. In this case, these lengths are just simple (piecewise-)linear combinations of the boundary lengths and the metric residue condition defining $\mathscr{B}(S \setminus P)$ just states that the boundaries that are glued together must have the same length. See Figure 1.

The space $\mathscr{H}^+(P)$ reduces to a sum of the twist spaces for each curve of P, and so Proposition 8.5 implies that $\mathcal{SH}^+(P)$ is a principal $\mathbb{R}^{3g-3}$–bundle over $\mathscr{B}(S \setminus P) \cong \mathbb{R}^{3g-3}_{>0}$. The transverse data recorded by this twist space then describes the signed distance between certain reference points in pairs of right-angled hexagons in $\widetilde{X}$ that are adjacent to the same curve of $\widetilde{P}$, which is the same as the twist parameter measured by the appropriate choice of Fenchel–Nielsen coordinates.[5]

We can similarly recognize $I_\lambda \colon \mathcal{MF}(P) \to \mathcal{SH}^+(P)$ as Dehn–Thurston coordinates. Now, from any integral point $\sigma \in \mathcal{SH}^+(P)$, we can construct a multicurve α with prescribed intersection and twisting parameters as follows: the weighted arc system describes how strands of α pass between and meet the

[5]Fenchel–Nielson coordinates always involve some choice of section of the space of twists over the length parameters, and so have only the structure of a principal $\mathbb{R}^{3g-3}$–bundle over $\mathbb{R}^{3g-3}_{+}$.

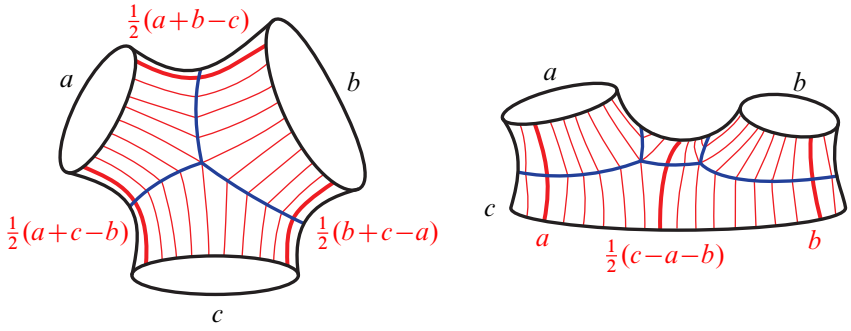


Figure 1: The orthogeodesic foliation on pairs of pants. Note that the weight of each bolded arc is a linear combination of the boundary lengths, whence the correspondence between shear-shape and Fenchel–Nielsen/Dehn–Thurston coordinates. If any of the weights is zero, the orthogeodesic foliation only picks out the two seams with nonzero weights.

components of P, while the transverse data recorded by $\mathcal{H}^+(P) \cong \mathbb{R}^P$ describes the extent that strands of α wrap around components of P. This procedure is clearly reversible and can easily be extended to transverse foliations using a family of standard train tracks on each pair of pants (see [Penner and Harer 1992, Section 2.6]). As in the hyperbolic case, one can easily pass between these coordinates and the standard Dehn–Thurston ones just by replacing the count of strands of α going from one boundary to the other with the total intersection of α with each boundary.

3 Outline of the paper

The rest of this paper is roughly divided into four parts, corresponding to the orthogeodesic foliation, shear-shape cocycles and shear-shape coordinates for flat and hyperbolic structures, as well as a collection of further directions for investigation, some of which have been completed while this article was in press (Section 16). While the constructions of I_λ and σ_λ both rely on foundational results established in the first two parts, we have attempted to direct the reader eager to understand our coordinates to the most important statements of these sections.

We expect that the reader is familiar with many of the standard constructions of Teichmüller theory, as well as the definitions of both the earthquake and horocycle flows; we recommend [Minsky and Weiss 2002, Section 4] for a particularly lucid overview of the relevant objects. We also refer the reader to [Casson and Bleiler 1988; Thurston 1979, Section 8] for more on laminations and to [Penner and Harer 1992] for a comprehensive introduction to train tracks.

Sections 4–6: the orthogeodesic foliation Cutting along a lamination results in a (possibly disconnected) hyperbolic surface Σ with crown boundary, and in Section 4 we recall some useful information about the Teichmüller spaces of such surfaces. One particularly important definition is that of the "metric residue"

of a crown end, which is a generalization of boundary length and plays an important role in cohomological constraints on the shape data of shear-shape cocycles (Lemma 7.9).

With these preliminaries established, in Section 5 we discuss in more detail the orthogeodesic foliation and the hyperbolic geometry of X in a neighborhood of λ. In this section we also give a geometric interpretation of the map in Corollary 1.6 that relates hyperbolic and extremal length.

The most important result of this part occupies Section 6, in which we show that the orthogeodesic foliation restricted to Σ completely determines its hyperbolic structure. More explicitly, dual to each compact edge of the spine of $\mathbb{O}_\lambda(X)$ is a packet of properly isotopic arcs joining nonasymptotic boundary components of Σ. By assigning geometric weights to each of these packets, we can therefore combinatorialize the restriction of $\mathbb{O}_\lambda(X)$ to Σ by a weighted filling arc system.

Using a geometric limit argument, in Theorem 6.4 we prove that the map which associates to a hyperbolic structure on Σ the associated arc system is a $\mathrm{Mod}(\Sigma)$–equivariant stratified real-analytic homeomorphism between $\mathcal{T}(\Sigma)$ and a certain type of arc complex for Σ, generalizing a theorem of [Luo 2007] for surfaces with totally geodesic boundary (see also [Mondello 2009b; Do 2008; Ushijima 1999]). Moreover, by construction, this map records both the combinatorial structure of the spine of $\mathbb{O}_\lambda(X)$ as well as the metric residue of the crowns of Σ.

Theorem 6.4 is used extensively throughout the paper in order to pass between the combinatorial data of a weighted arc system, the restriction of $\mathbb{O}_\lambda(X)$ to Σ, and the corresponding hyperbolic structure on Σ. The proof is independent of the main line of argument; as such, the reader is encouraged to understand the statement, but may wish only to skim the proof.

Sections 7–9: the space of shear-shape cocycles The second part of the paper is devoted to our construction of shear-shape cocycles for a given λ and an analysis of the space $\mathcal{SH}(\lambda)$ of all shear-shape cocycles. Upon reaching this section, the reader may find it useful to glance ahead to either Section 10 or Section 13 to instantiate our definitions.

After reviewing structural results on transverse cocycles, in Section 7 we give both cohomological and axiomatic definitions of shear-shape cocycles (Definitions 7.5 and 7.11, respectively), both predicated on some underlying weighted arc system on Σ. In Proposition 7.13 we prove these definitions agree. Using the cohomological description, we observe a constraint on the weighted arc systems that can underlie a shear-shape cocycle coming from metric residue conditions (Lemma 7.9); this can also be thought of as a generalization of the fact that one can only glue together totally geodesic boundary components of the same length (compare Lemma 13.1).

Letting $\mathcal{B}(S \setminus \lambda)$ denote the subvariety of the filling arc complex of Σ cut out by the aforementioned residue conditions, we show in Section 8 that the space $\mathcal{SH}(\lambda)$ of shear-shape cocycles forms a bundle of transverse cocycles over $\mathcal{B}(S \setminus \lambda)$ with some additional structure (Theorem 8.1) whose total space is

a cell of dimension $6g-6$ (Corollary 8.2). In this section we also introduce the Thurston intersection form on $\mathcal{SH}(\lambda)$ (Section 8.2) and prove that the positive locus $\mathcal{SH}^+(\lambda)$ it defines is itself a bundle over $\mathscr{B}(S\setminus\lambda)$ (Proposition 8.5).

Finally, in Section 9 we give train track coordinates for the space of shear-shape cocycles. The train tracks we use give a preferred decomposition of arcs on S into pieces that are measurable by shear-shape cocycles and as such give a useful way of specifying shear-shape cocycles by a finite amount of data. The weight space for a train track is also a natural model in which to consider local deformations of a shear-shape cocycle, a feature which we exploit in Section 14. In Section 9.3 we discuss how the piecewise-integral-linear structure induced by train track charts endows $\mathcal{SH}^+(\lambda)$ with a well-defined integer lattice and preferred measure in the class of Lebesgue.

The reader willing to accept the structure theorems can adequately navigate the remaining two parts of the paper using weight systems on (augmented) train tracks as a local description of the structure of shear-shape space.

Sections 10 and 11: coordinates for transverse foliations At this point, we have established the structure necessary to coordinatize foliations transverse to λ by shear-shape cocycles.

A measured foliation $\eta\in\mathcal{MF}(\lambda)$ determines a holomorphic quadratic differential $q=q(\eta,\lambda)\in\mathcal{F}^{uu}(\lambda)$ via Theorem 2.1, and we begin by specifying an arc system $\underline{\alpha}(q)$ that records the horizontal separatrices of q. We then build a train track τ carrying λ from a triangulation by saddle connections (Construction 10.4); augmenting τ by the arc system $\underline{\alpha}(q)$ then allows us to realize the periods of the triangulation as a (cohomological) shear-shape cocycle $I_\lambda(\eta)$. This identification also gives a useful formula for $I_\lambda(\eta)$ as a weight system on the augmented train track τ (Lemma 10.10).

We then show that one can rebuild q just from the train track weights defined by $I_\lambda(\eta)$; a similar (but more technical) argument then gives that $I_\lambda(\eta)\in\mathcal{SH}^+(\lambda)$ (Proposition 10.12). This reconstruction technique together with the structure of shear-shape space therefore allows to deduce that I_λ is a homeomorphism onto its image. At the end of this section, we explain how the work done in the fourth and final part of the paper implies that I_λ surjects onto $\mathcal{SH}^+(\lambda)$ (Theorem 10.15), and why we choose to prove surjectivity this way. See Remark 10.16 in particular.

Since I_λ essentially yields period coordinates, it is not surprising that (a variant of) Teichmüller geodesic flow is given in coordinates by dilation (Lemma 11.1), while the Teichmüller horocycle flow is translation by λ (Lemma 11.2). We also naturally recover the "tremor deformations" introduced in [Chaika et al. 2020] as translation by measures μ supported on λ that are not necessarily absolutely continuous with respect to λ (Definition 11.3). Figure 17 details a dictionary between the language of [Chaika et al. 2020] and our own.

Sections 12–15: coordinates for hyperbolic structures In the final part of the paper, we use the geometry of the orthogeodesic foliation to coordinatize hyperbolic structures via shear-shape cocycles.

From Theorem 6.4, we know that the combinatorialization of $\mathcal{O}_\lambda(X)$ on each subsurface $S \setminus \lambda$ by a weighted arc system completely encodes the geometry of the pieces. Cutting $X \setminus \lambda$ further along the orthogeodesic realization of each such arc, we obtain a family of (partially ideal) right-angled polygons. The orthogeodesic foliation equips each polygon with a natural family of basepoints, one on each of its sides adjacent to λ, that vary analytically in $\mathcal{T}(S \setminus \lambda)$. We are thus able to define a "shear" parameter between (some pairs of) degenerate polygons, and this shear data assembles together with the "shape" data on each subsurface to give instructions for gluing the polygonal pieces back together to obtain X.

In Section 12 we state the main Theorem 12.1, that the shear-shape coordinate map $\sigma_\lambda : \mathcal{T}(S) \to \mathcal{SH}^+(\lambda)$ is a homeomorphism, supply an outline of its proof, and derive some immediate corollaries. The construction of σ_λ is given in Section 13, where we formalize the discussion from the previous paragraph. We also prove that the central diagram (2) commutes (Theorem 13.13), which then implies that σ_λ takes hyperbolic length to the Thurston intersection form (Corollary 13.14).

Section 14 is the most technical part of the paper. In it, we define the "shape-shifting" cocycles (Proposition 14.26) along which a hyperbolic structure can be deformed (Theorem 15.1); these deformations are generalizations of Thurston's cataclysms or Bonahon's shear deformations. Although the construction of a shape-shifting deformation is rather involved, we attempt to keep the reader informed of the geometric intuition that guides the construction throughout. Finally, in Section 15 we assemble all of the necessary ingredients to prove Theorem 12.1. That the earthquake along λ is given by translation by λ in $\mathcal{SH}^+(\lambda)$ (Corollary 15.2) is an immediate consequence of the construction of shape-shifting deformations as generalizations of cataclysms. We then discuss how the action of dilation in coordinates can sometimes be identified with directed geodesics in Thurston's asymmetric metric (Propositions 15.12 and 15.18).

4 Crowned hyperbolic surfaces

When a hyperbolic surface is cut along a geodesic multicurve, the (completion of the) resulting space is a compact hyperbolic surface with compact, totally geodesic boundary. When the same surface is cut along a geodesic lamination, the (completion of the) complementary subsurface can have noncompact "crowned boundaries". This section collects results about hyperbolic structures on such "crowned surfaces" as well as the relationship between properties of the lamination and the topology of its complementary subsurfaces.

Remark 4.1 Throughout this section and the following, we reserve S to denote a closed surface. If λ is a geodesic lamination, then $S \setminus \lambda$ denotes the metric completion of the complementary subsurfaces to λ (with respect to some auxiliary hyperbolic metric); we will refer to the topological type of a component of $S \setminus \lambda$ by Σ. Hyperbolic metrics on S and Σ will be denoted by X and Y, respectively.

Hyperbolic crowns While less familiar than surfaces with boundary, crowned hyperbolic surfaces naturally arise by uniformizing surfaces with boundary and marked points on the boundary. They are also intricately related to meromorphic differentials on Riemann surfaces with high-order poles (see eg [Gupta 2021]).

A *hyperbolic crown* with c_k spikes is a complete, finite-area hyperbolic surface with geodesic boundary that is homeomorphic to an annulus with c_k points removed from one boundary component. In the hyperbolic metric, the circular boundary component corresponds to a closed geodesic and each interval of the other boundary becomes a bi-infinite geodesic running between ideal vertices; compare Figure 2.

In general, a *hyperbolic surface with crowned boundary* is a complete, finite-area hyperbolic surface with totally geodesic boundary; the boundary components are either compact or hyperbolic crowns. We record the topological type of a crowned surface of genus g with b closed boundary components and k crowns with $c_1, \dots, c_k$ many spikes as $\Sigma_{g,b}^{\{\underline{c}\}}$, where $\{\underline{c}\} = \{c_1, \dots, c_k\}$.

Remark 4.2 Ideal polygons may be considered as crowned surfaces of genus 0 with a single (crowned) boundary component. All of the results in this section hold for both crowned surfaces with nontrivial topology as well as for ideal polygons, but their proofs are slightly different. Our citations of [Gupta 2021] are all for the case when Σ is not an ideal polygon; for the corresponding statements for ideal polygons, see [Gupta 2021, Section 3.3] or [Han et al. 1995].

Every crowned surface Y with noncyclic (and nontrivial) fundamental group contains a "convex core" obtained by cutting off its crowns along a geodesic multicurve [Casson and Bleiler 1988, Lemma 4.4]. When Y has type $\Sigma_{g,b}^{\{\underline{c}\}}$, this core is a subsurface of genus g with $b+k$ closed boundary components. Since each crown with c_i spikes may be decomposed into c_i ideal hyperbolic triangles by introducing leaves wrapping around the totally geodesic boundary component, we have the following expression for the area:

$$\frac{1}{\pi}\operatorname{Area}(Y) = 4g - 4 + 2b + \sum_{i=1}^{k}(c_i + 2). \tag{4}$$

Note that one can triangulate an ideal polygon of c sides into $c-2$ ideal triangles, and so the above formula also holds for ideal polygons.

The metric residue While crown ends (and ideal polygons) do not have well-defined boundary lengths, one can define a natural generalization when there are an even number of spikes. This turns out to be a fundamental invariant that controls when crowns can be glued together along a lamination (Lemma 13.1).

Let $\mathscr{C}$ be a hyperbolic crown or an ideal polygon with c spikes, where c is even. One can then orient $\mathscr{C}$, that is, pick an orientation of the boundary leaves so that the orientations of asymptotic leaves agree. Truncating each spike of $\mathscr{C}$ along a horocycle based at the tip of the spike yields a surface with a boundary made up of horocyclic segments $h_1, \dots, h_c$ and geodesic segments $g_1, \dots, g_c$. See Figure 2.

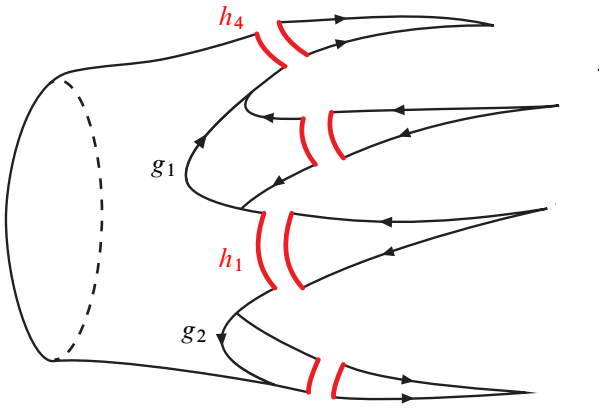


Figure 2: Truncating an (oriented) crown to compute its metric residue.

Definition 4.3 [Gupta 2021, Definition 2.9] Let $\mathscr{C}$ be either an oriented hyperbolic crown or an oriented ideal polygon with an even number of spikes. Then its *metric residue* $\operatorname{res}(\mathscr{C})$ is

$$\operatorname{res}(\mathscr{C}) = \sum_{i=1}^{c} \varepsilon_i \ell(g_i),$$

where ε_i is positive if the truncated crown lies on the left of g_i and negative if it lies on the right.

Since changing the truncation depth of a spike increases the length of two adjacent sides, the metric residue evidently does not depend on the choice of truncation [Gupta 2021, Lemma 2.10]. Observe also that flipping the orientation of $\mathscr{C}$ flips the sign of its metric residue.

Similarly, define the metric residue of an oriented totally geodesic boundary component β of Y to be $\pm\ell(\beta)$, where the sign depends on whether Y lies to the left of β (positive) or right (negative).

Deformation spaces of crowned surfaces We now record some useful facts about the Teichmüller spaces of crowned hyperbolic surfaces.

Given any crowned hyperbolic surface Y, one can obtain a natural compactification $\widehat{Y}$ by adding on an ideal vertex at the end of each spike of each crown. The corresponding (topological) surface $\widehat{\Sigma}_{g,b}^{\{\underline{c}\}}$ then has $b+k$ boundary components with c_i marked points on the $(b+i)^{\text{th}}$ boundary component. A *marking* of a crowned hyperbolic surface Y is a homeomorphism

$$f\colon \widehat{\Sigma}_{g,b}^{\{\underline{c}\}} \to \widehat{Y}$$

which takes boundary marked points to ideal vertices. We think of the boundary marked points as having distinct labels, so different identifications of the boundary points of $\widehat{\Sigma}_{g,b}^{\{\underline{c}\}}$ with the spikes of Y yield different markings. The Teichmüller space of a crowned hyperbolic surface $\Sigma_{g,b}^{\{\underline{c}\}}$ is then defined to be the space of all marked hyperbolic metrics on $\Sigma_{g,b}^{\{\underline{c}\}}$, up to isotopies which fix the totally geodesic boundary components pointwise and fix each ideal vertex of each crown.

As noted above, any crowned hyperbolic surface $\Sigma_{g,b}^{\{\underline{c}\}}$ contains an uncrowned subsurface which serves as its convex core. Therefore, the Teichmüller space of a crowned hyperbolic surface may be parametrized by the Teichmüller space of its convex core together with parameters describing each crown and how it is attached. A precise version of this dimension count is recorded below.

Lemma 4.4 [Gupta 2021, Lemma 2.16] *Let $\Sigma = \Sigma_{g,b}^{\{\underline{c}\}}$ be a crowned hyperbolic surface or an ideal polygon. Then $\mathcal{T}(\Sigma) \cong \mathbb{R}^d$, where*

$$d = 6g - 6 + 3b + \sum_{i=1}^{k}(c_i + 3). \tag{5}$$

Fixing the length of any closed boundary component of $\Sigma_{g,b}^{\{\underline{c}\}}$ cuts out a codimension 1 subvariety of $\mathcal{T}(\Sigma)$. Similarly, the subspace of surfaces with fixed metric residues at an even–spiked crown has codimension one. The following proposition ensures that the intersections of the level sets of length and metric residue are topologically just cells of the proper dimension:

Proposition 4.5 [Gupta 2021, Corollary 2.17] *Let $\Sigma = \Sigma_{g,b}^{\{\underline{c}\}}$ be a crowned surface or an ideal polygon. Let $\beta_1, \dots, \beta_b$ denote the closed boundary components of Σ and let $\mathcal{C}_1, \dots, \mathcal{C}_e$ denote the crown ends which have an even number of spikes. Fix an orientation of each crown end. Then, for any $(L_i) \in \mathbb{R}_{>0}^b$ and any $(R_j) \in \mathbb{R}^e$,*

$$\{(Y, f) \in \mathcal{T}(\Sigma) \mid \ell(\beta_i) = L_i \text{ and } \operatorname{res}(\mathcal{C}_j) = R_j \text{ for all } i, j\} \cong \mathbb{R}^{d-b-e}$$

where d is as in (5).

Topology When a crowned surface Σ comes from cutting a closed surface S along a geodesic lamination λ, we can relate the topology of λ to the topology of Σ.

Recall that the Euler characteristic of a lamination is defined to be alternating sum of the ranks of its Čech cohomology groups, viewing λ as a subset of S. Below, we compute the Euler characteristic of a geodesic lamination in terms of the topological type of its complementary subsurfaces.

Lemma 4.6 *Let λ be a geodesic lamination on S. Then the total number of spikes of $S \setminus \lambda$ equals $-2\chi(\lambda)$.*

We also record the corresponding formula for later use. Suppose that $\overline{S \setminus \lambda} = \Sigma_1 \cup \cdots \cup \Sigma_m$; then

$$\chi(\lambda) = -\frac{1}{2}\sum_{j=1}^{m}\sum_{i=1}^{k_j} c_i^j, \tag{6}$$

where $\{c_1^j, \dots, c_{k_j}^j\}$ denotes the crown type of Σ_j.

Proof Fix some train track τ which carries λ and has the same topological type; in Section 5.2 below, this is referred to as *snug* carrying of λ on τ. Lemma 13 of [Bonahon 1997b] states that, for any such train track, $\chi(\lambda) = \chi(\tau)$, and so it suffices to compute the Euler characteristic of τ.

Splitting the switches of τ if necessary, we may assume that τ is trivalent (observe that this operation preserves the Euler characteristic). Then each spike of $S \setminus \lambda$ corresponds to a unique switch of τ, and each switch corresponds to three half-edges, so

$$\#\text{spikes}(S \setminus \lambda) = \#\text{switches}(\tau) = \tfrac{2}{3} \cdot \#\text{edges}(\tau).$$

Plugging this into the formula $\chi(\tau) = \#\text{switches}(\tau) - \#\text{edges}(\tau)$ proves the claim. □

In general, the relationship between the boundary components of $S \setminus \lambda$ and λ can be rather involved. For example, one can construct a lamination on a closed surface of genus $g \geq 2$ consisting of three leaves, two of which are nonisotopic simple closed curves and one which spirals onto each of the closed leaves. In this scenario, there is not a precise correspondence between closed leaves of λ and totally geodesic boundary components of its complementary subsurface.

Note So that we do not have to deal with possible spiraling behavior of λ, we henceforth restrict our discussion to those laminations that support a measure.

5 The orthogeodesic foliation

In this section we construct the *orthogeodesic foliation* $\mathcal{O}_\lambda(X) \in \mathcal{MF}(\lambda)$ of a hyperbolic surface X with respect to λ and describe some of its basic properties.

5.1 The spine of a hyperbolic surface

We begin by describing the orthogeodesic foliation restricted to subsurfaces Y complementary to λ. Let Y be a finite-area hyperbolic surface with totally geodesic boundary, possibly with crowned boundary. As we are most interested in those Y coming from cutting a closed surface along a lamination, we also assume that Y has no annular cusps.

Definition 5.1 The orthogeodesic foliation $\mathcal{O}_{\partial Y}(Y)$ of Y is the (singular, piecewise-geodesic) foliation of Y whose leaves are fibers of the closest-point projection to ∂Y.

Near ∂Y, the leaves of $\mathcal{O}_{\partial Y}(Y)$ are geodesic arcs meeting ∂Y orthogonally. To understand the global structure of the foliation, however, we need to determine how the leaves extend into the interior of Y. In particular, we must understand the locus of points that are closest to multiple points of ∂Y.

To that end, for any point $x \in Y$, define the *valence* of x to be

$$\text{val}(x) := \#\{y \in \partial Y : d(x, y) = d(x, \partial Y)\}.$$

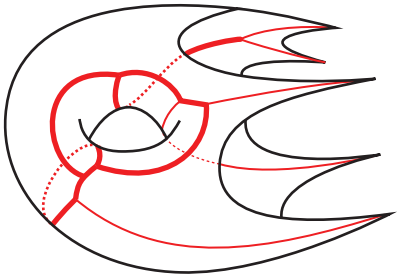


Figure 3: The spine of a hyperbolic surface with crowned boundary. Note that the finite core Sp^0 (represented in bold) contains a spine for the convex core of the surface.

The (*geometric*) *spine* $\mathsf{Sp}(Y)$ of Y is the set of points of Y with valence at least 2, and has a natural partition into subsets $\mathsf{Sp}_k(Y)$, where $x \in \mathsf{Sp}_k(Y)$ if it is equidistant from exactly k points in ∂Y. For the rest of the section we fix a hyperbolic surface Y and refer to $\mathsf{Sp}(Y)$ and $\mathsf{Sp}_k(Y)$ simply as Sp and Sp_k.

It is not hard to see that Sp is a properly embedded, piecewise-geodesic 1–complex with some nodes of valence 1 removed (equivalently, a ribbon graph with some half-infinite edges). Indeed, Sp decomposes into a *finite core* Sp^0 and a finite collection of open geodesic rays; since we assumed Y had no annular cusps, each ray corresponds with a spike of a crowned boundary component. See [Mondello 2009b, Section 2] for a discussion of the structure of the spine of a compact hyperbolic surface with geodesic boundary in which $\mathsf{Sp}^0 = \mathsf{Sp}$.

We record below a summary of this discussion; see also Figure 3.

Lemma 5.2 *The finite core Sp^0 is a piecewise–geodesically embedded graph, whose edges correspond to the components of Sp_2 with finite hyperbolic length and vertex set $\bigcup_{k\geq 3} \mathsf{Sp}_k$. Each geodesic ray of $\mathsf{Sp} \setminus \mathsf{Sp}^0$ exits a unique spike of Y.*

By definition, the orthogeodesic foliation $\mathcal{O}_{\partial Y}(Y)$ has k–pronged singular leaves emanating from $\bigcup_{k\geq 3} \mathsf{Sp}_k$ for $k \geq 3$. The nonsingular leaves of $\mathcal{O}_{\partial Y}(Y)$ glue along $\mathsf{Sp}_2(Y)$ (usually at an angle) and can be smoothed by an arbitrarily small isotopy supported near Sp. As the geometry of Sp interacts nicely with the leaves of $\mathcal{O}_{\partial Y}(Y)$, we generally prefer to think about $\mathcal{O}_{\partial Y}(Y)$ as a piecewise-geodesic singular foliation rather than as a smooth one. When convenient, we will pass freely between the orthogeodesic foliation and a smoothing.

We observe that there is also an isotopy supported in the ends of the spikes of Y and restricting to the identity on ∂Y that maps leaves of the orthogeodesic foliation to horocycles based at the tip of the spike. This equivalence between the orthogeodesic and horocyclic foliations in spikes is of vital importance in Sections 13–15 as it allows us to adapt many of Bonahon and Thurston's arguments to this setting.

Remark 5.3 One can check that, for regular ideal polygons, the isotopy in spikes extends to a global isotopy between the orthogeodesic foliation and the symmetric partial foliation by horocycles.

Following the leaves of the orthogeodesic foliation in the direction of Sp defines a deformation retraction of Y onto Sp; let $r : Y \to \mathsf{Sp}$ be the map fully collapsing Y onto Sp. For x and y in the same component of Sp_2, the leaves $r^{-1}(x)$ and $r^{-1}(y)$ of $\mathcal{O}_{\partial Y}(Y)$ are properly isotopic. We may therefore associate to each edge e of Sp_2 the (proper) isotopy class of $r^{-1}(x)$ for $x \in e$; we call this the *dual arc* α_e to e.

There is a distinguished representative of α_e that is geodesic and orthogonal to both ∂Y and e; compare Figure 7. By abuse of notation, we henceforth identify α_e with its orthogeodesic representative and define

$$\underline{\alpha}(Y) := \bigcup_{e \subset \mathsf{Sp}_2^0} \alpha_e.$$

Lemma 5.4 *The metric completion of the surface with corners $Y \setminus \underline{\alpha}(Y)$ is homeomorphic to a union of closed disks and closed disks with finitely many points on the boundary removed. That is, $\underline{\alpha}(Y)$ fills Y.*

Proof Each component of $Y \setminus \underline{\alpha}(Y)$ deformation retracts onto a component of the metric completion of $\mathsf{Sp} \setminus \underline{\alpha}(Y)$. By the duality of arcs and edges of Sp_2^0, each component of $\mathsf{Sp} \setminus \underline{\alpha}(Y)$ is contractible. □

The orthogeodesic foliation also comes with a natural transverse measure: the measure of an arc k transverse to (a smoothing of) $\mathcal{O}_{\partial Y}(Y)$ is defined on small enough transverse arcs k first by isotoping the arc into ∂Y transversely to $\mathcal{O}_{\partial Y}(Y)$ and then measuring the hyperbolic length there. Locally, the orthogeodesic foliation admits a reflection about each edge of Sp, so by restricting k to those leaves of $\mathcal{O}_{\partial Y}(Y)$ that intersect a given edge, we can use this symmetry to see that the measure of k is the same after a transverse isotopy onto either boundary component of Y. Extending to all transverse arcs by additivity defines a transverse measure on $\mathcal{O}_{\partial Y}(Y)$.

To each component e of Sp_2^0 we associate the length $c_e > 0$ of either component of $r^{-1}(e) \cap \partial Y$; the transverse measure of e is exactly c_e. Anticipating the contents of the next section (see eg Theorem 6.4), we define the formal sum

$$\underline{A}(Y) := \sum_{e \subset \mathsf{Sp}_2^0} c_e \alpha_e. \tag{7}$$

5.2 The orthogeodesic foliation

Now that we have described the orthogeodesic foliation on each component of $S \setminus \lambda$, we can glue these pieces together along the leaves of λ to get a foliation of S.

Construction 5.5 Let $X \in \mathcal{T}(S)$ and λ be a geodesic lamination on X. Cutting X open along λ taking the metric completion of each component, we obtain a union of hyperbolic surfaces with totally geodesic boundary (possibly with crowned boundary). On each such component Y, we construct the orthogeodesic foliation $\mathcal{O}_{\partial Y}(Y)$ as described in Section 5.1 above.

A standard fact from hyperbolic geometry [Canary et al. 2006, Lemma 5.2.6] shows that the line field defined by (a smoothing of) the orthogeodesic foliation forms a Lipschitz line field on $X \setminus \lambda$. Since λ has measure 0, this line field is integrable near λ, so the partial foliation defined on $X \setminus \lambda$ extends across the leaves of λ. This defines a measured foliation $\mathcal{O}_\lambda(X) \in \mathcal{MF}(S)$, and hence a map $\mathcal{O}_\lambda : \mathcal{T}(S) \to \mathcal{MF}(S)$.

Later, we prove in Lemma 5.8 that λ and $\mathcal{O}_\lambda(X)$ bind, allowing us to restrict the codomain of $\mathcal{O}_\lambda$ to $\mathcal{MF}(\lambda)$. Ultimately, our goal is to show that $\mathcal{O}_\lambda$ is a homeomorphism onto $\mathcal{MF}(\lambda)$.

Geometric train tracks We now consider the geometry of $\mathcal{O}_\lambda(X)$ in a neighborhood of λ. The following is a modification of an important construction of Thurston [1979, Chapter 8.9]:

Construction 5.6 Let $\epsilon > 0$ be small enough that the ϵ–neighborhood $\mathcal{N}_\epsilon(\lambda)$ is topologically stable. The orthogeodesic foliation $\mathcal{O}_\lambda(X)$ restricts to a foliation of $\mathcal{N}_\epsilon(\lambda)$ without singular points, and collapsing the leaves yields a quotient map $\pi : \mathcal{N}_\epsilon(\lambda) \to \tau$ where τ can be C^1–embedded in $\mathcal{N}_\epsilon(\lambda)$ as a train track carrying λ in X. By changing ϵ, we may assume that τ is trivalent.[6] Then $\tau = \tau(\lambda, X, \epsilon)$ is a *geometric train track*.

We sometimes refer to $\mathcal{N}_\epsilon(\lambda)$ as a *train track neighborhood* of λ and the leaves of $\mathcal{O}_\lambda(X)|_{\mathcal{N}_\epsilon(\lambda)}$ as *ties*. A train track neighborhood coming from Construction 5.6 is a union of bands and annuli foliated by ties glued together along the ties that collapse to switches of τ. We recall that, if λ meets every tie of τ and there is no path between spikes of $S \setminus \mathcal{N}_\epsilon(\lambda)$ that is contained in $\mathcal{N}_\epsilon(\lambda) \setminus \lambda$, then τ is said to *snugly* carry λ. Equivalently, τ snugly carries λ if and only if $S \setminus \lambda$ and $S \setminus \tau$ have the same topological type. With this definition, it is clear that the geometric train tracks constructed above always carry λ snugly.

Using the geometry of $\pi : \mathcal{N}_\epsilon(\lambda) \to \tau$, the branches of τ admit a well-defined notion of length. Indeed, let $b \subset \tau$ be a branch, and choose a lift $\tilde{b}$ to the universal cover $\widetilde{X}$. Let $\ell, \ell' \subset \widetilde{\lambda}$ be leaves of the elevation $\widetilde{\lambda}$ of λ to $\widetilde{X}$ that meet $\pi^{-1}(\tilde{b}) \subset \mathcal{N}_\epsilon(\widetilde{\lambda})$ in segments g and g'. Since $\mathcal{O}_\lambda(X)$ is equivalent to a horocyclic foliation in $\mathcal{N}_\epsilon(\lambda)$, transporting g along the leaves of $\mathcal{O}_{\widetilde{\lambda}}(\widetilde{X})$ near $\tilde{b}$ onto g' is isometric, so $\ell_X(g) = \ell_X(g')$. We may therefore define the *length* of b (along λ) as

$$\ell_X(b) := \ell_X(g)$$

for any g as above. Similarly, for any branch $b \subset \tau$, the ties of $\mathcal{N}_\epsilon(\lambda)$ collapsing to b all have the same integral with respect to λ. Define

$$\lambda(b) := \lambda(k)$$

for any tie $k \subset \mathcal{O}_\lambda(X)|_{\pi^{-1}(b)}$; this is equivalently the weight deposited by λ on b in its τ train track coordinates.

Lemma 5.7 *For any hyperbolic structure X and any measure λ' on λ, we have $i(\lambda', \mathcal{O}_\lambda(X)) = \ell_X(\lambda')$.*

Proof Using Construction 5.6, find a geometric train track $\pi : \mathcal{N}_\epsilon(\lambda) \to \tau$ snugly carrying λ on X. By definition, the intersection pairing is given by the integral over X of the product measure $d\lambda' \otimes d\mathcal{O}_\lambda(X)$,

[6] In the literature, trivalent train tracks are also called "generic".

whose support is contained entirely in the train track neighborhood $\mathcal{N}_\epsilon(\lambda)$. For each branch $b \subset \tau$, the integral of this measure on $\pi^{-1}(b)$ is just $\lambda'(b)\ell_X(b)$, so

$$\begin{aligned} i(\lambda', \mathbb{O}_\lambda(X)) &= \iint_X d\lambda' \otimes d\mathbb{O}_\lambda(X) = \iint_{\mathcal{N}_\epsilon(\lambda)} d\lambda' \otimes d\mathbb{O}_\lambda(X) \\ &= \sum_{b\subset\tau} \iint_{\pi^{-1}(b)} d\lambda' \otimes d\mathbb{O}_\lambda(X) = \sum_{b\subset\tau} \lambda'(b)\ell_X(b). \end{aligned}$$

On the other hand, $\ell_X(\lambda')$ is the integral over X of the measure $d\lambda' \otimes dl_{\lambda'}$, locally the product of the transverse measure λ' and 1–dimensional Lebesgue measure $l_{\lambda'}$ on the support of λ'. Since λ' is supported in λ, the integral of $d\lambda' \otimes dl_{\lambda'}$ is equal to the integral of $d\lambda' \otimes dl_\lambda$, and again the support of the product measure is contained in $\mathcal{N}_\epsilon(\lambda)$. On each thickened branch $\pi^{-1}(b) \subset \mathcal{N}_\epsilon(\lambda)$, the integral of $d\lambda' \otimes dl_\lambda$ is $\lambda'(b)\ell_X(b)$, giving

$$\ell_X(\lambda') = \sum_{b\subset\tau} \lambda'(b)\ell_X(b). \qquad \square$$

With this computation, we can now show that λ and $\mathbb{O}_\lambda(X)$ together bind S.

Lemma 5.8 *For any $X \in \mathcal{T}(S)$ and $\lambda \in \mathcal{ML}(S)$, we have $\mathbb{O}_\lambda(X) \in \mathcal{MF}(\lambda)$.*

Proof Suppose that η is an measured lamination such that $i(\eta, \lambda) = 0$; without loss of generality, we may assume that η is ergodic. Then one of two things must be true: either η is supported on λ or its support is disjoint from λ. In the first case, $i(\eta, \mathbb{O}_\lambda(X)) = \ell_X(\eta) > 0$ by Lemma 5.7.

If η is disjoint from λ then it is contained in a component Y of $\overline{X \setminus \lambda}$, and we need only show that $i(\eta, \mathbb{O}_\lambda(X)) > 0$. Scaling the measure of η as necessary, let us assume that $\ell_X(\eta) = \ell_Y(\eta) = 1$. Now we recall that the set of weighted simple closed curves is dense in the space of measured laminations on Y. By homogeneity and continuity of the intersection pairing, it therefore suffices to find some uniform $\epsilon > 0$ such that

$$i(\gamma, \mathbb{O}_\lambda(X)) \geq \epsilon \ell_X(\gamma)$$

for every simple closed curve $\gamma \subset Y$. Indeed, once we have demonstrated such a bound we may approximate η arbitrarily well by weighted curves $\gamma/\ell_X(\gamma)$ to deduce the desired bound on $i(\eta, \mathbb{O}_\lambda(X))$.

So let Y_0 be the convex hull of $r^{-1}(\mathsf{Sp}^0)$; Y_0 is compact and the inclusion of Y_0 into Y is a homotopy equivalence. Any simple closed geodesic γ in Y is contained in Y_0, and, since Y deformation retracts onto the component of Sp contained in Y, γ is homotopic to a concatenation of edges in Sp^0.

Give Sp^0 a metric making its edges e have length $c_e = i(e, \mathbb{O}_\lambda(X))$; then the inclusion $\mathsf{Sp}^0 \to Y_0$ with this metric induces an equivariant quasi-isometry on universal covers (this follows because they are both Gromov hyperbolic and $\pi_1(Y)$ acts cocompactly and properly discontinuous on each). The geodesic lengths of closed curves in Sp and in Y_0 are therefore comparable, so there is some $\epsilon > 0$ such that

$$i(\gamma, \mathbb{O}_\lambda(X)) = \ell_{\mathsf{Sp}^0}(\gamma) \geq \ell_X(\gamma)\epsilon,$$

demonstrating the desired uniform bound. $\square$

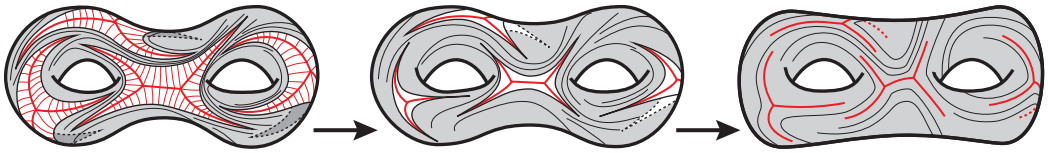


Figure 4: Inflating a lamination and deflating its complementary components.

5.3 Deflation

For a given pair $(X,\lambda)\in\mathcal{T}(S)\times\mathcal{ML}(S)$, the pair of laminations $\mathcal{O}_\lambda(X)$ and λ bind by Lemma 5.8. By Theorem 2.1, there is a unique quadratic differential $q=q(\mathcal{O}_\lambda(X),\lambda)$, holomorphic on some Riemann surface Z, whose real and imaginary foliations are $\mathcal{O}_\lambda(X)$ and λ, respectively. In this section we define a *deflation* map $\mathcal{D}_\lambda\colon X\to Z$ that allows us to make direct comparisons between the hyperbolic geometry of X and the singular flat geometry q. This discussion is further expanded on in [Calderon and Farre 2024a, Section 8].

An informal description of $\mathcal{D}_\lambda$ is that it "deflates" the subsurfaces of $X\setminus\lambda$, retracting them to Sp along the leaves of $\mathcal{O}_\lambda(X)$, while it "inflates" along the leaves of λ according to the transverse measure. The orthogeodesic foliation in a neighborhood of λ assembles into the vertical foliation of the resulting quadratic differential metric and $\mathcal{D}_\lambda$ maps $\mathsf{Sp}\subset X$ to the horizontal separatrices; compare Figure 4.

Remark 5.9 This heuristic description of $\mathcal{D}_\lambda$ can be made precise by grafting X along λ (see eg [Dumas 2009]) and then collapsing the hyperbolic pieces along the leaves of $\mathcal{O}_\lambda(X)$. In particular, $\mathcal{D}_\lambda$ is *not* the grafting map.

Proposition 5.10 *Given a marked hyperbolic structure*[7] $[f\colon S\to X]\in\mathcal{T}(S)$ *and* $\lambda\in\mathcal{ML}(S)$, *let* $[g\colon S\to Z]\in\mathcal{T}(S)$ *be the marked complex structure on which* $q(\mathcal{O}_\lambda(X),\lambda)$ *is holomorphic. There is a map*

$$\mathcal{D}_\lambda\colon X\to Z$$

that is a homotopy equivalence restricting to an isometry between Sp^0 *with its metric induced by integrating the edges against* $\mathcal{O}_\lambda(X)$ *and the graph of horizontal saddle connections of* $q(\mathcal{O}_\lambda(X),\lambda)$ *with the induced path metric. Moreover,* $\mathcal{D}_\lambda\circ f\sim g$ *and* $\mathcal{D}_{\lambda*}\mathcal{O}_\lambda(X)=\mathrm{Re}(q)$ *and* $\mathcal{D}_{\lambda*}\lambda=\mathrm{Im}(q)$ *as measured foliations.*

Proof Construction 5.6 supplies us with a geometric train track $\pi\colon\mathcal{N}_\epsilon(\lambda)\to\tau$. On the preimage $\pi^{-1}(b)$ of each closed branch b of τ we integrate the two measures $\mathcal{O}_\lambda(X)|_{\mathcal{N}_\epsilon(\lambda)}$ and λ giving $\pi^{-1}(b)$ the structure of a bifoliated Euclidean rectangle of length $\ell_X(b)$ and height $\lambda(b)$ These rectangles glue along their "short" sides $\{\pi^{-1}(s)\colon s \text{ is a switch of } \tau\}$ to give $\mathcal{N}_\epsilon(\lambda)$ the structure of a bifoliated Euclidean band complex.

[7]Throughout the paper we suppress markings in our notation, but reintroduce them here to state the proposition precisely.

The map π extends to a self–homotopy equivalence of X homotopic to the identity preserving the orthogeodesic foliation leafwise. This means that the boundary of $\mathcal{N}_\epsilon(\lambda)$ admits a natural retraction onto Sp by collapsing the leaves of the orthogeodesic foliation in the complement of $\mathcal{N}_\epsilon(\lambda)$, and we take the quotient generated by this equivalence relation to obtain a new surface Y with its complex structure described below.

On each rectangle $\pi^{-1}(b)$, the bifoliated Euclidean structure gives local coordinates to $\mathbb{C}$ away from the singular points of $\mathcal{O}_\lambda(X)$ locally mapping $\mathcal{O}_\lambda(X)$ to $|dx|$ and λ to $|dy|$, thought of as measured foliations on the plane. These coordinate patches glue together along the spine to give local coordinates away from the points of valence ≥ 3. Moreover, these charts preserve $|dx|$ and $|dy|$, so the transitions must be of the form $z \mapsto \pm z + \alpha$ for some $\alpha \in \mathbb{C}$. We have therefore built a Riemann surface Z equipped with a half-translation structure away from the vertices of Sp, which become cone points of cone angle equal to $\pi \cdot \mathrm{val}(v)$. Edges of Sp join vertices along horizontal trajectories representing all horizontal saddle connections on q; their lengths in the singular flat metric are given by the integral over $\mathcal{O}_\lambda(X)$. Thus $\mathcal{D}_\lambda$ induces an isometry of metric graphs, as claimed. □

This explicit description of the quadratic differential associated to the pair (X, λ) by the map $\mathcal{O}$ from the introduction will be useful in order to prove in Theorem 13.13 that (2) commutes.

6 Cellulating crowned Teichmüller spaces

We now define a certain arc complex which combinatorializes the structure the orthogeodesic foliation on complementary subsurfaces. The main result of this section is Theorem 6.4, which shows that this arc complex is equivariantly homeomorphic to the Teichmüller space of the complementary surface. In particular, this shows that the restriction of the orthogeodesic foliation to each component of $S \setminus \lambda$ completely determines the hyperbolic structure on that piece.

Before stating the theorem, we must first set up our combinatorial analogue for Teichmüller space. This appears as Definition 6.1 after a series of auxiliary constructions.

Suppose that $\Sigma = \Sigma_{g,b}^{\{c\}}$ is a finite-area hyperbolic surface with boundary and without annular cusps. A properly embedded arc $I \to \Sigma$ is *essential* if I cannot be isotoped (through properly embedded arcs) into $\partial\Sigma$ or into a spike. The *arc complex* $\mathscr{A}(\Sigma, \partial\Sigma)$ *of* Σ *rel boundary* is the (simplicial, flag) complex whose vertices are isotopy classes of simple essential arcs of Σ. Vertices span a simplex in $\mathscr{A}(\Sigma, \partial\Sigma)$ if and only if there exists a collection of pairwise disjoint representatives for each isotopy class. The *filling arc complex* $\mathscr{A}_{\mathrm{fill}}(\Sigma, \partial\Sigma)$ is the subset of $\mathscr{A}(\Sigma, \partial\Sigma)$ consisting only of those arc systems which cut Σ into a union of topological disks.

The geometric realization $|\mathscr{A}(\Sigma, \partial\Sigma)|$ of $\mathscr{A}(\Sigma, \partial\Sigma)$ is obtained by declaring every simplex to be a regular Euclidean simplex of the proper dimension; note that the topology of $|\mathscr{A}(\Sigma, \partial\Sigma)|$ obtained from the

metric structure is in general different from the standard simplicial topology (see eg [Bowditch and Epstein 1988]). The geometric realization $|\mathscr{A}_{\text{fill}}(\Sigma, \partial\Sigma)|$ is then the subspace of filling arc systems equipped with the subspace topology induced by the metric structure.

Definition 6.1 The *weighted filling arc complex* $|\mathscr{A}_{\text{fill}}(\Sigma, \partial\Sigma)|_{\mathbb{R}}$ *of* Σ *rel boundary* is the set of all weighted multiarcs of the form

$$\underline{A} = \sum c_i \alpha_i,$$

where $\underline{\alpha} = \bigcup \alpha_i \in \mathscr{A}_{\text{fill}}(\Sigma, \partial\Sigma)$ and $c_i > 0$ for all i.

Throughout, we will use α to denote a single arc, and $\underline{\alpha}$ to denote an (unweighted) multiarc. The symbol $\underline{A}$ will be reserved to denote a weighted multiarc.

Note If Σ is an ideal hyperbolic polygon, then the empty arc system fills Σ and we consider it as an element of $|\mathscr{A}_{\text{fill}}(\Sigma, \partial\Sigma)|_{\mathbb{R}}$. If Σ is not a polygon, then the empty arc system never fills.

So long as Σ is not an ideal polygon, $|\mathscr{A}_{\text{fill}}(\Sigma, \partial\Sigma)|_{\mathbb{R}}$ is just $|\mathscr{A}_{\text{fill}}(\Sigma, \partial\Sigma)| \times \mathbb{R}_{>0}$. When Σ is an ideal polygon, $|\mathscr{A}_{\text{fill}}(\Sigma, \partial\Sigma)|_{\mathbb{R}}$ is homeomorphic to the open cone on the filling arc complex

$$\big(|\mathscr{A}_{\text{fill}}(\Sigma, \partial\Sigma)| \times \mathbb{R}_{\geq 0}\big) / \big(|\mathscr{A}_{\text{fill}}(\Sigma, \partial\Sigma)| \times \{0\}\big).$$

See Figure 5, left, for an example of $|\mathscr{A}_{\text{fill}}(\Sigma, \partial\Sigma)|_{\mathbb{R}}$ in the case when Σ is an ideal pentagon.

Remark 6.2 The standard duality between arc systems and ribbon graphs (see eg [Mondello 2009a]) assigns to every $\underline{A} \in |\mathscr{A}_{\text{fill}}(\Sigma, \partial\Sigma)|_{\mathbb{R}}$ a metric ribbon graph spine for Σ (with some infinitely long edges if Σ has crowns). One could of course translate the cell structure of $|\mathscr{A}_{\text{fill}}(\Sigma, \partial\Sigma)|_{\mathbb{R}}$ into a cellulation of an appropriate space of marked metric ribbon graphs.

While the arc complex definition is more practical for our definition of shear-shape cocycles, the dual ribbon graph picture allows us to immediately understand how to record the geometry of the horizontal trajectories of a quadratic differential (see Section 10).

Combinatorial geometry Now that we have defined our combinatorial analogue of Teichmüller space, we can also define combinatorial notions of both length and metric residue.

Suppose that β is a compact boundary component of Σ and $\underline{A} \in |\mathscr{A}_{\text{fill}}(\Sigma, \partial\Sigma)|_{\mathbb{R}}$; then we define the $\underline{A}$–length $\ell_{\underline{A}}(\beta)$ of β to be the sum of the weights of the arcs of $\underline{A}$ incident to β (counted with multiplicity, so that, if both endpoints of $\underline{\alpha}$ lie on β, then its weight is counted twice).

Similarly, let $\mathscr{C}$ be an oriented crowned boundary component with an even number of spikes. Then the edges of $\mathscr{C}$ are partitioned into those that have the surface lying on their left and those which have the surface on their right; call these edges positively and negatively oriented, respectively. The $\underline{A}$–residue

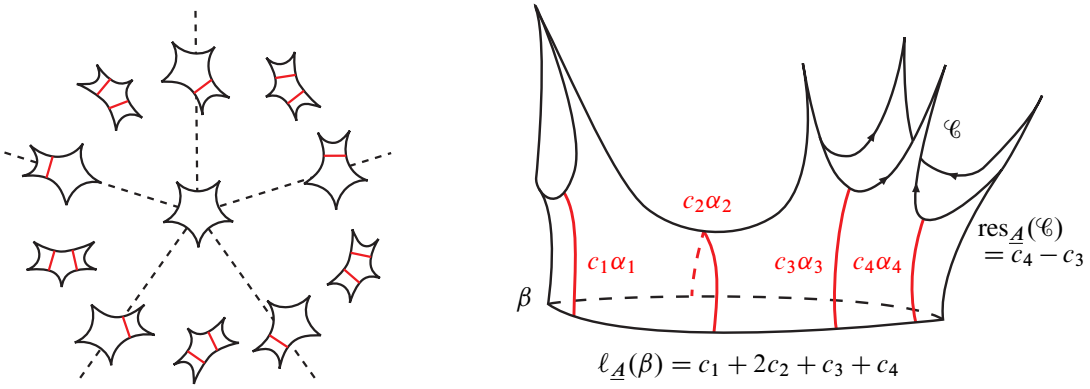


Figure 5: Arc complexes and combinatorial geometry. Left: the weighted arc complex of an ideal pentagon rel its boundary. Right: the combinatorial length and residue associated to a weighted filling arc system $\underline{A}$.

$\mathrm{res}_{\underline{A}}(\mathscr{C})$ of $\mathscr{C}$ is then defined to be the sum of the weights of the arcs incident to each positively oriented edge of $\mathscr{C}$ minus the sum of the weights of the arcs incident to the negatively oriented edges (where both sums are again taken with multiplicity). See Figure 5, right, for an example calculation.

We have now come to the most important object of this section, and a foundational result of this paper that allows us to pass between hyperbolic metrics, orthogeodesic foliations and metric graphs embedded in flat structures.

Construction 6.3 Let Y be a crowned hyperbolic surface. As discussed in Section 5.1, the orthogeodesic foliation determines a spine for Y together with a dual (filling) arc system $\underline{\alpha}(Y)$. Weighting each dual arc by integrating the measure induced by $\mathbb{O}_{\partial Y}(Y)$ over the corresponding edge of Sp (compare (7)) therefore defines a map

$$\underline{A}\colon \mathcal{T}(\Sigma) \to |\mathscr{A}_{\mathrm{fill}}(\Sigma, \partial\Sigma)|_{\mathbb{R}}.$$

When Σ has compact boundary, Luo [2007, Theorem 1.2 and Corollary 1.4] states that $\underline{A}(\,\cdot\,)$ is a $\mathrm{Mod}(\Sigma)$–equivariant stratified real-analytic homeomorphism; see also [Mondello 2009b; Do 2008; Ushijima 1999]. Our aim is to generalize Luo's theorem to surfaces with crowned boundary. While the arguments of [Luo 2007] can probably be adapted to this setting, we prefer to use some elementary hyperbolic geometry

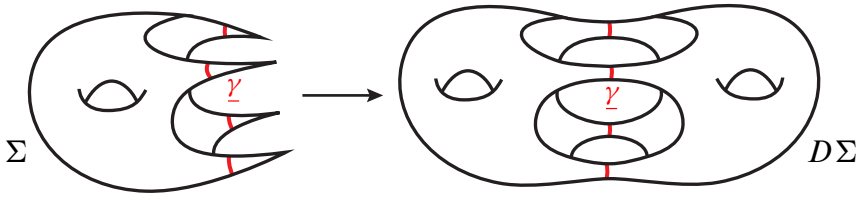


Figure 6: The truncation of a crowned surface Σ along $\underline{\gamma}$ and its double $D\Sigma$.

to realize $|\mathscr{A}_{\mathrm{fill}}(\Sigma, \partial\Sigma)|_{\mathbb{R}}$ as a subcomplex sitting "at infinity" of the weighted filling arc complex of a surface with compact boundary.

Theorem 6.4 *Let Σ be a crowned hyperbolic surface. Then the map*

$$\underline{A}\colon \mathcal{T}(\Sigma) \to |\mathscr{A}_{\mathrm{fill}}(\Sigma, \partial\Sigma)|_{\mathbb{R}}$$

is a $\mathrm{Mod}(\Sigma)$*–equivariant stratified real analytic homeomorphism. Moreover, let $\beta_1, \dots, \beta_b$ denote the closed boundary components of Σ and $\mathscr{C}_1, \dots, \mathscr{C}_e$ the crown ends which have an even number of spikes. Fix an orientation of each $\mathscr{C}_j$. Then the map above identifies the level sets*

$$\{(Y, f) \in \mathcal{T}(\Sigma) \mid \ell(\beta_i) = L_i, \mathrm{res}(\mathscr{C}_j) = R_j\} \cong \{\underline{A} \in |\mathscr{A}_{\mathrm{fill}}(\Sigma, \partial\Sigma)|_{\mathbb{R}} : \ell_{\underline{A}}(\beta_i) = L_i, \mathrm{res}_{\underline{A}}(\mathscr{C}_j) = R_j\}$$

for any $(L_i) \in \mathbb{R}^b_{>0}$ and any $(R_j) \in \mathbb{R}^e$.

The remainder of this section is devoted to deducing Theorem 6.4 from [Luo 2007, Theorem 1.2 and Corollary 1.4; Mondello 2009b, Section 2.4]. Our plan is to appeal to the aforementioned references to prove that, for a given maximal arc system $\underline{\alpha}$, the map $\underline{A}(\cdot)$ extends to a real analytic map $\underline{A}_{\underline{\alpha}}\colon \mathcal{T}(\Sigma) \to \mathbb{R}^{\underline{\alpha}}$ that agrees with $\underline{A}(\cdot)$ on the locus of hyperbolic surfaces whose spine has dual arc system contained in $\underline{\alpha}$ (Lemma 6.9). We show that $\underline{A}(\cdot)$ is a homeomorphism by building a continuous right inverse $Y\colon |\mathscr{A}_{\mathrm{fill}}(\Sigma, \partial\Sigma)|_{\mathbb{R}} \to \mathcal{T}(\Sigma)$; $Y(\underline{A})$ is obtained as a geometric limit metric on a larger compact surface with boundary as some arcs are pinched to spikes.

Endowing Σ with an auxiliary hyperbolic metric, we take Σ° to be the surface with geodesic and horocyclic boundary components obtained by truncating the tips of the spikes. Let $\underline{\gamma}$ be the union of horocyclic boundary components of Σ° and double Σ° along $\underline{\gamma}$ to obtain a (topological) surface $D\Sigma$ and an identification of Σ° with a subsurface of $D\Sigma$ taking $\partial\Sigma^\circ \setminus \underline{\gamma}$ into $\partial D\Sigma$; see Figure 6.

Let $\underline{A} = \sum c_i\alpha_i$ be a weighted filling arc system on Σ and let $\underline{\beta}$ be the mirror image of $\underline{\alpha}$ in $D\Sigma$, so that $\underline{\alpha} \cup \underline{\gamma} \cup \underline{\beta}$ is a filling arc system on $D\Sigma$. For each $t > 0$, define

$$\underline{B}_t = \sum c_i\beta_i + t\sum \gamma_i + \sum c_i\alpha_i \in |\mathscr{A}_{\mathrm{fill}}(D\Sigma, \partial D\Sigma)|_{\mathbb{R}}.$$

Since $D\Sigma$ is compact, we can apply [Luo 2007, Corollary 1.4], which states that there is a unique hyperbolic structure $X_t \in \mathcal{T}(D\Sigma)$ whose natural weighted arc system coincides with $\underline{B}_t$.

Remark 6.5 It will be convenient to assume that $\underline{\alpha}$ is maximal, formally adding arcs of weight 0 to $\underline{A}$ (and $\underline{B}_t$) as necessary.

Our goal is now to show that (X_t) converges as $t \to \infty$ to a surface $Y \in \mathcal{T}(\Sigma)$ such that $\underline{A}(Y) = \underline{A}$. The convergence is geometric: we take basepoints $x_t \in X_t$ lying outside of the "thin parts" of the subsurface corresponding to Σ° and extract a geometric limit of (X_t, x_t) as $t \to \infty$. The limit metric Y has spikes corresponding to $\underline{\gamma}$ and so defines a point in $\mathcal{T}(\Sigma)$. Moreover, Y inherits a filling arc system naturally identified with $\underline{\alpha}$, which is necessarily realized orthogeodesically.

We begin with an estimate on the lengths of orthogeodesic arcs.

Lemma 6.6 *If X is a hyperbolic metric on a compact surface with totally geodesic boundary and $\underline{A}(X) = \sum c_i \alpha_i$, then*

$$\min\left\{\log 3, 2\tanh^{-1}\left(\frac{\tanh(\log\sqrt{3})}{\cosh\left(\frac{1}{2}c_i\right)}\right)\right\} \leq \ell_X(\alpha_i) \leq \frac{2\pi}{c_i},$$

for each i.

Proof Any leaf of the orthogeodesic foliation properly homotopic to α_i has hyperbolic length at least $\ell_X(\alpha_i)$. Thus the embedded "collar" about α_i consisting of all leaves of the orthogeodesic foliation in the same homotopy class of α_i has area at least $c_i \ell_X(\alpha_i)$ (see Figure 7). On the other hand, the Gauss–Bonnet theorem bounds area of the collar above by 2π, so we get the bound

$$\ell_X(\alpha_i) \leq \frac{2\pi}{c_i}.$$

Now we would like to find a lower bound for $\ell_X(\alpha_i)$ in terms of c_i; for notational convenience we fix i and set $\alpha = \alpha_i$ and $c = c_i$. Assume that $\ell_X(\alpha) < \log 3$. Let H be a component of $X \setminus \underline{\alpha}$ meeting α; then there is a unique point $u \in H$ equidistant from all boundary components of X meeting H. There is also a universal lower bound to the distance from u to any such boundary component, given by $\log\sqrt{3}$, the radius of the circle inscribed in an ideal triangle. Thus the leaf of $\mathcal{O}_{\partial X}(X)$ through u has length at least $\log(3)$.

Since $\ell_X(\alpha) < \log 3$, there is a leaf of the orthogeodesic foliation parallel to α with length $\log 3$. Using a formula relating the lengths of the sides of a hyperbolic trirectangle [Buser 1992, Theorem 2.3.1], the distance c_0 from α and this leaf is given by

$$\tanh\left(\tfrac{1}{2}\ell_X(\alpha)\right) = \frac{\tanh(\log\sqrt{3})}{\cosh(c_0)}. \tag{8}$$

Now this expression is decreasing in c_0, and $x \mapsto \tanh^{-1}(x)$ is increasing. We have that $c > 2c_0$ by definition (see Figure 7), so the lemma follows. □

For any arc γ_i of $\underline{\gamma}$, some elementary estimates similar to those given in the proof of Lemma 6.6 (compare (8)) give $\ell_t(\gamma_i) = O(e^{-t/2})$. If α_i appears in $\underline{B}_t$ with coefficient $c_i = 0$, then Lemma 6.6 provides a lower bound of $\log 3$ for the length $\ell_t(\alpha_i)$ of α_i on X_t. We also have the following upper bound:

Lemma 6.7 *If $c_j = 0$ for some j, then, for t large enough,*

$$\log 3 \leq \ell_t(\alpha_j) \leq 2\sum c_i + 8\pi \sum \frac{1}{c_i} + |\underline{\gamma}| \log 144.$$

Proof We remove all arcs of $\underline{\alpha} \cup \underline{\gamma} \cup \underline{\beta}$ with positive weight from X_t and let H_t be (the metric completion of) the right-angled polygon component that contains α_j. Our strategy is to find a path of controlled length contained in ∂H_t joining the endpoints of α_j.

Notice that ∂H_t alternates between segments of ∂X_t and arcs of $\underline{\alpha} \cup \underline{\gamma} \cup \underline{\beta}$ with positive weight. From Lemma 6.6, the total length of segments coming from arcs of $\underline{\alpha} \cup \underline{\beta}$ is at most $8\pi \sum 1/c_i$, because each

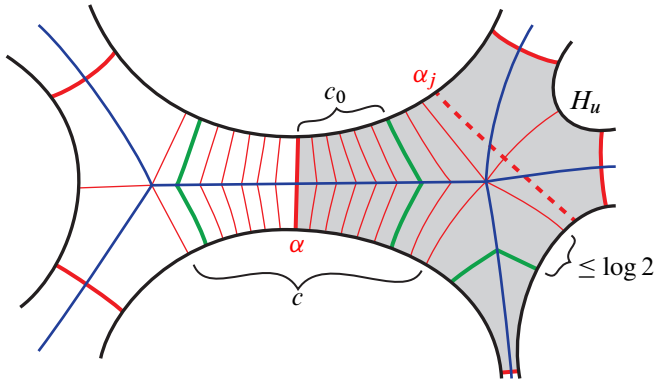


Figure 7: A foliated collar of width c about an orthogeodesic arc α. If the arc is shorter than $\log 3$, then there are (bold green) leaves of this collar of length equal to $\log 3$. For a very short arc γ_i, the distance between the longest leaf of its collar and the leaf of length $\log 3$ is at most $\log 2$. The dashed arc α_j has weight 0 and corresponds to one of two possible choices of maximal completion of $\underline{\alpha}$.

arc of $\underline{\alpha} \cup \underline{\beta}$ can appear at most two times on ∂H_t. Similarly, from the construction of our coordinate system, the total length of the segments coming from ∂X_t that correspond to collars of arcs in $\underline{\alpha} \cup \underline{\beta}$ is at most $2 \sum c_i$.

Suppose some arc γ_i of $\underline{\gamma}$ forms a segment of ∂H_t. The distance between the leaf of the orthogeodesic foliation parallel to γ_i with length $\log(3)$ and the singular, longest leaf parallel to γ_i has distance uniformly bounded above by $\log 2$ for large values of t (see Figure 7). Truncate H_t by removing the leaves of the orthogeodesic foliation parallel to γ_i with length at most $\log 3$ to obtain a new (nonconvex) geodesic polygon H_t°. An application of the collar lemma [Buser 1992, Theorem 4.1.1] to the double DX_t along its boundary shows that α_j does not enter the region of H_t that we removed.

Each arc γ_i of $\underline{\gamma}$ contributed at most $2t + O(e^{-t/2})$ to the length of ∂H_t. However, after truncating, each γ_i contributes at most $2(\log 2 + \log 3 + \log 2) = \log 144$ to the length of ∂H_t°. Putting together all of our estimates completes the proof. □

For each $\alpha_i \in \underline{\alpha}$ with positive coefficient c_i in $\underline{B}_t$, the orthogeodesic length $\ell_t(\alpha_i)$ of α_i on X_t is bounded above and below by the positive real numbers independent of t provided by Lemma 6.6. If $c_i = 0$ for some i, then Lemma 6.7 provides bounds on $\ell_t(\alpha_i)$ independent of t. Therefore, there exists a subsequence t_k tending to infinity such that $(\ell_{t_k}(\alpha_i))$ converges to a positive number ℓ_i for each i, while $\ell_t(\gamma_i) = O(e^{-t/2})$ for each $\gamma_i \in \underline{\gamma}$.

The metric completion of $X_{t_k} \setminus (\underline{\alpha} \cup \underline{\gamma} \cup \underline{\beta})$ is a collection of hyperbolic right-angled hexagons, each with three nonadjacent sides that correspond to arcs of $\underline{\alpha} \cup \underline{\gamma} \cup \underline{\beta}$. The lengths of these sides determine uniquely an isometry class of right-angled hexagons, which we have just proved converge to (degenerate) right-angled hexagons in which the edges corresponding to arcs of γ become spikes in the limit. The

(degenerate) right-angled hexagons glue along $\underline{\alpha}$ to form a complete hyperbolic surface Y homeomorphic to Σ with a maximal filling arc system labeled by $\underline{\alpha}$ and realized orthogeodesically on Y. That is, we have constructed a surface $Y(\underline{A}) = Y \in \mathcal{T}(\Sigma)$.

Lemma 6.8 $$\underline{A}(Y(\underline{A})) = \underline{A}.$$

Proof By construction, the length of the projection of every edge of the spine of X_t dual to an arc of $\underline{\alpha}$ was constant along the sequence (X_{t_k}) converging geometrically to $Y(\underline{A})$. The lemma follows. □

In order to show that the inverse $Y(\cdot)$ is well defined, we will need the following statement, which refines the relationship between the coefficients of $\underline{B}_t$ and the lengths of its arcs.

Let $\underline{\delta} = \underline{\alpha} \cup \underline{\gamma} \cup \underline{\beta}$ denote the support of $\underline{B}_t$. According to [Luo 2007, Theorem 1.2], the lengths of the closest-point projections of the edges of the spine dual to the arcs of $\underline{\delta}$ (ie the coefficients of the weighted arc system) extend to an analytic local diffeomorphism $\underline{B}_{\underline{\delta}} \colon \mathcal{T}(D\Sigma) \to \mathbb{R}^{\underline{\delta}}$ whose image is a convex cone with finitely many sides.[8] Now we show that analyticity extends to infinity.

Lemma 6.9 *For each maximal filling arc system $\underline{\alpha}$ defining a cell of full dimension in $\mathscr{A}_{\mathrm{fill}}(\Sigma, \partial\Sigma)$, there is an analytic map*

$$\underline{A}_{\underline{\alpha}} \colon \mathcal{T}(\Sigma) \to \mathbb{R}^{\underline{\alpha}}$$

such that, if the spine of $Y \in \mathcal{T}(\Sigma)$ has dual arc system contained in $\underline{\alpha}$, then $\underline{A}_{\underline{\alpha}}(Y) = \underline{A}(Y)$.

Proof The orthogeodesic length functions associated to our maximal arc system $\underline{\delta} = \underline{\alpha} \cup \underline{\gamma} \cup \underline{\beta}$ on $D\Sigma$ form an analytic parametrization of $\mathcal{T}(D\Sigma)$, which we denote by $\ell_{\underline{\delta}} \colon \mathcal{T}(D\Sigma) \to \mathbb{R}^{\underline{\delta}}_{>0}$. We have a commutative diagram of analytic embeddings

(9)
$$\begin{array}{ccc} \mathbb{R}^{\underline{\delta}}_{>0} & \xrightarrow{\underline{B}_{\underline{\delta}} \circ \ell_{\underline{\delta}}^{-1}} & \mathbb{R}^{\underline{\delta}} \\ & {\scriptstyle \ell_{\underline{\delta}}} \nwarrow \quad \nearrow {\scriptstyle \underline{B}_{\underline{\delta}}} & \\ & \mathcal{T}(D\Sigma) & \end{array}$$

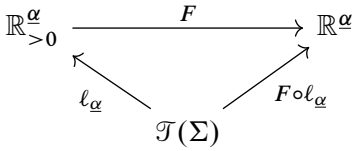
An explicit formula for $\underline{B}_{\underline{\delta}} \circ \ell^{-1}$ can be recovered from [Mondello 2009b, Section 2.4], which produces an analytic mapping $G \colon \mathbb{R}^{\underline{\alpha}\cup\underline{\beta}}_{>0} \to \mathbb{R}^{\underline{\alpha}\cup\underline{\beta}}$ that describes how $\underline{B}_{\underline{\delta}}$ behaves when the arcs corresponding to $\underline{\gamma}$ have length close to 0. More precisely, let $\pi_{\underline{\alpha}\cup\underline{\beta}} \colon \mathbb{R}^{\underline{\delta}} \to \mathbb{R}^{\underline{\alpha}\cup\underline{\beta}}$ be the coordinate projection. Then, for $x_{\underline{\delta}} = (x_{\underline{\alpha}}, x_{\underline{\gamma}}, x_{\underline{\beta}}) \in \mathbb{R}^{\underline{\alpha}\cup\underline{\beta}}_{>0} \times \mathbb{R}^{\underline{\gamma}}_{\geq 0}$, we have

(10)
$$\pi_{\underline{\alpha}\cup\underline{\beta}} \circ B_{\underline{\delta}} \circ \ell_{\underline{\delta}}^{-1}(x_{\underline{\delta}}) = G(x_{\underline{\alpha}}, x_{\underline{\beta}}) + E$$

uniformly on compact subsets of $\mathbb{R}^{\underline{\alpha}\cup\underline{\beta}}_{>0} \times \mathbb{R}^{\underline{\gamma}}_{\geq 0}$, where E is a vector whose entries are all of order $O(\max_{\gamma\in\underline{\gamma}}\{x_\gamma^2\})$.

[8] The "projection length" associated to each arc of $\underline{\delta}$ (called the "radius coordinate" in [Luo 2007] and the "width" in [Mondello 2009b]) is positive when that arc is dual to an edge of the spine of a surface $X \in \mathcal{T}(D\Sigma)$.

Restricting to the locus of symmetric surfaces $\{X \in \mathcal{T}(D\Sigma) : \ell_{\alpha_i}(X) = \ell_{\beta_i}(X) \text{ for all } i\}$, the map G therefore induces an analytic map $F : \mathbb{R}^{\underline{\alpha}}_{>0} \to \mathbb{R}^{\underline{\alpha}}$. Again, we have an analytic parametrization $\ell_{\underline{\alpha}} : \mathcal{T}(\Sigma) \to \mathbb{R}^{\underline{\alpha}}_{>0}$ by length functions and a diagram

(11)


So take $\underline{A}_{\underline{\alpha}} = F \circ \ell_{\underline{\alpha}}$; it follows from the definitions that, if the dual arc system to the spine of a surface $Y \in \mathcal{T}(\Sigma)$ is contained in $\underline{\alpha}$, then $\underline{A}_{\underline{\alpha}}(Y) = \underline{A}(Y)$. □

A priori, $Y(\underline{A})$ depends on the subsequence X_{t_k} converging geometrically to $Y(\underline{A})$. However:

Lemma 6.10 *The limit $Y(\underline{A})$ does not depend on choice of subsequence X_{t_k}, ie $X_t \to Y$. Moreover, $Y : |\mathscr{A}_{\mathrm{fill}}(\Sigma, \partial\Sigma)|_{\mathbb{R}} \to \mathcal{T}(\Sigma)$ is continuous.*

Proof Throughout this proof, we let $\pi := \pi_{\underline{\alpha} \cup \underline{\beta}}$ be the coordinate projection from the proof of Lemma 6.9.

Let $s > 0$ and $X_{t,s} \in \mathcal{T}(D\Sigma)$ be the surface obtained from X_t by keeping all lengths of arcs of $\underline{\alpha} \cup \underline{\beta}$ fixed and taking $\ell_{\gamma_i}(X_{t,s}) := \ell_{\gamma_i}(X_{t+s})$ for each $\gamma_i \in \underline{\gamma}$. Note that $\ell_{\gamma_i}(X_{t+s}) = O(e^{-(s+t)/2})$. By construction of $X_{t,s}$, the lengths of arcs of $\underline{\alpha} \cup \underline{\beta}$ agree with those of X_t, so (10) gives

$$\pi(\underline{B}_{\underline{\delta}}(X_t)) - \pi(\underline{B}_{\underline{\delta}}(X_{t,s})) = O(e^{-(s+t)}).$$

Recall that $\pi(\underline{B}_{\underline{\alpha}}(X_t)) = \pi(\underline{B}_t)$ is constant for all $t > 0$, so that

$$\pi(\underline{B}_{\underline{\delta}}(X_{s+t})) - \pi(\underline{B}_{\underline{\delta}}(X_{t,s})) = O(e^{-(s+t)})$$

as well. Since $\underline{B}_{\underline{\delta}}$ is open analytic, and $\{\pi(\ell_{\underline{\delta}}(X_t)) : t > 0\} \subset \mathbb{R}^{\underline{\alpha} \cup \underline{\beta}}_{>0}$ lies in a compact set (Lemmas 6.6 and 6.7), we can adjust the lengths of arcs α_i and β_i of $\underline{\alpha} \cup \underline{\beta}$ in $X_{t,s}$ by $O(e^{-(s+t)})$ to obtain X_{s+t}. Thus, for any $t_k \to \infty$, the lengths $(\ell_{t_k}(\underline{\alpha} \cup \underline{\beta}))$ form a Cauchy sequence, and hence converge. Thus any two subsequential geometric limits (with basepoints away from the spikes of the subsurface associated with Σ°) coincide, which proves that $Y(\underline{A})$ is well defined.

To see that $Y(\cdot)$ is continuous, let $\underline{A}_k \to \underline{A}$; by passing to a subsequence, we may assume that $\underline{A}_k$ are in the closure of the cell associated to a maximal filling arc system $\underline{\alpha}$. Let $\overline{\underline{A}}_k$ and $\overline{\underline{A}}$ be the mirror images (with corresponding weights) of $\underline{A}_k$ and $\underline{A}$ in $D\Sigma$, respectively. We build two families of approximating surfaces $X_k, X_k^k \in \mathcal{T}(D\Sigma)$ corresponding to the weighted arc systems

$$\overline{\underline{A}} + k \sum \gamma_i + \underline{A} \quad \text{and} \quad \overline{\underline{A}}_k + k \sum \gamma_i + \underline{A}_k$$

on $D\Sigma$, respectively. By [Luo 2007, Theorem 1.2] (alternatively the proof of Lemma 6.9), each X_k^k is close to X_k in $\mathcal{T}(D\Sigma)$; hence, X_k^k and X_k have the same geometric limit $Y(\underline{A}) \in \mathcal{T}(\Sigma)$, which is what we wanted to show. □

We now have all of the pieces in place to complete the proof of Theorem 6.4.

Proof of Theorem 6.4 By Lemma 6.10, $Y(\cdot)$ is well defined and continuous, and, by Lemma 6.8, $Y(\cdot)$ is a right inverse to $\underline{A}(\cdot)$; in particular, $Y(\cdot)$ is injective. For a given maximal arc system $\underline{\alpha}$, the open orthant $U_{\underline{\alpha}} = \mathbb{R}^{\underline{\alpha}}_{>0} \subset \mathbb{R}^{\underline{\alpha}}$ is identified with the interior of a top-dimensional cell of $|\mathscr{A}_{\mathrm{fill}}(\Sigma, \partial\Sigma)|_{\mathbb{R}}$. Some of the hyperplanes in $\partial\mathbb{R}^{\underline{\alpha}}_{\geq 0}$ are identified with the interior of cells associated with nonmaximal filling arc systems contained in $\underline{\alpha}$; let $\overline{U_{\underline{\alpha}}}$ denote the closure of $U_{\underline{\alpha}}$ in $|\mathscr{A}_{\mathrm{fill}}(\Sigma, \partial\Sigma)|_{\mathbb{R}}$.

Then $Y(\cdot)$ defines a continuous bijection $\overline{U_{\underline{\alpha}}} \to \overline{Y(U_{\underline{\alpha}})}$, and this identification is homeomorphic, because $\underline{A}_{\underline{\alpha}}$ supplies an analytic inverse on $\overline{Y(U_{\underline{\alpha}})}$, by Lemma 6.9. Since these homeomorphisms glue along the combinatorics of $|\mathscr{A}_{\mathrm{fill}}(\Sigma, \partial\Sigma)|_{\mathbb{R}}$, the map $Y(\cdot)$ is the desired global homeomorphic inverse to $\underline{A}(\cdot)$.

Again by Lemma 6.9, $A(\cdot)$ is analytic restricted to the relative interior of the image under $Y(\cdot)$ of each cell of $|\mathscr{A}_{\mathrm{fill}}(\Sigma, \partial\Sigma)|_{\mathbb{R}}$, demonstrating the stratified real analytic structure. That level sets of the residue functions are mapped to one another is an exercise in unpacking the definitions. □

7 Transverse and shear-shape cocycles

We now define the main protagonists of this paper, the *shear-shape cocycles* on a measured lamination. In Section 7.2, we give a first definition of shear-shape cocycles in terms of the cohomology of an augmented neighborhood of λ, twisted by its local orientation (Definition 7.5). While this definition has technical merit (and exactly parallels the construction of period coordinates for quadratic differentials, a fact which we exploit in Section 10), it is impractical to use. We rectify this deficiency in Section 7.3 by giving a second formulation which parallels Bonahon's axiomatic approach to transverse cocycles (compare Definitions 7.4 and 7.11). The main result of this section, Proposition 7.13, proves that these two definitions agree.

The reader may find it helpful to consult Sections 10 or 13 while digesting these definitions so as to have a concrete model of shear-shape cocycles in mind.

7.1 Transverse cocycles

As shear-shape cocycles generalize Bonahon's transverse cocycles, we begin by recalling two equivalent definitions of transverse cocycles for geodesic laminations which we generalize in Sections 7.2 and 7.3.

Remark 7.1 We have chosen to present transverse cocycles in a way that anticipates our construction of shear-shape cocycles. The reader is advised that our treatment is ahistorical, and in particular omits the fascinating (and quite subtle) relationship between transverse cocycles and transverse Hölder distributions. For more on this correspondence, see [Bonahon 1997a; 1997b; 1996].

The first definition we consider is cohomological. Let λ be a measured lamination on S; then an orientation of λ is a continuous choice of orientation of the leaves of λ. If N is any snug neighborhood of λ, then one may take a corresponding (snug) neighborhood $\widehat{N}$ of the orientation cover $\widehat{\lambda}$ of λ. Let ι be the covering involution of $\widehat{N} \to N$, and let $H^1(\widehat{N}, \partial\widehat{N}; \mathbb{R})^-$ denote the -1 eigenspace for the action of ι^*,

Definition 7.2 With all notation as above, a *transverse cocycle* for λ is an element of $H^1(\widehat{N}, \partial\widehat{N}; \mathbb{R})^-$. We use $\mathcal{H}(\lambda)$ to denote the set of all transverse cocycles for λ.

With the definition above it is clear that $\mathcal{H}(\lambda)$ is a vector space, and, if λ is a union of sublaminations $\lambda_1, \dots, \lambda_L$, then the space of transverse cocycles splits as

$$\mathcal{H}(\lambda) = \bigoplus_{l=1}^{L} \mathcal{H}(\lambda_l).$$

We record the dimension of $\mathcal{H}(\lambda)$ below.

Lemma 7.3 [Bonahon 1997b, Theorem 15] *The space of transverse cocycles forms a vector space of real dimension $-\chi(\lambda) + n_0(\lambda)$, where $n_0(\lambda)$ is the number of orientable components of λ.*

When working with individual transverse cocycles, the above definition is rather unwieldy. Instead, it is often more useful to think of a transverse cocycle as a function on actual arcs instead of on homology classes.

Definition 7.4 Let $\lambda \in \mathcal{ML}(S)$. A *transverse cocycle* σ for λ is a function which assigns to every arc k transverse to λ a real number $\sigma(k)$ such that:

(H0) **Support** If k does not intersect λ, then $\sigma(k) = 0$.

(H1) **Transverse invariance** If k and k' are isotopic transverse to λ, then $\sigma(k) = \sigma(k')$.

(H2) **Finite additivity** If $k = k_1 \cup k_2$, where k_i have disjoint interiors, then $\sigma(k) = \sigma(k_1) + \sigma(k_2)$.

The reader familiar with train tracks will recognize that these rules resemble those governing weight systems on train tracks; see Section 9 for a continuation of this discussion.

We direct the reader to [Bonahon 1997b] or [Bonahon 1996, Section 3] for a proof of the equivalence of Definitions 7.2 and 7.4 (our proof of Proposition 7.13, the corresponding statement for shear-shape cocycles, can also be adapted to prove this equivalence).

7.2 Shear-shape cocycles as cohomology classes

Our first definition of a shear-shape cocycle is as a cohomology class on an appropriate augmented orientation cover, paralleling Definition 7.2. This viewpoint allows us to deduce global structural results about spaces of shear-shape cocycles (Lemma 7.8) and also reveals implicit constraints on the structure of individual shear-shape cocycles (Lemma 7.9).

Suppose that $\underline{\alpha}$ is a filling arc system for $S \setminus \lambda$. For each arc $\alpha_i \in \underline{\alpha}$, choose an arc t_i which meets α_i exactly once and is disjoint from $\lambda \cup \underline{\alpha} \setminus \{\alpha_i\}$. We call such an arc t_i a *standard transversal to* α_i. Compare Figure 9. An *orientation* of $\lambda \cup \underline{\alpha}$ is a continuous orientation of the leaves of λ together with a choice of orientation on each t_i such that t_i can be isotoped transverse to α_i into λ so that the orientations agree. Most pairs $\lambda \cup \underline{\alpha}$ are not orientable, but each has an *orientation double cover* $\widehat{\lambda} \cup \widehat{\underline{\alpha}}$ (the reader should have in mind the orientation cover of a quadratic differential). We note that if $\lambda \cup \underline{\alpha}$ is orientable then λ itself must be.

Consider a snug neighborhood $\mathcal{N}_\epsilon(\lambda)$ of λ on some hyperbolic surface X; since $X \setminus \lambda$ and $X \setminus \mathcal{N}_\epsilon(\lambda)$ have the same topological type, we can identify the arc system $\underline{\alpha}$ as an arc system on $X \setminus \mathcal{N}_\epsilon(\lambda)$. In particular, taking a small neighborhood $\mathcal{N}_\epsilon(\underline{\alpha})$ of $\underline{\alpha}$, there is a correspondence between complementary components of $X \setminus (\lambda \cup \underline{\alpha})$ and $X \setminus \mathcal{N}_\epsilon(\lambda \cup \underline{\alpha})$. We will refer to any neighborhood $N_{\underline{\alpha}}$ of $\lambda \cup \underline{\alpha}$ whose complementary components have the same topological type as $X \setminus (\lambda \cup \underline{\alpha})$ as a *snug neighborhood.*

Now let $N_{\underline{\alpha}}$ be a snug neighborhood of $\lambda \cup \underline{\alpha}$; then the cover $\widehat{\lambda} \cup \widehat{\underline{\alpha}} \to \lambda \cup \underline{\alpha}$ extends to a covering $\widehat{N}_{\underline{\alpha}} \to N_{\underline{\alpha}}$ with covering involution ι. By definition of the orientation cover, each standard transversal t_i lifts to a pair of distinguished homology classes

$$t_i^{(1)}, t_i^{(2)} \in H_1(\widehat{N}_{\underline{\alpha}}, \partial \widehat{N}_{\underline{\alpha}}; \mathbb{R})$$

such that $\iota_* t_i^{(1)} = -t_i^{(2)}$.

The odd cocycles $H^1(\widehat{N}_{\underline{\alpha}}, \partial \widehat{N}_{\underline{\alpha}}; \mathbb{R})^-$ for the covering involution ι^* now provide a local cohomological model for the space of shear-shape cocycles on λ. Observe that, for each i and each $\sigma \in H^1(\widehat{N}_{\underline{\alpha}}, \partial \widehat{N}_{\underline{\alpha}}; \mathbb{R})^-$,

$$\sigma(t_i^{(1)}) = -\iota^* \sigma(t_i^{(1)}) = -\sigma(\iota_* t_i^{(1)}) = \sigma(t_i^{(2)}).$$

Definition 7.5 Let $\lambda \in \mathcal{ML}(S)$. A *shear-shape cocycle* for λ is a pair $(\underline{\alpha}, \sigma)$ where $\underline{\alpha} = \sum \alpha_i$ is a filling arc system on $S \setminus \lambda$ and $\sigma \in H^1(\widehat{N}_{\underline{\alpha}}, \partial \widehat{N}_{\underline{\alpha}}; \mathbb{R})^-$ is such that the values $\sigma(t_i^{(j)})$ are all positive.[9]

Let $\Sigma_1 \cup \cdots \cup \Sigma_m$ denote the components of $S \setminus \lambda$; then we define the *weighted arc system underlying* σ,

$$\underline{A} := \sum \sigma(t_i^{(j)}) \alpha_i \in \prod_{j=1}^{m} |\mathscr{A}_{\mathrm{fill}}(\Sigma_j, \partial \Sigma_j)|_{\mathbb{R}}.$$

We denote the set of all shear-shape cocycles for λ by $\mathcal{SH}(\lambda)$, the set of all shear-shape cocycles with underlying arc system $\underline{\alpha}$ by $\mathcal{SH}^\circ(\lambda; \underline{\alpha})$, and the set of all shear-shape cocycles with underlying weighted arc system $\underline{A}$ by $\mathcal{SH}(\lambda; \underline{A})$. Often, we will leave the arc system implicit and just say that σ is a shear-shape cocycle for λ.

[9] By Poincaré–Lefschetz duality, we have a linear isomorphism $H^1(\widehat{N}_{\underline{\alpha}}, \partial \widehat{N}_{\underline{\alpha}}; \mathbb{R}) \cong H_1(\widehat{N}_{\underline{\alpha}}; \mathbb{R})$ mapping the odd cocycles for ι^* to the odd cycles for ι_*. Compare with [Bonahon and Dreyer 2017, Sections 4.1 and 4.4], where a theory of (appropriately generalized) transverse (co)cycles are applied to give shear-type coordinates for some higher-rank Teichmüller spaces.

Remark 7.6 By Theorem 6.4, a filling weighted arc system $\underline{A}$ is the same data as a marked hyperbolic structure on each component of $S \setminus \lambda$. In Sections 12–15, we prove that (so long as σ satisfies a positivity condition) these metrics glue together to give a complete hyperbolic metric on S.

Our definition of shear-shape cocycle a priori depends on the choice of auxiliary neighborhood $N_{\underline{\alpha}}$ of $\lambda \cup \underline{\alpha}$. However, it is not hard to see that:

Lemma 7.7 *The spaces of shear-shape cocycles defined by different snug neighborhoods are linearly isomorphic. Moreover, any two choices of snug neighborhoods define the same underlying weighted arc system.*

Proof Given two nested, snug neighborhoods $N'_{\underline{\alpha}} \subset N_{\underline{\alpha}}$ there is a deformation retraction of $N_{\underline{\alpha}}$ onto $N'_{\underline{\alpha}}$ (this comes from the assumption of snugness). This induces an isomorphism

$$H^1(\widehat{N}_{\underline{\alpha}}, \partial \widehat{N}_{\underline{\alpha}}; \mathbb{R}) \cong H^1(\widehat{N}'_{\underline{\alpha}}, \partial \widehat{N}'_{\underline{\alpha}}; \mathbb{R}) \tag{12}$$

which also identifies the -1 eigenspaces of the covering involution. Therefore, we may identify the shear-shape cocycles defined by $N_{\underline{\alpha}}$ with those defined by $N'_{\underline{\alpha}}$. To see that the weights on $\underline{\alpha}$ do not depend on the choice of $N_{\underline{\alpha}}$, we note that the deformation retraction of $N_{\underline{\alpha}}$ onto $N'_{\underline{\alpha}}$ takes standard transversals to standard transversals, and hence the value of the cocycle on the transversals does not change as we change neighborhoods.

Now, given any two snug neighborhoods $N_{\underline{\alpha}}$ and $N'_{\underline{\alpha}}$ of $\lambda \cup \underline{\alpha}$, one may take a common refinement $N''_{\underline{\alpha}}$ of $N_{\underline{\alpha}}$ and $N'_{\underline{\alpha}}$ and apply (12) to deduce that the spaces of shear-shape cocycles defined by $N_{\underline{\alpha}}$ and $N'_{\underline{\alpha}}$ are linearly isomorphic and define the same underlying arc system. □

In view of this lemma, throughout the sequel we will change the neighborhood $N_{\underline{\alpha}}$ carrying σ at will.

As the orientation cover of λ naturally embeds into $\widehat{N}_{\underline{\alpha}}$, we may identify $\mathcal{H}(\lambda)$ with a subspace of $H^1(\widehat{N}_{\underline{\alpha}}, \partial \widehat{N}_{\underline{\alpha}}; \mathbb{R})$. Since any element of $\mathcal{H}(\lambda)$ evaluates to 0 on each standard transversal, we can add and subtract transverse cocycles from shear-shape cocycles without changing the underlying weighted arc system. We therefore have the following analogue of Lemma 7.3:

Lemma 7.8 *Let $\underline{A}$ be the weighted arc system underlying some shear-shape cocycle. Then $\mathcal{SH}(\lambda; \underline{A})$ is an affine space modeled on the vector space $\mathcal{H}(\lambda)$. In particular,* $\dim_{\mathbb{R}}(\mathcal{SH}(\lambda; \underline{A})) = -\chi(\lambda) + n_0(\lambda)$.

Homological constraints on residues When λ is orientable (or, more generally, contains orientable components), there are homological constraints governing which weighted arc systems may underlie a shear-shape cocycle. Passing between arc systems and hyperbolic structures on complementary subsurfaces (via Theorem 6.4), these homological constraints govern when two structures can be glued together along λ.

For example, if λ is a simple closed curve then in order to glue a hyperbolic structure on $S \setminus \lambda$ along λ, the lengths of the boundary components must have equal length. Tracing through the combinatorialization by weighted arc systems, this implies that the $\underline{A}$–length of the boundary components must be the same. The following lemma generalizes this observation to the case when $S \setminus \lambda$ has crowned boundary (compare Lemma 13.1 for a similar discussion using hyperbolic geometry):

Lemma 7.9 *Suppose that σ is a shear-shape cocycle for λ with underlying weighted arc system $\underline{A}$, and let μ be an orientable component of λ. Then the sum of the (signed) residues of the boundary components incident to μ is 0.*

Proof For any component μ of λ, let $\partial(\mu)$ denote the boundary components (either closed or crowned) resulting from cutting along μ. For the purposes of this proof, let $\alpha(\mu)$ denote the subarc system of $\underline{\alpha}$ consisting of those arcs with endpoints on μ.

Pick an orientation on μ; this induces an orientation on each boundary component $\mathcal{C} \in \partial(\mu)$, and hence gives the metric residue of each such $\mathcal{C}$ a definite choice of sign. Since we are eventually going to prove that the sum of these residues is 0, it does not matter which orientation of μ we pick.

As μ is orientable, picking an orientation on μ is also equivalent to picking one of the lifts $\hat{\mu}$ of μ in the orientation cover $\widehat{\lambda} \cup \widehat{\underline{\alpha}}$. Let $\widehat{\alpha(\mu)}$ denote the set of all lifts of arcs of $\alpha(\mu)$ which meet $\hat{\mu}$. Then, since severing $\widehat{\alpha(\mu)}$ disconnects $\hat{\mu}$ from the rest of $\widehat{\lambda} \cup \widehat{\underline{\alpha}}$, there is a relation

$$\sum_{\hat{\alpha}_i \in \widehat{\alpha(\mu)}} \varepsilon_i \hat{t}_i = 0 \quad \text{in } H_1(\widehat{N}_{\underline{\alpha}}, \partial \widehat{N}_{\underline{\alpha}}; \mathbb{Z}),$$

where ε_i is 1 if $\hat{\alpha}_i$ is on the left-hand side of $\hat{\mu}$ and -1 if $\hat{\alpha}_i$ is on the right-hand side, and $\hat{t}_i$ is the (relative homology class of the) oriented standard transversal corresponding to $\hat{\alpha}_i$. See Figure 8.

Therefore, for any cohomology class $\sigma \in H^1(\widehat{N}_{\underline{\alpha}}, \partial \widehat{N}_{\underline{\alpha}}; \mathbb{Z})$, and in particular any shear-shape cocycle,

$$\sum_{\hat{\alpha}_i \in \widehat{\alpha(\mu)}} \varepsilon_i \sigma(\hat{t}_i) = 0. \tag{13}$$

Now ε_i is positive when the arc is on the left–hand side of $\hat{\mu}$, or equivalently (equipping $\mu \subset S$ with the corresponding orientation) when $S \setminus \lambda$ is on the left–hand side of μ. Similarly, ε_i is negative when the complementary subsurface lies to the right of μ. Unraveling the definitions and partitioning the arcs of $\alpha(\mu)$ into their incident boundary components, (13) is equivalent to the statement that

$$\sum_{\mathcal{C} \in \partial(\mu)} \operatorname{res}_{\underline{A}}(\mathcal{C}) = \sum_{\alpha_i \in \alpha(\mu)} \varepsilon_i c_i = 0,$$

which is what we wanted to prove. □

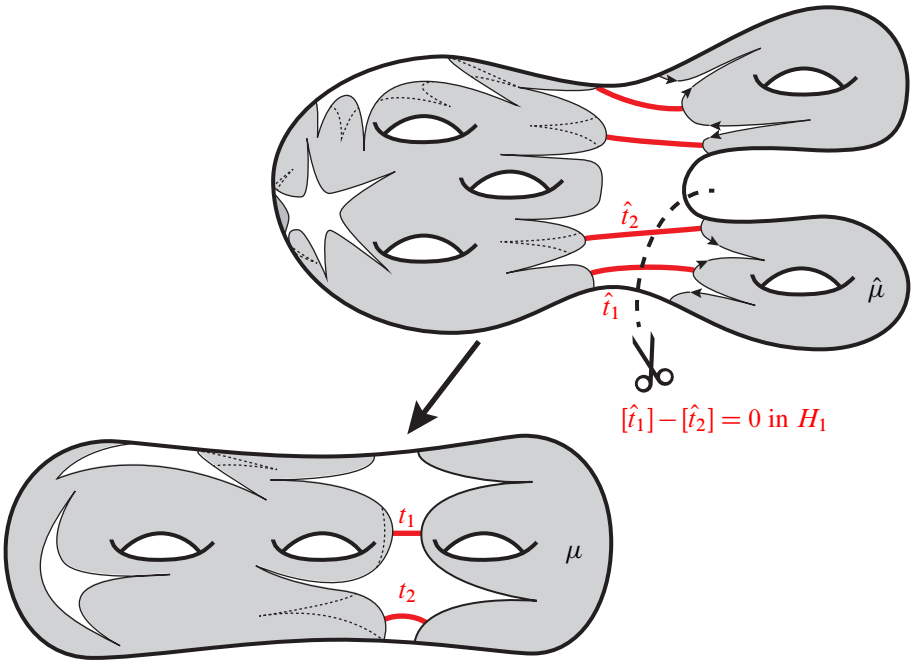


Figure 8: Severing ties with one of the lifts $\hat{\mu}$ of an orientable component μ of λ. This partition induces a relation in homology, and hence a restriction on shear-shape cocycles. The top surface contains $\hat{\lambda}$ while the bottom contains λ; the shaded regions are neighborhoods of these laminations.

7.3 Shear-shape cocycles as functions on arcs

In analogy with Definition 7.4, we can also view shear-shape cocycle as functions on transverse arcs which satisfy certain properties. While this definition is more involved, it is more convenient for the calculations of Sections 13–15 and better reflects the process of "measuring" arcs by a shear-shape cocycle.

As indicated by Lemma 7.9, we must first cut out the space of all possible weighted arc systems underlying a shear-shape cocycle. Denote the complementary subsurfaces of $\lambda \in \mathcal{ML}(S)$ by $\Sigma_1, \dots, \Sigma_m$, and set

$$\mathscr{B}(S \setminus \lambda) := \left\{ \underline{A} \in \prod_{j=1}^{m} |\mathscr{A}_{\mathrm{fill}}(\Sigma_j, \partial\Sigma_j)|_{\mathbb{R}} \;\middle|\; \sum_{\mathscr{C} \in \partial(\mu)} \mathrm{res}_{\underline{A}}(\mathscr{C}) = 0 \text{ for all orientable components } \mu \subset \lambda \right\},$$

where we recall that $\partial(\mu)$ denotes the set of boundary components of $S \setminus \lambda$ resulting from cutting along μ.

By Theorem 6.4, we can reinterpret $\mathscr{B}(S \setminus \lambda)$ as the set of all hyperbolic structures on $S \setminus \lambda$ such that the metric residues of the boundary components resulting from any orientable component μ of λ sum to zero. We note that when each component of λ is nonorientable, $\mathscr{B}(S \setminus \lambda)$ is just the product of the Teichmüller spaces of the complementary subsurfaces. When λ is a simple closed curve, $\mathscr{B}(S \setminus \lambda)$ consists of those metrics on $S \setminus \lambda$ where the two boundary components have the same length.

Using this reinterpretation together with Lemma 4.4, $\mathscr{B}(S \setminus \lambda)$ is topologically just a cell:

Lemma 7.10 *Let $\lambda \in \mathcal{ML}(S)$ with $S \setminus \lambda = \Sigma_1 \cup \cdots \cup \Sigma_m$. Then $\mathscr{B}(S \setminus \lambda) \cong \mathbb{R}^d$, where*

$$d = -n_0(\lambda) + \sum_{j=1}^{m} \dim(\mathcal{T}(\Sigma_j)),$$

where $n_0(\lambda)$ is the number of orientable components of λ.

Proof Let $\mu_1, \ldots, \mu_{n_0(\lambda)}$ denote the orientable components of λ and fix an arbitrary orientation on each. Then the lemma follows from the observation that $\mathscr{B}(S \setminus \lambda)$ is a fiber bundle over

$$\prod_{i=1}^{n_0(\lambda)} \left\{ (R_k^i) \in \mathbb{R}^{|\partial(\mu_i)|} \;\middle|\; \sum_k R_k^i = 0 \right\}$$

with fibers equal to

$$\left\{ [Y, f] \in \prod_{j=1}^{m} \mathcal{T}(\Sigma_j) \;\middle|\; \operatorname{res}(\mathscr{C}_k) = R_k^i \text{ for each } \mathscr{C}_k \in \partial(\mu_i) \right\}.$$

By Proposition 4.5, the fibers are each homeomorphic to $\mathbb{R}^d$, where

$$d = \left(\sum_{j=1}^{m} \dim(\mathcal{T}(\Sigma_j)) \right) - \left(\sum_{i=1}^{n_0(\lambda)} |\partial(\mu_i)| \right).$$

Totaling the dimensions of base and fiber gives the desired result. □

We can now present our second definition of shear-shape cocycles.

Definition 7.11 Let $\lambda \in \mathcal{ML}(S)$. A *shear-shape cocycle* for λ is a pair $(\sigma, \underline{A})$ where $\underline{A}$ is a weighted filling arc system

$$\underline{A} = \sum_{i=1}^{n} c_i \alpha_i \in \mathscr{B}(S \setminus \lambda)$$

and σ is a function which assigns to every arc k transverse to λ and disjoint from $\underline{\alpha} := \bigcup \alpha_i$ a real number $\sigma(k)$, satisfying the following axioms:

(SH0) **Support** If k does not intersect λ, then $\sigma(k) = 0$.

(SH1) **Transverse invariance** If k and k' are isotopic through arcs transverse to λ and disjoint from $\underline{\alpha}$, then $\sigma(k) = \sigma(k')$.

(SH2) **Finite additivity** If $k = k_1 \cup k_2$, where k_i have disjoint interiors, then $\sigma(k) = \sigma(k_1) + \sigma(k_2)$.

(SH3) **$\underline{A}$–compatibility** Suppose that k is isotopic rel endpoints and transverse to λ to some arc which may be written as $t_i \cup \ell$, where t_i is a standard transversal and ℓ is disjoint from $\underline{\alpha}$. Then the loop $k \cup t_i \cup \ell$ encircles a unique point p of $\lambda \cap \underline{\alpha}$, and

$$\sigma(k) = \sigma(\ell) + \varepsilon c_i,$$

where ε denotes the winding number of $k \cup t_i \cup \ell$ about p (where the loop is oriented so that the edges are traversed k then t_i then ℓ). See Figure 9.

While axiom (SH3) may seem convoluted upon first inspection, its entire effect is to prescribe how the value $\sigma(k)$ evolves as an endpoint of k passes through an arc of $\underline{\alpha}$. The sign change records whether the map induced by $k = t_i \cup \ell$ from the oriented simplex into S is orientation-preserving or -reversing.

Remark 7.12 In Section 9 (Proposition 9.5 in particular), we show that there exists a choice of "smoothing" for $\underline{\alpha}$ which resolves condition (SH3) into an additivity condition. This is equivalent to prescribing that an arc k may only be dragged over a point of $\lambda \cap \underline{\alpha}$ in one direction.

The equivalence between Definitions 7.5 and 7.11 is essentially the same as the equivalence of the cohomological and axiomatic definitions of transverse cocycles [Bonahon 1996, pages 248–249]. However, the $\underline{A}$–compatibility condition (axiom (SH3)) contributes new technical difficulties, and so we have included a full proof for completeness.

Proposition 7.13 *The cohomological and axiomatic definitions of shear-shape cocycles agree.*

Proof Suppose first that σ is a cohomological shear-shape cocycle, that is, a cohomology class of the orientation cover $\widehat{N}_{\underline{\alpha}}$ of $N_{\underline{\alpha}}$ that is anti-invariant under the covering involution and that gives positive weight to the canonical lifts of the standard transversals of each arc of a filling arc system $\underline{\alpha}$. We begin by building from σ a function f_σ; the basic idea is to restrict an arc to a neighborhood of λ, resulting in a relative homology class, and to set f_σ to be σ evaluated on this class.

Suppose that k is any arc transverse to λ and disjoint from $\underline{\alpha}$. Choose a small neighborhood $N_{\underline{\alpha}}$ of $\lambda \cup \underline{\alpha}$ so that k meets $\partial N_{\underline{\alpha}}$ transversely and $\partial k \cap N_{\underline{\alpha}} = \varnothing$; then $k|_{N_{\underline{\alpha}}}$ is a union of arcs with endpoints on $\partial N_{\underline{\alpha}}$. Each arc k_i of $k|_{N_{\underline{\alpha}}}$ has two distinguished, oriented lifts $k_i^{(1)}$ and $k_i^{(2)}$ to $\widehat{N}_{\underline{\alpha}}$ that cross $\widehat{\lambda}$ from right to left. As in Section 7.2, these distinguished lifts satisfy

$$\iota_*([k_i^{(1)}]) = -[k_i^{(2)}] \tag{14}$$

in $H_1(\widehat{N}_{\underline{\alpha}}, \partial\widehat{N}_{\underline{\alpha}}; \mathbb{Z})$, where ι is the covering involution of $\widehat{N}_{\underline{\alpha}} \to N_{\underline{\alpha}}$. In particular $\sigma([k_i^{(1)}]) = \sigma([k_i^{(2)}])$ since σ is anti-invariant under ι. We therefore set

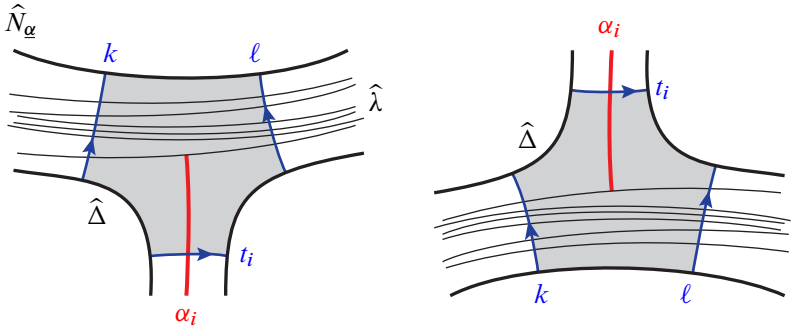
$$f_\sigma(k) := \sigma([k]),$$

where $[k]$ is the homology class of either lift of $k|_{N_{\underline{\alpha}}}$ to $\widehat{N}_{\underline{\alpha}}$.

We now prove that f_σ satisfies the axioms of Definition 7.11:

(SH0) If k does not intersect λ, then $k|_{N_{\underline{\alpha}}}$ is empty and $[k] = 0$, implying $f_\sigma(k) = 0$.

(SH1) If k and k' are isotopic through arcs transverse to λ and disjoint from $\underline{\alpha}$, then $k|_{N_{\underline{\alpha}}}$ and $k'|_{N_{\underline{\alpha}}}$ are properly isotopic. One can lift this isotopy to the orientation cover to deduce that $[k] = [k']$ for the correct choice of lifts, so $f_\sigma(k) = f_\sigma(k')$.



$[k] = [t_i] + [\ell]$ $\qquad$ $[k] + [t_i] = [\ell]$

Figure 9: Possible configurations of the disk $\widehat{\Delta}$ and the corresponding homological relations.

(SH2) Suppose that $k = k_1 \cup k_2$; then, so long as $N_{\underline{\alpha}}$ is small enough, it is clear that $k|_{N_{\underline{\alpha}}} = k_1|_{N_{\underline{\alpha}}} \cup k_2|_{N_{\underline{\alpha}}}$. Therefore, since a lift of $k|_{N_{\underline{\alpha}}}$ consists of the union of lifts of $k_1|_{N_{\underline{\alpha}}}$ and $k_2|_{N_{\underline{\alpha}}}$, we see that $[k] = [k_1] + [k_2]$, and hence the corresponding equality of f_σ values also holds.

(SH3) Finally, suppose that k is isotopic (rel endpoints and transverse to λ) to $\ell \cup t_i$. Without loss of generality, we assume that the restriction of each of k, ℓ and t_i to N is a single properly embedded arc (if not, simply break the arcs into smaller pieces and apply (SH1) and (SH2) repeatedly). We also assume the restrictions are all disjoint (even at their endpoints), appealing to (SH1) as necessary.

The isotopy between k and $\ell \cup t_i$ induces a map from a disk Δ to $N_{\underline{\alpha}}$ such that $\partial\Delta \subset \partial N_{\underline{\alpha}} \cup k \cup \ell \cup t_i$. Refining $N_{\underline{\alpha}}$, isotoping the arcs, and homotoping the map as necessary, we may assume that Δ embeds into $N_{\underline{\alpha}}$, and therefore must occur in one of the configurations shown in Figure 9.

Now choose one of the lifts $\widehat{\Delta} \subset \widehat{N}_{\underline{\alpha}}$ of Δ; this choice specifies lifts of the arcs k, ℓ and t_i and therefore (after equipping the lifts with their canonical orientations) relative homology classes $[k]$, $[\ell]$ and $[t_i]$. As these lifts together with $\partial\widehat{N}_{\underline{\alpha}}$ bound the disk $\widehat{\Delta}$, we therefore get

$$[k] = [\ell] \pm [t_i],$$

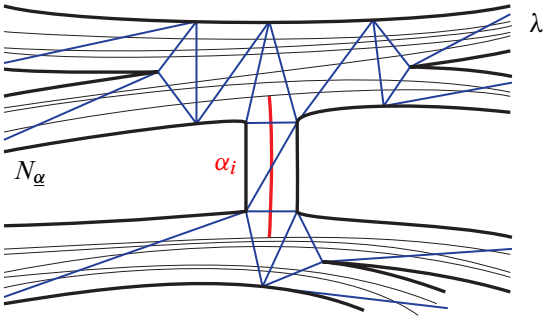


Figure 10: A triangulation of a (snug) neighborhood of $\lambda \cup \underline{\alpha}$. Axioms (SH0)–(SH3) imply that $\sigma(\partial\Delta) = 0$ for each triangle Δ in the triangulation, ie σ is a cocycle.

where the sign is determined by the relative configuration of the arcs. Inspection of Figure 9 reveals that the sign coincides with the winding number of the loop $k \cup t_i \cup \ell$ about p.

Now suppose that $(\sigma, \underline{A})$ is an axiomatic shear-shape cocycle in the sense of Definition 7.11. Pick a snug neighborhood $N_{\underline{\alpha}}$ of $\lambda \cup \underline{\alpha}$; our task is to show that the function $k \mapsto \sigma(k)$ is indeed a cocycle (on the orientation cover, and is anti-invariant under the covering involution).

We first show that σ naturally defines a cochain on $\widehat{N}$ relative to $\partial\widehat{N}_{\underline{\alpha}}$ which is anti-invariant by ι^*. Recall that any arc in the orientation cover comes with a canonical orientation. We may then assign to any oriented arc $\hat{k}$ properly embedded in $\widehat{N}_{\underline{\alpha}}$ the value $\pm\sigma(k)$, where k is the image of $\hat{k}$ under the covering projection and where the sign is positive if $\hat{k}$ is oriented according to the canonical orientation and negative otherwise. To the (canonically oriented lifts of the) standard transversals t_i we assign the value c_i. Anti-invariance then follows by construction (compare (14)).

To see that this cochain is actually a cocycle, we show that it evaluates to 0 on every boundary. For the purposes of this argument, it will be convenient to realize $H^1(\widehat{N}_{\underline{\alpha}}, \partial\widehat{N}_{\underline{\alpha}}; \mathbb{R})$ in terms of simplicial (co)homology. The neighborhood $N_{\underline{\alpha}}$ may be triangulated as depicted in Figure 10 (compare [Sözen and Bonahon 2001, Figure 1]). In such a triangulation, each point of $\lambda \cap \underline{\alpha}$ and each switch of $N_{\underline{\alpha}}$ corresponds to a unique triangle, while the remaining branches each contribute a rectangle which is in turn subdivided into two triangles. This triangulation clearly lifts to an (ι–invariant) triangulation of $\widehat{N}_{\underline{\alpha}}$.

It therefore suffices to prove that, for each oriented triangle Δ of $\widehat{N}_{\underline{\alpha}}$, we have $\sigma(\partial\Delta) = 0$. There are three types of triangles, each of which corresponds to a different axiom of Definition 7.11:

- If Δ is (the lift of) a triangle coming from a subdivision of a branch, then one if its sides does not intersect λ and is thus assigned the value 0 by (SH0). The other two sides are isotopic rel λ, cross λ with different orientations, and are assigned the same value by (SH1). Therefore $\sigma(\partial\Delta) = 0$. Similarly, if Δ comes from a neighborhood of α, then the edges transverse to α are assigned the arc weight c_i (with opposite signs) while the other edge gets zero weight, so $\sigma(\partial\Delta) = 0$.
- Now suppose Δ is (the lift of) a triangle corresponding to a switch of $N_{\underline{\alpha}}$ with $\partial\Delta = k_1 + k_2 - k$. Then, since the concatenation of k_1 and k_2 is isotopic transverse to λ to $-k$, axiom (SH2) implies
$$\sigma(k_1) + \sigma(k_2) - \sigma(k) = 0$$
and again $\sigma(\partial\Delta) = 0$.
- Finally, suppose that Δ is (the lift of) a triangle corresponding to a point of $\lambda \cap \underline{\alpha}$, so $\partial\Delta$ is some signed combination of the (canonically oriented) lifts of arcs k, ℓ and t, where t is a standard transversal and k is isotopic rel endpoints and transverse to λ to $\ell \cup t$. Without loss of generality we assume that Δ is positively oriented; then, depending on the configuration of k, t and ℓ, we have either
$$\ell - k + t = 0 \quad \text{or} \quad \ell - t - k = 0$$
(as in Figure 9). In either case, axiom (SH3) implies that $\sigma(\partial\Delta) = 0$.

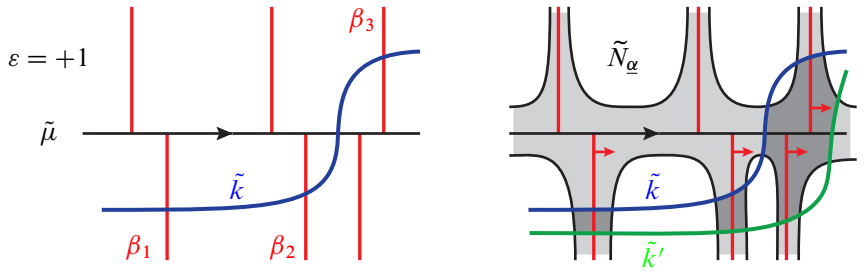


Figure 11: Since k makes progress around μ in the positive direction, $\varepsilon = +1$.

We have therefore shown that $\sigma(\partial\Delta) = 0$ for every triangle of a triangulation and hence σ is indeed a 1–cocycle on $\widehat{N}_{\underline{\alpha}}$ rel boundary, finishing the proof of the lemma. □

Measuring arcs along curves We will also want to associate a number $\sigma(k)$ to certain arcs k that have nonempty intersection with $\underline{\alpha}$; this quantity should be invariant under suitable isotopy transverse to λ respecting the combinatorics of intersections with $\underline{\alpha}$.

So suppose $\mu \subset \lambda$ is an isolated leaf, ie a simple closed curve. We say that an arc k transverse to $\lambda \cup \underline{\alpha}$ and contained in an annular neighborhood of μ is *nonbacktracking* if any lift $\tilde{k}$ of k to the universal cover intersects the entire preimage $\tilde{\mu}$ of μ exactly once and $\tilde{k}$ crosses each lift of an arc of $\underline{\alpha}$ at most once.

If k is a nonbacktracking arc, then one may orient k and give μ the orientation that makes k start to the right of μ. Record the sequence of arcs $\beta_1, \dots, \beta_m$ crossed by k, in order (note that arcs of $\underline{\alpha}$ may repeat in this sequence). Then, up to isotopy, we may assume that k is a concatenation of standard transversals $t_1, \dots, t_m$ together with a small segment k_0 disjoint from $\underline{\alpha}$ crossing μ from right to left. Compare Figure 11.

Since k is nonbacktracking, the points $\beta_1 \cap \mu, \dots, \beta_m \cap \mu$ make progress around μ in either the positive direction or the negative direction. Take $\varepsilon = +1$ in the former case and $\varepsilon = -1$ in the latter, then define

$$\sigma(k) := \sigma(k_0) + \varepsilon \sum_{j=1}^{m} c_j, \tag{15}$$

where c_j is the weight corresponding to the arc β_j. Note that the value of ε only depends on k and not on its orientation, as reversing its orientation also reverses the orientation of μ.

Lemma 7.14 *Suppose that k and k' are nonbacktracking arcs transverse to $\lambda \cup \underline{\alpha}$ contained in an annular neighborhood of a simple closed curve component μ of λ. If there exist lifts $\tilde{k}$ and $\tilde{k}'$ to $\widetilde{S}$ whose endpoints lie in the same component of $\widetilde{S} \setminus (\tilde{\lambda} \cup \underline{\tilde{\alpha}})$ and k is isotopic to k' transverse to λ, then $\sigma(k) = \sigma(k')$.*

Proof Fix a snug neighborhood $N_{\underline{\alpha}}$ of $\lambda \cup \underline{\alpha}$; then we need only show that $k|_{N_{\underline{\alpha}}}$ and $k'|_{N_{\underline{\alpha}}}$ define homologous cycles in the orientation cover.

We can find an isotopy $[0,1]^2 \to \widetilde{S}$ between lifts of k and k' (transverse to λ) that leaves the endpoints in the same component of $\widetilde{S} \setminus \widetilde{N}_{\underline{\alpha}}$. Such an isotopy then descends to S under the covering projection. The intersection of the image of each transverse arc with $N_{\underline{\alpha}}$ defines a cycle in the relative homology group, and this family of cycles is constant along the isotopy.

Since μ is orientable, an annular neighborhood of μ lifts homeomorphically to $\widehat{N}_{\underline{\alpha}}$, as do k and k'. Therefore, the isotopy between k and k' (and the homology between their restrictions) also lifts to the orientation cover $\widehat{N}_{\underline{\alpha}}$, showing that the (lifts of the) restrictions of k and k' are homologous there as well. Compare Figure 11. □

8 The structure of shear-shape space

In this section, we investigate the global structure of the space of shear-shape cocycles. Whereas Bonahon's transverse cocycles assemble into a vector space, the space $\mathcal{SH}(\lambda)$ of all shear-shape cocycles is more complex when λ is not maximal, forming an principal $\mathcal{H}(\lambda)$–bundle over $\mathcal{B}(S \setminus \lambda)$ (Theorem 8.1).

After understanding the structure of shear-shape space, we define an intersection form on $\mathcal{SH}(\lambda)$ (Section 8.2) and use it to specify the "positive locus" $\mathcal{SH}^+(\lambda)$ (Definition 8.4), which we show in Sections 10–15 serves as a global parametrization of both $\mathcal{MF}(\lambda)$ and $\mathcal{T}(S)$.

8.1 Bundle structure

Lemma 7.8 of the previous section parametrizes all shear-shape cocycles which are compatible with a given weighted arc system. In this section, we analyze how these parameter spaces piece together to get a global description of the space of all shear-shape cocycles for a fixed lamination.

Let G be a topological group. A principal G–bundle is a fiber bundle whose fibers are equipped with a transitive, continuous G–action with trivial point stabilizers together with a bundle atlas whose transition functions are continuous maps into G. We remind the reader that a principal G–bundle does not typically have a natural "zero section", but, instead, any local section of the bundle defines an identification of the fibers with G via the G–action. Moreover, any two sections define local trivializations of the bundle that differ by an element of G in each fiber.

Theorem 8.1 *Let $\lambda \in \mathcal{ML}(S)$. The space $\mathcal{SH}(\lambda)$ forms a principal $\mathcal{H}(\lambda)$–bundle over $\mathcal{B}(S \setminus \lambda)$ whose fiber over $\underline{A} \in \mathcal{B}(S \setminus \lambda)$ is $\mathcal{SH}(\lambda; \underline{A})$.*

Proof There is an obvious map from $\mathcal{SH}(\lambda)$ to $\mathcal{B}(S \setminus \lambda)$ given by remembering only the values $\sigma(t_i)$ of transversals to the arcs. For a given choice σ_0 in the fiber $\mathcal{SH}(\lambda; \underline{A})$ over $\underline{A}$, Lemma 7.8 identifies $\mathcal{SH}(\lambda; \underline{A})$ with $\mathcal{H}(\lambda)$ via the assignment $\sigma \mapsto \sigma - \sigma_0$.

For any filling arc system $\underline{\alpha}$ of $S \setminus \lambda$, the space $\mathcal{SH}^\circ(\lambda;\underline{\alpha})$ of shear-shape cocycles with underlying arc system $\underline{\alpha}$ is naturally identified with the open orthant

$$\{\sigma \in H^1(\widehat{N}_{\underline{\alpha}}, \partial \widehat{N}_{\underline{\alpha}}; \mathbb{R})^- : \sigma(t_i^{(j)}) > 0 \text{ for all } i, j = 1, 2\}, \tag{16}$$

where $N_{\underline{\alpha}}$ is a snug neighborhood of $\lambda \cup \underline{\alpha}$ on S.

Consider the open cell $\mathscr{B}^\circ(\underline{\alpha}) \subset \mathscr{B}(S \setminus \lambda)$ defined as all those weighted arc systems with support equal to a maximal arc system $\underline{\alpha}$. Using cohomological coordinates (16) for $\mathcal{SH}^\circ(\lambda;\underline{\alpha})$, we can find a continuous section σ of $\mathcal{SH}^\circ(\lambda;\underline{\alpha}) \to \mathscr{B}^\circ(\underline{\alpha})$. Then

$$\phi_\sigma : \mathscr{B}^\circ(\underline{\alpha}) \times \mathcal{H}(\lambda) \to \mathcal{SH}^\circ(\lambda;\underline{\alpha}), \quad (\underline{A}, \eta) \mapsto \sigma(\underline{A}) + \eta,$$

is a homeomorphism preserving fibers of the natural projections. For another choice of section σ',

$$\phi_\sigma^{-1}(\phi_{\sigma'}(\underline{A}, \eta)) = (\underline{A}, \eta + \sigma'(\underline{A}) - \sigma(\underline{A})).$$

Evidently, the map $\underline{A} \mapsto \sigma'(\underline{A}) - \sigma(\underline{A}) \in \mathcal{H}(\lambda)$ is continuous.

If $N'_{\underline{\alpha}}$ is another snug neighborhood of $\lambda \cup \underline{\alpha}$, then $N_{\underline{\alpha}}$ and $N'_{\underline{\alpha}}$ share a common deformation retract. The composition of the linear isomorphisms induced on cohomology by inclusion of the deformation retract preserves the orthants defined as in (16) as well as fibers of projection to $\mathscr{B}(S \setminus \lambda)$. This proves that the principal $\mathcal{H}(\lambda)$–structure of the bundle lying over $\mathscr{B}^\circ(\underline{\alpha})$ does not depend on the snug neighborhood whose cohomology coordinatizes $\mathcal{SH}^\circ(\lambda;\underline{\alpha})$.

To show that the principal $\mathcal{H}(\lambda)$–bundle structures over all cells of $\mathscr{B}(S \setminus \lambda)$ glue together nicely, we find a continuous section of $\mathcal{SH}(\lambda) \to \mathscr{B}(S \setminus \lambda)$ near any given weighted arc system $\underline{A}$. Indeed, if $\underline{\alpha} \subset \underline{\beta}$, then inclusion $N_{\underline{\alpha}} \hookrightarrow N_{\underline{\beta}}$ of snug neighborhoods defines a map on cohomology. This map restricts to a linear isomorphism on the kernel of the evaluation map on the transversals to $\underline{\beta} \setminus \underline{\alpha}$. Thus, the closure

$$\mathcal{SH}(\lambda;\underline{\beta}) = \bigcup_{\substack{\underline{\beta} \supseteq \underline{\alpha} \\ \underline{\alpha} \text{ fills } S \setminus \lambda}} \mathcal{SH}^\circ(\lambda;\underline{\alpha}) \tag{17}$$

of $\mathcal{SH}^\circ(\lambda;\underline{\beta})$ in $\mathcal{SH}(\lambda)$ may be realized as an orthant in $H^1(\widehat{N}_{\underline{\beta}}, \partial \widehat{N}_{\underline{\beta}}; \mathbb{R})^-$ with some open and closed faces; one of the closed faces corresponds to $\mathcal{SH}^\circ(\lambda;\underline{\alpha})$.[10]

Since the complex $\mathscr{A}_{\text{fill}}(S \setminus \lambda)$ is locally finite, there are only finitely many arcs $\beta_1, \dots, \beta_k$ disjoint from $\underline{\alpha}$. Let $U \subset \mathscr{B}(S \setminus \lambda)$ be a small neighborhood of $\underline{A}$ and σ be a continuous section of $\mathcal{SH}(\lambda;\underline{\alpha}) \to \mathscr{B}^\circ(\underline{\alpha}) \cap U$. For each i, after including $\mathcal{SH}^\circ(\lambda;\underline{\alpha})$ as a face of $\mathcal{SH}(\lambda;\underline{\alpha} \cup \beta_i)$, we may extend σ continuously on $U \cap \mathscr{B}^\circ(\underline{\alpha} \cup \beta_i)$. Continuing this process, eventually extending σ to higher-dimensional cells meeting U, we end up with a continuous section $U \to \mathcal{SH}(\lambda)$, as claimed. As before, trivializations defined by two different sections differ by a continuous function $U \to \mathcal{H}(\lambda)$; this completes the proof of the theorem. □

[10] When every component of $S \setminus \lambda$ is simply connected, the empty set is a filling arc system. When this is the case, $\mathscr{B}^\circ(\varnothing)$ is identified with a point, while $\mathcal{SH}(\lambda;\varnothing) = \mathcal{H}(\lambda)$.

Since every bundle over a contractible base is trivial, this implies that:

Corollary 8.2 *Shear-shape space $\mathcal{SH}(\lambda)$ is homeomorphic to $\mathbb{R}^{6g-6}$.*

Proof Let $\Sigma_1, \dots, \Sigma_m$ denote the complementary components of λ, where Σ_j has genus g_j with b_j closed boundary components and k_j crowns of types $\{c_1^j, \dots, c_{k_j}^j\}$. By Lemmas 7.10 and 4.4, $\mathscr{B}(S \setminus \lambda)$ is homeomorphic to a cell of dimension

$$-n_0(\lambda) + \sum_{j=1}^{m} \dim(\mathcal{T}(\Sigma_j)) = -n_0(\lambda) + \sum_{j=1}^{m} \left(6g_j - 6 + 3b_j + \sum_{i=1}^{k_j} (c_i^j + 3) \right)$$

Lemmas 7.8 and 4.6 together imply that $\mathcal{SH}(\lambda; \underline{A})$ is an affine $\mathcal{H}(\lambda)$–space of dimension

$$n_0(\lambda) - \chi(\lambda) = n_0(\lambda) + \frac{1}{2} \sum_{j=1}^{m} \sum_{i=1}^{k_j} c_i^j$$

Putting these dimension counts together via Theorem 8.1, $\mathcal{SH}(\lambda)$ is homeomorphic to a cell of dimension

$$\sum_{j=1}^{m} \left(6g_j - 6 + 3b_j + \frac{3}{2} \sum_{i=1}^{k_j} (c_i^j + 2) \right) = \frac{3}{2\pi} \sum_{j=1}^{m} \operatorname{Area}(\Sigma_j) = \frac{3}{2\pi} \operatorname{Area}(S) = 6g - 6,$$

where the first equality follows from (4). □

8.2 Intersection forms and positivity

Now that we have a global description of shear-shape space, we restrict our attention to a certain positive locus $\mathcal{SH}^+(\lambda)$ inside $\mathcal{SH}(\lambda)$. The main result of this section is Proposition 8.5, in which we identify $\mathcal{SH}^+(\lambda)$ as an affine cone bundle over $\mathscr{B}(S \setminus \lambda)$.

Positive transverse cocycles We begin by recalling the definition of positivity for transverse cocycles, as developed in [Bonahon 1996, Section 6]. Fixing some $\lambda \in \mathcal{ML}(S)$, we recall that a transverse cocycle for λ may be identified with a relative cohomology class of the orientation cover $\widehat{N}$ of a snug neighborhood N of λ (Definition 7.2). The intersection pairing of $\widehat{N}$ therefore induces a antisymmetric bilinear pairing

$$\omega_{\mathcal{H}} \colon \mathcal{H}(\lambda) \times \mathcal{H}(\lambda) \to \mathbb{R},$$

called the *Thurston intersection/symplectic form*. This form is nondegenerate when λ is maximal and, more generally, when λ cuts S into polygons each with an odd number of sides [Penner and Harer 1992, Section 3.2].

Each transverse measure for λ is in particular a transverse cocycle. Using the intersection form one can therefore define a positive cone $\mathcal{H}^+(\lambda)$ inside $\mathcal{H}(\lambda)$ with respect to the (nonatomic) measures supported on λ. Write

$$\lambda = \lambda_1 \cup \cdots \cup \lambda_L \cup \gamma_1 \cup \cdots \cup \gamma_M,$$

where the γ_m are all weighted simple closed curves and the λ_l are minimal measured sublaminations whose supports are not simple closed curves. Then set

$$\mathscr{H}^+(\lambda) := \left\{ \rho \in \mathscr{H}(\lambda) \;\middle|\; \omega_{\mathscr{H}}(\rho, \mu) > 0 \text{ for all } \mu \in \bigcup_{l=1}^{L} \Delta(\lambda_l) \right\}, \tag{18}$$

where $\Delta(\lambda_l)$ denotes the collection of measures supported on λ_l.

The reason for this involved definition is that the Thurston form is identically 0 exactly when the underlying lamination is a multicurve. Therefore, if the support of λ contains a simple closed curve γ, the pairing of γ with every transverse cocycle supported on λ is 0.[11]

On the other hand, so long as λ is not a multicurve then the Thurston form is not identically 0. In fact, the cone $\mathscr{H}^+(\lambda)$ splits as a product

$$\mathscr{H}^+(\lambda) = \bigoplus_{l=1}^{L} \mathscr{H}^+(\lambda_l) \oplus \bigoplus_{m=1}^{M} \mathscr{H}(\gamma_m).$$

As λ supports at most $3g-3$ (projective classes of) ergodic measures, each $\mathscr{H}^+(\lambda_l)$ is a cone with a side for each (projective class of) ergodic measure supported on λ_l.

When λ is a multicurve, there are no λ_l's and so the condition of (18) is empty. As such, in this case the space of positive transverse cocycles is the entire twist space:

$$\mathscr{H}^+(\gamma_1 \cup \cdots \cup \gamma_M) = \mathscr{H}(\gamma_1 \cup \cdots \cup \gamma_M) = \bigoplus_{m=1}^{M} \mathscr{H}(\gamma_m) \cong \mathbb{R}^M.$$

Therefore, no matter whether γ is a multicurve or not, the space $\mathscr{H}^+(\lambda)$ is a convex cone of full dimension (where we expand our definition of "cone" to include the entire vector space).

Positive shear-shape cocycles We now repeat the above discussion for shear-shape cocycles. By Definition 7.5, any shear-shape cocycle $(\sigma, \underline{\alpha})$ may be identified with a relative cohomology class of the orientation cover $\widehat{N}_{\underline{\alpha}}$ of a neighborhood $N_{\underline{\alpha}}$ of $\lambda \cup \underline{\alpha}$. As above, the intersection pairing of $\widehat{N}$ then defines a pairing between any two shear-shape cocycles with underlying arc system contained inside $\underline{\alpha}$. However, if the underlying arc systems of $\sigma, \rho \in \mathcal{SH}(\lambda)$ are not nested, then there is no obvious way to pair the two cocycles.

While it does not make sense to pair two arbitrary shear-shape cocycles, we can always pair shear-shape cocycles with transverse cocycles. Recall from (the discussion before) Lemma 7.8 that $\mathscr{H}(\lambda)$ naturally embeds as a subspace of the cohomology of the neighborhood $\widehat{N}_{\underline{\alpha}}$ defining a shear-shape cocycle and

[11] This is because the components of the orientation cover are all annuli, whose first (co)homologies all have rank 1. For noncurve laminations, the homology has higher rank and so can support a nonzero intersection form.

may be identified with the kernel of the evaluation map on transversals to $\underline{\alpha}$. Therefore, the intersection pairing on $\widehat{N}_{\underline{\alpha}}$ gives rise to a function

$$\omega_{\mathcal{SH}}\colon \mathcal{SH}(\lambda) \times \mathcal{H}(\lambda) \to \mathbb{R},$$

which we also refer to as the *Thurston intersection form*. Throughout the paper, we will differentiate between the different intersection forms by indicating their domains in subscript.

We record some of the relevant properties of $\omega_{\mathcal{SH}}$ below:

Lemma 8.3 *The Thurston intersection form $\omega_{\mathcal{SH}}$ is a $\mathrm{Mod}(S)[\lambda]$–invariant continuous pairing which is homogeneous in the first factor and linear in the second. Moreover, for any $\underline{A} \in \mathscr{B}(S \setminus \lambda)$ and $\rho \in \mathcal{H}(\lambda)$, the function*

$$\omega_{\mathcal{SH}}(\cdot\,, \rho)\colon \mathcal{SH}(\lambda; \underline{A}) \to \mathbb{R}$$

is an affine homomorphism inducing $\omega_{\mathcal{H}}(\cdot\,, \rho)$ on the underlying vector space $\mathcal{H}(\lambda)$.

Proof We begin by showing that the form is actually well defined. Suppose first that $\underline{\alpha}$ is maximal; then, since the (homological) intersection form is natural with respect to deformation retracts, and any two snug neighborhoods of $\lambda \cup \underline{\alpha}$ share a common deformation retract, the form does not depend on the choice of neighborhood.

Now suppose that $\underline{\beta}$ is a filling arc system that is a subsystem of two different maximal arc systems $\underline{\alpha}_1$ and $\underline{\alpha}_2$. Then one can take a snug neighborhood $N_{\underline{\beta}}$ of $\lambda \cup \underline{\beta}$ which includes into neighborhoods N_i of $\lambda \cup \underline{\alpha}_i$ for $i = 1, 2$. Now, since the (homological) intersection form is also natural with respect to inclusions, the Thurston form must be as well. Therefore, for any $\sigma \in \mathcal{SH}(\lambda; \underline{\beta})$ and $\rho \in \mathcal{H}(\lambda)$ it does not matter if we compute $\omega_{\mathcal{SH}}(\sigma, \rho)$ in $N_{\underline{\beta}}$, N_1, or N_2.

Now that we have established that $\omega_{\mathcal{SH}}$ is well defined, the other properties follow readily from properties of the (homological) intersection form. Since the homological intersection pairing is linear in each coordinate, $\omega_{\mathcal{SH}}$ is in particular linear in the second coordinate. Similarly, for any $\underline{A} \in \mathscr{B}(S \setminus \lambda)$ and any two $\sigma_1, \sigma_2 \in \mathcal{SH}(\lambda; \underline{A})$, we know that $\sigma_1 - \sigma_2$ is a transverse cocycle, and again by linearity of the homological intersection form we get that

$$\omega_{\mathcal{SH}}(\sigma_1, \rho) - \omega_{\mathcal{SH}}(\sigma_2, \rho) = \omega_{\mathcal{H}}(\sigma_1 - \sigma_2, \rho)$$

for all $\rho \in \mathcal{H}(\lambda)$. Thus $\omega_{\mathcal{SH}}$ is affine on each $\mathcal{SH}(\lambda; \underline{A})$.

Finally, to see that the map $\omega_{\mathcal{SH}}(\cdot\,, \rho)$ is continuous for a fixed ρ, we recall that for any maximal arc system $\underline{\alpha}$, the space $\mathcal{SH}^\circ(\lambda; \underline{\alpha})$ of shear-shape cocycles with underlying arc system $\underline{\alpha}$ may be realized as an open orthant in cohomological coordinates (16), and this parametrization extends to its closure $\mathcal{SH}(\lambda; \underline{\alpha})$.

Since the intersection pairing on cohomology is continuous, for each maximal arc system $\underline{\alpha}$ the function $\omega_{\mathcal{SH}}(\cdot, \rho)$ is continuous on $\mathcal{SH}(\lambda; \underline{\alpha})$. But now, since we have checked that the value of $\omega_{\mathcal{SH}}(\cdot, \rho)$ does not actually depend on the neighborhood, it agrees on the overlaps of closures $\mathcal{SH}(\lambda; \underline{\alpha})$ for maximal $\underline{\alpha}$. Therefore, since the cell structure of $\mathscr{B}(S \setminus \lambda)$ is locally finite, we may glue together the functions $\omega_{\mathcal{SH}}(\cdot, \rho)$ (which are continuous on each $\mathcal{SH}(\lambda; \underline{\alpha})$) to get a globally continuous function on $\mathcal{SH}(\lambda)$. □

With this intersection form in hand, we may now define a positive locus with respect to the set of measures supported on λ.

Definition 8.4 The space of *positive shear-shape cocycles* $\mathcal{SH}^+(\lambda)$ is the set

$$\mathcal{SH}^+(\lambda) = \{\sigma \in \mathcal{SH}(\lambda) : \omega_{\mathcal{SH}}(\sigma, \mu) > 0 \text{ for all } \mu \in \Delta(\lambda)\}.$$

Observe the difference between the definition above and the one appearing in (18): any positive shear-shape cocycle must also pair positively with all simple closed curves γ_m appearing in the support of λ. The essential difference between the two cases is that additional branches of $\tau_{\underline{\alpha}}$ coming from the underlying arc system allows a shear-shape cocycle to meet each γ_m without being completely supported on γ_m. Indeed, one can check that the contribution to the Thurston form coming from the intersection of $\underline{\alpha}$ with a simple closed curve component of λ is always positive (compare (20)). In particular, the positivity condition is automatically fulfilled for any measure supported on a curve component of λ.

On each cohomological chart (16) or (17) it is clear that $\mathcal{SH}^+(\lambda)$ is an open cone cut out by finitely many linear inequalities (one for each ergodic measure supported on λ, plus positivity of arcs weights). However, this does not yield a global description of $\mathcal{SH}^+(\lambda)$. In order to get one, we must show that the linear subspaces cut out by the positivity conditions intersect the $\mathcal{H}(\lambda)$ fibers transversely.

Proposition 8.5 *The space $\mathcal{SH}^+(\lambda)$ is an affine cone bundle over $\mathscr{B}(S \setminus \lambda)$ with fibers isomorphic to $\mathcal{H}^+(\lambda)$.*

By an affine cone bundle, we mean that there is a (nonunique) section $\sigma_0 \colon \mathscr{B}(S \setminus \lambda) \to \mathcal{SH}(\lambda)$ such that

$$\mathcal{SH}^+(\lambda) \cap \mathcal{SH}(\lambda; \underline{A}) = \sigma_0(\underline{A}) + \mathcal{H}^+(\lambda)$$

for every $\underline{A} \in \mathscr{B}(S \setminus \lambda)$. Moreover, any two such sections differ by a continuous map $\mathscr{B}(S \setminus \lambda) \to \mathcal{H}(\lambda)$.

Proof Choose mutually singular ergodic measures $\mu_1, \dots, \mu_N, \gamma_1, \dots, \gamma_M$ on λ that span $\Delta(\lambda)$, where the supports of the μ_n are noncurve laminations and the γ_m are all simple closed curves. Pick an arbitrary $\sigma \in \mathcal{SH}(\lambda; \underline{A})$, and define

$$C(\sigma) := \{\rho \in \mathcal{H}(\lambda) \mid \omega_{\mathcal{H}}(\rho, \mu_n) > -\omega_{\mathcal{SH}}(\sigma, \mu_n) \text{ for all } n = 1, \dots, N\}.$$

By linearity of $\omega_{\mathcal{H}}$ on $\mathcal{H}(\lambda)$, together with the fact that the pairing $\omega_{\mathcal{H}}(\cdot, \mu_n)$ is not identically 0 since the support of μ_n is not a simple closed curve, this is an intersection of N affine half-spaces which do

not depend on our choice of ergodic measures μ_i in their projective classes. Again by linearity, this is just a translate of $\mathcal{H}^+(\lambda)$ and hence is a cone of full dimension.

Now, since $\omega_{\mathcal{SH}}(\cdot, \mu_j)$ is an affine map on $\mathcal{SH}(\lambda; \underline{A})$ for each j,

$$\sigma + C(\sigma) = \{\eta \in \mathcal{SH}(\lambda; \underline{A}) \mid \omega_{\mathcal{SH}}(\eta, \mu_n) > 0 \text{ for all } n = 1, \dots, N\} = \mathcal{SH}^+(\lambda) \cap \mathcal{SH}(\lambda; \underline{A})$$

is an affine cone of full dimension (where the last equality holds because the positive discussion is automatically fulfilled for each γ_m). It is a further consequence of affinity that this identification does not depend on the choice of σ. The bundle structure then follows from continuity of $\omega_{\mathcal{SH}}$. □

9 Train track coordinates for shear-shape space

In this section, we introduced train track charts for shear-shape cocycles. In Section 9.1, we recall Bonahon's realization of transverse cocycles to a lamination in the weight space of a train track that snugly carries it. In Section 9.2, we reinterpret the cohomological coordinate charts (16) for $\mathcal{SH}^\circ(\lambda; \underline{\alpha})$ by "smoothing" $\lambda \cup \underline{\alpha}$ onto a train track $\tau_{\underline{\alpha}}$ (Construction 9.3) and realizing $\mathcal{SH}^\circ(\lambda; \underline{\alpha})$ as an orthant in the weight space of $\tau_{\underline{\alpha}}$ (Proposition 9.5). This construction also has the added benefit of converting axiom (SH3) of Definition 7.11 into a simpler additivity condition; this is convenient for computations and provides an explicit formula (20) for the Thurston intersection pairing. We rely on this formula in Section 10.2 to show that foliations transverse to λ define positive shear-shape cocycles (Proposition 10.12).

Later, in Section 9.3, we explain how the PIL structure of $\mathcal{SH}(\lambda)$ is manifest in train track coordinates and provides a canonical measure in the class of Lebesgue. When λ is maximal, this measure is a constant multiple of the symplectic volume element induced by $\omega_{\mathcal{H}}$. Finally, in Section 9.4 we consider how train track charts facilitate an interpretation of $\mathcal{SH}(\lambda)$ as organizing the fragments of the cotangent space to $\mathcal{ML}$ at λ.

Remark 9.1 We advise the reader that two different types of train tracks appear below: those which carry transverse cocycles for λ and give coordinates on the fiber $\mathcal{SH}(\lambda; \underline{A})$, and those which carry shear-shape cocycles and give coordinates on the total space $\mathcal{SH}(\lambda)$.

9.1 Train track coordinates for transverse cocycles

We begin by recalling how transverse cocycles can be parametrized by weight systems on (snug) train tracks. The advantage of these coordinates is that they determine the cocycle with only finitely many values (a main benefit of the cohomological Definition 7.2), but do so using unoriented arcs on the surface, not the orientation cover (a main benefit of the axiomatic Definition 7.4).

Let τ be a train track snugly carrying a geodesic lamination λ and σ a transverse cocycle, thought of as a function on transverse arcs. For each branch b of τ, pick a tie t_b. Then one can assign to b the

weight $\sigma(t_b)$; by axiom (H1) this value does not depend on the choice of tie, and by axiom (H2) these weights necessarily satisfy the switch conditions. Therefore, any transverse cocycle can be represented by a weight system on τ, and in fact this map is an isomorphism.

Proposition 9.2 [Bonahon 1997b, Theorem 11] *Let τ be a train track snugly carrying a geodesic lamination λ. Then the map $\sigma \mapsto \{\sigma(t_b)\}_{b\in b(\tau)}$ is a linear isomorphism between $\mathcal{H}(\lambda)$ and $W(\tau)$, the space of all (real) weights on τ satisfying the switch conditions.*

On a given train track snugly carrying λ, the Thurston intersection form $\omega_{\mathcal{H}}$ is easily computable in terms of the weight systems. To wit, if $\sigma, \rho \in \mathcal{H}(\lambda)$ then their intersection is equal to

$$\omega_{\mathcal{H}}(\sigma,\rho) = \frac{1}{2}\sum_s \begin{vmatrix} \sigma(r_s) & \sigma(\ell_s) \\ \rho(r_s) & \rho(\ell_s) \end{vmatrix}, \tag{19}$$

where the sum is over all switches s of τ, and r_s and ℓ_s are the half-branches which leave s from the right and the left, respectively. Compare [Penner and Harer 1992, Section 3.2].

9.2 Train track coordinates for shear-shape cocycles

In order to imitate the above construction for shear-shape cocycles, we first must explain how to build a train track from λ and a filling arc system $\underline{\alpha}$ on its complement.

Suppose that τ carries λ snugly; then the complementary components of $\tau \cup \underline{\alpha}$ correspond to those of $\lambda \cup \underline{\alpha}$. A *smoothing of* $\tau \cup \underline{\alpha}$ is a train track $\tau_{\underline{\alpha}}$ which is obtained by choosing tangential data at each of the points of $\tau \cap \underline{\alpha}$ and isotoping each arc of $\underline{\alpha}$ to meet τ along the prescribed direction. Each component of $S \setminus \tau$ inherits an orientation from S, which in turn gives an orientation to the boundary (of the metric completion) of each subsurface. A smoothing $\tau_{\underline{\alpha}}$ is *standard* if for each switch of $\tau_{\underline{\alpha}}$ with an incoming half branch corresponding to an arc $\alpha_i \in \underline{\alpha}$, the incoming tangent vector to α_i is pointing in the positive direction with respect to the boundary orientation of the component of $S \setminus \tau$ containing α_i; see Figure 12.

Recall (Construction 5.6) that a geometric train track τ constructed from a hyperbolic structure $X \in \mathcal{T}(S)$, $\lambda \in \mathcal{ML}(S)$, and $\epsilon > 0$ is obtained as the leaf space of the orthogeodesic foliation restricted to an ϵ–neighborhood of λ in X (for small enough values of ϵ).

Construction 9.3 (geometric standard smoothings) Let $\lambda \in \mathcal{ML}(S)$ and X be a hyperbolic metric on S. Let $\underline{\alpha}$ be a filling arc system in $S \setminus \lambda$, realized orthogeodesically on X. For small enough $\epsilon > 0$, $\underline{\alpha} \cap \mathcal{N}_\epsilon(\lambda)$ lies in a finite collection of leaves of $\mathcal{O}_\lambda(X)$ and so each end of each arc of $\underline{\alpha}$ defines a point in the quotient $\tau = \mathcal{N}_\epsilon(\lambda)/\sim$, where $\sim$ is the equivalence relation induced by collapsing the leaves of $\mathcal{O}_\lambda(X)|_{\mathcal{N}_\epsilon(\lambda)}$.

The geometric standard smoothing $\tau_{\underline{\alpha}}$ is then obtained by attaching $\underline{\alpha}$ onto the geometric train track τ at these points and smoothing in the standard way.

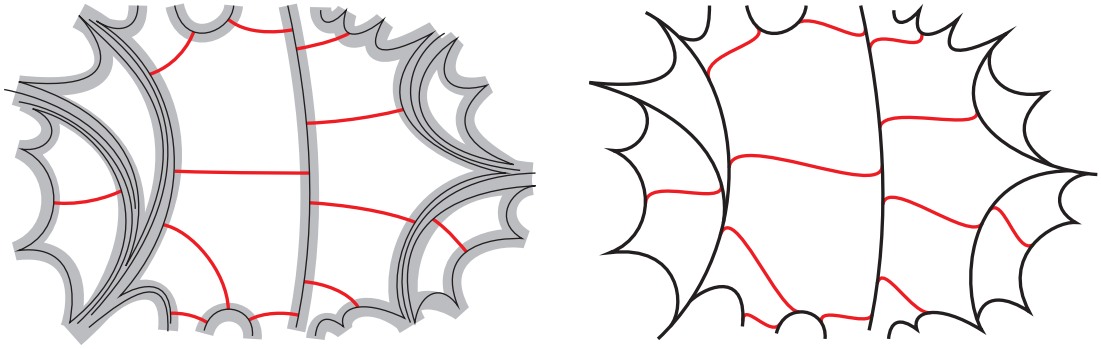


Figure 12: Left: a geometric train track neighborhood of $\widetilde{\lambda}$ together with an arc system $\underline{\widetilde{\alpha}}$. Right: the (preimage of the) standard smoothing $\tau_{\underline{\alpha}}$.

Since $\underline{\alpha}$ is filling, the components of $X \setminus (\lambda \cup \underline{\alpha})$ are topological disks. In a geometric standard smoothing $\tau_{\underline{\alpha}}$, each complementary disk incident to an arc α of $\underline{\alpha}$ has at least one spike corresponding to an end of that α. Since no arc of $\underline{\alpha}$ joins asymptotic geodesics of λ, the complementary polygons all have at least three spikes and so $\tau_{\underline{\alpha}}$ is indeed a train track.

Remark 9.4 A geometric standard smoothing keeps track of the intersection pattern of λ with $\underline{\alpha}$ on "either side" of τ, and the endpoints of $\underline{\alpha}$ on a geometric train track $\tau_\epsilon \subset X$ constructed from λ by a parameter $\epsilon > 0$ as in Construction 9.3 are stable as $\epsilon \to 0$.

A standard smoothing $\tau_{\underline{\alpha}}$ is reminiscent of the construction of completing λ to a maximal lamination λ' by "spinning" the arcs of $\underline{\alpha}$ around the boundary geodesics of complementary subsurfaces to λ in the positive direction to obtain spiraling isolated leaves of λ' in bijection with the arcs of $\underline{\alpha}$. In Proposition 9.5 below, we observe that, by smoothing $\underline{\alpha}$ onto τ in a standard way, axiom (SH3) allows us to assign weights to the branches of $\tau_{\underline{\alpha}}$ in such a way that the switch conditions are satisfied. Thus, for a shear-shape cocycle carried by $\tau_{\underline{\alpha}}$, the weights deposited on the branches $\underline{\alpha} \subset \tau_{\underline{\alpha}}$ encode "shape" data, rather than "shear" data. As such, we do not think of a standard smoothing as corresponding to the completion of λ to a maximal lamination λ'.

Proposition 9.5 *Every shear-shape cocycle $(\sigma, \underline{A}) \in \mathcal{SH}(\lambda)$ may be represented by a weight system $w_{\underline{\alpha}}(\sigma)$ on a standard smoothing $\tau_{\underline{\alpha}}$ that also carries λ. Moreover, the map $\sigma \mapsto w_{\underline{\alpha}}(\sigma)$ extends to a linear isomorphism*

$$H^1(\widehat{N}_{\underline{\alpha}}, \partial \widehat{N}_{\underline{\alpha}}; \mathbb{R})^- \cong W(\tau_{\underline{\alpha}}),$$

where $N_{\underline{\alpha}}$ is a neighborhood of $\lambda \cup \underline{\alpha}$, $\widehat{N}_{\underline{\alpha}}$ is its orientation cover and $H^1(\widehat{N}_{\underline{\alpha}}, \partial \widehat{N}_{\underline{\alpha}}; \mathbb{R})^-$ is the -1 eigenspace for the covering involution ι^*.

In particular, this isomorphism realizes $\mathcal{SH}(\lambda; \underline{\alpha})$ and $\mathcal{SH}^+(\lambda, \underline{\alpha})$ as convex cones (with some open and some closed faces) inside $W(\tau_{\underline{\alpha}})$.

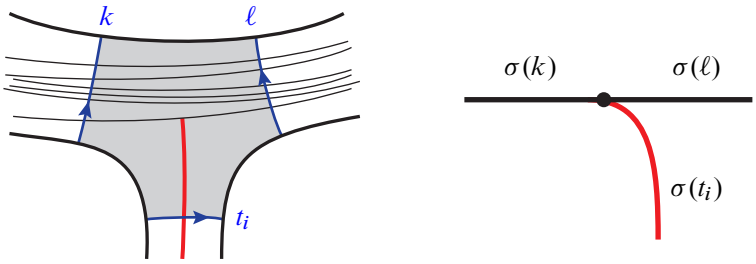


Figure 13: A standard smoothing of a geometric train track. The equation in homology encoded by axiom (SH3) becomes an additivity condition on the train track.

Proof Let $\tau_{\underline{\alpha}}$ be a standard smoothing of $\tau \cup \underline{\alpha}$ and for each branch b of $\tau_{\underline{\alpha}}$, let t_b denote a tie transverse to b. Evaluating a shear-shape cocycle σ on t_b yields an assignment of weights

$$w_{\underline{\alpha}}(\sigma) : b \to \sigma(t_b).$$

By axiom (SH1) of Definition 7.11, this weight system does not depend on the choice of tie.

To check that $w_{\underline{\alpha}}(\sigma)$ satisfies the switch conditions, we observe that there are two types of switches of $\tau_{\underline{\alpha}}$: those that come from switches of τ and those that come from smoothings of points of $\lambda \cap \underline{\alpha}$. Axiom (SH2) implies that the switch condition holds at each of the former, while axiom (SH3) together with our choice of smoothing ensures that $w_{\underline{\alpha}}(\sigma)$ satisfies the switch conditions at each of the latter. Compare Figure 13.

We note that this discussion does not rely on the positivity of σ on standard transversals, and so can be repeated to realize an arbitrary element of $H^1(\widehat{N}_{\underline{\alpha}}, \partial \widehat{N}_{\underline{\alpha}}; \mathbb{R})^-$ as a weight system on $\tau_{\underline{\alpha}}$. □

Let $\underline{A} = \sum c_i \alpha_i$; then on any smoothing $\tau_{\underline{\alpha}}$ the identification of Proposition 9.5 restricts to an isomorphism

$$\mathcal{SH}(\lambda; \underline{A}) \cong \{w \in W(\tau_{\underline{\alpha}}) : w(b_i) = c_i\},$$

where b_i is the branch of $\tau_{\underline{\alpha}}$ corresponding to α_i. Indeed, these coordinates together with the parametrization of transverse cocycles by weight systems on $\tau \prec \tau_{\underline{\alpha}}$ (Proposition 9.2) give another proof that the difference of any two shear-shape cocycles compatible with a given $\underline{A} \in \mathcal{B}(S \setminus \lambda)$ is a transverse cocycle (Lemma 7.8).

Remark 9.6 The metric residue condition (Lemma 7.9) is still visible in train track coordinates, though it is somewhat obscured. Indeed, suppose that λ contains an orientable component carried on a component ζ of the geometric train track τ; fix an arbitrary orientation of ζ.

Take a geometric standard smoothing $\tau_{\underline{\alpha}}$ of $\tau \cup \underline{\alpha}$. Reversing the tangential information as necessary, we can then construct a (nonstandard) smoothing of $\tau \cup \underline{\alpha}$ so that every arc of $\underline{\alpha}$ is a small branch entering ζ according to the orientation. Moreover, by reversing the sign of the weight on each arc which

has its smoothing data modified, this nonstandard smoothing still carries shear-shape cocycles as a weight systems. But then, by conservation of mass, the total sum of the weights on the branches entering ζ must be 0. Hence, in this setting, the metric residue condition manifests as a condition embedded in the recurrence structure of smoothings.

The extended intersection form on $\mathcal{SH}(\lambda)$ also has a nice formula in terms of train tracks. Let τ be a (trivalent) train track snugly carrying λ and let $\tau_{\underline{\alpha}}$ be a standard smoothing of $\tau \cup \underline{\alpha}$; then, for $\sigma \in \mathcal{SH}(\lambda)$ and $\rho \in \mathcal{H}(\lambda)$,

$$\omega_{\mathcal{SH}}(\sigma, \rho) = \frac{1}{2} \sum_s \begin{vmatrix} \sigma(r_s) & \sigma(\ell_s) \\ \rho(r_s) & \rho(\ell_s) \end{vmatrix}, \tag{20}$$

where the sum is over all switches s of $\tau_{\underline{\alpha}}$, and r_s and ℓ_s are the right and left small half-branches, respectively. The proof of this formula is the same as that of (19) and is therefore omitted; the only thing to note in this case is that the value does not change if one completes $\underline{\alpha}$ by adding in arcs of zero weight.

9.3 Piecewise-integral-linear structure

A piecewise-linear manifold is said to be *piecewise-integral-linear* or *PIL* with respect to a choice of charts if the transition functions are invertible piecewise-linear maps with integral coefficients. The track charts that we have constructed from standard smoothings in this section endow each cell $\mathcal{SH}(\lambda; \underline{\alpha})$ with a PIL structure which clearly extends over all of $\mathcal{SH}(\lambda)$ (compare [Penner and Harer 1992, Section 3.1]).

The points of the integer lattice in $W(\tau_{\underline{\alpha}})$ are invariant under coordinate transformation; thus, the *integer points* $\mathcal{SH}_{\mathbb{Z}}(\lambda) \subset \mathcal{SH}(\lambda)$ are well defined.

The PIL structure defined by train track charts gives a canonical measure $\mu_{\mathcal{SH}}$ in the class of the $(6g-6)$–dimensional Lebesgue measure on $\mathcal{SH}(\lambda)$. Namely, if $B \subset \mathcal{SH}(\lambda)$ is a Borel set, then

$$\mu_{\mathcal{SH}}(B) := \lim_{R \to \infty} \frac{\#\, R \cdot B \cap \mathcal{SH}_{\mathbb{Z}}}{R^{6g-6}}. \tag{21}$$

Since the symplectic intersection form $\omega_{\mathcal{SH}}$ is constant (19) in a train track chart, the volume element defined by the $(3g-3)$–fold wedge product $\wedge\omega_{\mathcal{SH}}$ is a constant multiple of $\mu_{\mathcal{SH}}$ on each chart.

We note that $\mathcal{B}(S \setminus \lambda)$ is cut out of $|\mathcal{A}_{\text{fill}}(S \setminus \lambda)|$ by linear equations with integer coefficients, as is each cell of $|\mathcal{A}_{\text{fill}}(S \setminus \lambda)|$. Therefore, the integer lattice $\mathcal{SH}_{\mathbb{Z}}(\lambda)$ restricts to a integer lattice in the bundle $\mathcal{SH}(\lambda; \underline{\alpha})$ over every cell $\mathcal{B}(\underline{\alpha})$. Thus we obtain a natural volume element on the bundle over the k–skeleton of $\mathcal{B}(S \setminus \lambda)$ whenever it is not empty.

9.4 Duality in train track coordinates

We now take a moment to discuss shear-shape coordinates from the point of view of train track weight spaces; this discussion is motivated by that in [Thurston 1986], and is meant to clarify how shear-shape cocycles fit into the broader theory of train tracks.

We begin by recalling the analogy between shear coordinates for Teichmüller space and the "horospherical coordinates" for hyperbolic space. As observed by Thurston [1986, page 42], projecting the Lorentz model

$$\mathbb{H}^n = \{x_1^2 + \cdots + x_n^2 - x_{n+1}^2 = -1 \mid x_{n+1} > 0\}$$

to $\langle x_1, \ldots, x_n \rangle$ along a family of parallel light rays gives a parametrization for $\mathbb{H}^n$ in terms of a half-space. In these coordinates, horospheres based at the boundary point $\xi \in \partial_\infty \mathbb{H}^n$ corresponding to the choice of light ray are mapped to affine hyperplanes and geodesics from ξ are mapped to rays from the origin.[12]

When λ is maximal and uniquely ergodic, Bonahon and Thurston's shear coordinates similarly realize $\mathcal{T}(S)$ as the space of positive transverse cocycles $\mathscr{H}^+(\lambda)$, in which planes parallel to the boundary are level sets of the hyperbolic length of λ and rays through the origin are Thurston geodesics. Equivalently, if τ is a train track carrying λ, then shear coordinates identify $\mathcal{T}(S)$ as a half-space inside $W(\tau)$.

However, shear coordinates are no longer induced by a global projection. Instead, as noted by Thurston, they can be thought of as a map that takes a hyperbolic structure X to (the 1–jet of) its length function with respect to a given lamination. Shear coordinates are then a map not into $W(\tau)$ but into its dual space $W(\tau)^*$ (which can be identified with $W(\tau)$ via the nondegenerate Thurston symplectic form). The image cone is then the positive dual[13] of the cone of measures on λ.

This formalism then indicates how shear coordinates generalize to maximal but nonuniquely ergodic laminations. The map is the same, but now the positive dual of $\Delta(\lambda)$ has angles obtained from the intersection of hyperplanes, one for each ergodic measure on λ. Rays in the cone still correspond to geodesics, and affine planes parallel to the bounding planes correspond with the level sets of hyperbolic length of the ergodic measures on λ.

Our shear-shape coordinates come into play when λ is not maximal. In this case, one can go through the above steps for each maximal train track τ, obtained from a snug train tack carrying λ by adding finitely many branches. Since λ is carried on a proper subtrack of τ its cone of measures lives in a proper subspace $E \subset W(\tau)$. Taking the positive dual of $\Delta(\lambda)$ and applying the isomorphism $W(\tau) \cong W(\tau)^*$ induced by the Thurston form then realizes Teichmüller space as a cone C in $W(\tau)$. By definition, $C \cap E$ is exactly $\mathscr{H}^+(\lambda)$, and one can check this demonstrates C is an affine $\mathscr{H}^+(\lambda)$–bundle.

However, the base of this bundle structure is not canonically determined, in part because $E < W(\tau)$ is generally not symplectic. Moreover, the same hyperbolic structure is parametrized by elements in many different maximal completions, and to achieve $\mathrm{Mod}(S)$–equivariance one needs to understand how to compare coordinates for different completions. Shear-shape space is designed to solve both of these

[12] This coordinate system is in some sense dual to the paraboloid model of [Thurston 1997, Problem 2.3.13]. Horospherical coordinates place an observer looking out from the center of a family of expanding horospheres, whereas the paraboloid model places an observer at another boundary point looking in.

[13] That is, those elements of $W(\tau)^*$ which pair positively with every element in $\Delta(\lambda)$ via the intersection form.

problems, picking out geometrically meaningful completions and gluing together the corresponding cones all while preserving the bundle structure.

Indeed, the shear-shape coordinates defined in Section 13 associate to each hyperbolic structure a natural finite set of completions (corresponding to standard smoothings of snug train tracks plus geometric arc systems) together with a weight system on each completion. The discussion of this section (Proposition 9.5 especially) then implies that the associated shear-shape cocycle is independent of the choice of completion, and that the corresponding train track charts glue together according to the combinatorics of $\mathscr{B}(S \setminus \lambda)$. In this picture, level sets of the hyperbolic length now correspond to bundles over $\mathscr{B}(S \setminus \lambda)$ whose fibers are affine subspaces parallel to the boundary of $\mathscr{H}^+(\lambda)$, while rays in $\mathscr{SH}^+(\lambda)$ correspond to scaling both the coordinate in $\mathscr{B}(S \setminus \lambda)$ as well as the coordinate in $\mathscr{H}^+(\lambda)$.

10 Shear-shape coordinates for transverse foliations

We now show how the familiar period coordinates for a stratum of quadratic differentials can be re-interpreted as shear-shape coordinates. The main construction of this section is that of the map

$$I_\lambda : \mathscr{F}^{uu}(\lambda) \to \mathscr{SH}(\lambda)$$

which records the vertical foliation of a quadratic differential and should be thought of as a joint extension of [Mirzakhani 2008, Theorem 6.3; Minsky and Weiss 2014, Theorem 1.2].

The idea is straightforward: Given some quadratic differential $q \in \mathscr{F}^{uu}(\lambda)$, the complement $S \setminus Z(q)$ of its zeros deformation retracts onto a neighborhood $N_{\underline{\alpha}(q)}$ of $\lambda \cup \underline{\alpha}(q)$ for some filling arc system $\underline{\alpha}(q)$ (whose topological type reflects the geometry of q). We may therefore identify the period coordinates of q as a relative cohomology class in (the orientation cover of) $N_{\underline{\alpha}(q)}$ with complex coefficients. The imaginary part of this class corresponds to λ, while its real part is the desired shear-shape cocycle $I_\lambda(q)$.

The only obstacle to this plan is in showing that $S \setminus Z(q)$ can actually be identified with a neighborhood of $\lambda \cup \underline{\alpha}(q)$. To overcome this, we recall first in Section 10.1 how to reconstruct the topology of $S \setminus \lambda$ from the horizontal separatrices of q; this guarantees that all relevant objects have the correct topological types. We then describe in Section 10.2 how to build from $S \setminus Z(q)$ a train track $\tau_{\underline{\alpha}}$ snugly carrying $\lambda \cup \underline{\alpha}(q)$ (Lemma 10.6); this in particular allows us to identify $S \setminus Z(q)$ as a neighborhood of $\lambda \cup \underline{\alpha}(q)$. We may then define $I_\lambda(q)$ using the strategy outlined above and identify it as a weight system on $\tau_{\underline{\alpha}}$ (Lemma 10.10).

Section 10.3 contains a discussion of the global properties of the map I_λ: piecewise-linearity, injectivity, and its behavior with respect to the intersection pairing. In this section, we also record Theorem 10.15, which states that I_λ is a homeomorphism onto $\mathscr{SH}^+(\lambda)$. For purposes of convenience, the proof of this theorem is deduced from our later (logically independent) work on shear-shape coordinates for hyperbolic structures (Sections 12–15). See Remark 10.16.

10.1 Separatrices and arc systems

Given a quadratic differential with $|\mathrm{Im}(q)| = \lambda$, our first task towards realizing $|\mathrm{Re}(q)|$ as a shear-shape cocycle is to build a filling arc system $\underline{\alpha}(q)$ on $S \setminus \lambda$ that encodes the horizontal separatrices of q. We begin by recalling how to recover the topology of $S \setminus \lambda$ from the realization of λ as a measured foliation on q.

Recall that a *boundary leaf* ℓ of a component of $S \setminus \lambda$ is a complete geodesic contained in its boundary. Note that infinite boundary leaves of $S \setminus \lambda$ are in one-to-one correspondence with leaves of λ which are isolated on one side, while finite boundary leaves (ie closed boundary components) are in two-to-one correspondence with closed leaves of λ.[14]

The corresponding notion for measured foliations is that of *singular leaves*. Let $\mathcal{F}$ be a measured foliation on S and $\widetilde{\mathcal{F}}$ denote its full preimage to $\widetilde{S}$ under the covering projection; then a bi-infinite geodesic path of horizontal separatrices ℓ is a singular leaf of $\widetilde{\mathcal{F}}$ if, for every saddle connection s of ℓ, the separatrices adjacent to s leave from the same side of ℓ (ie always from the left or always from the right); see [Levitt 1983, Figure 2].

There is a fundamental correspondence between boundary leaves of a lamination and singular leaves of a foliation, which we record below. Heuristically, collapsing the complementary regions of a lamination yields a foliation; the deflation map of Section 5.3 is a geometric realization of this phenomenon. Again, compare [Levitt 1983, Figure 2] as well as [Minsky 1992, Lemma 2.1].

Lemma 10.1 *Let λ be a measured lamination on S and let $\mathcal{F}$ be a measure-equivalent measured foliation. Then there is a one-to-one, $\pi_1(S)$–equivariant correspondence between the boundary leaves of $\widetilde{S} \setminus \widetilde{\lambda}$ and singular leaves of $\widetilde{\mathcal{F}}$. Moreover, singular leaves of $\widetilde{\mathcal{F}}$ that share a common separatrix correspond to boundary leaves of the same component of $\widetilde{S} \setminus \widetilde{\lambda}$.*

This lemma in particular allows us to read off the topological type of $S \setminus \lambda$ from the horizontal separatrices of q. Set $\Xi(q)$ to be the union of the horizontal separatrices of q, equipped with the path metric. This 1–complex also comes equipped with a ribbon structure (that is, a cyclic ordering of the edges incident to each vertex) and, by thickening each component of $\Xi(q)$ according to this ribbon structure, $\Xi(q)$ can be regarded as a spine for the components of $S \setminus \lambda$.

Our construction of $\underline{\alpha}(q)$ then records the dual arc system to the spine $\Xi(q)$ of $S \setminus \lambda$.

Construction 10.2 Let q be a quadratic differential on S with $|\mathrm{Im}(q)| = \lambda$. By the correspondence of Lemma 10.1, each horizontal separatrix of q corresponds to a pair of boundary leaves of the same component of $S \setminus \lambda$. Each infinite separatrix corresponds to a pair of asymptotic boundary leaves, while nonasymptotic boundary leaves are glued along horizontal saddle connections. Dual to each horizontal saddle connection of $\Xi(q)$ is a proper isotopy class of arcs on $S \setminus \lambda$, and we set $\underline{\alpha}(q)$ to be the union of all of these arcs.

[14] This is true because we have insisted that λ support a measure, and so no nonclosed leaf may be isolated from both sides.

Since $\Xi(q)$ is a spine for $S \setminus \lambda$ and $\underline{\alpha}(q)$ consists of arcs dual to its compact edges, we quickly see that:

Lemma 10.3 *The arcs of $\underline{\alpha}(q)$ are disjoint and fill $S \setminus \lambda$.*

Proof Each component of $\widetilde{S} \setminus \widetilde{\lambda}$ has a deformation retract onto the universal cover $\widetilde{\Xi}$ of a component of $\Xi(q)$. In particular, as the interiors of the edges of $\widetilde{\Xi}$ are disjoint, duality implies that the arcs of $\widetilde{\underline{\alpha}}(q)$ can all be realized disjointly. As this picture is invariant under the covering transformation, this implies that the arcs are disjoint downstairs in $S \setminus \lambda$.

Similar considerations also imply that the arc system is filling: let Σ be a component of $S \setminus \lambda$ with universal cover $\widetilde{\Sigma}$ with spine $\widetilde{\Xi}$. By construction, the edges of $\widetilde{\underline{\alpha}}(q)$ in $\widetilde{\Sigma}$ are dual to the edges of $\widetilde{\Xi}$. Since $\Xi(q)$ is a spine for $S \setminus \lambda$, any loop in Σ is homotopic to a union of saddle connections, implying that any nontrivial loop must pass through an edge of $\underline{\alpha}(q)$. Hence, $\underline{\alpha}(q)$ fills $S \setminus \lambda$. □

10.2 Period coordinates as shear-shape cocycles

Now that we understand the relationship between λ and the horizontal data of q, it is easy to build objects $\mathsf{T}^* \setminus \mathsf{H}^*$ and T^* on q of the same topological type as λ and $\lambda \cup \underline{\alpha}(q)$. However, it is not immediate to actually identify these objects as neighborhoods of λ and $\lambda \cup \underline{\alpha}(q)$. Below, we deduce this from the stronger statement that they admit smoothings onto train tracks snugly carrying λ and $\lambda \cup \underline{\alpha}(q)$; compare [Mirzakhani 2008, Sections 5.2 and 5.3].

Construction 10.4 (train tracks from triangulations) Let H denote the set of all horizontal saddle connections on q and let T be a triangulation of q containing H. Let T^* be the 1–skeleton of the dual complex to T and let H^* denote the edges of T^* dual to H. Note that T^* is trivalent by definition.

Let Δ denote a triangle of T with dual vertex v_Δ in T^*. Using the $|q|$–geometry of Δ we may assign tangential data to v_Δ as follows (compare Figures 14 and 15):

- If no edge of Δ is horizontal, then a unique edge e has largest (magnitude of) imaginary part. Assign tangential data to v_Δ so that the dual edge to e is a large half-branch.
- Otherwise, some edge of Δ is horizontal and the other two edges have the same imaginary parts. In this case, we choose tangential data so that the horizontal edge corresponds to a small half-branch and leaves the large half-branch from the right, as seen by the large half-branch.

We denote the resulting train track by $\tau_{\underline{\alpha}}$. The subgraph $\mathsf{T}^* \setminus \mathsf{H}^*$ can also be converted into a train track τ by deleting the branches of $\tau_{\underline{\alpha}}$ dual to H.

Remark 10.5 The edges of H^* correspond to the arcs of $\underline{\alpha}(q)$ and $\tau_{\underline{\alpha}}$ is a standard smoothing of $\tau \cup \underline{\alpha}(q)$. Our convention for "standard" ensures that additivity in period coordinates corresponds to additivity in train track coordinates.

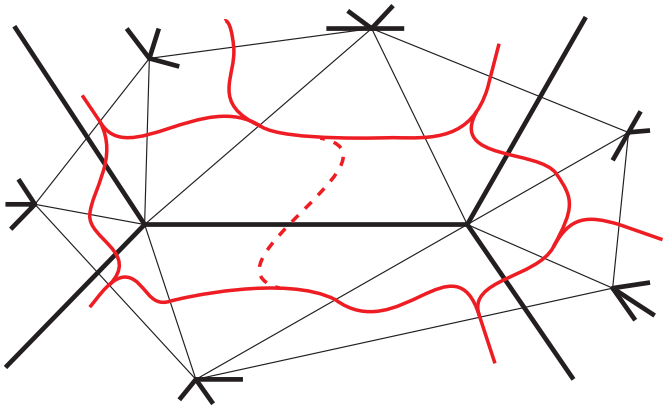


Figure 14: An example of the train track $\tau_{\underline{\alpha}}$ around a saddle connection. The thick black lines are stems of horizontal separatrices of q while the light black lines are nonhorizontal edges of the triangulation T. The dashed line is a branch of $\tau_{\underline{\alpha}} \setminus \tau$.

By construction, the graph T^* (equivalently, the train track $\tau_{\underline{\alpha}}$) is a deformation retract of $S \setminus Z(q)$. Similarly, $\mathsf{T}^* \setminus \mathsf{H}^*$ (and τ) are deformation retracts of the complement of the horizontal saddle connections. Together with our discussion above, this implies that τ has the same topological type as λ and $\tau_{\underline{\alpha}}$ has the same topological type as $\lambda \cup \underline{\alpha}(q)$.

In order to actually realize these objects as neighborhoods of λ, we observe that we can build an explicit carrying map from (a foliation measure equivalent to) λ onto τ.

Lemma 10.6 *The train track τ carries λ snugly. The weight system on τ that specifies λ is exactly the (magnitude of) the imaginary parts of the periods of the edges of T.*

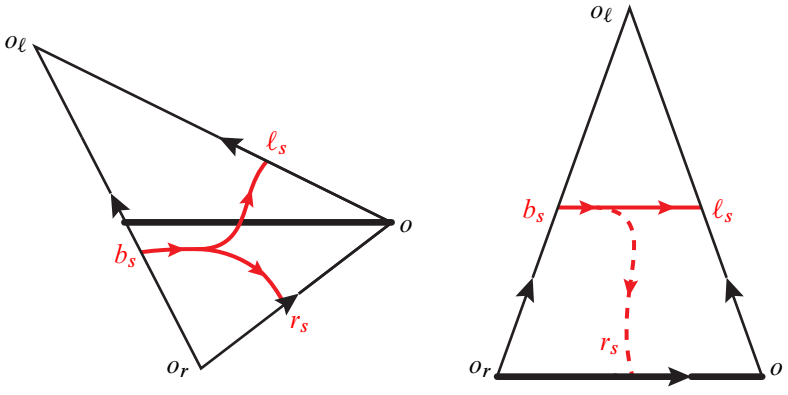


Figure 15: Local pictures of the different types of switches of $\tau_{\underline{\alpha}}$. Here we have illustrated the images of each triangle under the holonomy map. The orientation of each edge should be interpreted as indicating the value of $[\cdot]_+$, so that the edge vector is exactly the complex weight assigned to the dual branch of $\tau_{\underline{\alpha}}$. The graphical conventions of this figure mirror those of Figure 14.

Proof Let all notation be as above and let $\mathcal{F}$ denote the (singular) horizontal foliation of q.

One can directly build a homotopy of the nonsingular leaves of $\mathcal{F}$ onto τ: in a neighborhood of each edge e of $\mathsf{T} \setminus \mathsf{H}$ there is a homotopy of the leaves of $\mathcal{F}$ onto the branch of τ dual to e. Now any leaf of $\mathcal{F}$ which passes through a triangle Δ of T does so (locally) only twice and must pass through the side of Δ with the largest imaginary part, which corresponds to a large half-branch of τ. The complement of the separatrix meeting the vertex of Δ opposite to the side with largest imaginary part separates the (locally) nonsingular leaves of $\mathcal{F}$ passing through Δ into two packets that can be homotoped onto τ, respecting the smooth structure at the switch dual to Δ; compare Figure 15.

Now the horizontal foliation $\mathcal{F}$ of q is measure equivalent to λ, and so as τ carries $\mathcal{F}$ it carries λ (snugness follows as τ and λ have the same topological type). The statement about the weight system follows from our description of the carrying map. □

Now that we have identified τ as a snug train track carrying λ, we may in turn identify a neighborhood of $\lambda \cup \underline{\alpha}(q)$ with (a thickened neighborhood of) $\tau_{\underline{\alpha}}$. With this correspondence established, we may now define $I_\lambda(q)$ as the image of the real part of the period coordinates of q under the natural isomorphism on cohomology.

Construction 10.7 (definition of $I_\lambda(q)$) Let S, λ, q, $\underline{\alpha}(q)$ and $\tau_{\underline{\alpha}}$ be as above, Set $M_{\underline{\alpha}}$ to be a thickened neighborhood of $\tau_{\underline{\alpha}}$ (in the flat metric defined by q) and let $N_{\underline{\alpha}}$ be a snug neighborhood of $\lambda \cup \underline{\alpha}(q)$ (taken in some auxiliary hyperbolic metric). Perhaps by shrinking $N_{\underline{\alpha}}$, we may assume it embeds into $M_{\underline{\alpha}}$ as a deformation retract (this follows by snugness).

Now $\tau_{\underline{\alpha}}$ is itself a deformation retract of $S \setminus Z(q)$, so the inclusion $M_{\underline{\alpha}} \hookrightarrow S \setminus Z(q)$ is a homotopy equivalence; composing inclusions $N_{\underline{\alpha}} \hookrightarrow M_{\underline{\alpha}} \hookrightarrow S \setminus Z(q)$ and lifting to the orientation covers yields the isomorphism

$$H^1(\widehat{S}, Z(\sqrt{q}); \mathbb{C}) \xrightarrow{j^*} H^1(\widehat{N}_{\underline{\alpha}}, \partial \widehat{N}_{\underline{\alpha}}; \mathbb{C}), \tag{22}$$

where the hats denote the corresponding orientation covers. As the composite retraction respects the covering involution ι, this isomorphism also identifies -1 eigenspaces for ι^*. We therefore define

$$I_\lambda(q) = \mathrm{Re}(j^* \,\mathrm{Per}(q)),$$

where $\mathrm{Per}(q)$ are the period coordinates for q, and where the real part is taken relative to the natural splitting $\mathbb{C} = \mathbb{R} \oplus i\mathbb{R}$.

Remark 10.8 From the above construction, a basis consisting of branches for the weight space of $\tau_{\underline{\alpha}}$ (equivalently a basis for $H_1(\widehat{N}_{\underline{\alpha}}, \partial \widehat{N}_{\underline{\alpha}}; \mathbb{Z})$ of dual arcs) picks out a basis for $H_1(\widehat{S}, Z(\sqrt{q}); \mathbb{Z})$. Moreover, each relative cycle is realized geometrically as a saddle connection (as opposed to concatenations, thereof).

To see that $I_\lambda(q)$ is indeed a shear-shape cocycle, we need only observe that the values on standard transversals to $\underline{\alpha}(q)$ are all positive. This follows essentially by definition of the orientation cover and construction of $\underline{\alpha}(q)$. To wit: if α is an arc of $\underline{\alpha}(q)$ dual to a saddle connection s, and t is a standard transversal to α, then the canonical lifts of t are mapped to those of s under the isomorphism (22). As the periods of $\sqrt{q}$ increase as you move along the (oriented) horizontal foliation of $(\widehat{S}, \sqrt{q})$, this implies that the value of $I_\lambda(q)$ on either of the lifts of t is exactly the length of the saddle connection s.

Therefore, the weighted arc system underlying $I_\lambda(q)$ is none other than

$$\underline{A}(q) := \sum_{\alpha \in \underline{\alpha}(q)} c_\alpha \alpha,$$

where c_α is the $|q|$–length of the horizontal saddle connection dual to the arc α.

Remark 10.9 Naturality of all of the isomorphisms involved quickly implies that this construction does not depend on the choice of initial triangulation T. Indeed, suppose that T_1 and T_2 are two triangulations giving rise to train tracks τ_1 and τ_2 and hence shear-shape cocycles σ_1 and σ_2. Since both τ_i carry $\lambda \cup \underline{\alpha}(q)$ snugly, Lemma 10.6 implies that they have a common refinement τ. Lifting the inclusions

$$N(\tau \cup \underline{\alpha}(q)) \hookrightarrow N(\tau_i \cup \underline{\alpha}(q)) \hookrightarrow S \setminus Z(q)$$

to their orientation covers and drawing the appropriate commutative diagram of cohomology groups, the shear-shape cocycles built from each T_i coincide as weight systems on the common refinement τ.

For use in the sequel, we record below the weight systems on $\tau_{\underline{\alpha}}$ corresponding to λ and $I_\lambda(q)$. The proof follows by combining the constructions above with the discussion in Section 9 and is therefore left to the scrupulous reader. See also Figure 15.

For a complex number z, define

$$[z]_+ = \begin{cases} z & \text{if } \arg(z) \in [0, \pi), \\ -z & \text{if } \arg(z) \in [\pi, 2\pi). \end{cases}$$

Observe that $[z]_+ = [-z]_+$ for all $z \in \mathbb{C}$.

Lemma 10.10 *Let all notation be as above and, for each edge e of T, let b_e denote the branch of $\tau_{\underline{\alpha}}$ dual to it. Then the assignment*

$$b_e \mapsto \left[\int_e \sqrt{q} \right]_+$$

defines a complex weight system $w(q)$ on $\tau_{\underline{\alpha}}$ satisfying the switch conditions. Moreover,

$$\mathrm{Im}(w(q)) = \lambda \quad \textit{and} \quad \mathrm{Re}(w(q)) = I_\lambda(q).$$

10.3 Global properties of the coordinatization

In this section, we show that the map I_λ defined above gives a global coordinatization of $\mathcal{MF}(\lambda) \cong \mathcal{F}^{uu}(\lambda)$. First, we record certain global properties of this map; as it is defined by reinterpreting period coordinates as shear-shape cocycles, it preserves many of the structures imposed by period coordinates.

For example, it follows by construction that I_λ respects the stratification of each space. That is, if $q \in \mathcal{QT}(k_1, \dots, k_n) \cap \mathcal{F}^{uu}(\lambda)$, then the spine dual to $\underline{\alpha}(q)$ has vertices of valence $k_1 + 2, \dots, k_n + 2$. In a similar vein, since both $\mathcal{F}^{uu}(\lambda)$ and $\mathcal{SH}(\lambda)$ have local cohomological coordinates (which induce PIL structures) we can deduce the following:

Lemma 10.11 *For any $\lambda \in \mathcal{ML}(S)$, the map I_λ is $\mathrm{Mod}(S)[\lambda]$–equivariant and PIL.*[15]

Proof Equivariance follows from the naturality of our construction: all combinatorial data (arc systems, train tracks, etc) can be pulled back to a reference surface equipped with λ, so changing the marking by an element of $\mathrm{Mod}(S)[\lambda]$ acts by transforming the combinatorial data on the reference surface.

The piecewise-linear structure on $\mathcal{F}^{uu}(\lambda)$ (respectively $\mathcal{SH}(\lambda)$) is given by period coordinates (respectively cohomological coordinates in a neighborhood/train track coordinates) and so the map is by construction piecewise-linear. Integrality comes from the fact that a homotopy equivalence induces an isomorphism on cohomology with $\mathbb{Z}$–coefficients, and hence takes integral points to integral points. □

The Thurston intersection pairing gives us a powerful tool to understand constraints on the image of I_λ; in particular, $I_\lambda(q)$ must be a *positive* shear-shape cocycle. Indeed, the tangential structure of the train track $\tau_{\underline{\alpha}}$ at each switch provides us with an identification of each triangle Δ of T with an oriented simplex. With respect to this orientation, we can compute the area of Δ by taking (one half of) the cross product of two of its sides. Comparing the formula for the cross product with the Thurston intersection pairing (20) then allows us to see that the intersection of λ and $I_\lambda(q)$ is exactly the area of q; compare [Mirzakhani 2008, Lemma 5.4].

Proposition 10.12 *For all $\eta \in \mathcal{MF}(\lambda)$ and all $\mu \in \Delta(\lambda)$,*

$$\omega_{\mathcal{SH}}(I_\lambda(\eta), \mu) = i(\eta, \mu).$$

In particular, $I_\lambda(\mathcal{MF}(\lambda)) \subseteq \mathcal{SH}^+(\lambda)$.

The proof of this proposition is made technical by the fact that if μ and $\mu' \in \Delta(\lambda)$ are ergodic but not projectively equivalent then they are mutually singular. To deal with this difficulty, we build a flat structure on the subsurface filled by μ by integrating against $\lambda + t\mu$ and $I_\lambda(\eta)$ for small t. The triangulation T then induces a combinatorially equivalent triangulation of this new flat structure by saddle connections, allowing us to compare the area of this new flat metric (computed via cross products) with the Thurston form on our original train track $\tau_{\underline{\alpha}}$. This inverse construction will also be used in the proof of Proposition 10.14.

Proof We begin by observing that since $\mu \in \Delta(\lambda)$, there is a union of minimal components of the horizontal foliation of $q(\eta, \lambda)$ that supports μ. Call this subfoliation $\mathcal{F}$ and let Y denote the subsurface filled by $\mathcal{F}$ on $q(\eta, \lambda)$. Note that ∂Y must be a union of horizontal saddle connections, and hence is contained in any triangulation T used to define $\tau_{\underline{\alpha}}$. In particular, $\mathsf{T}|_Y$ is a triangulation of Y.

[15] We recall that a PL map between PIL manifolds is itself PIL if it sends integral points to integral points.

Since η and λ are realized transversely on $q(\eta, \lambda)$ and this specific realization of η is nonatomic (as any closed leaves of η have become vertical cylinders), we can compute the intersection number between η and any measure μ supported on $\mathcal{F}$ as

$$i(\eta, \mu) = \int_S \eta \times \mu = \int_Y \eta \times \mu. \tag{23}$$

We now build a new flat structure on Y whose conical singularities coincide with those of Y; the salient feature is that $\tau|_Y$ can be straightened out to a triangulation by saddle connections on the new singular flat structure that reflects the geometry of $\lambda + t\mu$. To construct the new singular flat structure, we build charts from a neighborhood of each triangle $\Delta \subset \tau|_Y$ to $\mathbb{C}$ and describe the transitions.

Each triangle Δ of τ is dual to a switch s with an edge that is dual to a large half-branch b incident to s. Orient $\tau_{\underline{\alpha}} \cap \Delta$ so that a train traveling along b toward s is moving in the positive direction. The other edges r and ℓ of Δ are dual to the half-branches of $\tau_{\underline{\alpha}}$ to the right and left of s, respectively. The vertices o_r and o_ℓ are adjacent to r and ℓ, respectively, and the vertex o is opposite b; see Figure 15. On the interior of each triangle Δ, we orient the leaves of $\mathcal{F}$ parallel to b. The leaves of η are given the orientation such that the ordered basis of tangent vectors to λ and η at each point agree with the underlying orientation of S. With this orientation, the measures η and λ induce smooth real 1–forms $d\eta$ and $d\lambda$ that look locally like dx and dy, respectively (as opposed to $|dx|$ and $|dy|$, respectively).

Restricted to the interior of Δ, the local orientation of the leaves of η also gives the measure μ the structure of a measurable 1–form that we call $d\mu$. Spreading out the measure on a closed leaf of μ over the horizontal cylinder of λ corresponding to its support as necessary, we get that the map

$$F_t \colon \Delta \to \mathbb{C}, \quad p \mapsto \int_{\gamma_p} d\eta + i\, d(\lambda + t\mu),$$

obtained by integrating along a path γ_p from o_r to p is isometric along leaves of $\mathcal{F}$ and nondecreasing along leaves of η. We compute

$$F_t(o) = I_\lambda(\eta)(r) + i(\lambda + t\mu)(r) \quad \text{and} \quad F_t(o_\ell) = I_\lambda(\eta)(b) + i(\lambda + t\mu)(b).$$

Transverse invariance and additivity of μ gives

$$F_t(o_\ell) - F_t(o) = I_\lambda(\eta)(\ell) + i(\lambda + t\mu)(\ell). \tag{24}$$

Since the pair $(F_0(o), F_0(o_\ell))$ forms a positively ordered basis for $\mathbb{C}$ (or, equivalently, since the triangle Δ is positively oriented), the pair $(F_t(o), F_t(o_\ell))$ is also positively oriented for small enough t. Let Δ'_t be the convex hull of $(F_t(o_r), F_t(o), F_t(o_\ell))$.

The area of Δ'_t may now be computed as half the cross product of $F_t(o)$ and $F_t(o_\ell)$. Using (24) and linearity of the cross product,

$$\operatorname{Area}(\Delta'_t) = \frac{1}{2}\begin{vmatrix} I_\lambda(\eta)(r) & I_\lambda(\eta)(\ell) \\ \lambda + t\mu(r) & \lambda + t\mu(\ell) \end{vmatrix} > 0. \tag{25}$$

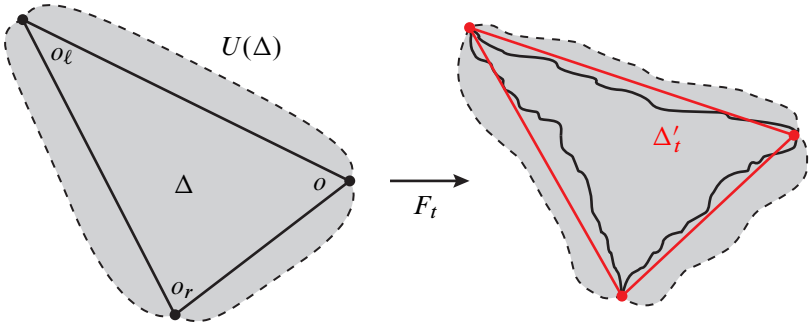


Figure 16: Integrating against η and $\lambda + t\mu$ defines a new flat structure on triangles. These charts piece together to give a new half-translation structure on the subsurface filled by μ.

Now, for each Δ and any small enough t, the map F_t may be extended to an open set $U(\Delta)$ in $Y \setminus Z(q)$ that contains Δ (minus its vertices) and is such that, for every $p \in U(\Delta)$, there is a unique nonsingular $|q|$–geodesic segment γ_p joining o_r to p. We claim that, moreover, we may choose $U(\Delta)$ so that $\Delta'_t \subset F(U(\Delta))$; see Figure 16.

If not, there is some vertex v of $\mathsf{T}_Y \setminus \Delta$ such that $F_t(v) \in \Delta'_t \setminus F_t(\Delta)$. Indeed, by construction, $U(\Delta)$ is a star-shaped neighborhood about the vertex o_r of Δ, so there is a saddle connection joining o_r to v. This saddle connection passes through or shares a vertex of an edge e of Δ. Moreover, we may find v such that the triangle Δ_v formed by e and v is singularity free and contained in $U(\Delta)$. But now, the straightening Δ'_v of $F_t(\Delta_v)$ in $\mathbb{C}$ lies inside Δ'_t with the wrong orientation since $F_t(v)$ lies between $F_t(e)$ and the corresponding edge of Δ'_t. This is a contradiction to the fact that F_t is nondecreasing along leaves of η, or, alternatively, to the fact that the straightenings Δ'_t are all positively oriented for small enough t. So we may assume that $\Delta'_t \subset F(U(\Delta))$.

If $\Delta_1 \subset \mathsf{T}_Y$ shares an edge with Δ, then the construction of the map F_t on Δ_1 agrees with F_t on $U(\Delta) \cap U(\Delta_1)$ up to multiplication by ± 1 (depending on the configuration of the switches dual to Δ and Δ_1) and translation by the period of the arc connecting the basepoints o_r of each triangle. Thus these triangles glue up to a half-translation structure on $Y \setminus Z$ equipped with a triangulation by saddle connections corresponding to $\mathsf{T}|_Y$.

In our new flat structure on Y, $\lambda + t\mu$ is measure equivalent to the horizontal foliation and (the restriction of) η is equivalent to the vertical foliation. Hence, we obtain, for any t small enough, that

$$\int_Y \eta \times (\lambda + t\mu) = \sum_{\Delta \in \mathsf{T}|_Y} \operatorname{Area}(\Delta'_t) = \sum_{\Delta \in \mathsf{T}|_Y} \frac{1}{2} \begin{vmatrix} I_\lambda(\eta)(r) & I_\lambda(\eta)(\ell) \\ \lambda + t\mu(r) & \lambda + t\mu(\ell) \end{vmatrix} = \omega_{\mathcal{FH}}(I_\lambda(\eta), \lambda + t\mu),$$

where the second equality follows from (25) and the third from (20). Combining this with formula (23) and the linearity of the Thurston intersection form (Lemma 8.3),

$$i(\eta, \mu) = \frac{1}{t}\left(\int_Y \eta \times \lambda + t\mu - \int_Y \eta \times \lambda\right) = \frac{1}{t}\big(\omega_{\mathcal{FH}}(I_\lambda(\eta), \lambda + t\mu) - \omega_{\mathcal{FH}}(I_\lambda(\eta), \lambda)\big) = \omega_{\mathcal{FH}}(I_\lambda(\eta), \mu),$$

completing the proof of the proposition. □

From the proof of Proposition 10.12, we can also extract the following, which allows us to reconstruct a (triangulated) quadratic differential from a sufficiently positive shear-shape cocycle, inverting Construction 10.4.

Lemma 10.13 *Let τ be a train track snugly carrying λ and let $\tau_{\underline{\alpha}}$ be a standard smoothing of $\lambda \cup \underline{\alpha}$. Suppose that $\sigma \in \mathcal{SH}(\lambda)$ is represented by a weight system on $\tau_{\underline{\alpha}}$ such that, at every switch s of $\tau_{\underline{\alpha}}$, the contribution*

$$\frac{1}{2}\begin{vmatrix} \sigma(r_s) & \sigma(\ell_s) \\ \lambda(r_s) & \lambda(\ell_s) \end{vmatrix}$$

of s to $\omega_{\mathcal{SH}}(\sigma, \lambda)$ is positive. Then there exists a quadratic differential $q \in \mathcal{F}^{uu}(\lambda)$ such that $I_\lambda(q) = \sigma$ and the dual triangulation to $\tau_{\underline{\alpha}}$ is realized by saddle connections on q.

Proof The assumption that the contribution at each switch is positive implies that the basis $(F(o), F(o_\ell))$ is positively oriented at each switch, and so we can build a positively oriented triangle Δ with the prescribed periods. These glue together into the desired quadratic differential. □

In particular, we can locally invert I_λ by building a quadratic differential out of triangles whose edges have specified periods, so I_λ is injective.

Proposition 10.14 *For any $\lambda \in \mathcal{ML}(S)$, the map I_λ is a homeomorphism onto its image.*

Proof To see that I_λ is injective, we observe that Lemma 10.13 provides a (left) inverse map Δ_λ to I_λ. Indeed, suppose that $\sigma = I_\lambda(q)$ for some q and pick a triangulation T as in Construction 10.4; let $\tau_{\underline{\alpha}}$ denote the dual train track. Applying Lemma 10.13 then constructs a quadratic differential q' on which each edge of T is realized as a saddle connection. Since q and q' have the same periods with respect to the same geometric triangulation, they must be equal.

To prove that I_λ is continuous, we first observe that I_λ is by definition continuous on the closure $\mathcal{SH}(\lambda; \underline{\alpha}(q))$ of any cell, as it is induced by a continuous mapping on the level of cohomology. In general, we need only exploit this fact together with a standard reformulation of sequential continuity: a function $f : X \to Y$ is continuous if and only if every convergent sequence $x_n \to x$ has a subsequence x_{n_k} such that $f(x_{n_k}) \to f(x)$.

So let $q_n \to q \in \mathcal{F}^{uu}(\lambda)$. The polyhedral structure of $\mathcal{SH}(\lambda)$ is locally finite, so, for n large enough, $I_\lambda(q_n)$ is contained in a finite union of cells. After passing to a subsequence q_{n_k}, we may assume that q_{n_k} all share the same underlying (maximal) arc system $\underline{\beta}$ completing $\underline{\alpha}$. In particular, $I_\lambda(q_{n_k}) \in \mathcal{SH}(\lambda; \underline{\beta})$ for all k and so $I_\lambda(q_{n_k}) \to I_\lambda(q)$ follows from continuity on cells. Therefore I_λ is a continuous injective map between Euclidean spaces of the same dimension (Proposition 8.5 and Corollary 8.2) and so invariance of domain guarantees it is a homeomorphism onto its image. □

The image of I_λ In light of Lemma 10.13, to show that I_λ surjects onto $\mathcal{SH}^+(\lambda)$ it would suffice to show that every positive shear-shape cocycle can be realized as a weight system on a train track where every switch contributes positively to the intersection form. However, it is rather complicated to show that every positive shear-shape cocycle admits such a representation (see the discussion in Remark 10.16).

Instead, we deduce this fact using the commutativity of (2) and the results appearing in Sections 12–15 coordinatizing hyperbolic structures by shear-shape cocycles. We emphasize, however, that Theorem 10.15 is logically independent from the work done in Sections 12–15 that leads to its proof. We include the statement here (as opposed to after Section 15) to provide some closure to our discussion of the parametrization of $\mathcal{MF}(\lambda)$ by shear-shape cocycles.

Theorem 10.15 *The map $I_\lambda : \mathcal{F}^{uu}(\lambda) \to \mathcal{SH}^+(\lambda)$ is a homeomorphism.*

Proof In Section 13, we define the geometric shear-shape cocycle $\sigma_\lambda(X) \in \mathcal{SH}(\lambda)$ associated to a hyperbolic metric $X \in \mathcal{T}(S)$ and show (Theorem 13.13) that $\sigma_\lambda(X) = I_\lambda(\mathbb{C}_\lambda(X))$. In Section 15, we prove Theorem 12.1, which states that the map $\sigma_\lambda : \mathcal{T}(S) \to \mathcal{SH}^+(\lambda)$ is a homeomorphism. In particular, σ_λ is surjective and hence so is I_λ. Together with Proposition 10.14, this implies the theorem. □

Remark 10.16 If λ is a maximal lamination, one can deduce surjectivity of I_λ by appealing to the theory of "tangential coordinates" for measured foliations transverse to λ. In general, given τ snugly carrying λ, tangential coordinates can be constructed as a quotient of $\mathbb{R}^{b(\tau)}$ by a vector subspace spanned by vectors that model the change of length of branches of a train track on either side of a switch after a small "fold" or "unzip". When λ is maximal, there is a linear isomorphism from shear coordinates to tangential coordinates via the symplectic pairing $\omega_{\mathcal{H}}$; we refer the interested reader to [Thurston 1986, Section 9] or [Penner and Harer 1992, Section 3.4] for details.

The transverse weights defined by the measure of λ on τ together with positive[16] tangential data give τ the structure of a bifoliated Euclidean band complex. If the tangential data satisfy a collection of triangle-type inequalities, this band complex can be "zipped up" to obtain a bifoliated flat surface with conical singularities. When defined, the linear transformation mapping tangential coordinates to shear coordinates preserves the intersection number, and hence positivity.

A standard positivity argument (see [Thurston 1979, Proposition 9.7.6] or [Thurston 1986, Theorem 9.3]) shows that any tangential data with positive intersection with λ has a positive representative, and hence defines a foliation transverse to λ. In particular, the map from $\mathcal{MF}(\lambda)$ to the space of tangential coordinates with positive intersection with λ is surjective. As the space of tangential coordinates with positive intersection is isomorphic to $\mathcal{H}^+(\lambda)$, this completes the proof of surjectivity in the maximal case.

This being considered, even in the case when λ is maximal "it is harder to see the [positivity] inequalities satisfied by the shear coordinates [than the tangential coordinates]" [Thurston 1986, page 45] and it is not

[16] Here positive means that there is a representative of the tangential data that is positive on each branch of τ.

clear how to run the "standard positivity argument" without passing through tangential coordinates. We have therefore chosen to prove Theorem 10.15 in a way that avoids developing a theory of tangential coordinates dual to shear-shape cocycles. Instead, we take advantage of the relationship between the Thurston intersection form on $\mathcal{SH}(\lambda)$ and the length of λ on a given hyperbolic surface, as exploited in the proof of Theorem 12.1 (see in particular Claim 15.8).

11 Flat deformations in shear-shape coordinates

The identification of Section 10 between periods of saddle connections and the values of the shear-shape cocycle $I_\lambda(q)$ immediately allows us to transport certain flows on $\mathcal{F}^{uu}(\lambda)$ to shear-shape space. Moreover, Theorem 10.15 affords a new perspective on the "tremor deformations" of [Chaika et al. 2020] (see Definition 11.3).

The horizontal stretch We begin by observing that the space $\mathcal{SH}^+(\lambda)$ carries a natural $\mathbb{R}_{>0}$–action given by scaling both the underlying arc system $\underline{A}$ and the values assigned to test arcs (equivalently, the corresponding cohomology class or the weights on a train track realization). Using our correspondence between period coordinates and shear-shape cocycles (Lemma 10.10), this dilation expands the real part of each period, so the corresponding flat deformation is just a horizontal stretch.[17]

Lemma 11.1 *Let $q \in \mathcal{F}^{uu}(\lambda)$; then*

$$I_\lambda\left(\begin{pmatrix} e^t & 0 \\ 0 & 1 \end{pmatrix} q\right) = e^t I_\lambda(q) \tag{26}$$

for all $t \in \mathbb{R}$.

In particular, our coordinatization linearizes the expansion of the strong unstable foliation under the Teichmüller geodesic flow.

Horocycle flow and tremors We now consider the horocycle flow on $\mathcal{F}^{uu}(\lambda)$, which is just the restriction of the standard horocycle flow h_s to the strong unstable leaf. An easy computation shows that, for every saddle connection e of q, one has

$$\left[\int_e \sqrt{h_s q}\right]_+ = \left(\operatorname{Re}\left[\int_e \sqrt{q}\right]_+ + s \operatorname{Im}\left[\int_e \sqrt{q}\right]_+\right) + i \operatorname{Im}\left[\int_e \sqrt{q}\right]_+ \tag{27}$$

(here we have invoked the $[\cdot]_+$ function to avoid fussing over square roots and orientations).

With the help of Lemma 10.10 we may translate this into the language of transverse and shear-shape cocycles to observe:

Lemma 11.2 *The map I_λ takes horocycle flow to translation by λ in a time-preserving way. In symbols,*

$$I_\lambda(h_s q) = I_\lambda(q) + s\lambda.$$

[17] This is just the Teichmüller geodesic flow normalized so that the horizontal foliation remains constant. Applying the standard geodesic flow takes $(I_\lambda(q), \lambda)$ to $(e^{t/2} I_\lambda(q), e^{-t/2}\lambda)$.

More generally, we can perform a similar deformation for *any* measure μ supported on λ, resulting in the *tremor flow* along μ. First defined by Chaika, Smillie and Weiss in the context of abelian differentials, the *tremor* $\mathrm{trem}_\mu(q)$ of a quadratic differential $q = q(\eta, \lambda)$ by a measure $\mu \in \Delta(\lambda)$ is the unique quadratic differential specified by shearing η by μ and leaving λ fixed. Why this makes sense (note that η and μ may not fill S) and why it can be continued for all time present significant technical challenges in [Chaika et al. 2020, Sections 4 and 13]. However, when considered in our coordinates (and restricted to a leaf of the unstable foliation), tremors become quite simple.

For a given lamination λ, let $|\Delta(\lambda)|_\pm$ denote the vector space of all signed transverse measures on λ; this is naturally a vector subspace of $\mathscr{H}(\lambda)$ of dimension at most $3g-3$ with basis consisting of the length 1 (with respect to some auxiliary hyperbolic metric) ergodic measures on λ.

Definition 11.3 Let $q \in \mathscr{F}^{uu}(\lambda)$ and let $\mu \in |\Delta(\lambda)|_\pm$. Then the tremor $\mathrm{trem}_\mu(q)$ of q along μ is the unique quadratic differential specified by

$$I_\lambda(\mathrm{trem}_\mu(q)) = I_\lambda(q) + \mu. \tag{28}$$

Note that the fact that $I_\lambda(q) + \mu \in \mathscr{SH}^+(\lambda)$ follows by affinity of the Thurston form (Lemma 8.3).

Remark 11.4 Technically, the deformation considered above is a "nonatomic tremor" in the language of [Chaika et al. 2020]. One can also consider "atomic tremors", which transform q by twisting along certain admissible loops of horizontal saddle connections.

In shear-shape coordinates, these admissible loops correspond to certain simple closed curves in the complementary subsurfaces. Atomic tremors are then realized by appropriately shearing the underlying arc system $\underline{A}(q)$ along the curves and transporting the transverse cocycle using the affine connection coming from train track coordinates. Of course, one can also define tremors along more complicated laminations contained in $S \setminus \lambda$.

For the convenience of the reader familiar with the terminology of [Chaika et al. 2020], we have included a dictionary which translates between our notation and theirs (at least when the horizontal lamination is filling — when it is not, one must replace $\Delta(\lambda)$ with a subset of the zero set of λ and take more care). See Figure 17.

We can now immediately deduce certain properties of the tremor map from the structure of $\mathscr{SH}^+(\lambda)$ and the intersection pairing. While we will not use these results in the sequel, we have chosen to include them in order to demonstrate the utility of our new perspective on these deformations. For example, using our coordinates one can easily deduce that (nonatomic) tremors leave horizontal data invariant and hence can be continued indefinitely while remaining in the same stratum.

Lemma 11.5 *For any $q \in \mathscr{F}^{uu}(\lambda)$ and $\mu \in |\Delta(\lambda)|_\pm$, the tremor path $\mathrm{trem}_{t\mu}(q)$ is defined for all time and is completely contained in $\mathscr{SH}(\lambda; \underline{A}(q))$. In particular, $\{\mathrm{trem}_{t\mu}(q)\}$ always remains in the same stratum.*

shear-shape cocycles	foliation cocycles
$\Delta(\lambda)$	C_q^+
$\lvert\Delta(\lambda)\rvert_\pm$	$\mathcal{T}_q$
$\omega_{\mathcal{SH}}(\eta,\mu) = i(\mu_+,\eta) - i(\mu_-,\eta)$	signed mass $L_q(\mu)$
$i(\mu_+,\eta) + i(\mu_-,\eta)$	total variation $\lvert L\rvert_q(\mu)$

Figure 17: Translating between our language of shear-shape cocycles and the "foliation cocycles" of [Chaika et al. 2020]. Throughout, we assume that $q = q(\eta,\lambda)$ where λ is filling (or, equivalently, q has no loops of horizontal saddle connections). We have written a signed transverse measure μ as $\mu = \mu_+ - \mu_- \in \lvert\Delta(\lambda)\rvert_\pm$, where $\mu_\pm \in \Delta(\lambda)$.

Remark 11.6 The above lemma is one specific instance of a much more general phenomenon. The global description of $\mathcal{F}^{uu}(\lambda)$ afforded by shear-shape coordinates allows one to formulate a general criterion for extending affine period geodesics, a topic which the authors hope to address in future work.

Using our interpretation of tremors as translation, it is similarly easy to describe how tremors interact with other flat deformations. Compare with [Chaika et al. 2020, Propositions 6.1 and 6.5]. We leave proofs to the reader, as they follow immediately from (28) and (26).

Lemma 11.7 *Let $q \in \mathcal{F}^{uu}(\lambda)$. Then, for any $\mu \in \lvert\Delta(\lambda)\rvert_\pm$ and for $g_t = \left(\begin{smallmatrix} e^{t/2} & 0 \\ 0 & e^{-t/2}\end{smallmatrix}\right)$,*

$$g_t \operatorname{trem}_\mu(q) = \operatorname{trem}_{e^{t/2}\mu}(g_t(q)).$$

Additionally, for any $\mu_1, \mu_2 \in \lvert\Delta(\lambda)\rvert_\pm$,

$$\operatorname{trem}_{\mu_1}(q)\operatorname{trem}_{\mu_2}(q) = \operatorname{trem}_{\mu_1+\mu_2}(q) = \operatorname{trem}_{\mu_2}(q)\operatorname{trem}_{\mu_1}(q).$$

In particular, tremors commute with the horocycle flow.

12 Shear-shape coordinates for hyperbolic metrics

We now parametrize hyperbolic structures on S by shear-shape cocycles for a measured geodesic lamination λ. With respect to the Lebesgue measure on $\mathcal{ML}(S)$, the generic lamination cuts a hyperbolic surface into ideal triangles. As all ideal triangles are isometric, Bonahon and Thurston's shearing coordinates need only take into account the "shear" between pairs of complementary triangles to describe a hyperbolic structure. As our objective is to generalize these coordinates to laminations with arbitrary topology, we must therefore combine the data of the geometry of hyperbolic metrics in complementary subsurfaces with the shearing data between them. The shear-shape space $\mathcal{SH}(\lambda)$ is well suited to this task.

In Sections 13–15, we explain how to associate a "geometric shear-shape cocycle" to a hyperbolic metric and prove that the space of positive shear-shape cocycles coordinatizes Teichmüller space:

Theorem 12.1 *The map $\sigma_\lambda : \mathcal{T}(S) \to \mathcal{SH}^+(\lambda)$ that associates to a hyperbolic metric its geometric shear-shape cocycle is a stratified real-analytic homeomorphism.*

As detailed in the introduction, combining this theorem with Theorems 2.1 and 10.15 implies that the orthogeodesic foliation map $\mathbb{O}_\lambda$ is a homeomorphism, and consideration of the earthquake/horocycle flows in $\mathcal{SH}^+(\lambda)$ coordinates then proves the conjugacy on slices (Theorem D).

We remark that the stratified regularity of σ_λ and $\mathbb{O}_\lambda$ is the best one can expect, since the adjacency of strata of differentials is not analytic (as there are multiple inequivalent ways to "break up a zero"). Compare with [Dumas 2015, Theorem D], in which it is shown that, for a fixed Riemann surface Z, the identification $Q(Z) \cong \mathcal{ML}$ guaranteed by the Hubbard–Masur theorem [1979] is stratified real-analytic.

Fixed complementary subsurfaces By definition (see Section 13.2), the weighted arc system $\underline{A}(X)$ underlying $\sigma_\lambda(X)$ exactly identifies the geometry of $X \setminus \lambda$ via Theorem 6.4. Setting

$$\mathcal{T}(S; \underline{A}) := \{X \in \mathcal{T}(S) : \underline{A}(X) = \underline{A}\},$$

Theorem 12.1 therefore implies that $\mathcal{T}(S; \underline{A})$ is nonempty if and only if $\underline{A} \in \mathscr{B}(S \setminus \lambda)$.

Remark 12.2 The authors do not know a proof of this fact that does not factor through Theorem 12.1 except in some special cases (for example, when the complement of λ is polygonal, or when λ is a union of simple closed curves).[18]

In fact, since $\mathcal{SH}^+(\lambda)$ is an affine cone bundle over $\mathscr{B}(S \setminus \lambda)$ (Proposition 8.5), we see that:

Corollary 12.3 *For each $\underline{A} \in \mathscr{B}(S \setminus \lambda)$, the set $\mathcal{T}(S; \underline{A})$ is a real-analytic submanifold of $\mathcal{T}(S)$ and the restriction of σ_λ to*

$$\mathcal{T}(S; \underline{A}) \to \mathcal{SH}^+(\lambda; \underline{A}) \cong \mathcal{H}^+(\lambda)$$

is a real-analytic homeomorphism.

In this setting, the correspondence between $\mathcal{T}(S; \underline{A})$ and $\mathcal{H}^+(\lambda)$ is a natural generalization of shear coordinates, since the complementary subsurfaces to λ are always isometric. In fact, the shape-shifting deformations built to deform X by some $\mathfrak{s} \in \mathcal{H}(\lambda)$ (see the proof sketch of Theorem 12.1 just below) restrict to cataclysms/shear maps in the sense of [Bonahon 1996, Section 5]. In particular, if $\mathfrak{s}$ represents a measure supported on λ, then the shape-shifting deformation determined by $\mathfrak{s}$ is part of an earthquake in $\mathfrak{s}$ (Corollary 15.2); if $\mathfrak{s}$ is a multiple of $\sigma_\lambda(X)$, the shape-shifting transformation can sometimes be identified with part of a (generalized) stretch ray (Propositions 15.12 and 15.18).

In addition to being nonempty, $\mathcal{T}(S; \underline{A})$ is structurally rich; the authors hope to explore this space further in future work. Of particular interest is the (degenerate) Weil–Petersson pairing on this locus and its relation with the Thurston symplectic form and Masur–Veech measures.

[18] One can of course complete λ to a maximal lamination and then specify the shear coordinates on each of the added leaves, but then one must be very careful to ensure that these shears satisfy the relations coming from the metric residue condition. The argument then requires an involved computation with train tracks carrying the completed lamination.

A sketch of the proof Since the proof of Theorem 12.1 spans several sections (two of which consist of involved constructions of the relevant objects), we devote the remainder of this section to a broad-strokes outline of the arguments involved. Our exposition throughout these sections is mostly self-contained, but we sometimes refer to [Bonahon 1996] for proofs and to [Thurston 1986] for inspiration.

We begin in Section 13 by defining the map σ_λ. Under the correspondence established in Theorem 6.4, we associate to X the weighted arc system $\underline{A}(X)$ recording the hyperbolic structure on $X \setminus \lambda$. We cut X along the (ortho)geodesic realization of $\lambda \cup \underline{\alpha}$ into a union of (degenerate) right-angled polygons, and measure the shear between certain pairs of polygons. We then argue using train tracks that it suffices to record the shearing data of $\sigma_\lambda(X)$ on short enough arcs k transverse to λ and disjoint from $\underline{\alpha}(X)$. The value of $\sigma_\lambda(X)$ on short k may then be defined by isotoping k to a path connecting vertices of the spine Sp and built of segments alternating between leaves of λ and of $\mathbb{C}_\lambda(X)$, then measuring the total (signed) length along λ. These measurements are equivalent to Bonahon and Thurston's method of measuring shears (via the horocyclic foliation) when k is short enough, but cannot be globally derived from theirs due to obstructions coming from complementary subsurfaces.

The proof that σ_λ is a homeomorphism then follows the same general steps as appear in [Bonahon 1996]. After proving that σ_λ is injective and lands inside $\mathcal{SH}^+(\lambda)$ (Proposition 13.12 and Corollary 13.14), we then show that it is open (Theorem 15.1) and proper. Since $\mathcal{SH}^+(\lambda)$ is a cell (Proposition 8.5), invariance of domain then implies that σ_λ must be a homeomorphism.

Our proof of injectivity mirrors that of [Bonahon 1996, Theorem 12] with an additional invocation of Theorem 6.4. For properness we mostly appeal to [Bonahon 1996, Theorem 20] but need to discuss complications that arise from the piecewise-linear structure of shear-shape space. Similarly, our broad-strokes strategy to prove openness parallels that of [Bonahon 1996, Section 5], in that we build a "shape-shifting cocycle" $\varphi_\mathfrak{s}$ for all small-enough deformations $\mathfrak{s}$ of $\sigma_\lambda(X)$ (see Section 14). Deforming X by postcomposing its charts to $\mathbb{H}^2$ with $\varphi_\mathfrak{s}$ then yields a surface $X_\mathfrak{s}$ with $\sigma_\lambda(X_\mathfrak{s}) = \sigma_\lambda(X) + \mathfrak{s}$.

It is in the construction of $\varphi_\mathfrak{s}$, performed in Section 14, where our discussion truly diverges from [Bonahon 1996; Thurston 1986]. When λ is maximal, one can specify $\varphi_\mathfrak{s}$ by shearing X along the leaves of λ (ie performing a cataclysm). Even in the maximal case this procedure is delicate, hinging on the convergence of infinite products of small Möbius transformations (compare Section 14.2). In the nonmaximal case, we must also simultaneously account for the changing shapes of complementary subsurfaces (which also introduces extra complications into the shearing deformations since the shapes of spikes are changing). See the introduction to Section 14 for a more granular description of the construction of $\varphi_\mathfrak{s}$.

13 Measuring hyperbolic shears and shapes

In this section, we take our first steps towards proving Theorem 12.1 by describing how to associate to any hyperbolic surface X a *geometric shear-shape cocycle* $\sigma_\lambda(X)$ in a natural way; this yields the map

$$\sigma_\lambda : \mathcal{T}(S) \to \mathcal{SH}(\lambda).$$

After fixing some notational conventions that we will use throughout the sequel, we define $\sigma_\lambda(X)$ by first specifying its underling arc system $\underline{A}(X)$ in a variety of equivalent ways. After doing so, we define the shear between "nearby" hexagons analogously to Bonahon and Thurston; placing all of this data onto a standard smoothing $\tau_{\underline{\alpha}}$ of a geometric train track is therefore enough to specify $\sigma_\lambda(X)$ (Lemma 13.6).

We then show that the data of shears between any two nearby hexagons can be recovered from the weight system on $\tau_{\underline{\alpha}}$, even if those hexagons are not "visible" to $\tau_{\underline{\alpha}}$ (Lemma 13.9). This in particular implies that our choice of $\tau_{\underline{\alpha}}$ does not actually matter, and hence $\sigma_\lambda(X)$ is well defined.

We then conclude the section by proving some initial properties of σ_λ. Proposition 13.12 shows that the map is injective following an argument of Bonahon, while in Theorem 13.13 we show that our map captures the geometry of the orthogeodesic foliation.

13.1 Preliminaries and notation

In this section, we discuss the geometry of a geodesic lamination on a hyperbolic surface and fix notation in preparation for our definition of the geometric shear-shape cocycle of a hyperbolic structure.

Throughout, we use the symbol λ to refer to both the measured lamination λ and its support, realized geodesically with respect to any number of hyperbolic metrics. We reserve S to denote a topological surface and Σ the topological type of a component of $S \setminus \lambda$, while X and Y will denote their hyperbolic incarnations. We also adopt the following family of notational conventions: the expression $g \subset \lambda$ means that g is a leaf of λ, and $Y \subset X \setminus \lambda$ means that Y is a component of (the metric completion of) $X \setminus \lambda$, etc. The notation of [Bonahon 1996] is used as inspiration, since we will make direct appeals to the results therein. However, our situation requires more care, since we have more objects to keep track of. A key difference is that we will focus not on the relative shear between complementary subsurfaces of $X \setminus \lambda$, but on the relative positioning of pairs of boundary leaves of λ, equipped with a natural collection of basepoints determined by the orthogeodesic foliation.

Hexagons Given $X \in \mathcal{T}(S)$ and $\lambda \in \mathcal{ML}(S)$, realize λ geodesically on X. Construct the orthogeodesic foliation $\mathcal{O}_\lambda(X)$ on X with piecewise-geodesic spine Sp and dual arc system $\underline{\alpha} = \underline{\alpha}(X)$, realized orthogeodesically with respect to X and λ. The union $\lambda_{\underline{\alpha}} = \lambda \cup \underline{\alpha}$ is a geometric object on X that fills; that is, the metric completion of $X \setminus \lambda_{\underline{\alpha}}$ is a union of geometric pieces that are topological disks, possibly with some points on the boundary removed corresponding to spikes. We lift the situation to universal covers $\widetilde{\lambda}_{\underline{\alpha}} \subset \widetilde{X}$, where we have also the full preimages $\widetilde{\mathsf{Sp}}$, $\widetilde{\lambda}$, $\widetilde{\underline{\alpha}}$, etc of various objects.

Let $\mathcal{H}$ be the vertex set of $\widetilde{\mathsf{Sp}}$; we will sometimes refer to $v \in \mathcal{H}$ as a *hexagon*. Indeed, to v there is associated a component H_v of $\widetilde{X} \setminus \widetilde{\lambda}_{\underline{\alpha}}$ which is generically a degenerate right-angled hexagon, though H_v may also be a regular ideal or right-angled polygon, for example. We reiterate that, by abuse of terminology, *any complementary component H_v of $\widetilde{X} \setminus \widetilde{\lambda}_{\underline{\alpha}}$ is called a hexagon*, no matter its shape.

If $\{H_v : v \in \mathcal{H}\}$ contains components that are not degenerate right-angled hexagons in the usual sense, then $\underline{\alpha}$ corresponds to a simplex of $\mathscr{A}_{\mathrm{fill}}(S \setminus \lambda)$ of nonmaximal dimension (or the empty set, if λ is filling and $\underline{\alpha}$ is empty). One may always include $\underline{\alpha}$ in a maximal arc system $\underline{\beta}$, which necessarily defines a simplex of full dimension. The complementary components of $\widetilde{X} \setminus \widetilde{\lambda}_{\underline{\beta}}$ are now degenerate right-angled hexagons in the usual sense, and gluing them in pairs along $\underline{\beta} \setminus \underline{\alpha}$ gives the more general "hexagons" of $\widetilde{X} \setminus \widetilde{\lambda}_{\underline{\alpha}}$. We will often tacitly choose and work with a maximal arc system containing the original when convenient.

Pointed geodesics We now define a natural family of basepoints associated to boundary leaves of $\widetilde{\lambda}$. For $v \in \mathcal{H}$ and its associated hexagon H_v, define the λ–*boundary* $\partial_\lambda H_v$ of H_v to be the set of leaves of $\widetilde{\lambda}$ that meet ∂H_v.

For $v \in \mathcal{H}$ and g a leaf of $\partial_\lambda H_v$, define p_v to be the orthogonal projection of v to g. Observe that v and p_v lie along the same (singular) leaf of $\mathcal{O}_\lambda(X)$. The orientation of S gives H_v an orientation and hence orients ∂H_v; this yields an orientation-preserving, isometric identification of (g, p_v) with $(\mathbb{R}, 0)$. We refer to points on a based geodesic by their signed distance to the basepoint, so that 0 refers to p_v while $\pm x$ refer to the points at signed distance $\pm x$ from p_v.

For a pair $v \neq w \in \mathcal{H}$ not in the same component of $\widetilde{\mathsf{Sp}}$, there is a unique geodesic $g_v^w \in \partial_\lambda H_v$ that separates v from w. Symmetrically, there is such a pointed geodesic $g_w^v \in \partial_\lambda H_w$ separating w from v. Note that $g_v^w = g_w^v$ occurs if and only if this leaf is isolated, and, by the assumption that λ is measured, projects to a simple closed curve component of λ. Even in this case, the points p_v and p_w are in general different.

13.2 The shear-shape cocycle of a hyperbolic structure

Our first task towards defining the geometric shear-shape cocycle $\sigma_\lambda(X)$ of a hyperbolic structure X is to construct a weighted filling arc system $\underline{A}(X) \in \mathscr{B}(S \setminus \lambda)$ which records the shapes of the complementary subsurfaces.

With the technology we have developed up to this point, we now have many ways of constructing $\underline{A}(X)$, all of which are easily seen to be equivalent:

- To each $\alpha \in \underline{\alpha}(X)$, we associate the weight $c_\alpha := i(\mathcal{O}_\lambda(X), e_\alpha)$, where e_α is the edge of Sp dual to α. Equivalently, c_α is the length of the projection of e_α to either of the two leaves of λ to which it is closest. Then set $\underline{A}(X) = \sum c_\alpha \alpha$.
- Each component $Y \subset X \setminus \lambda$ is naturally endowed with a hyperbolic structure; by Theorem 6.4 this metric corresponds to a weighted filling arc system in $|\mathscr{A}_{\mathrm{fill}}(Y, \partial Y)|_{\mathbb{R}}$, and we let $\underline{A}(X)$ denote the union of these arc systems over all components of $X \setminus \lambda$.
- Let q be the quadratic differential with $|\mathrm{Re}(q)| = \mathcal{O}_\lambda(X)$ and $|\mathrm{Im}(q)| = \lambda$; then set $\underline{A}(X) = \underline{A}(q)$.

The final definition together with the results of Section 10 implies that $\underline{A}(X) \in \mathscr{B}(S \setminus \lambda)$ for every hyperbolic structure X on S. In the interest of providing the reader with geometric intuition for this condition, we have included an alternative, purely hyperbolic-geometric proof of this fact below.

Lemma 13.1 *With notation as above,* $\underline{A}(X) \in \mathscr{B}(S \setminus \lambda)$.

Proof By Theorem 6.4, it suffices to show that for each minimal, orientable component μ of λ, the sum of the metric residues of the crown ends of $X \setminus \lambda$ incident to μ is 0. If μ is a simple closed curve, then the metric residue is just equal to the (signed) lengths of the boundary components resulting from cutting along μ, which clearly must match.

So assume that μ is not a closed curve and pick an orientation. Construct a geometric train track τ snugly carrying μ as in Construction 5.6; then τ inherits an orientation from the inclusion of μ and so has well-defined left- and right-hand sides. As in Section 5.2, every branch b of τ has a well-defined length along λ which we denote by $\ell_\tau(b) > 0$. At each switch s of τ, let h_s be the leaf of the horocyclic foliation of $\mathcal{N}_\epsilon(\mu)$ projecting to s. By assumption of snugness, the spikes of $S \setminus \tau$ correspond with the spikes of $S \setminus \mu$, so the union of the h_s truncates each spike of each crown end incident to μ by h_s.

Each crown incident to μ inherits an orientation from the chosen orientation on μ, and we now compute the total metric residue with respect to these orientations and the truncations induced by the h_s. Recall that the metric residue of an oriented crown $\mathscr{C}$ is the alternating sum of the lengths of the geodesic boundary segments running between the truncation horospheres (Definition 4.3). Each such geodesic segment defines a cooriented trainpath $(b_1 \cdots b_n, \pm)$ in τ (ie a trainpath and a distinguished side, left or right, corresponding to $+$ and $-$, respectively) which runs along the entirety of a smooth component of the boundary of $X \setminus \tau$. Using this identification, we may compute that the corresponding contribution to the total metric residue is given by $\pm\sum_i \ell_\tau(b_i)$.

Finally, we observe that every branch of τ is a subpath of exactly two smooth boundary edges of $X \setminus \tau$ (corresponding to its left and right sides). Therefore, the sum of the metric residues of all of the crown ends incident to μ is the sum of the contributions of the corresponding cooriented trainpaths, which is necessarily 0 since each branch is counted twice, once with positive and once with negative sign. Thus $\underline{A}(X) \in \mathscr{B}(S \setminus \lambda)$. □

Shears between nearby hexagons Our second step towards defining $\sigma_\lambda(X)$ is to determine how to record shearing data between two hexagons that lie in different components of $\widetilde{X} \setminus \widetilde{\lambda}$ yet are close enough together. Except for sign conventions (see Remark 13.3), our discussion is essentially identical to Bonahon's definition [1996, Section 2] of shearing between the plaques of a maximal lamination. Our restriction to pairs of nearby hexagons reflects the fact that if two hexagons are far apart, a path connecting them may meet a subsurface of $\widetilde{X} \setminus \widetilde{\lambda}$ in a variety of ways.

Given $v, w \in \mathcal{H}$, consider the associated pointed geodesics $(g_v^w, p_v) \in \partial_\lambda H_v$ closest to H_w and $(g_w^v, p_w) \in \partial_\lambda H_w$ closest to H_v. We say that the geodesic segment $k_{v,w} \subset \widetilde{X}$ joining p_v to p_w is a *simple piece* if $k_{v,w}$ projects to a simple geodesic segment in X and $k_{v,w}$ bounds a spike in every hexagon that it crosses. That is, if $k_{v,w}$ crosses H_u for some $u \in \mathcal{H}$, then $k_{v,w} \cap H_u$ bounds a triangle in H_u, two sides of which lie on asymptotic leaves g_u^v and g_u^w defining a spike of $\widetilde{\lambda}$. If $k_{v,w}$ is a simple piece, then we say that (v, w) is a *simple pair*.

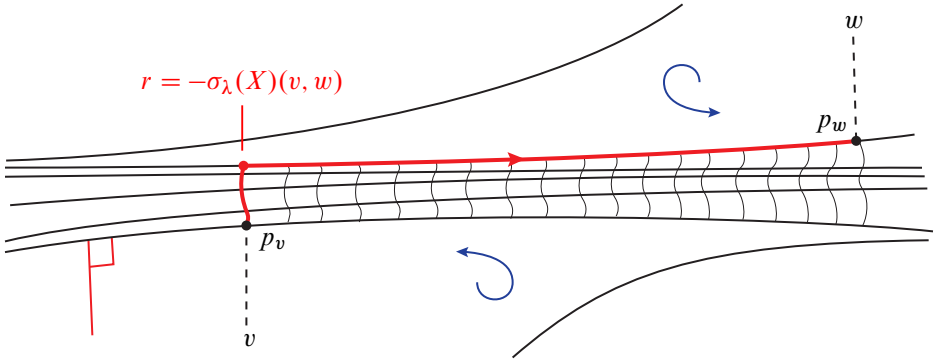


Figure 18: Computing the shears between two nearby hexagons v and w. In this example, $r < 0$, so $\sigma_\lambda(X)(v, w) > 0$.

We observe that, if $v, w \in \mathcal{H}$ are close enough together and lie in different components of $\widetilde{\mathsf{Sp}}$, then (v, w) is a simple pair. The exact value of "close enough" is unimportant, but we note that it suffices for $d(H_u, H_v)$ to be smaller than the length of the shortest arc of $\underline{\alpha}(X)$.

Now, following [Bonahon 1996, Section 2], let $\Lambda_{v,w}$ be the leaves of $\widetilde{\lambda}$ that separate g_v^w from g_w^v, equipped with the linear order $<$ induced by traversing $k_{v,w}$ from p_v to p_w. Since (v, w) is a simple pair, the subset of those leaves that are also the boundary of a complementary component of $\widetilde{X} \setminus \widetilde{\lambda}$ come in pairs that are asymptotic in one direction. The partial horocyclic foliations on the wedges bounded by pairs of asymptotic boundary leaves extend across the leaves of $\Lambda_{v,w}$, foliating the region bounded by g_v^w and g_w^v. In particular, the leaf of the horocyclic foliation containing p_v meets g_w^v (and the leaf containing p_w meets g_v^w).

Since the orthogeodesic foliation is equivalent to the horocyclic foliation in spikes, for any simple pair (v, w), the leaf of $\mathcal{O}_{\Lambda_{v,w}}(\widetilde{X})$ containing p_v meets g_w^v (and the leaf containing p_w meets g_v^w). In fact, simplicity implies that $\mathcal{O}_{\Lambda_{v,w}}(\widetilde{X})$ foliates the "quadrilateral" bounded by g_w^v, g_v^w and the two leaves of $\mathcal{O}_{\Lambda_{v,w}}(\widetilde{X})$ containing p_v and p_w.

Definition 13.2 Suppose that (v, w) is a simple pair of hexagons. Using the orientation conventions of Section 13.1, identify the corresponding pointed geodesics (g_v^w, p_v) and (g_w^v, p_w) with $(\mathbb{R}, 0)$. Now since the hexagons are close enough, the singular leaf of $\mathcal{O}_{\Lambda_{v,w}}(\widetilde{X})$ containing p_v meets g_w^v in some point $r \in \mathbb{R}$, and we set $\sigma_\lambda(X)(v, w) = -r$. See Figure 18.

It is not hard to see that $\sigma_\lambda(X)(v, w)$ remains the same if we flip the roles of v and w. Indeed, following along the leaves of the orthogeodesic foliation defines an orientation reversing isometry from a subsegment of g_v^w to a subsegment of g_w^v that takes $t \mapsto r - t$. In particular, p_v maps to a point on g_w^v that is positioned r signed units away from p_w, and so $\sigma_\lambda(X)(v, w) = \sigma_\lambda(X)(w, v)$.

Remark 13.3 Our choice to set $\sigma_\lambda(X)(v,w) = -r$ instead of $+r$ records "how far along g_w^v you must travel from r to get to p_w". Though this convention is the opposite of what appears in [Bonahon 1996], it allows us to combine the data of $\sigma_\lambda(X)(v,w)$ and $\underline{A}(X)$ into a system of train track weights on a standard smoothing (Construction 13.5). Our convention also parallels our choice of $[\cdot]_+$ function when measuring periods of a quadratic differential (Lemma 10.10), which makes the relationship between the hyperbolic geometry of (X,λ) and the flat geometry of $q(\mathbb{O}_\lambda(X),\lambda)$ more transparent.

Below, we give an elementary estimate that will be used in the proof of Proposition 13.12; compare with [Bonahon 1996, Lemma 8].

Lemma 13.4 *Suppose that (v,w) is a simple pair of hexagons. Let (g_v^w, p_v) and (g_w^v, p_w) be the associated pointed geodesics. Then the geodesic segment $k_{v,w}$ joining p_v to p_w satisfies*

$$|\sigma_\lambda(X)(k_{v,w})| \le \ell(k_{v,w}).$$

Proof As (v,w) is simple, the partial orthogeodesic foliation $\mathbb{O}_{\Lambda_{v,w}}(\widetilde{X})$ foliates the region U bounded by g_w^v, g_v^w and the two leaves of $\mathbb{O}_{\Lambda_{v,w}}(\widetilde{X})$ containing p_v and p_w. This foliation gives rise to a 1–Lipschitz retraction π from U to g_w^v defined by following the leaves of the orthogeodesic foliation to g_w^v. The image $\pi(k_{v,w})$ is then equal to the segment of g_w^v joining p_w to the point labeled by $\sigma_\lambda(X)(v,w)$, which has length $|\sigma_\lambda(X)(v,w)|$. The lemma follows. □

Hyperbolic shearing as train track weights Now that we have explained how to record the shapes of $X \setminus \lambda$ (Lemma 13.1) and the shears between nearby hexagons (Definition 13.2), we can package this information together to define the *geometric shear-shape cocycle* $\sigma_\lambda(X)$ of a hyperbolic structure X.

Below, we realize the shape and shear information specified above as a weight system on a standard smoothing of a geometric train track carrying λ; this strategy allows us to specify $\sigma_\lambda(X)$ by a finite collection of information. Once we show that the weights are well defined and satisfy the switch conditions, we then invoke Proposition 9.5 to interpret this weight system as an (axiomatic) shear-shape cocycle (see Definition 13.8). This reinterpretation in turn makes it apparent that our initial choice of train track does not matter.

Using Construction 5.6, choose a geometric train track $\tau \subset X$ that carries λ snugly and let $\tau_{\underline{\alpha}}$ be a standard smoothing of $\tau \cup \underline{\alpha}(X)$ (see Construction 9.3). Note that the components of $\widetilde{X} \setminus \tilde{\tau}_{\underline{\alpha}}$ are in bijection with the set of hexagons $\mathcal{H}$, and that the assumption that τ carries λ snugly ensures that if two hexagons correspond to adjacent components of $\widetilde{X} \setminus \tilde{\tau}_{\underline{\alpha}}$ then they either share an edge of $\widetilde{\underline{\alpha}}$ or form a simple pair. We recall that two hexagons form a simple pair if the geodesic connecting their basepoints passes only through spikes of $S \setminus \lambda$.

Construction 13.5 Fix $\tau_{\underline{\alpha}} \subset X$ as above. We then associate a weight system $w(X) \in \mathbb{R}^{b(\tau_{\underline{\alpha}})}$ as follows:

- To each branch corresponding to $\alpha \in \underline{\alpha}$, assign the weight $c_\alpha = i(\mathbb{O}_\lambda(X), e_\alpha)$, where e_α is the edge of Sp dual to α.

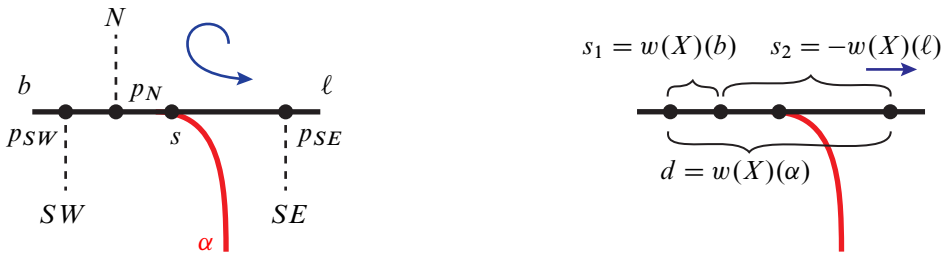


Figure 19: Left: a local picture of τ near s. Right: case (3). The switch condition is satisfied because $s_1 = d - s_2$.

- For each branch $b \subset \tau_{\underline{\alpha}}$ that dos not correspond to an arc of $\underline{\alpha}$, choose a lift $\tilde{b} \in \tilde{\tau}_{\underline{\alpha}}$. Let $v, w \in \mathcal{H}$ denote the vertices of $\widetilde{\mathsf{Sp}}$ corresponding to the hexagons adjacent to $\tilde{b}$, and set $w(X)(b) = \sigma_\lambda(X)(v, w)$.

Lemma 13.6 *Let X, λ, $\underline{\alpha}$ and $\tau_{\underline{\alpha}}$ be as above. Then the edge weights $w(X) \in \mathbb{R}^{b(\tau_{\underline{\alpha}})}$ given by Construction 13.5 satisfy the switch conditions.*

Proof Reference to Figure 19 will provide clarity throughout. We note that $\tau_{\underline{\alpha}}$ is generically trivalent, but may be 4–valent if there are arcs $\alpha_1, \alpha_2 \in \underline{\alpha}$ whose endpoints on λ lie on a common leaf of the orthogeodesic foliation. We give an argument only for the trivalent switches of $\tau_{\underline{\alpha}}$, and leave it to the reader to make the necessary adjustments for 4–valent switches (the statement for 4–valent switches can also be deduced by continuity).

Let s be a trivalent switch; then standing at s and looking into the spike, there are small half-branches exiting s on our right and left; call these r and ℓ, respectively. By our convention on standard smoothings, every half-branch of $\tau_{\underline{\alpha}}$ corresponding to an arc of $\underline{\alpha}$ is a right small half-branch.

If no branch of s corresponds to an arc of $\underline{\alpha}$, then the arguments appearing in [Bonahon 1996, Section 2] imply that the weights satisfy the switch conditions, because the orthogeodesic foliation is equivalent to the horocycle foliation in near s. See also [Papadopoulos 1991, Section 6] for a discussion more similar in spirit to ours.

Otherwise, the right small half-branch r is labeled by some $\alpha \in \underline{\alpha}$. Let b be the large half-branch incident to s. Give names also to the hexagons incident to s and their distinguished points on b or ℓ by projection; they are N, SE, and $SW \in \mathcal{H}$, and p_N, p_{SE} and p_{SW}, respectively, where b and ℓ form part of the boundary of N, ℓ and r form part of the boundary of SE, and r and b form part of the boundary of SW. See Figure 19.

Now take $d = d(p_{SW}, p_{SE})$, which is equal to $w(X)(r) = c_\alpha > 0$ by definition. Define also

$$s_1 := |w(X)(b)| = d_\tau(p_{SW}, p_N) \quad \text{and} \quad s_2 := |w(X)(\ell)| = d_\tau(p_N, p_{SE}).$$

Here d_τ is understood to mean the distance between leaves of the orthogeodesic foliation near τ, measured along any leaf of λ (see Section 5.2 for an explanation of why this value is well defined).

There are three kinds of configurations for the projection points p_{SW}, p_N and p_{SE} that determine the signs of $w(X)(b)$ and $w(X)(\ell)$:

(1) The point p_N precedes both p_{SW} and p_{SE} with respect to the orientation of τ on induced by H_N, so that
$$w(X)(b) = -s_1 \quad \text{and} \quad w(X)(\ell) = -s_2 \quad \text{with } s_2 > s_1.$$
In this case, $d = s_2 - s_1$ and so $d - s_2 = -s_1$, which is exactly the switch condition.

(2) Both p_{SW} and p_{SE} precede p_N, so that
$$w(X)(b) = s_1 \quad \text{and} \quad w(X)(\ell) = s_2 \quad \text{with } s_1 > s_2.$$
This possibility gives that $d = s_1 - s_2$ and so $d + s_2 = s_1$.

(3) The point p_{SW} precedes p_N, which in turn precedes p_{SE}, so that $w(X)(b) = s_1$ and $w(X)(\ell) = -s_2$. In this case, $d = s_1 + s_2$ and so $d - s_2 = s_1$, which is again the switch condition.

Therefore, no matter the configuration of points p_N, p_{SW} and p_{SE}, the switch conditions are fulfilled at s, completing the proof of the lemma. □

Remark 13.7 Importantly, $w(X)$ is generally *not* the same as the weight system coming from the shear coordinates of a completion of λ (unless λ was maximal to begin with).

Invoking Proposition 9.5 and Lemma 13.1, the weight system $w(X)$ defines a shear-shape cocycle with underlying arc system $\underline{A}(X)$.

Definition 13.8 The *geometric shear-shape cocycle* $(\sigma_\lambda(X), \underline{A}(X))$ of a hyperbolic metric X is the unique shear-shape cocycle for λ corresponding to the weight system $w(X)$ of Construction 13.5.

The rule that assigns to a hyperbolic structure its geometric shear-shape cocycle therefore defines a map
$$\sigma_\lambda : \mathcal{T}(S) \to \mathcal{SH}(\lambda), \quad X \mapsto \sigma_\lambda(X),$$
which is the subject of the rest of this article.

Train track independence We have employed the language of train tracks for convenience — the ties of a train track are a useful class of measurable arcs in the sense that they can be made transverse to λ and disjoint from $\underline{\alpha}$ (or record the weight associated to an arc of $\underline{\alpha}$). However, Construction 13.5 and Definition 13.8 a priori depend on the choice of geometric train track $\tau_{\underline{\alpha}}$ carrying λ.

Now that we have identified the weight system $w(X)$ with the shear-shape cocycle $\sigma_\lambda(X)$, however, we can invoke both the axiomatic and cohomological interpretations (Definitions 7.5 and 7.11) to see that

the value of $\sigma_\lambda(X)$ on any arc k transverse to λ but disjoint from $\underline{\alpha}$ does not depend on the choice of geometric train track. Indeed, let k be any such arc; then by transverse invariance (axiom (SH1)) we may replace k with a concatenation of short geodesics, all of which are transverse to λ but disjoint from $\underline{\alpha}$. By additivity (axiom (SH2)), it therefore suffices to show that the value of $\sigma_\lambda(X)$ on any short geodesic disjoint from $\underline{\alpha}$ does not depend on the train track.

Lemma 13.9 *Let k be a short enough geodesic segment on X that is transverse to λ. Lift k to an arc $\tilde{k}$ on $\widetilde{X}$ and let v and w be the hexagons containing the endpoints of $\tilde{k}$; then*

$$\sigma_\lambda(X)(k) = \sigma_\lambda(X)(v, w),$$

where on the left $\sigma_\lambda(X)$ represents the axiomatic shear-shape cocycle and on the right $\sigma_\lambda(X)$ represents the shear between nearby hexagons (Definition 13.2). In particular, $\sigma_\lambda(X)(k)$ does not depend on the choice of train track employed in Definition 13.8.

In fact, the conclusion of this lemma holds for *all* simple pairs.

Proof So long as k is short enough (shorter than all arcs of $\underline{\alpha}(X)$), (v, w) is a simple pair. Using axiom (SH1), we may therefore isotope k through arcs transverse to λ but disjoint from $\underline{\alpha}$ to an arc k', defined to be the concatenation of $k_{v,w}$, the geodesic connecting the points p_v and p_w on the boundary geodesics g_v^w and g_w^v, together with segments of the orthogeodesic foliation inside each hexagon H_v and H_w.

Let τ be a geometric train track snugly carrying λ defined with parameter ϵ; then the collapse map $\pi\colon \mathcal{N}_\epsilon(\lambda) \to \tau$ takes k' to a train route on τ, and hence on $\tau_{\underline{\alpha}}$. Orient k' (and hence also the train route $\pi(k')$) so that it travels from v to w. Let $v = u_1, u_2, \dots, u_N = w$ denote the sequence of hexagons corresponding to regions of $\widetilde{X} \setminus \tilde{\tau}_{\underline{\alpha}}$ bordering this train route, so that the regions corresponding to u_i and u_{i+1} both meet the same subsegment of $\pi(k')$. Let p_i denote their corresponding projections onto λ. Note that, since $\pi(k')$ is carried on $\tau \prec \tau_{\underline{\alpha}}$, no pair of subsequent hexagons u_i and u_{i+1} lies in the same component of $\widetilde{\mathsf{Sp}}$. This plus the construction of the train track implies that (u_i, u_{i+1}) is a simple pair, and we can measure the shear $\sigma_\lambda(X)(u_i, u_{i+1})$ (up to sign) as the distance along the train track between $\pi(p_i)$ and $\pi(p_{i+1})$.

Now, given $\tau_{\underline{\alpha}}$ carrying λ, we observe that k' also determines a (pair of) relative cycle(s) in the corresponding (orientation cover of the) ϵ–neighborhood of $\lambda_{\underline{\alpha}}$. The value $\sigma_\lambda(X)(k) = \sigma_\lambda(X)(k')$ is then equal to the value of the cohomological shear-shape cocycle evaluated on either of the oriented lifts $\hat{k}'$ of k' which cross the lift of λ with positive local orientation. We may therefore express

$$[\hat{k}'] = [t_1] - [t_2] + [t_3] - \cdots \pm [t_{N-1}],$$

where t_i is a (lift of a) tie corresponding to the branch of the train track connecting the regions corresponding to u_i and u_{i+1}, lifted to the orientation cover to have positive intersection with λ. See Figure 20.

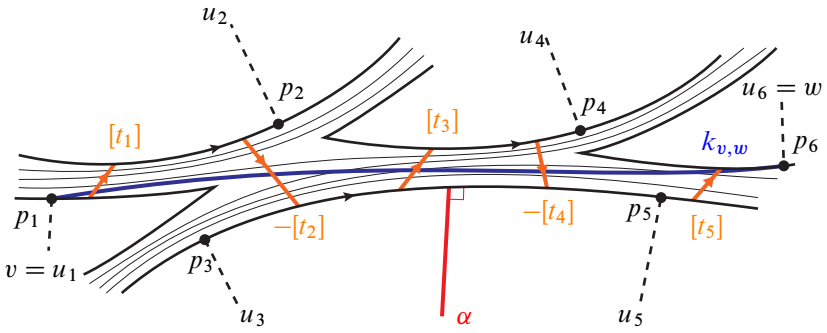


Figure 20: Measuring the shear of a small arc using a geometric train track. By isotoping k to a proper arc in the geometric train track neighborhood and then expressing its relative homology class as a sum of the branches, we can compute its shear as the alternating sum of shears between adjacent hexagons.

But now, by construction, $\sigma_\lambda(X)$ evaluated on $[t_i]$ is just the shear $\sigma_\lambda(X)(u_i, u_{i+1})$. In turn, this shear is equal to the signed distance along the train track between $\pi(p_i)$ and $\pi(p_{i+1})$ (where the sign is determined by the local orientation of λ). Combining this with the expression for $[\hat{k}']$ above, $\sigma_\lambda(X)(k)$ is exactly equal to the signed distance along the train track between $\pi(p_1)$ and $\pi(p_N)$, which is the shear $\sigma_\lambda(X)(v, w)$. □

We note that, in the proof above, the cohomological interpretation of shear-shape cocycles provides a convenient workaround for the obstacle that the train route with dual transversals $t_1, \dots, t_{N-1}$ is not in general isotopic to k through arcs transverse to λ. Regardless, the relative homology class defined by $k' \cap \mathcal{N}_\epsilon(\lambda)$ is homologous to a linear combination of $\{t_i\}$ in the orientation cover of $\mathcal{N}_\epsilon(\lambda)$.

Remark 13.10 The lemma above can also be proved by splitting any two geometric train tracks to a common subtrack [Penner and Harer 1992, Theorem 2.3.1]. Each splitting sequence can then be realized in the orthogeodesically foliated neighborhood $\mathcal{N}_\epsilon(\lambda) \subset X$ by cutting along compact paths in the spine associated to a spike, as in [Zhu and Bonahon 2004, Section 3]. Splits induce maps on weight spaces, and so Lemma 13.9 is essentially equivalent to the statement that Construction 13.5 is compatible with splitting and collapsing. See also [Bonahon 1997b, Lemma 6].

The cocycle as a map on pairs It will be convenient to repackage the data provided by $\sigma_\lambda(X)$ in yet another form, which also explains our choice of notation in Definition 13.2.

If $v, w \in \mathcal{H}$ can be joined by a Lipschitz continuous segment $k_{v,w}$ which is transverse to λ, disjoint from $\underline{\alpha}$, and meets no leaf of $\widetilde{\lambda}$ twice, then we say that (v, w) is a *transverse pair* and that $k_{v,w}$ is a *transversal*. If (v, w) is a transverse pair, we say that r is *between* v and w if there is a transversal $k_{v,w}$ that decomposes as a concatenation of transversals $k_{v,w} = k_{v,r} \cdot k_{r,w}$. Finally, we define

$$\sigma_\lambda(X)(v, w) := \sigma_\lambda(X)(k_{v,w})$$

and declare that $\sigma_\lambda(X)(v,v) = 0$. Observe that, if (v,w) is a simple pair, then this agrees with our definition of the shear between nearby hexagons (Definition 13.2).

Lemma 13.11 *The shear-shape cocycle $\sigma_\lambda(X)$ defines a map on transverse pairs that satisfies*:

(1) ***π_1–invariance*** *For each $\gamma \in \pi_1(X)$, we have $\sigma_\lambda(X)(\gamma v, \gamma w) = \sigma_\lambda(X)(v,w)$.*

(2) ***Finite additivity*** *If (v,w) is a transverse pair and r is between v and w, then*

$$\sigma_\lambda(X)(v,w) = \sigma_\lambda(X)(v,r) + \sigma_\lambda(X)(r,w).$$

(3) ***Symmetry*** $\sigma_\lambda(X)(v,w) = \sigma_\lambda(X)(w,v)$.

The proof of this lemma is simply a consequence of unpacking the definitions and showing that two different choices of transversals give the same shear values; the latter statement is just a repeated application of axiom (SH3).

13.3 Injectivity and positivity

We now record some initial structural properties of the map σ_λ defined above. In particular, we demonstrate that σ_λ is injective and interacts coherently with the orthogeodesic foliation map $\mathbb{O}_\lambda$ and the shear-shape coordinatization I_λ of transverse foliations.

Observe that injectivity of σ_λ is equivalent to the statement that if two hyperbolic structures have the same complementary subsurfaces and same gluing data along λ, then they must be isometric. As the horocyclic and orthogeodesic foliations are equivalent in spikes of complementary subsurfaces, the proofs of [Bonahon 1996, Lemma 11 and Theorem 12] may be invoked mutatis mutandis. We outline this argument below for the convenience of the reader, and direct them to [Bonahon 1996] for a more thorough discussion of the estimates involved. We remark that this strategy also appears in the proof of Proposition 15.12, where we use it to piece together Lipschitz-optimal homeomorphisms along λ.

Proposition 13.12 *The map $\sigma_\lambda : \mathcal{T}(S) \to \mathcal{SH}(\lambda)$ is injective.*

Sketch of proof Fix homeomorphisms $(\widetilde{S}, \widetilde{\lambda})$ with $(\widetilde{X}_i, \widetilde{\lambda})$ that lift the markings $S \to X_i$ and are such that each component $\widetilde{\Sigma} \subset \widetilde{S} \setminus \widetilde{\lambda}$ maps homeomorphically to a component $\widetilde{Y}_i \subset \widetilde{X}_i \setminus \widetilde{\lambda}$ for $i = 1, 2$.

Suppose that $\sigma_\lambda(X_1) = \sigma_\lambda(X_2)$; then in particular $\underline{A}(X_1) = \underline{A}(X_2)$ and so, by Theorem 6.4, the complementary subsurfaces $\overline{X_1 \setminus \lambda}$ and $\overline{X_2 \setminus \lambda}$ are isometric. Therefore, for a given component $\Sigma \subset S \setminus \lambda$, we can find an $\pi_1(\Sigma)$ equivariant isometry $\varphi_\Sigma \colon \widetilde{Y}_1 \to \widetilde{Y}_2$. Define $\varphi \colon \widetilde{X}_1 \setminus \lambda \to \widetilde{X}_2 \setminus \lambda$ to be the union of these maps on each complementary component; by construction, φ is an isometry.

We need to show that φ extends to a $\pi_1(S)$–equivariant isometry $\varphi \colon \widetilde{X}_1 \to \widetilde{X}_2$. To prove this, we apply the arguments of [Bonahon 1996, Lemma 11], which we summarize presently. The first step is to construct

a locally Lipschitz continuous extension of φ; this step employs the length bound of Lemma 13.4 and some elementary hyperbolic geometry, and the arguments of the first ten paragraphs of [Bonahon 1996, Lemma 11] may be applied verbatim.

As in Bonahon's original proof, we now show that φ is actually 1–Lipschitz, given that it is locally Lipschitz. We first show that φ does not increase the length of leaves of the orthogeodesic foliation.

Given any segment ℓ of a leaf of the orthogeodesic foliation $\widetilde{\mathcal{C}_\lambda(X_1)}$, the length of ℓ restricted to any hexagon H_u where $u \in \mathcal{H}$ is completely determined by the isometry type of H_u and the distance along $\widetilde{\lambda}$ from $p_u \in \partial_\lambda H_u$. As $\sigma_\lambda(X_1)$ determines the shape of $X_1 \setminus \lambda$, we can recover this information and hence determine the length of $\ell \cap H_u$ just from the data of $\sigma_\lambda(X_1)$.

From $\sigma_\lambda(X_2) = \sigma_\lambda(X_1)$, we deduce that the length of ℓ in any hexagon of $\widetilde{X}_1$ is equal to the length of $\varphi(\ell)$ in the corresponding hexagon of $\widetilde{X}_2$. Moreover, since φ is locally Lipschitz, the 1–dimensional Lebesgue measure of $\varphi(\ell) \cap \varphi(\widetilde{\lambda})$ is at most the 1–dimensional Lebesgue measure of $\ell \cap \widetilde{\lambda}$. By a now-classical fact, the latter is zero [Birman and Series 1985]; hence, so is the former. Therefore, the length of ℓ in X_1 is equal to the length of ℓ in X_2.

Now there is a path joining any two points in $\widetilde{X}_1$ built from geodesic segments and segments of leaves of the orthogeodesic foliation. The argument above shows that φ preserves the lengths of such paths, so φ is globally 1–Lipschitz. The construction is completely symmetric, so φ^{-1} is 1–Lipschitz as well. Now every 1–Lipschitz homeomorphism between metric spaces with 1–Lipschitz inverse is necessarily an isometry, and equivariance of φ is immediate from the construction. Therefore X_1 and X_2 must be isometric. □

The diagram commutes We have now developed sufficient technology to prove that the geometric shear-shape cocycle of a hyperbolic metric is the same as the shear-shape cocycle associated to its orthogeodesic foliation. In other words, diagram (2) commutes. Compare with [Mirzakhani 2008, Proposition 6.1].

Theorem 13.13 *For all $\lambda \in \mathcal{ML}$ and all $X \in \mathcal{T}(S)$, we have $\sigma_\lambda(X) = I_\lambda \circ \mathcal{C}_\lambda(X)$.*

Proof Fix a standard smoothing $\tau_{\underline{\alpha}}$ of a geometric train track τ for λ on X. Our approach is to compute both $\sigma_\lambda(X)(b)$ and $I_\lambda \circ \mathcal{C}_\lambda(X)(b)$ for each branch b of $\tau_{\underline{\alpha}}$. These numbers will coincide, so, by Proposition 9.5, $\sigma_\lambda(X) = I_\lambda \circ \mathcal{C}_\lambda(X)$.

Let $\mathsf{T}_X \subset X$ be the piecewise-geodesic triangulation of X whose vertices are the vertices of Sp, so that there is an edge between $v, w \in \mathsf{Sp}$ if the corresponding regions of $X \setminus \tau_{\underline{\alpha}}$ share a branch. This recipe generically yields a triangulation, but may have quadrilaterals in the case that two points of $\underline{\alpha}(X) \cap \lambda$ lie on the same leaf of $\mathcal{C}_\lambda(X) \cap \mathcal{N}_\epsilon(\lambda)$. In this case, we may either choose a smaller initial neighborhood to define our geometric train track so that this does not occur, or these points correspond to arcs that meet an isolated leaf of λ on either side; in the latter case, choose either diagonal that crosses the quadrilateral to include into T_X. Observe that each edge of T_X is either transverse to $\mathcal{C}_\lambda(X)$ or a segment of a leaf (on the off chance that two adjacent regions have exactly 0 shear between them).

Let $q = q(\mathcal{C}_\lambda(X), \lambda)$, and recall that Proposition 5.10 provides a homotopy equivalence $\mathcal{D}_\lambda : X \to q$ in the correct homotopy class satisfying $\mathcal{D}_{\lambda *}\mathcal{C}_\lambda(X) = V(q)$ and $\mathcal{D}_{\lambda *}\lambda = H(q)$ both leafwise and measurably. Furthermore, $\mathcal{D}_\lambda$ maps T_X to a (topological) triangulation of q with vertices at its zeros. It therefore remains to show that $\sigma_\lambda(X)$ evaluated on a branch of $\tau_{\underline{\alpha}}$ is the same as $I_\lambda(q)$ evaluated on the dual edge of this triangulation.

Now, by definition, $\underline{A}(X) = \underline{A}(q)$, so consider a branch b of $\tau_{\underline{\alpha}}$ not corresponding to an arc of the arc system. Dual to b there is an edge e of the triangulation $\mathcal{D}_\lambda(\mathsf{T}_X)$ which is transverse to the orthogeodesic foliation $\mathcal{C}_\lambda(X)$ on q (since T_X was transverse to $\mathcal{C}_\lambda(X)$ on X). Up to sign, the value of $I_\lambda(q)$ on b is the magnitude of the real part of the period of e, which is just the geometric intersection number $i(\mathcal{C}_\lambda(X), e)$ by transversality.

On the other hand, $\sigma_\lambda(X)(b)$ is equal to the shear between the two hexagons on either side of b. This in turn is equal to the geometric intersection number $i(\mathcal{C}_\lambda(X), k_{v,w})$ up to sign, where $k_{v,w}$ is the geodesic connecting the vertices p_v and p_w of $\lambda \cap \underline{\alpha}(X)$. Since $\mathcal{D}_\lambda$ takes $k_{v,w}$ to an arc transversely isotopic to e, we have $|\sigma_\lambda(X)(b)| = |I_\lambda(q)(b)|$.

Finally, to show that the signs are equal, fix matching orientations on $k_{v,w}$ and e. These induce local orientations on the leaves of λ such that the algebraic intersection of λ with $k_{v,w}$, (respectively e) is positive. In turn, this induces a local orientation on the leaves of $\mathcal{C}_\lambda(X)$ near $k_{v,w}$ (respectively e) and our sign conventions are equivalent to stipulating that the sign is positive if $k_{v,w}$ (respectively e) crosses $\mathcal{C}_\lambda(X)$ from left to right and negative if it crosses from right to left (compare [Mirzakhani 2008, Section 5.2]). In particular, the signs agree and so $\sigma_\lambda(X)(b) = I_\lambda(q)(b)$ for all branches b, completing the proof of the theorem. □

Corollary 13.14 *For all* $\mu \in \Delta(\lambda)$,

$$\omega_{\mathcal{SH}}(\sigma_\lambda(X), \mu) = i(\mathcal{C}_\lambda(X), \mu) = \ell_X(\mu) > 0.$$

In particular, $\sigma_\lambda(\mathcal{T}(S)) \subseteq \mathcal{SH}^+(\lambda)$.

Proof The first equality is a direct consequence of Theorem 13.13 and Proposition 10.12. The second equality was proved in Lemma 5.7. □

14 Shape-shifting cocycles

In the previous section, we explained how to associate to each hyperbolic structure X a shear-shape cocycle $\sigma_\lambda(X)$. In this one, we explain how to upgrade a small deformation $\mathfrak{s}$ of the cocycle into a deformation of the hyperbolic structure; this is eventually used to prove that $\sigma_\lambda : \mathcal{T}(S) \to \mathcal{SH}^+(\lambda)$ is open (Theorem 15.1). The main issue that we need to overcome is that we must simultaneously change the geometry of the nonrigid components of $X \setminus \lambda$ while shearing these subsurfaces along one another.

The goal of this section is therefore to build, for every small enough deformation $\mathfrak{s}$ of $\sigma_\lambda(X)$, a $\pi_1(S)$–equivariant *shape-shifting cocycle* that records how to adjust the relative position of geodesics of λ,

$$\varphi_{\mathfrak{s}}\colon \partial_\lambda \mathscr{H} \times \partial_\lambda \mathscr{H} \to \mathrm{Isom}^+ \widetilde{X},$$

where $\partial_\lambda \mathscr{H} := \{(h_v, p_v) \subset \partial_\lambda H_v : v \in \mathscr{H}\}$ is the set of boundary geodesics of $\widetilde{\lambda}$ equipped with basepoints obtained from projections of the vertices of $\widetilde{\mathsf{Sp}}$. See Proposition 14.26.

In Section 15.1, we explain how to modify the developing map $\widetilde{X} \to \mathbb{H}^2$ according to $\varphi_{\mathfrak{s}}$, resulting in a new (equivariant) hyperbolic structure $X_{\mathfrak{s}}$ with geometric shear-shape cocycle $\sigma_\lambda(X) + \mathfrak{s}$ (Lemma 15.6). By fixing a pointed geodesic $(h_v, p_v) \in \partial_\lambda \mathscr{H}$ we identify $\mathrm{Isom}^+(\widetilde{X})$ with $T^1\widetilde{X}$, so that the projection of $\{\varphi_{\mathfrak{s}}((h_v, p_v), (h_w, p_w)) \mid (h_w, p_w) \in \partial_\lambda \mathscr{H}\}$ to $\widetilde{X}$ is then the geodesic realization of $\widetilde{\lambda}$ in the new metric $\widetilde{X}_{\mathfrak{s}}$.

When the deformation $\mathfrak{s}$ preserves $\underline{A}(X)$, the cocycle $\varphi_{\mathfrak{s}}$ corresponds to a cataclysm map: the complementary components of $\widetilde{X} \setminus \widetilde{\lambda}$ are sheared along the leaves of $\widetilde{\lambda}$ and map isometrically into the deformed surface $X_{\mathfrak{s}}$. When $\mathfrak{s}$ alters $\underline{A}(X)$, we must shear the complementary subsurfaces while also simultaneously changing their shape, introducing complications not present in Bonahon and Thurston's original considerations.

Deforming the cocycle We first make explicit what we mean by a deformation of a shear-shape cocycle; we quantify what we mean by "small" in Section 14.2.

Observe that, if σ and σ' in $\mathcal{SH}^+(\lambda)$ are close, then, by Proposition 8.5, their underlying weighted arc systems $\underline{A}$ and $\underline{A}'$ are close in $\mathscr{B}(S \setminus \lambda)$. In particular, the corresponding unweighted arc systems $\underline{\alpha}$ and $\underline{\alpha}'$ must both live in some common top-dimensional cell of $\mathscr{B}(S \setminus \lambda)$, ie must both be contained in some common maximal arc system $\underline{\beta}$. Let τ be some snug train track for λ and let $\tau_{\underline{\beta}}$ be a standard smoothing of $\tau \cup \underline{\beta}$. By Proposition 9.5, we may then identify σ and σ' as weight systems on $\tau_{\underline{\beta}}$; the difference $\sigma - \sigma' \in W(\tau_{\underline{\beta}})$ is then a deformation of σ.

In general, if $(\sigma, \underline{A}) \in \mathcal{SH}^+(\lambda)$ and $\underline{\beta}$ is any maximal arc system containing the support of $\underline{A}$, then the deformations we consider in this section are those $\mathfrak{s} \in W(\tau_{\underline{\beta}})$ such that $\sigma + \mathfrak{s} \in W(\tau_{\underline{\beta}})$ corresponds to a positive shear-shape cocycle. Passing between equivalent definitions of shear-shape cocycles, we may also think of $\mathfrak{s}$ as a "shear-shape cocycle with negative arc weights". The underlying weighted arc system of any deformation $\mathfrak{s}$ will be denoted by $\mathfrak{a}$; while its coefficients are not necessarily positive, they will satisfy the zero total residue condition of (13) by construction.

By Theorem 6.4, the arc system $\underline{A} + \mathfrak{a}$ gives each component of $S \setminus \lambda$ a new complete hyperbolic metric Y with (noncompact) totally geodesic boundary. Since the supports of $\underline{A}$ and $\underline{A} + \mathfrak{a}$ are both contained inside some common maximal $\underline{\beta}$, one may set up a correspondence between the complementary components of $X \setminus \lambda_{\underline{\alpha}}$ with the components of $Y \setminus \mathrm{supp}(\underline{A} + \mathfrak{a})$ (adding in weight 0 edges as necessary).

A blueprint To help guide the reader through this rather intricate construction, we include here a top-level overview of the necessary steps, together with an outline of the section. Briefly, our strategy is

to explicitly define $\varphi_{\mathfrak{s}}$ on two types of pairs of pointed geodesics: the "simple pairs" between which the orthogeodesic foliation is comparable to the horocyclic, and the pairs which live in the boundary of a common subsurface. Piecing together these basic deformations then allows us to define $\varphi_{\mathfrak{s}}$ on arbitrary pairs of pointed geodesics.

Our construction of $\varphi_{\mathfrak{s}}$ for simple pairs parallels Bonahon's construction [1996, Section 5] of shear maps, and as such requires a detailed analysis of the geometry of the spikes of $\widetilde{X} \setminus \widetilde{\lambda}$. We therefore devote Section 14.1 to recording a number of useful notions and estimates from [Bonahon 1996]. In this section, we also introduce the "injectivity radius of X along λ", which measures the length of the shortest curve carried on a maximal snug train track for λ and plays a crucial role in our convergence estimates.

After these preliminary considerations, we turn in Section 14.2 to the actual construction of $\varphi_{\mathfrak{s}}$ on simple pairs. As in [Bonahon 1996], the map is defined by adjusting the lengths of countably many horocyclic arcs in an appropriate neighborhood of λ, compensating for changing shears between hexagons. Unlike in [Bonahon 1996], we must also adjust the arcs to account for the changing shapes of each of the spikes (as we are deforming the complementary subsurfaces). Convergence of the resulting infinite product of parabolic transformations is delicate; our approach follows [Bonahon 1996, Section 5] with influence from the more geometric approach of [Thurston 1986]. An accessible treatment of Thurston's construction of "cataclysm coordinates" can be found in [Papadopoulos and Théret 2007, Section 3.5].

We then turn in Sections 14.3 and 14.4 to defining $\varphi_{\mathfrak{s}}$ on pairs of geodesics in the boundary of the same hexagon or the same complementary subsurface, respectively. It is here that our work significantly differs from that of Bonahon and Thurston. In these sections we also develop the idea of "sliding" a deformed complementary subsurface along the original; this viewpoint allows us to easily demonstrate a number of otherwise nontrivial relations between Möbius transformations (see Propositions 14.18, 14.19 and 14.24).

Finally, in Section 14.5 we build the shape-shifting cocycle $\varphi_{\mathfrak{s}}$ from these pieces; the cocycle relation (Proposition 14.26) then follows from the cocycle relations for pieces and the separation properties of $\widetilde{\lambda}$.

Note Throughout this section and the next, we consider isometries via their action on a pointed geodesic, and compositions should be read from right to left.

14.1 Geometric control in the spikes

We first record some useful definitions and associated geometric estimates. These estimates play a crucial role in establishing convergence of the infinite products appearing in Section 14.2. Many of our definitions follow Bonahon's, but in order to contend with the fact that the complementary subsurfaces of λ are not always isometric, we must relate certain constants to the geometry of λ on X (see Lemma 14.5, in particular).

Our discussion will take place with certain data fixed. Choose a hyperbolic surface $X \in \mathcal{T}(S)$ and a measured lamination $\lambda \in \mathcal{ML}(S)$. Let $\epsilon > 0$ be small enough that an ϵ–geometric train track τ on X

carries λ snugly. The standard smoothing $\tau_{\underline{\alpha}}$ for the arc system $\underline{\alpha} = \underline{\alpha}(X)$ provides us with a vector space $W(\tau_{\underline{\alpha}})$ that models $\mathcal{SH}(\lambda; \underline{\alpha})$. With $\tau_{\underline{\alpha}}$ fixed, we endow the vector space of weights on branches of $\tau_{\underline{\alpha}}$ with the sup norm $\|\cdot\|_{\tau_{\underline{\alpha}}}$, and restrict this norm to the weight space $W(\tau_{\underline{\alpha}})$.

Let k_b be an oriented geodesic transverse to a branch $b \in \tau$ that also avoids $\underline{\alpha}$. Following Bonahon, we define the *divergence radius* or *depth* $r_b(d) \in \mathbb{Z}_{>0}$ of a component d of $k_b \setminus \lambda$ to be "how long the leaves of λ incident to d track each other", as viewed by τ.

More precisely, lift everything to the universal cover $\widetilde{X}$. By convention, set $r_b(d) = 1$ if d contains one of the endpoints of k_b. Otherwise, d is contained in a spike of H_v for some $v \in \mathcal{H}$, ie d connects a pair of asymptotic geodesics g_d^- and g_d^+. The divergence radius $r_b(d)$ is then the largest integer $r \geq 1$ such that $\pi(g_d^+)$ and $\pi(g_d^-)$ successively cross the same sequence of branches

$$b_{-r+1}, b_{-r+2}, \ldots, b_0, \ldots, b_{r-2}, b_{r-1}$$

of $\tilde{\tau}$, where b_0 is the lift of b meeting $\tilde{k}_b$ and $\pi : \mathcal{N}_\epsilon(\tilde{\lambda}) \to \tilde{\tau}$ is the collapse map. By equivariance, $r_b(d)$ is clearly independent of the choice of lift $\tilde{k}_b$ of k_b.

Remark 14.1 After projecting back down to $\tau \subset X$, either $b_{-r+1} \cdots b_0$ or $b_0 \cdots b_{r-1}$ defines a train route γ_d in τ that starts at b and terminates by "opening up" into the projection of H_v in X. That is, the geodesics g_d^+ and g_d^- diverge from each other (at scale ϵ) at the terminus of γ_d.

Now there are boundedly many spikes of $X \setminus \lambda$, and for each $r \geq 1$ each spike may contain at most one component $d \subset k_b \setminus \lambda$ with depth exactly r. This gives us the following bound:

Lemma 14.2 [Bonahon 1996, Lemma 4; Sözen and Bonahon 2001, Lemma 5] *For any branch b of τ and any transversal k_b, the number of components d of $k_b \setminus \lambda$ with $r_b(d) = r$ is at most $6|\chi(S)|$.*

The train track interpretation of the depth of a segment also allows us to bound the value of a shear-shape cocycle $\mathfrak{s}$ in terms of its weights on a snug train track and the depth of its endpoints.

More specifically, for each component d of $k_b \setminus \lambda$, let k_b^d be the subarc of k_b joining the initial point of k_b to any point of d. Then, for any combinatorial deformation $\mathfrak{s}$ and b a branch of $\tau_{\underline{\alpha}}$, there is an explicit formula for $\mathfrak{s}(k_b^d)$ as a linear function of the weights of $\mathfrak{s}$ on $\tau_{\underline{\alpha}}$ with at most $r_b(d)$ terms [Bonahon 1997b, Lemma 6]. Conceptually, this formula arises by splitting $\tau_{\underline{\alpha}}$ open along the spike s containing d, until d is "visible" in some new track $\tau'_{\underline{\alpha}}$ carried by $\tau_{\underline{\alpha}}$ (see also the proof of Lemma 13.9).

The exact expression for $\mathfrak{s}(k_b^d)$ will not be important for us; instead, we record the following estimate, which follows by considering the growth of edge weights upon splitting.

Lemma 14.3 [Bonahon 1996, Lemma 6; Sözen and Bonahon 2001, Lemma 6] *Let k_b be a transversal of a branch b. Then*

$$|\mathfrak{s}(k_b^d)| \leq \|\mathfrak{s}\|_{\tau_{\underline{\alpha}}} r_b(d)$$

for every $\mathfrak{s} \in \mathcal{SH}(\lambda; \underline{\alpha})$ and every component d of $k_b \setminus \lambda$.

We remark that our definitions of $\|\cdot\|_{\tau_{\underline{\alpha}}}$ and $r_b(\cdot)$ make the bound given in Lemma 14.3 hold without a topological multiplicative factor, as in [Bonahon 1996].

Geometric estimates on depth The depth of a component d of $k_b \setminus \lambda$ is proportional to the distance from a lift $\tilde{d}$ to the vertex $u \in \mathcal{H}$ inside the corresponding spike. The constant of proportionality in turn depends on how quickly the spike of H_u containing $\tilde{d}$ returns to k_b on X; we now identify a quantity that will allow us to estimate this constant.

Let k be any geodesic arc transverse to λ such that each lift $\tilde{k}$ to $\widetilde{X}$ bounds a spike in every hexagon that it crosses; equivalently, the endpoints of $\tilde{k}$ lie in a simple pair of hexagons. As in Section 13.2, it suffices for k to be shorter than the shortest arc of $\underline{\alpha}(X)$. Now, for each leaf g of $\widetilde{\lambda}$, there is a bound $R_k(g) > 0$ for the distance in g between intersections of g with different lifts $\tilde{k}_1$ and $\tilde{k}_2$ of k. Indeed, any two lifts of k meeting g differ by a deck transformation $\gamma \in \pi_1(X)$ determined by a path in X that traces along the projection of a segment in g and then closes up along k.

We then define the *injectivity radius of X along λ* to be

$$\operatorname{inj}_\lambda(X) := \inf_{k \pitchfork \lambda} \inf_{g \subset \widetilde{\lambda}} R_k(g),$$

where the infimum is taken over all transverse arcs k whose endpoints lie in a simple pair of hexagons.

Equivalently, the injectivity radius of λ may also be computed by taking an ϵ such that the geometric train track $\tau_{\max}$ built from $\mathcal{N}_\epsilon(\lambda)$ is snug and such that, for all $\epsilon' > \epsilon$, the train track built from $\mathcal{N}_\epsilon(\lambda)$ is the same (not just equivalent) to $\tau_{\max}$, as follows.[19]

For each branch of $\tau_{\max}$, choose a tie t_b (that is, a leaf of the orthogeodesic (or horocyclic) foliation restricted to $\mathcal{N}_\epsilon(\lambda)$ that is transverse to b). The injectivity radius along λ is then equal to the infimum of the recurrence times of λ to any t_b. Using the "length along a geometric train track" function $\ell_{\tau_{\max}}$ defined in Section 5.2, we may therefore write

$$\operatorname{inj}_\lambda(X) = \inf_{\gamma \prec \tau_{\max}} \ell_{\tau_{\max}}(\gamma), \tag{29}$$

where the infimum is taken over all simple closed curves γ carried on the train track $\tau_{\max}$.

Remark 14.4 The length of the hyperbolic systole of X is clearly a lower bound for $\operatorname{inj}_\lambda(X)$, which is therefore positive. However, $\operatorname{inj}_\lambda(X)$ can be much larger than the length of the systole.

For example, if λ does not fill the surface then there can be a disjoint curve of arbitrarily small length. In addition, X may have a very short curve γ transverse to λ, and if λ does not twist around γ, then $\operatorname{inj}_\lambda(X)$ is necessarily very large.

We can now relate the geometry of small arcs to their depth and injectivity radius along λ.

[19] Any ϵ sufficiently close to the supremum of ϵ for which $\mathcal{N}_\epsilon(\lambda)$ is snug satisfies these conditions.

Lemma 14.5 [Bonahon 1996, Lemmas 3 and 5; Sözen and Bonahon 2001, Lemma 4] *Given a branch b of a geometric train track τ constructed from λ on X and a short transversal k_b, there exists $B > 0$ such that the following holds. For every component d of $k_b \setminus \lambda$ with depth $r_b(d)$,*

$$\ell_X(d) \leq Be^{-D_\lambda(X) r_b(d)},$$

where $D_\lambda(X) = \mathrm{inj}_\lambda(X)/9|\chi(S)|$.

Proof The idea is the same as in the references, but our constants are different. Small geodesic arcs meeting a spike s of a hexagon H_v transversely and far away from the vertex v look like horocycles, which have length that decays exponentially in distance from v. Therefore, we just need to give a lower bound for the distance between d and $v \in H_v$ along the spike s in terms of $\mathrm{inj}_\lambda(X)$ and the topological complexity of S.

Consider the train path γ_d starting at b that defines $r_b(d)$. By definition, γ_d traverses exactly $r_b(d)$ branches of τ (counted with multiplicity). Now γ_d decomposes as a concatenation of maximal sub–train paths with embedded interiors, each forming a *simple loop* in τ.[20]

The depth $r_b(d)$ is thus bounded above by the number of consecutive simple loops in γ_d times the size of the longest simple loop in τ. The size of a simple loop in τ is in turn bounded above by the number of branches of τ, which is at most $9|\chi(S)|$. Finally, since each simple loop in γ_d is carried on $\tau \prec \tau_{\max}$, it must have length at least $\mathrm{inj}_\lambda(X)$ by (29).

Putting the above estimates together, the distance between v and d in H_v is at least

$$\mathrm{inj}_\lambda(X) \cdot \#\{\text{simple loops in } \gamma_d\} \geq \frac{\mathrm{inj}_\lambda(X) r_b(d)}{\text{size of the longest simple loop in } \tau} \geq \frac{\mathrm{inj}_\lambda(X)}{9|\chi(S)|} r_b(d),$$

and the lemma follows. □

14.2 Shape-shifting in the spikes

Our discussion now begins to diverge from [Bonahon 1996]. While pairs of asymptotic geodesics are all isometric, the spikes of $X \setminus \lambda$ come with extra decoration, namely, a choice of horocycle at each cusp (equivalently, basepoints which lie on a common leaf of the orthogeodesic foliation). In this section, we explain how to use these decorations to define the shape-shifting cocycle $\varphi_{\mathfrak{s}}$ on pairs of basepointed geodesics coming from simple pairs of hexagons.

We remind the reader that X, λ and $\tau_{\underline{\alpha}}$ are fixed so that geometric objects like geodesic segments, hexagons, arcs of $\underline{\alpha}(X)$, etc are understood to live in and be realized (ortho)geodesically on X. Throughout this section we will fix $\underline{A} = \underline{A}(X)$ and use it to denote both a weighted arc system and the induced metric on $S \setminus \lambda$. Finally, we recall that $\mathfrak{s}$ is a combinatorial deformation of $\sigma_\lambda(X)$ which changes $\underline{A}$ by $\mathfrak{a}$; we will refer to the deformed hyperbolic structure on $S \setminus \lambda$ by $\underline{A} + \mathfrak{a}$ and its hexagonal pieces by G_u for $u \in \mathcal{H}$.

[20] A simple loop on a train track is a carried curve which traverses each branch at most once.

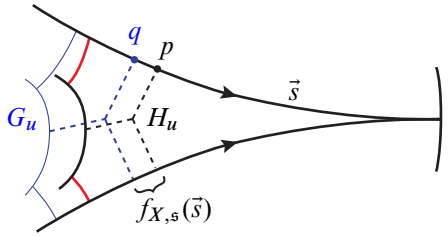


Figure 21: Superimposing hexagons to measure the difference in the shapes of their spikes.

Shapes of spikes The group $\mathsf{PSL}_2(\mathbb{R})$ acts transitively on pairs of asymptotic geodesics but, having done so, cannot further act on the family of horocycles based at the spike. To measure this failure, we associate below a geometric parameter which records the placement of basepoints in each spike.

Suppose that $u \in \mathcal{H}$ is a hexagon of $\widetilde{X} \setminus \widetilde{\lambda}_{\underline{\alpha}}$ and s is a spike of H_u, that is, a pair of asymptotic geodesics g and g'. Both g and g' come with basepoints p and p' obtained by projecting u to these geodesics. We then associate to s the number $h_{\underline{A}}(s)$ which measures the length of either of the orthogeodesic leaves connecting u to p or p':

$$h_{\underline{A}}(s) := d(p,u) = d(p',u).$$

Our notation reflects the fact that this function clearly depends only on the geometry of $X \setminus \lambda$ and not the shearing along λ. The reader familiar with the literature will observe that this parameter is essentially an orthogeodesic version of the "sharpness functions" appearing in [Thurston 1986].

In order to measure the difference in sharpness functions between the realizations of s in $\underline{A}$ and in the deformed metric $\underline{A} + \mathfrak{a}$, we superimpose the hexagons H_u and G_u and measure the distance between their boundary basepoints.

More concretely, choose an arbitrary orientation $\vec{s}$ of the spike s and fix realizations of both H_u and G_u inside $\mathbb{H}^2$. As $\mathsf{PSL}_2(\mathbb{R})$ acts simply transitively on triples in $\partial\mathbb{H}^2$, there is a unique isometry that takes the realization of s in G_u to its realization in H_u. The vertex u of Sp is realized in both H_u and G_u; let p and q denote the projections of these points to one of the boundary geodesics g of s. See Figure 21.

Lemma 14.6 *With all notation as above, the signed distance from q to p along g is*

$$f_{X,\mathfrak{s}}(\vec{s}) := \varepsilon \log\left(\frac{\tanh h_{\underline{A}+\mathfrak{a}}(s)}{\tanh h_{\underline{A}}(s)}\right) \in \mathbb{R}, \tag{30}$$

where $\varepsilon = +1$ if $\vec{s}$ is oriented towards the shared ideal endpoint, and $\varepsilon = -1$ otherwise.

The parameter $f_{X,\mathfrak{s}}(\vec{s})$ plays a crucial role below in our definition of the shape-shifting map on spikes. In our convergence estimates, we will also need to consider the parameter

$$\|\mathfrak{s}\|_{\vec{s}} := \max_s |f_{X,\mathfrak{s}}(\vec{s})| < \infty, \tag{31}$$

which quantifies the maximum distance that the deformation $\mathfrak{s}$ moves a basepoint in a spike.

Proof We compute in the upper half-plane; up to isometry, we may assume that s is bound by the imaginary axis $V = i\mathbb{R}_{>0}$ and its translate $2 + V$; the spine of the orthogeodesic foliation in this spike is a subsegment of the vertical line $1 + V$. With this choice fixed, the projections p and q of u to V may be identified with ie^a and ie^b for some a and b, respectively. Without loss of generality, we may also assume that V is oriented upwards (towards ∞); the opposite choice of orientations simply reverses all signs at the end of the computation.

Now, for $t \geq 0$, the path $t \mapsto \tanh t + i \operatorname{sech} t$ is the unit-speed parametrization of the orthogeodesic emanating from V at i. Observe that the isometry $z \mapsto e^a z$ stabilizes V and takes this segment to an orthogeodesic segment emanating from $ie^a = p$ which is distance a from i. Since the orthogeodesic segment through p meets the spine $1 + V$ after traveling distance $h_{\underline{A}}(s)$ (by definition), this implies that

$$e^{-a} = \tanh h_{\underline{A}}(s).$$

Similarly, $e^{-b} = \tanh h_{\underline{A}+\mathfrak{a}}(s)$. Together, these imply that

$$\frac{\tanh h_{\underline{A}+\mathfrak{a}}(s_u)}{\tanh h_{\underline{A}}(s_u)} = e^{a-b}.$$

Taking logarithms, we see that $a - b$ is the signed distance from q to p along V, as claimed. □

Remark 14.7 By Theorem 6.4, the parameter $f_{X,\mathfrak{s}}(\vec{s})$ varies analytically in $\mathfrak{a}$ (and hence $\mathfrak{s}$).

Orientation conventions We now specialize to the case where (v, w) is a simple pair of hexagons with associated oriented geodesic $k_{v,w}$ running between p_v^w on g_v^w (the projection of v to the boundary leaf of $\partial_\lambda H_v$ closest to w) and p_w^v on g_w^v.

Each leaf $g \subset \widetilde{\lambda}$ crossed by $k_{v,w}$ inherits an orientation by declaring that *turning right* onto g while traveling from v to w along $k_{v,w}$ is the positive direction. We remark that if $k_{v,w}$ crosses a hexagon H_u, then the induced orientation of g_u^w, the geodesic in $\partial_\lambda H_u$ closest to w, is the *opposite* of the orientation of g_u^w induced as a part of the boundary of H_u. On the other hand, the two orientations on g_u^v induced by $k_{v,w}$ and coming from H_u agree. This is an artifact of our sign convention for measuring shears; see Remark 13.3.

If g is a complete oriented geodesic in the hyperbolic plane and $t \in \mathbb{R}$, we let T_g^t be the hyperbolic isometry stabilizing g and acting by oriented translation distance t along g. The opposite orientation of g will be denoted by $\bar{g}$, so that $T_{\bar{g}}^t = T_g^{-t}$.

For an oriented spike $\vec{s} = (g_u^v, g_u^w)$, its opposite orientation is $\overleftarrow{s} = (\bar{g}_u^w, \bar{g}_u^v)$. In particular, we note that, if $\vec{s}$ is an oriented spike of H_u crossed by $k_{v,w}$, then $\overleftarrow{s}$ is an oriented spike crossed by $k_{w,v} = \bar{k}_{v,w}$.

Shape-shifting in spikes Suppose (v, w) is a simple pair and suppose u is between v and w. Let $\vec{s} = (g_u^v, g_u^w)$ be the spike of u crossed by $k_{v,w}$ with basepoints p_v and p_w. We define the elementary

shaping transformation $A(\vec{s}) \in \mathrm{Isom}^+(\widetilde{X}) = \mathsf{PSL}_2\,\mathbb{R}$ determined by X, $\mathfrak{s}$ and s to be

$$A(\vec{s}) := T_{g_u^v}^{f_{X,\mathfrak{s}}(\vec{s})} \circ T_{g_u^w}^{-f_{X,\mathfrak{s}}(\vec{s})}. \tag{32}$$

Ultimately, the element $A(\vec{s})$ will be the value of the shape-shifting cocycle $\varphi_{\mathfrak{s}}$ on the pair (g_u^v, g_u^w); see just below for an explanation of how we think of $A(\vec{s})$ as "changing the shape" of s.

Observe that $A(\vec{s})$ is a parabolic transformation preserving the common ideal endpoint of s. A familiar computation shows that in the spike determined by g_u^v and $A(\vec{s})g_u^w$, the orthogeodesics emanating from p_v and $A(\vec{s})p_w$ meet at a point distance $h_{\underline{A}+\mathfrak{a}}(s)$ from each (supposing that the deformation is small enough).

To the oriented spike $\vec{s}$ of u, we also associate the *elementary shape-shift*

$$\varphi(\vec{s}) := T_{g_u^v}^{\mathfrak{s}(v,u)} \circ A(\vec{s}) \circ T_{g_u^w}^{-\mathfrak{s}(v,u)} = T_{g_u^v}^{\mathfrak{s}(v,u)+f_{X,\mathfrak{s}}(\vec{s})} \circ T_{g_u^w}^{-(\mathfrak{s}(v,u)+f_{X,\mathfrak{s}}(\vec{s}))}, \tag{33}$$

where we recall that the value $\mathfrak{s}(v,u)$ is obtained by thinking of $\mathfrak{s}$ as a function on transverse pairs (à la Lemma 13.11). Note that $\varphi(s)$ depends on our reference point v: whereas $A(\vec{s})$ is eventually identified as a value of the shape-shifting cocycle $\varphi_{\mathfrak{s}}$, the elementary shape-shifts $\varphi(\vec{s})$ are only building blocks for values of $\varphi_{\mathfrak{s}}$.

For the opposite orientation $\overleftarrow{s} = (\bar{g}_u^w, \bar{g}_u^v)$, we check

$$A(\overleftarrow{s}) = T_{\bar{g}_u^w}^{f_{X,\mathfrak{s}}(\overleftarrow{s})} T_{\bar{g}_u^v}^{-f_{X,\mathfrak{s}}(\overleftarrow{s})} = T_{g_u^w}^{f_{X,\mathfrak{s}}(\vec{s})} T_{g_u^v}^{-f_{X,\mathfrak{s}}(\vec{s})} = A(\vec{s})^{-1}. \tag{34}$$

Since $\mathfrak{s}(v,u) = \mathfrak{s}(u,v)$, we may similarly observe that $\varphi(\overleftarrow{s}) = \varphi(\vec{s})^{-1}$.

Take $\mathcal{H}_{v,w}$ to be the set of hexagons between v and w equipped with the linear order $u_1 < u_2$ induced by the orientation of $k_{v,w}$. Let $\underline{\mathcal{H}} \subset \mathcal{H}_{v,w}$ be any finite subset and order its elements $\underline{\mathcal{H}} = \{u_1, \dots, u_n\}$. For short, we denote hexagons by $H_i := H_{u_i}$, spikes by $s_i := \vec{s}_{u_i}$, geodesics by $g_i^v := g_{u_i}^v$, etc.

To the finite ordered set $\underline{\mathcal{H}}$ we associate the product

$$\varphi_{\underline{\mathcal{H}}} := \varphi(s_1) \circ \cdots \circ \varphi(s_n) \circ T_{g_w^v}^{\mathfrak{s}(v,w)} \in \mathrm{Isom}^+(\widetilde{X}). \tag{35}$$

The goal of the rest of the section is then to extract a meaningful limit from $\varphi_{\underline{\mathcal{H}}}$ as $\underline{\mathcal{H}}$ increases to $\mathcal{H}_{v,w}$. Ultimately, this limit is how we will define the shape-shifting cocycle $\varphi_{\mathfrak{s}}$ on the boundary geodesics g_v^w and g_w^v corresponding to the simple pair (v,w).

Remark 14.8 In the case that λ is maximal, each H_i is an ideal triangle and so $\underline{A} = \underline{A} + \mathfrak{a} = \varnothing$. In this case, each spike parameter $f_{X,\mathfrak{s}}(s_i)$ is 0 and we recover the formula from [Bonahon 1996, page 255].

Geometric explanation of (33) Before proving convergence, however, let us explain the intuition behind the formulas above. In order to interpret $A(\vec{s})$ and $\varphi(\vec{s})$ as deformations of the hyperbolic structure X, we will switch our viewpoint to think of them as values of a deformation cocycle, and so as affecting the placement of pointed geodesics relative to each other. For brevity, let $f_{X,\mathfrak{s}}(\vec{s}) = t$.

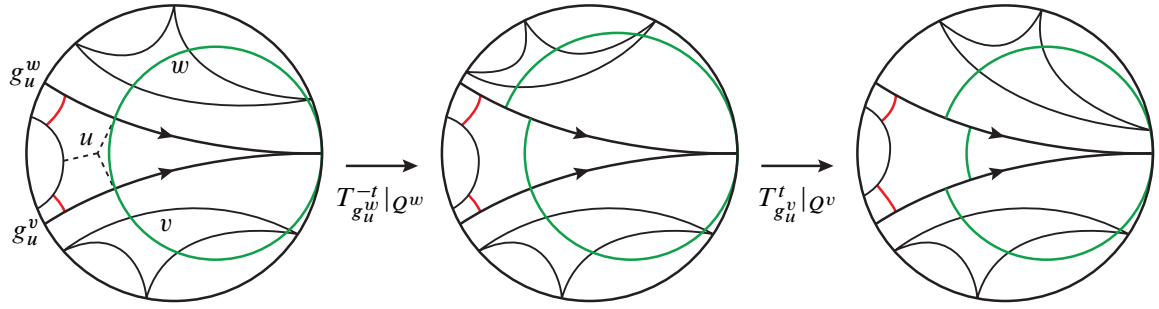


Figure 22: The effect of $A(\vec{s})$ when considered as a composition of left and right earthquakes.

Let us focus first on the shaping transformation $A(\vec{s})$. The oriented spike $\vec{s}$ in the hexagon H_u is formed by two pointed geodesics (g_u^v, p_u^v) and (g_u^w, p_u^w). Fixing our viewpoint at (g_u^v, p_u^v), we may think of $A(\vec{s})$ as deforming $\widetilde{X}$ by holding (g_u^v, p_u^v) fixed and identifying (g_u^w, p_u^w) with $A(\vec{s})\cdot(g_u^v, p_u^v)$. This has the overall effect of "widening" the spike s so that its sharpness parameter increases from $h_{\underline{A}}$ to $h_{\underline{A}+\mathfrak{a}}$.

If instead we fix our basepoint to be outside of H_u, say at the basepoint p_v^w on $g_v^w \subset \partial_\lambda H_v$, then this transformation can viewed as a composition of left and right earthquakes. Let Q^w and Q^v denote the half-spaces to the left of the oriented geodesics g_u^w and g_u^v, respectively. Note that $Q^w \subset Q^v$. The deformation $A(\vec{s})$ may then be thought of as first transforming all geodesics of $\widetilde{\lambda}$ that lie in Q^w by $T^{-t}_{g_u^w}$; this has the effect of breaking $\widetilde{X}$ open along g_u^w and sliding Q_w to the left by distance t along g_u^w while keeping $\widetilde{X} \setminus Q^w$ fixed. The deformation then further transforms all geodesics in Q^v by $T^{t}_{g_u^v}$; this is equivalent to the right earthquake with fault locus g_u^v that slides Q^v to the right while keeping $\widetilde{X} \setminus Q^v$ fixed. The cumulative effect is then that the spike s has been "pushed" in the direction of $\vec{s}$ by distance t. See Figure 22.

Remark 14.9 We give one final interpretation of $A(\vec{s})$ as "sliding G_u along H_u" in the proof of Proposition 14.19 below (see also Figure 26), once we have set up the framework to understand the utility of this viewpoint.

In particular, note that the shear from H_v to H_u measured from p_v^w to the image of p_u^v under this composition of earthquakes has increased by $t = f_{X,\mathfrak{s}}(\vec{s})$. Therefore, if we let q_u^v denote the basepoint on g_u^v corresponding to the hexagon G_u, then the shear from H_v to H_u measured from p_v^w to the image of q_u^v under the deformation is exactly the original shear $\sigma_\lambda(X)(v,u)$ between v and u.

The elementary shape-shift $\varphi(\vec{s})$ can be interpreted in much the same way, but now the spike should be pushed distance $f_{X,\mathfrak{s}}(\vec{s}) + \mathfrak{s}(v,u)$, so that the resulting shear (measured between p_v^w and the image of q_u^v) is exactly $\sigma_\lambda(X)(v,u) + \mathfrak{s}(v,u)$.

Finally, the composition (35) can be thought of as a composition of the operations described above (read from right to left). Therefore, $\varphi_{\underline{\mathcal{H}}}$ first performs a right earthquake along g_v^w by $\mathfrak{s}(v,w)$, then performs

an elementary shape-shift to pushing the spike s_n by $\mathfrak{s}(v,u_n)+f_{X,\mathfrak{s}}(s_n)$, then performs a shape-shift for s_{n-1}, etc. Observe that, if q_i^v denotes the basepoint in g_i^v corresponding to G_{u_i}, then, by construction, the shear between v and each u_i measured from p_v^w to the image of q_i^v under the composite deformation is exactly the desired shear $\sigma_\lambda(X)(v,u_i)+\mathfrak{s}(v,u_i)$.

Assuming the convergence of $\varphi_{\underline{\mathcal{H}}}$ to a limit $\varphi_{v,w}$ (a step performed just below), the placement of $\varphi_{\underline{\mathcal{H}}}(g_w^v,p_w^v)$ limits to that of $\varphi_{v,w}(g_w^v,p_w^v)$. This in turn will be the placement of the geodesic (g_w^v,p_w^v) relative to (g_v^w,p_v^w) straightened in the deformed surface $\widetilde{X}_{\mathfrak{s}}$; see Lemma 15.6.

Convergence We now consider the limiting behavior of $\varphi_{\underline{\mathcal{H}}}$ as $\underline{\mathcal{H}}\to\mathcal{H}_{v,w}$; that a limit exists is almost exactly the content of [Bonahon 1996, Lemma 14]. We give a proof here for convenience of the reader and to make sure that we are extracting the correct radius of convergence, ie that the modifications in the cusps actually do not affect the radius of convergence (even though there are countably many contributions from changing the shape of each spike).

Recall from Lemma 14.5 that the function $D_\lambda(X)=\operatorname{inj}_\lambda(Y)/9|\chi(S)|$ gives a bound for the rate of decay of the length of a piece of a leaf of $\mathcal{C}_\lambda(X)$ in terms of its divergence radius.

Lemma 14.10 (compare [Bonahon 1996, Lemma 14]) *If $\|\mathfrak{s}\|_{\tau_{\underline{\alpha}}}<D_\lambda(X)$, then $\varphi_{\underline{\mathcal{H}}}$ converges to a well-defined isometry $\varphi_{v,w}$ as $\underline{\mathcal{H}}$ tends to $\mathcal{H}_{v,w}$.*

Definition 14.11 The limiting isometry $\varphi_{v,w}$ is called the *shape-shifting map* for the simple pair (v,w).

Remark 14.12 After combining all of our deformations in Section 14.5, the shape-shifting map $\varphi_{v,w}$ will be identified as the value of the shape-shifting cocycle $\varphi_{\mathfrak{s}}$ on the pair (g_v^w,g_w^v). However, due to the asymmetry of our current definition, it is not clear that $\varphi_{v,w}^{-1}=\varphi_{w,v}$. See Corollary 14.14.

Proof of Lemma 14.10 For brevity, we set $D=D_\lambda(X)$ for the remainder of the proof.

Identify $\widetilde{X}$ with $\mathbb{H}^2$ and $\mathrm{Isom}^+(\widetilde{X})$ with the unit tangent bundle $T^1\mathbb{H}^2$ so that the identity I is the vector over $p_v\in\widetilde{X}$ that is tangent to g_v^w and pointed in the positive direction with the orientation on g_v^w induced by $k_{v,w}$. Equip $T^1\mathbb{H}^2$ with a left-invariant metric d that is right-invariant with respect to the stabilizer of p_v. Finally, for $A\in\mathrm{Isom}^+(\widetilde{X})$, let $\|A\|:=d(I,A)$, so that $\|AB\|\le\|A\|+\|B\|$ holds by the triangle inequality.

We first show that $\varphi_{\underline{\mathcal{H}}}$ stays in a compact set in $\mathrm{Isom}^+(\widetilde{X})$. Using boundedness, we then show that any sequence $\underline{\mathcal{H}}\to\mathcal{H}$ is in fact Cauchy with respect to d, and hence converges.

We start by bounding the lengths of segments of the form $k_{v,w}\cap H_u$, where $u\in\mathcal{H}_{v,w}$. To this end, construct a geometric train track τ from λ on X, and assume that the projection of $k_{v,w}$ meets τ transversely. Subdivide $k_{v,w}$ into arcs $k_1,\dots,k_m$ whose projections meet τ once in branches $b_1,\dots,b_m$. For all but finitely many $u\in\mathcal{H}_{v,w}$, we have $k_{v,w}\cap H_u\subset k_j\setminus\widetilde{\lambda}$ for some $j=1,\dots,m$.

If $d \subset k_j \setminus \widetilde{\lambda}$, we set $r(d)$ to be the depth $r_{b_j}(d)$ of d with respect to b_j, and set $r(d) = 1$ otherwise. By Lemma 14.5, there is $B > 0$ such that for all $u \in \mathscr{H}_{v,w}$,

$$\ell(k_{v,w} \cap H_u) \leq Be^{-Dr(k_{v,w} \cap H_u)}.$$

With this estimate in mind, our next task is to give a uniform bound on $\|\varphi_{\underline{\mathscr{H}}} \circ T_{g_w^v}^{-\mathfrak{s}(v,w)}\|$ for all finite $\underline{\mathscr{H}} \subset \mathscr{H}_{v,w}$. For each i, let $\gamma_i \in \mathrm{Isom}^+(\widetilde{X})$ be the isometry corresponding to the tangent vector over $k_{v,w} \cap g_i^v$ pointing toward the positive endpoint of g_i^v. Unpacking definitions, we may therefore write the shape-shift $\varphi(s_i)$ as

$$\varphi(s_i) = \gamma_i T_{g_v^w}^{\mathfrak{s}(v,u_i)+f_{X,\mathfrak{s}}(s_i)} T_{h_i}^{-(\mathfrak{s}(v,u_i)+f_{X,\mathfrak{s}}(s_i))} \gamma_i^{-1},$$

where $h_i := \gamma_i^{-1} g_i^w$.

An explicit computation (in the upper half-plane model, say) shows that

$$\begin{aligned}\|T_{g_v^w}^{\mathfrak{s}(v,u_i)+f_{X,\mathfrak{s}}(s_i)} T_{h_i}^{-(\mathfrak{s}(v,u_i)+f_{X,\mathfrak{s}}(s_i))}\| &\leq (e^{|\mathfrak{s}(v,u_i)+f_{X,\mathfrak{s}}(s_i)|} - 1)\ell(k_{v,w} \cap H_i) \\ &\leq Be^{|\mathfrak{s}(v,u_i)+f_{X,\mathfrak{s}}(s_i)|-Dr(k_{v,w} \cap H_i)}.\end{aligned}$$

By Lemma 14.3 and the triangle inequality,

$$|\mathfrak{s}(v,u_i) + f_{X,\mathfrak{s}}(s_i)| \leq \|\mathfrak{s}\|_{\tau_{\underline{\alpha}}} r(k_{v,w} \cap H_i) + \|\mathfrak{s}\|_{\vec{s}}$$

and so we conclude that

$$\|\gamma_i^{-1}\varphi(s_i)\gamma_i\| \leq B' e^{r(k_{v,w} \cap H_i)(\|\mathfrak{s}\|_{\tau_{\underline{\alpha}}} - D)}$$

for $B' = Be^{\|\mathfrak{s}\|_{\vec{s}}}$.

Notice now that conjugation by γ_i changes the reference point of our calculation at a distance in the plane at most $\ell(k_{v,w})$, so the effect of $\gamma_i^{-1}\varphi(s_i)\gamma_i$ on $g_i^v \cap k_{v,w}$ is a displacement by $e^{\ell(k_{v,w})}$ times the quantity indicated above. Since this is independent of $\underline{\mathscr{H}}$,

$$\|\varphi(s_i)\| = O(e^{r(k_{v,w} \cap H_i)(\|\mathfrak{s}\|_{\tau_{\underline{\alpha}}} - D)}) \tag{36}$$

for any spike s_i corresponding to any hexagon u between v and w.

Expanding out $\varphi_{\underline{\mathscr{H}}}$ in terms of the $\varphi(s_i)$ (see (35)),

$$\|\varphi_{\underline{\mathscr{H}}} \circ T_{g_w^v}^{-\mathfrak{s}(v,w)}\| = \left\| \prod_{i=1}^{n} \varphi(s_i) \right\| \leq \sum_{i=1}^{n} \|\varphi(s_i)\| = O\left(\sum_{i=1}^{n} e^{r(k_{v,w} \cap H_i)(\|\mathfrak{s}\|_{\tau_{\underline{\alpha}}} - D)} \right).$$

Since there is a uniformly bounded number of gaps with given depth (Lemma 14.2), the last expression is bounded by the sum of at most $6|\chi(S)|$ many geometric series which are convergent so long as $\|\mathfrak{s}\|_{\tau_{\underline{\alpha}}} < D$. Therefore, there is a compact set K in $\mathrm{Isom}^+(\widetilde{X})$ such that $\varphi_{\underline{\mathscr{H}}} \in K$ for any finite subset $\underline{\mathscr{H}} \subset \mathscr{H}_{v,w}$.

Now that we have shown the family of isometries $\{\varphi_{\underline{\mathscr{H}}}\}$ to be uniformly bounded, we can show that any sequence of refinements is in fact Cauchy. So suppose that $\underline{\mathscr{H}}_n$ increases to $\mathscr{H}_{v,w}$ and $|\underline{\mathscr{H}}_n| = n$. By construction, we may therefore write

$$\varphi_n = \psi\psi' \quad \text{and} \quad \varphi_{n+1} = \psi\varphi(s_u)\psi',$$

where $\underline{\mathcal{H}}_{n+1} = \underline{\mathcal{H}}_n \cup \{u\}$ and $\psi, \psi' \in K$. Writing $\varphi_{n+1} = \psi\psi'\varphi(s_u)[\varphi(s_u)^{-1}, \psi'^{-1}]$, we have

$$d(\varphi_n, \varphi_{n+1}) = \|\varphi(s_u)[\varphi(s_u)^{-1}, \psi'^{-1}]\|.$$

The zeroth-order term in the Taylor expansion near the identity for the function $X \mapsto \|[X, \psi'^{-1}]\|$ is 0, because $[I, \psi'^{-1}] = I$. Since ψ'^{-1} stays in a compact set,

$$\|[\varphi(s_u)^{-1}, \psi'^{-1}]\| = O\big(\|\varphi(s_u)\|\big)$$

(see [Thurston 1997, Theorem 4.1.6] or [Gelander 2014, Lecture 2, Lemma 2.1]).

Combining this estimate with the triangle inequality and (36),

$$d(\varphi_n, \varphi_{n+1}) = O\big(\|\varphi(s_u)\|\big) = O(e^{r(k_{v,w} \cap H_u)(\|\mathfrak{s}\|_{\tau_{\underline{\alpha}}} - D)}).$$

Now there are at most $6|\chi(S)|$ many $u \in \mathcal{H}_{v,w}$ with $r(k_{v,w} \cap H_u) = r$ (Lemma 14.2), so as $n \to \infty$ we must have that $r \to \infty$, and hence $d(\varphi_n, \varphi_{n+1}) \to 0$. Moreover, since this goes to 0 exponentially quickly, the sequence is in fact Cauchy. □

Shape-shifting as a limit of signed earthquakes Here we give a different description of the shape-shifting map which forgoes approximations by "pushing spikes" in favor of approximations by left and right simple earthquakes (compare [Epstein and Marden 2006, Section III]). While this reformulation is symmetric and geometrically meaningful, it comes at the cost of restricting which approximating sequences $\underline{\mathcal{H}} \to \mathcal{H}$ actually yield convergent sequences of deformations $\varphi_{\underline{\mathcal{H}}}$. See also the remark at the top of page 261 in [Bonahon 1996].

Let (v, w) be a simple pair and fix a geometric train track τ snugly carrying λ. So long as τ is built from a small enough neighborhood, we may assume that the geodesic $k_{v,w}$ is transverse to the branches of τ. Then, for each integer $r \geq 0$, let $\mathcal{H}^r_{v,w}$ denote the set of hexagons such that $k_{v,w} \cap H_u$ has depth at most r with respect to the branches of τ. Order

$$\mathcal{H}^r_{v,w} = (u_0 = v, u_1, \dots, u_n, u_{n+1} = w),$$

and, for each $i = 0, \dots, n$, choose a geodesic h_i that separates the interior of H_i from the interior of H_{i+1}. Orient each h_i so that it crosses $k_{v,w}$ from left to right and set

$$\varphi^r_{v,w} = T^{\mathfrak{s}(u_0,u_1)}_{h_0} \circ A(s_1) \circ T^{\mathfrak{s}(u_1,u_2)}_{h_1} \circ A(s_2) \circ \cdots \circ A(s_n) \circ T^{\mathfrak{s}(u_n,u_{n+1})}_{h_n}. \tag{37}$$

We now wish to show that $\varphi^r_{v,w} \to \varphi_{v,w}$ as $r \to \infty$. As in the case of $\varphi_{\underline{\mathcal{H}}} \to \varphi_{v,w}$, this argument will parallel that of [Bonahon 1996], with the extra complicating factor of the adjustments $A(s_i)$ to the shape of cusps.

The interpretation of (37) as a deformation cocycle is now similar to that of (35), but is now a combination of spike-shaping transformations together with simple earthquakes.

Let us give a description of the action of this deformation on the pointed geodesic (g_w^v, p_w^v) in $\partial_\lambda H_w$ closest to v. Reading the formula from right to left, we can obtain $\varphi^r_{v,w}(g_w^v, p_w^v)$ by first breaking $\widetilde{X}$ along $h_n = g_w^v$ and sliding the closed half-space containing H_w signed distance $\mathfrak{s}(u_n, u_{n+1})$, keeping the open half-space containing H_v fixed. Applying the spike shaping transformation $A(s_n)$ then preserves the natural basepoints p_n^v and p_n^w but increases the sharpness parameter $h_{\underline{A}}(s_n)$, making it match that of the spike in the hexagon G_{u_n} in the deformed metric $\underline{A} + \mathfrak{a}$. We then simply continue moving from w to v (ie backwards along $k_{v,w}$), alternating between signed earthquakes in the h_i and shaping the next spike until we reach g_v^w. Note that, unlike the deformations associated to $\varphi_{\mathcal{H}}$, each step of the process requires only local information about the spike s_i and the shear between u_i and u_{i+1}.

Lemma 14.13 [Bonahon 1996, Lemma 16] *So long as* $\|\mathfrak{s}\|_{\tau_{\underline{\alpha}}} < \frac{1}{2} D_\lambda(X)$, *we have* $\lim_{r\to\infty} \varphi^r_{v,w} = \varphi_{v,w}$.

Proof Using additivity, $\mathfrak{s}(u_i, u_{i+1}) = \mathfrak{s}(v, u_{i+1}) - \mathfrak{s}(v, u_i)$, we observe that

$$\varphi^r_{v,w} = (T_{h_0}^{\mathfrak{s}(v,u_1)} A(s_1) T_{h_1}^{-\mathfrak{s}(v,u_1)})(T_{h_1}^{\mathfrak{s}(v,u_2)} A(s_2) T_{h_2}^{-\mathfrak{s}(v,u_2)}) \cdots (T_{h_{n-1}}^{\mathfrak{s}(v,u_n)} A(s_n) T_{h_n}^{-\mathfrak{s}(v,u_n)}) T_{h_n}^{\mathfrak{s}(v,w)}. \tag{38}$$

So $\varphi^r_{v,w}$ is obtained from $\varphi_{\mathcal{H}^r_{v,w}}$ by replacing each term of the form

$$\varphi(s_i) = T_{g_i^v}^{\mathfrak{s}(v,u_i)} A(s_i) T_{g_i^w}^{-\mathfrak{s}(v,u_i)}$$

with

$$\phi(s_i) := T_{h_{i-1}}^{\mathfrak{s}(v,u_i)} A(s_i) T_{h_i}^{-\mathfrak{s}(v,u_i)}$$

and $T_{g_w^v}^{\mathfrak{s}(v,w)}$ with $T_{h_n}^{\mathfrak{s}(v,w)}$.

The basic estimate we need is approximately how close $\varphi(s_i)$ is to $\phi(s_i)$ in $\mathrm{Isom}^+(\widetilde{X})$ as r tends to infinity. For this we will want to understand how closely h_{i-1} approximates g_i^v near its intersection with $k_{v,w}$, as well as for g_i^w and h_i.

By construction, h_i must be between g_i^w and g_{i+1}^v for each $i = 1, \dots, n$ and h_0 is between g_v^w and g_1^v. But g_i^w and g_{i+1}^v follow the same edge path of length $2r$ in $\tau \subset \tau_{\underline{\alpha}}$, for otherwise there would be another $u \in \mathcal{H}^r_{v,w}$ such that H_u separates H_i from H_{i+1}. Thus h_i follows the same edge path and fellow travels g_i^w and g_{i+1}^v for length at least $O(2rD_\lambda(X))$; using negative curvature, h_i is $O(e^{-D_\lambda(X)r})$ close to both g_i^w and g_{i+1}^v near $k_{v,w}$.

From closeness of these geodesics from the previous paragraph (and our estimates for $\|\varphi(s_i)\|$ from Lemma 14.10) it is possible to obtain the basic estimate

$$\|\phi(s_i)^{-1}\varphi(s_i)\| = O\big(\exp\big(\|\mathfrak{s}\|_{\tau_{\underline{\alpha}}} r(k_{v,w} \cap H_i) - rD_\lambda(X)\big)\big),$$

which is small when $\|\mathfrak{s}\|_{\tau_{\underline{\alpha}}} < D_\lambda(X)$. Notice that we have also used the fact that the adjustment parameter associated to each spike $\mathfrak{s}(s_i)$ is uniformly bounded; that said, even if it grew linearly in r we would obtain the same estimate (up to a multiplicative factor).

The rest of the argument ensuring that $\varphi^r_{v,w}$ and $\varphi_{\mathcal{H}^r_{v,w}}$ have the same limit as long as $\|\mathfrak{s}\|_{\tau_{\underline{\alpha}}} < \frac{1}{2} D_\lambda(X)$ follows [Bonahon 1996, Lemma 16] and is omitted. We remark that the factor of $\frac{1}{2}$ appearing at the end is a relic of the techniques used in [Bonahon 1996, Lemma 16]. □

The following simple fact was not apparent from the definition of $\varphi_{v,w}$ due to its lack of symmetry. Fortunately, the approximation of $\varphi_{v,w}$ by $\varphi^r_{v,w}$ gives us a symmetric description of $\varphi_{v,w}$.

Corollary 14.14 *If (v, w) is simple and $\|\mathfrak{s}\|_{\tau_{\underline{\alpha}}} < \frac{1}{2} D_\lambda(X)$, then*

$$\varphi_{w,v} = \varphi_{v,w}^{-1}.$$

Proof We observe first that $\mathscr{H}^r_{v,w} = \mathscr{H}^r_{w,v}$, so that each term of $\varphi^r_{v,w}$ appears in $\varphi^r_{w,v}$ with the opposite orientation. Now, by (34), the inverse of the shaping transformation of an oriented spike is equal to the shaping transformation of the same spike with opposite orientation. Therefore, $\varphi^r_{v,w} = (\varphi^r_{w,v})^{-1}$ for all r, and the equality holds as we take $r \to \infty$. □

14.3 Shape-shifting in hexagons

In this section, we explain how to define the shape-shifting cocycle $\varphi_{\mathfrak{s}}$ on pairs of basepointed geodesics that lie in the boundary of a common hexagon; this will encode the change in hyperbolic structure on $X \setminus \lambda$.

While in this setting we do not have to worry about delicate convergence results, we must be more diligent about recording the placement of basepoints on each geodesic of $\partial_\lambda H_u$. Moreover, the cocycle condition (Propositions 14.18 and 14.19) only becomes apparent once we reinterpret the shaping deformations defined below as "sliding the deformed hexagon along the original".

Throughout this section, we have extended both $\underline{A}$ and $\underline{A} + \mathfrak{a}$ to some common maximal arc system $\underline{\alpha}$ by adding in arcs of weight 0 as necessary. We remind the reader that $\mathfrak{s}(\alpha)$ denotes the coefficient of α in $\mathfrak{a}$.

Notation and orientations Let $H_u \subset \widetilde{X} \setminus \widetilde{\lambda}_{\underline{\alpha}}$ be a nondegenerate right-angled hexagon and enumerate the λ–boundary components of H_u as $\partial_\lambda H_u = \{(h_1, p_1), (h_2, p_2), (h_3, p_3)\}$, cyclically ordered about u. Let $\alpha_i \in \widetilde{\underline{\alpha}}$ be the orthogeodesic arc opposite to h_i, and denote by p_{ij} the vertex of H_u meeting both h_i and α_j. See Figure 23. If H_u is a degenerate hexagon (ie a pentagon with one ideal vertex or a quadrilateral with two), then we label only those points and geodesics which appear in its boundary.

Each choice of orientation $\vec{\alpha}_1$ of α_1 induces orientations of h_2 and h_3 such that α_1 leaves from the left-hand side of h_j and arrives on the right-hand side of h_k for $\{j, k\} = \{2, 3\}$; an example is pictured in Figure 23. Observe that the opposite orientation $\overleftarrow{\alpha}_1$ induces the opposite orientations on h_2 and h_3. Throughout this section, we also adopt similar conventions for each orientation of α_2 and α_3.

Recall that (by Theorem 6.4) the deformation $\mathfrak{s}$ induces a new metric on $\widetilde{X} \setminus \widetilde{\lambda}$ denoted by $\underline{A} + \mathfrak{a}$ and which contains a hexagon G_u corresponding to H_u. The corresponding basepointed λ–boundary geodesics and vertices of G_u will be denoted by (g_i, q_i) and q_{ij}, respectively. We adopt similar orientation conventions as above for the realizations of α_j in G_u.

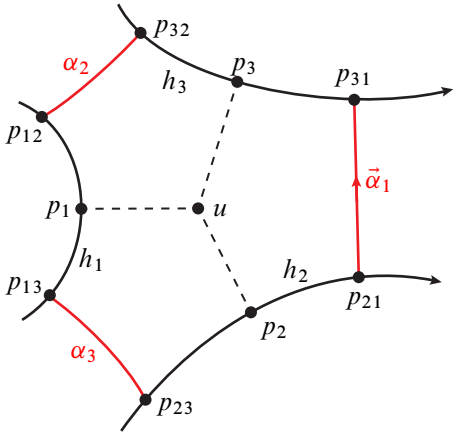


Figure 23: Distinguished points on a hexagon H_u and induced orientations on $h_2, h_3 \in \partial_\lambda H_u$.

Shapes of hexagons Paralleling our discussion for spikes, we first need to define geometric parameters that measure the shape of the hexagon as well as the difference of the placements of the basepoints p_i and q_i on the geodesics h_i and g_i. For concreteness, we only consider α_1 below; the parameters for α_2 and α_3 are defined symmetrically.

We begin by associating to α_1 the parameter

$$\ell_{\mathfrak{s}}(\alpha_1) := \ell_{\underline{A}+\mathfrak{a}}(\alpha_1) - \ell_{\underline{A}}(\alpha_1) \in \mathbb{R},$$

which measures the difference in the hyperbolic length of α_1 in the metric determined by $\underline{A} + \mathfrak{a}$ versus in the original metric $\underline{A}$ induced by X.

Now fix an orientation $\vec{\alpha}_1$ of α_1; as above, this induces orientations of the geodesics h_2, h_3, g_2 and g_3. Let $d_{\underline{A}}(\vec{\alpha}_1, u)$ be the signed distance from p_2 to p_{21} on h_2;[21] the local symmetry of the orthogeodesic foliation implies that $d_{\underline{A}}(\vec{\alpha}_1, u)$ can also be computed as the signed distance from p_3 to p_{31} on h_3. Define similarly $d_{\underline{A}+\mathfrak{a}}(\vec{\alpha}_1, u)$ as the distance from q_2 to q_{21} on g_2 (equivalently, the signed distance from q_3 to q_{31} on g_3).

To all of this information, we associate the parameter

$$f_{X,\mathfrak{s}}(\vec{\alpha}_1, u) := d_{\underline{A}+\mathfrak{a}}(\vec{\alpha}_1, u) - d_{\underline{A}}(\vec{\alpha}_1, u) \in \mathbb{R},$$

which measures the difference in how far u is from α_1 in G_u versus in H_u. More precisely, considering H_u and G_u in the hyperbolic plane, we can use an element of $\mathsf{PSL}_2(\mathbb{R})$ to line up (h_2, p_{21}) with (g_2, q_{21}) so that the basepoints and orientations agree. The parameter $f_{X,\mathfrak{s}}(\vec{\alpha}_1, u)$ then measures the distance from q_2 to p_2 along $h_2 = g_2$. See Figure 24. Of course, symmetry shows that it is equivalent to align (h_3, p_{31}) with (g_3, q_{31}) and measure the signed distance from q_3 to p_3 along $h_3 = g_3$.

[21]The parameter $d_{\underline{A}}(\vec{\alpha}_1, u)$ is called the "t–coordinate" of the arc α_1 in the hexagon H_u in [Luo 2007]. See also [Mondello 2009b, Proposition 2.10], where a formula is given in terms of the lengths of $\{\alpha_i : i = 1, 2, 3\}$.

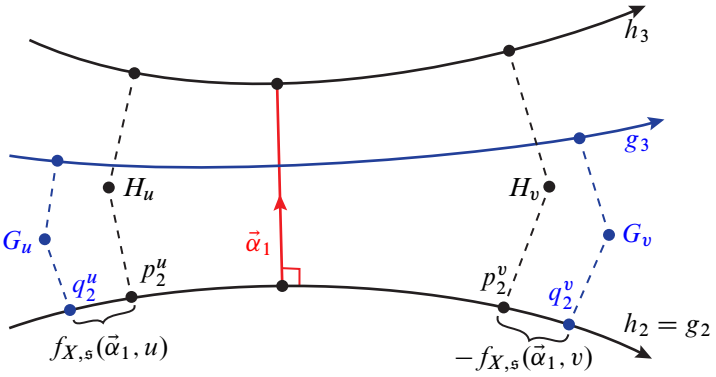


Figure 24: The parameter $f_{X,\mathfrak{s}}(\vec{\alpha}_1, u)$ for two adjacent hexagons. We have decorated the basepoints on $h_2 = g_2$ with a superscript to emphasize their dependence on the hexagon.

Note that reversing orientations reverses signs, so that $d_{\underline{A}}(\vec{\alpha}_1, u) = -d_{\underline{A}}(\overleftarrow{\alpha}_1, u)$ and hence

$$f_{X,\mathfrak{s}}(\vec{\alpha}_1, u) = -f_{X,\mathfrak{s}}(\overleftarrow{\alpha}_1, u).$$

The parameters associated to the hexagons which border a given arc are related in the following way:

Lemma 14.15 *Let α be any edge of $\widetilde{\underline{\alpha}}$ and let H_u and H_v be its adjoining hexagons. Then*

$$f_{X,\mathfrak{s}}(\vec{\alpha}, u) + f_{X,\mathfrak{s}}(\overleftarrow{\alpha}, v) = \mathfrak{s}(\alpha),$$

where the orientation $\vec{\alpha}$ is chosen so that u is on its left (equivalently $\overleftarrow{\alpha}$ is oriented so that v is on its left).

Proof The proof is an exercise in unpacking the definitions and being careful with orientations; compare Figure 24. Let p_2^u and p_2^v denote the projections of u and v to either of geodesics common to $\partial_\lambda H_u$ and $\partial_\lambda H_v$, and let q_2^u and q_2^v play similar roles for G_u and G_v.

We can then write

$$\begin{aligned} f_{X,\mathfrak{s}}(\vec{\alpha}, u) + f_{X,\mathfrak{s}}(\overleftarrow{\alpha}, v) &= f_{X,\mathfrak{s}}(\vec{\alpha}, u) - f_{X,\mathfrak{s}}(\vec{\alpha}, v) \\ &= d_{\underline{A}+\mathfrak{a}}(\vec{\alpha}, u) - d_{\underline{A}}(\vec{\alpha}, u) - d_{\underline{A}+\mathfrak{a}}(\vec{\alpha}, v) + d_{\underline{A}}(\vec{\alpha}, v) \\ &= d_h(p_2^u, p_2^v) - d_g(q_2^u, q_2^v) = \mathfrak{s}(\alpha), \end{aligned}$$

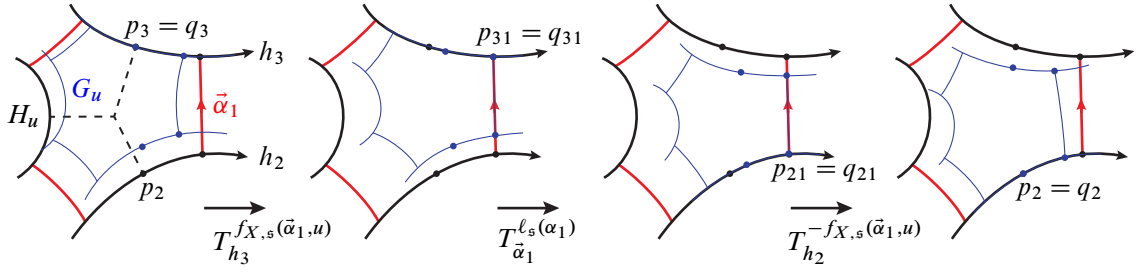


Figure 25: How the shaping transformation $A(\vec{\alpha}_1, u)$ slides G_u along H_u.

where we recall that $\mathfrak{s}(\alpha)$ denotes the coefficient of α in $\mathfrak{a}$ and where d_h and d_g represent the signed distance measured along h_2 and g_2, equipped with the orientation induced by $\vec{\alpha}$. □

Remark 14.16 Using Theorem 6.4 and some hyperbolic trigonometry, one may show that $f_{X,\mathfrak{s}}(\vec{\alpha}_1, u)$ depends analytically on both $\underline{A}$ and $\mathfrak{a}$ for fixed α_1 and u.

Shaping hexagons To the hexagon H_u and oriented arc $\vec{\alpha}_1$ in its boundary, we associate the *shaping transformation* $A(\vec{\alpha}_1, u)$ given by

$$A(\vec{\alpha}_1, u) := T_{h_2}^{-f_{X,\mathfrak{s}}(\vec{\alpha}_1,u)} \circ T_{\vec{\alpha}_1}^{\ell_{\mathfrak{s}}(\alpha_1)} \circ T_{h_3}^{f_{X,\mathfrak{s}}(\vec{\alpha}_1,u)} \in \mathrm{Isom}^+(\widetilde{X}), \tag{39}$$

where $T_{\vec{\alpha}_1}$ denotes translation along the complete oriented geodesic extending $\vec{\alpha}_1$. The shaping transformation is explicitly constructed so that, if H_u and G_u are superimposed with $(h_2, p_2) = (g_2, q_2)$, then

$$A(\vec{\alpha}_1, u)(h_3, p_3) = (g_3, q_3).$$

This claim is not immediately apparent from the expression of (39), but is easy to verify once we reinterpret $A(\vec{\alpha}_1, u)$ as "sliding G_u along H_u".

To wit, suppose that we superimpose H_u and G_u so that $(h_3, p_3) = (g_3, q_3)$. Now consider what happens as we apply $A(\vec{\alpha}_1, u)$ to G_u while holding H_u fixed; the first term $T_{h_3}^{f_{X,\mathfrak{s}}(\vec{\alpha}_1,u)}$ translates G_u along h_3 so that $q_{31} = p_{31}$, and the right angle formed by α_1 and g_3 in G_u lines up with the same angle in H_u. The transformation $T_{\vec{\alpha}_1}^{\ell_{\mathfrak{s}}(\alpha_1)}$ then slides $T_{h_3}^{f_{X,\mathfrak{s}}(\vec{\alpha}_1,u)} G_u$ along α_1 so that $(h_2, q_{21}) = (g_2, p_{21})$. Finally, $T_{h_2}^{-f_{X,\mathfrak{s}}(\vec{\alpha}_1,u)}$ slides $T_{\vec{\alpha}_1}^{\ell_{\mathfrak{s}}(\alpha_1)} T_{h_3}^{f_{X,\mathfrak{s}}(\vec{\alpha}_1,u)} G_u$ along $h_2 = g_2$ so that q_2 lines up with p_2. See Figure 25.

Summarizing, we have shown that $A(\vec{\alpha}_1, u)$ takes a superimposition of G_u on H_u with $(h_3, p_3) = (g_3, q_3)$ to another superimposition with $(h_2, p_2) = (g_2, q_2)$. In particular, this implies that applying $A(\vec{\alpha}_1, u)$ to (h_3, p_3) takes it to the position of (g_3, q_3) in the latter placement of G_u, which is what we claimed.

Remark 14.17 An elementary hyperbolic geometry argument similar to that in the proof of Lemma 14.6 shows that if α_1 in X degenerates to an oriented spike $\vec{s}$ then the corresponding geometric parameter $f_{X,\mathfrak{s}}(\vec{\alpha}_1, u)$ limits to the parameter $f_{X,\mathfrak{s}}(\vec{s})$. In particular, along this degeneration the corresponding hexagon-shaping transformation $A(\vec{\alpha}_1, u)$ converges to the spike-shaping transformation $A(\vec{s})$.

A cocycle condition for hexagons A number of relations hold between the shaping transformations for different arcs and different orientations; eventually, these relations are what ensure that the deformations we are currently building package together into an honest cocycle (see Proposition 14.26).

First, we observe that reversing the orientation of α_1 inverts the shaping transformation:

$$\begin{aligned} A(\overleftarrow{\alpha}_1, u) &= T_{\overleftarrow{h}_3}^{-f_{X,\mathfrak{s}}(\overleftarrow{\alpha}_1,u)} \circ T_{\overleftarrow{\alpha}_1}^{\ell_{\mathfrak{s}}(\alpha_1)} \circ T_{\overleftarrow{h}_2}^{f_{X,\mathfrak{s}}(\overleftarrow{\alpha}_1,u)} \\ &= T_{h_3}^{-f_{X,\mathfrak{s}}(\vec{\alpha}_1,u)} \circ T_{\vec{\alpha}_1}^{-\ell_{\mathfrak{s}}(\alpha_1)} \circ T_{h_2}^{f_{X,\mathfrak{s}}(\vec{\alpha}_1,u)} = A(\vec{\alpha}_1, u)^{-1}. \end{aligned} \tag{40}$$

Now suppose that H_u is on the left and H_v is on the right of the oriented arc $\vec{\alpha}_1$. Combining the relation of Lemma 14.15 with the definition of the shaping transformation, we obtain

$$A(\vec{\alpha}_1, u) = T_{h_2}^{\mathfrak{s}(\alpha_1)} \circ A(\vec{\alpha}_1, v) \circ T_{h_3}^{-\mathfrak{s}(\alpha_1)}. \tag{41}$$

This equation is used frequently in Section 14.4.

Finally, a beautiful and important relationship holds among the three shaping transformations in a single right-angled hexagon. Our proof utilizes the "sliding" viewpoint explained above; the statement seems difficult to prove just by writing down a string of Möbius transformations.

Proposition 14.18 *Let $u \in \mathcal{H}$ be a nondegenerate right-angled hexagon with boundary arcs $\vec{\alpha}_1$, $\vec{\alpha}_2$ and $\vec{\alpha}_3$, oriented so that H_u lies to the left of each $\vec{\alpha}_i$. Then*

$$A(\vec{\alpha}_3, u) \circ A(\vec{\alpha}_2, u) \circ A(\vec{\alpha}_1, u) = 1.$$

A similar statement clearly holds for any cyclic permutation of $(3, 2, 1)$.

Proof In order to prove the lemma, we superimpose G_u on top of H_u so that $(g_3, q_3) = (h_3, p_3)$. Holding H_u fixed, the first shaping transformation $A(\vec{\alpha}_1, u)$ slides G_u along h_3, then along α_1, then along h_2 so that (g_2, q_2) lines up with (h_2, p_2). The second shaping transformation $A(\vec{\alpha}_2, u)$ then acts on this translated copy of G_u by sliding it along h_2, then α_2, then h_1 so that $(g_1, q_1) = (h_1, p_1)$. Finally, the last term slides $A(\vec{\alpha}_1, u)A(\vec{\alpha}_2, u)G_u$ along the edges of H_u so that (g_3, q_3) returns to (h_3, p_3) (with the same orientation).

Therefore, since $A(\vec{\alpha}_1, u) \circ A(\vec{\alpha}_2, u) \circ A(\vec{\alpha}_3, u)$ preserves a unit tangent vector and $\mathrm{Isom}^+(\widetilde{X})$ acts simply transitively on $T^1\widetilde{X}$, the composition of the three shaping transformations must be trivial. □

A similar result holds for degenerate right-angled hexagons, with the hexagon-shaping transformation replaced with the corresponding spike-shaping transformation.

Proposition 14.19 *Suppose that $u \in \mathcal{H}$ is a pentagon with two orthogeodesic arcs α_1 and α_2 and one spike s, labeled so that (α_1, α_2, s) runs counterclockwise around u. Orient each α_j so that H_u is on its left and orient s so that it is pointing towards the ideal vertex of H_u. Then*

$$A(\vec{s}) \circ A(\vec{\alpha}_2, u) \circ A(\vec{\alpha}_1, u) = 1.$$

Similarly, if $u \in \mathcal{H}$ is a quadrilateral with one orthogeodesic edge α and two spikes s_1 and s_2 (labeled so that (α, s_1, s_2) is read counterclockwise), then

$$A(\vec{s}_2) \circ A(\vec{s}_1) \circ A(\vec{\alpha}, u) = 1,$$

where all orientation conventions are as above.

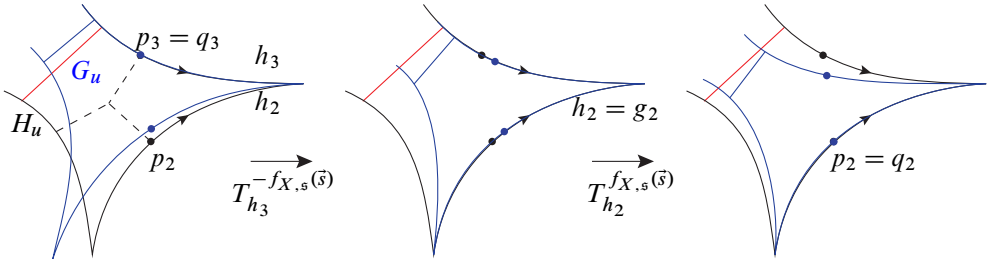


Figure 26: Interpreting the spike-shaping transformation $A(\vec{s})$ as sliding G_u along H_u. Note that, in this picture, $f_{X,\mathfrak{s}}(\vec{s}) < 0$.

Proof We only explain how to interpret the spike-shaping transformation $A(\vec{s})$ in our "sliding" framework; once we have done so, the rest of the proof is completely analogous to that of Proposition 14.18.

So let $\vec{s}$ be a spike of H_u, oriented as described; suppose that its left and right boundary geodesics are h_3 and h_2. Recall that $A(\vec{s})$ is constructed so that, if we superimpose G_u and H_u with $(g_2, q_2) = (h_2, p_2)$, then $A(\vec{s})(h_3, p_3) = (g_3, q_3)$. This can equivalently be interpreted by superimposing G_u on H_u with $(g_3, q_3) = (h_3, p_3)$; then applying the shaping transformation to G_u while leaving H_u fixed takes G_u to another superimposition where $(g_2, q_2) = (h_2, p_2)$. See Figure 26. □

14.4 Shape-shifting along the spine

In this section we package together the hexagon-shaping deformations defined in (39) into deformations of entire complementary subsurfaces of $\widetilde{X} \setminus \widetilde{\lambda}$. As always, we will exhibit this deformation by explaining how to adjust the positions of the pointed geodesics in the boundary of each component of $\widetilde{X} \setminus \widetilde{\lambda}$ relative to one another. This in turn requires some bookkeeping of orientations and a liberal application of the cocycle relation (Propositions 14.18 and 14.19).

Throughout this section, we fix some component Y of $\widetilde{X} \setminus \widetilde{\lambda}$. We remind the reader that the deformation $\mathfrak{s}$ induces a new hyperbolic structure $\underline{A} + \mathfrak{a}$ on Y whose hexagons and basepointed geodesics correspond to those of Y.

Hexagonal hulls and induced orientations Suppose that $v, w \in \mathcal{H}$ are distinct hexagons of Y. Since the corresponding component of $\widetilde{\mathsf{Sp}}$ is a tree it contains a unique oriented nonbacktracking edge path $[v, w]$ joining v to w. We then define the *hexagonal hull* $H(v, w)$ of (v, w) to be the union of all of the hexagons corresponding to the vertices of $[v, w]$. Define also the *truncated hexagonal hull* $\widehat{H}(v, w)$ by truncating each spike of $H(v, w)$ by the horocycle through the basepoints that are closest to the ideal vertex. Note that both $H(v, w)$ and $\widehat{H}(v, w)$ come with ($\pi_1(Y)$–equivariant) collections of basepoints on their boundaries obtained by projecting each of the vertices of $[v, w]$ onto the associated boundary geodesics.

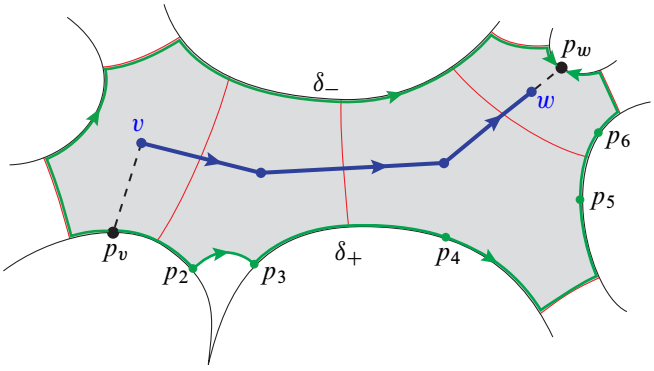


Figure 27: The truncated hexagonal hull (shaded) of the path $[v, w]$ and the induced orientations on the paths $\delta_\pm$ from p_v to p_w in its boundary.

Now, for any $(h_v, p_v) \in \partial_\lambda H_v$ and $(h_w, p_w) \in \partial_\lambda H_w$, we have that $\partial \widehat{H}(v, w) \setminus \{p_v, p_w\}$ consists of two paths $\delta_\pm$. We orient each of $\delta_\pm$ so that they both travel from p_v to p_w. See Figure 27.

With this induced orientation, the path δ_+ passes through a sequence of basepoints

$$p_v = p_1, p_2, \dots, p_{n+1} = p_w.$$

We then associate a shaping transformation A_i to each subsequent pair of basepoints as follows:

- If p_i and p_{i+1} are in different hexagons, then they must lie on the same geodesic h_i of ∂Y and correspond to two hexagons H_i and H_{i+1} both adjacent to an arc α_i. In this case, define $A_i = T_{h_i}^{\mathfrak{s}(\alpha_i)}$, where h_i is given the orientation induced by δ_+ and where we recall that $\mathfrak{s}(\alpha_i)$ is the coefficient of α_i in $\mathfrak{a}$.
- If p_i and p_{i+1} are in the same hexagon H_{u_i} but do not lie on a common spike, then necessarily they lie on geodesics connected by some arc α_i. In this case, define $A_i = A(\vec{\alpha}_i, u_i)$ where the orientation on α_i is induced from δ_+.
- If p_i and p_{i+1} lie on a common spike s_i, then we define $A_i = A(\vec{s}_i)$, where the orientation on s_i is such that the horocyclic segment of δ_+ cutting off s_i runs from the left of one of the oriented geodesics to the right of the other.

Finally, we then combine all of this information to define the shape-shifting transformation

$$A(\delta_+) := A_1 \circ A_2 \circ \cdots \circ A_n, \tag{42}$$

where we recall that we are multiplying from right to left. Define $A(\delta_-)$ analogously; the point is, however, that the choice of $\pm$ does not matter.

Lemma 14.20 $$A(\delta_-) = A(\delta_+).$$

Definition 14.21 We call $\varphi_{p_v, p_w} := A(\delta_+) = A(\delta_-)$ the *shape-shifting map* for $((h_v, p_v), (h_w, p_w))$.

Proof The proof follows by induction on the length of $[v, w]$. If $[v, w]$ has length 0, ie $v = w$, then this statement is exactly the content of the cocycle relation for hexagons (Propositions 14.18 and 14.19).

Now suppose that $[v, w]$ has length n and let u be the penultimate vertex in $[v, w]$. Let α denote the arc separating u from w, and choose the orientation $\vec{\alpha}$ so that u lies on its left. Up to relabeling, we may assume that the orientation of δ_+ agrees with the orientation of $\partial\widehat{H}(v, w)$. Denote by $(h_u^{\pm}, p_u^{\pm}) \in \partial Y$ the last basepoints of H_u visited by $\delta_\pm$ and let $\gamma_\pm$ denote the subpaths of $\delta_\pm$ from p_v to $p_u^{\pm}$ in the boundary of the truncated hexagonal hull $\widehat{H}(v, u)$. Define $A(\gamma_\pm)$ analogously to $A(\delta_\pm)$. Then we may write

$$\begin{aligned} A(\delta_+)A(\delta_-)^{-1} &= A(\gamma_+)T_{h_u^+}^{\mathfrak{s}(\alpha)} B_1 B_2 T_{h_u^-}^{-\mathfrak{s}(\alpha)} A(\gamma_-)^{-1} \\ &= (A(\gamma_+)A(\vec{\alpha}, u)A(\gamma_-)^{-1}) \cdot A(\gamma_-)(A(\vec{\alpha}, u)^{-1} T_{h_u^+}^{\mathfrak{s}(\alpha)} B_1 B_2 T_{h_u^-}^{-\mathfrak{s}(\alpha)}) A(\gamma_-)^{-1}, \end{aligned}$$

where B_1 and B_2 are the shaping transformations corresponding to arcs and spikes of w that are different from α (labeled counterclockwise from α), oriented either so that w lies on the left of the arc or so that the spike points into the common ideal endpoint.

Now observe that $A(\gamma_+)A(\vec{\alpha}, u)A(\gamma_-)^{-1}$ is trivial by the inductive hypothesis, as it corresponds to the comparison between the two possible definitions of φ_{p_v, p_u^-}. We also note that

$$A(\vec{\alpha}, u)^{-1} T_{h_u^+}^{\mathfrak{s}(\alpha)} B_1 B_2 T_{h_u^-}^{-\mathfrak{s}(\alpha)}$$

is conjugate to

$$T_{h_u^-}^{-\mathfrak{s}(\alpha)} A(\vec{\alpha}, u)^{-1} T_{h_u^+}^{\mathfrak{s}(\alpha)} B_1 B_2 = A(\bar{\alpha}, w) B_1 B_2 = 1,$$

where the first equality follows from (41) (note the reversals in orientations of $h_u^{\pm}$) and the second follows from the cocycle relation (Propositions 14.18 and 14.19). Therefore, the entire term $A(\delta_+)A(\delta_-)^{-1}$ is trivial, which is what we wanted to show. □

Remark 14.22 The above statement can also be proven by interpreting $A(\delta_\pm)$ in terms of sliding. In particular, let Z denote the $\pi_1(Y)$–equivariant hyperbolic structure on Y corresponding to the weighted arc system $\underline{A} + \mathfrak{a}$. Then, superimposing Z on Y so that $(g_w, q_w) = (h_w, p_w)$, one can consecutively apply the shaping transformations A_i to Z while keeping Y fixed.

Doing so, A_n moves Z so that $(g_n, q_n) = (h_n, p_n)$, then $A_{n-1} \circ A_n$ moves Z so that $(g_{n-1}, q_{n-1}) = (h_{n-1}, p_{n-1})$, etc. At the end of this process, we have applied $A(\delta_+)$ to Z and by construction, the pointed geodesic (g_v, q_v) matches up with (h_v, p_v). Since the final positioning of Z is the same relative to Y whether we used $A(\delta_+)$ or $A(\delta_-)$, this allows us to conclude that the two compositions define the same element.

Remark 14.23 While we used the distinguished boundary paths $\delta_\pm$ to define the shape-shifting map, one could in fact use *any* path from p_v to p_w in $Y \cup \widetilde{\underline{\alpha}}$. In this case, one must take more care to enumerate basepoints so that p_i and p_{i+1} always either lie on the same geodesic or in the same hexagon.

Observe that reversing the orientation $\overline{[v, w]} = [w, v]$ also reverses the sequence $p_{n+1}, \dots, p_1$ of basepoints that the boundary paths $\overline{\delta}_\pm$ meet. Since flipping the order of p_i and p_{i+1} inverts each of the A_i transformations defined above, we therefore discover that $\varphi_{p_w, p_v} = \varphi^{-1}_{p_v, p_w}$.

In a similar vein, it is not hard to see that the shape-shifting maps satisfy a cocycle relation.

Proposition 14.24 *For any triple of pointed geodesics (h_u, p_u), (h_v, p_v) and (h_w, p_w) of ∂Y,*

$$\varphi_{p_u, p_v} \circ \varphi_{p_v, p_w} \circ \varphi_{p_w, p_u} = 1.$$

Proof This follows immediately from the definitions when v lies on either of the paths $\delta_\pm$ from u to w.

Otherwise, note that the intersection of paths $[u, v] \cap [v, w] \cap [w, u]$ is a point $x \in \mathcal{H}$. Choosing a basepoint $p_x \in \partial_\lambda H_x$, compute the shape-shifting transformations using the boundary arcs that pass through x. Then we may express $\varphi_{p_u, p_v} = \varphi_{p_u, p_x} \circ \varphi_{p_x, p_v}$ and, using the observation about inverses above, we realize that

$$\varphi_{p_u, p_v} \circ \varphi_{p_v, p_w} \circ \varphi_{p_w, p_u} = \varphi_{p_u, p_x} \circ (\varphi_{p_x, p_v} \circ \varphi_{p_v, p_x}) \circ (\varphi_{p_x, p_w} \circ \varphi_{p_w, p_x}) \circ \varphi_{p_x, p_u} = 1. \qquad \square$$

14.5 The shape-shifting cocycle

We now combine the shape-shifting maps for simple pairs (Definition 14.11) with those for complementary subsurfaces (Definition 14.21) into the promised shape-shifting cocycle (Proposition 14.26), which is well defined as long as the combinatorial deformation $\mathfrak{s}$ is small enough. As usual, we construct a geometric train track τ from λ on X such that the weight space of $\tau_{\underline{\alpha}}$ provides a notion of size for $\mathfrak{s}$.

Admissible routes For $v, w \in \mathcal{H}$ and Y a component of $\widetilde{X} \setminus \widetilde{\lambda}$, we say that Y is *thick* with respect to v and w if either

(1) Y contains v and/or w, or

(2) v and w lie in different components of $\widetilde{X} \setminus Y$ and the boundary leaves of Y closest to v and w are not asymptotic.

Observe that, in the first case, there is either no or one boundary geodesic of Y separating v from w (depending on whether v and w are both in Y or not), while, in the second, there are exactly two boundary components of Y separating v from w.

Now let $v, w \in \mathcal{H}$ be any pair of distinct hexagons that do not lie in the same component of $\widetilde{X} \setminus \widetilde{\lambda}$ and let (h_v, p_v) and (h_w, p_w) be a pointed geodesic in $\partial_\lambda H_v$ and $\partial_\lambda H_w$. Then there is a unique (possibly empty) sequence $h_1, \dots, h_n$ of boundary geodesics of thick subsurfaces separating p_v from p_w, ordered by proximity to v (with h_1 closest).[22] If one of the h_i lies in the boundary of two complementary subsurfaces (so corresponds to a lift of a curve component of λ), then we record it twice, once for each of the adjoining

[22] This sequence is necessarily finite, as the distance that any geodesic travels in a thick subsurface is bounded below by the shortest arc of $\underline{\alpha}$ (compare the discussion of "close enough" pairs of hexagons in Section 13.2).

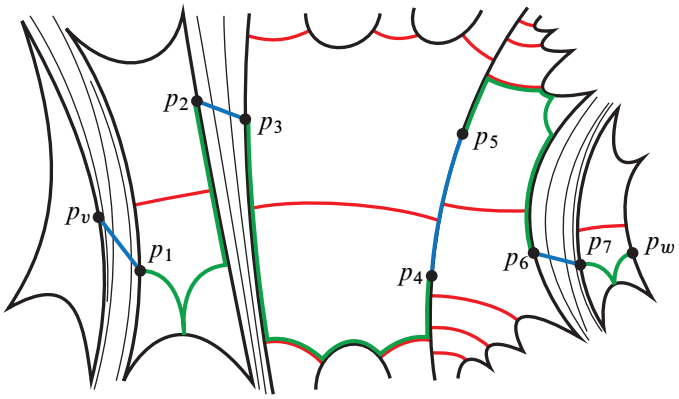


Figure 28: Thick subsurfaces between v and w and an admissible route from p_v to p_w. We have highlighted a path from p_v to p_w through the p_i; each subpath from p_i to p_{i+1} specifies a factor in the shape-shifting transformation.

subsurfaces. Additionally, if either h_v or h_w is a boundary geodesic separating v from w, then we do not record it as one of the h_i. See Figure 28.

We now define an *admissible route* from p_v to p_w to be any sequence of basepoints

$$p_v = p_0, \quad p_1 \in h_1, \quad \dots, \quad p_n \in h_n, \quad p_{n+1} = p_w$$

coming from the projections of the central vertices u_i of hexagons H_{u_i} to $h_i \in \partial_\lambda H_{u_i}$. If any geodesic $h_i = h_{i+1}$ is repeated, then we require that v and u_i lie on one side of h_i and that w and u_{i+1} lie on the other. Observe that the sequence of pairs (u_i, u_{i+1}) necessarily alternates between simple pairs/pairs sharing a boundary geodesic and pairs which lie in the same (thick) subsurface.

Shape-shifting along admissible routes To any admissible route we can then define a shape-shifting transformation by concatenating the shape-shifting transformations for subsequent pairs:

$$\varphi_{p_v,p_w} := \varphi_{p_0,p_1} \circ \cdots \circ \varphi_{p_n,p_{n+1}}, \tag{43}$$

where $\varphi_{p_i,p_{i+1}}$ is as in Definition 14.11 if (u_i, u_{i+1}) is simple and as in Definition 14.21 if u_i and u_{i+1} lie in the same subsurface. If $h_i = h_{i+1}$, then we orient h_i to the right as seen from u_i and set $\varphi_{p_i,p_{i+1}} = T_{h_i}^{s(u_i,u_{i+1})}$ (recall that we can associate a shear value to the pair (u_i, u_{i+1}) by (15)).

Lemma 14.25 *The shape-shifting map φ_{p_v,p_w} is independent of the choice of admissible route (as long as it is defined).*

Proof Since the h_i are uniquely determined, it suffices to change one point at a time.

So suppose that p_i and p_i' are both basepoints on the geodesic h_i; we then demonstrate the equality

$$\varphi_{p_{i-1},p_i} \circ \varphi_{p_i,p_{i+1}} = \varphi_{p_{i-1},p_i'} \circ \varphi_{p_i',p_{i+1}},$$

from which the lemma follows. Orient h_i so that it runs to the right as seen from v or u_{i-1}.

Without loss of generality, we may assume that the hexagons u_i and u_{i+1} lie in the same subsurface. Otherwise, the hexagons u_{i-1} and u_i lie in the same subsurface and so (p_i, p_{i+1}) is either simple or the points lie on the same isolated leaf. If this happens, we prove that

$$\varphi_{p_{i+1},p_i} \circ \varphi_{p_i,p_{i-1}} = \varphi_{p_{i+1},p_i'} \circ \varphi_{p_i',p_{i-1}},$$

which is equivalent to the equation above since each of the shape-shifting factors inverts when one flips the order of the points.

We first consider the shape-shifting transformations coming from comparing p_i or p_i' with p_{i+1}. By our reduction above, u_i and u_i' lie in the same thick subsurface Y. Let $\alpha_1, \dots, \alpha_m$ denote the arcs of $\widetilde{\underline{\alpha}} \cap Y$ encountered when traveling from p_i' to p_i along h_i; then our definition of shape-shifting in subsurfaces associates the transformation

$$\varphi_{p_i',p_i} = T_{h_i}^{\varepsilon_1 \sum_{j=1}^m \mathfrak{s}(\alpha_j)},$$

where $\varepsilon_1 = 1$ if h_i is oriented from p_i' to p_i and -1 otherwise. Combining this equation with the subsurface cocycle relation (Proposition 14.24),

$$\varphi_{p_i',p_{i+1}} = \varphi_{p_i',p_i} \circ \varphi_{p_i,p_{i+1}} = T_{h_i}^{\varepsilon_1 \sum_{j=1}^m \mathfrak{s}(\alpha_j)} \circ \varphi_{p_i,p_{i+1}}. \tag{44}$$

We now turn our attention to the transformation $\varphi_{p_{i-1},p_i'}$. Consider first the case when (p_{i-1}, p_i) is simple; since p_i and p_i' both lie on h_i, this implies that (p_{i-1}, p_i') is also simple. Moreover, since the geodesics $\mathcal{H}_{i-1,i}$ that separate p_{i-1} from p_i are the same that separate p_{i-1} from p_i', we may write

$$\varphi_{p_{i-1},p_i} = \lim_{\underline{\mathcal{H}} \to \mathcal{H}_{v,w}} \varphi(s_1) \circ \cdots \circ \varphi(s_n) \circ T_{h_i}^{\mathfrak{s}(u_{i-1},u_i)}$$

and similarly for $\varphi_{p_{i-1},p_i'}$. In particular, each approximation for φ_{p_{i-1},p_i} differs from the approximation for $\varphi_{p_{i-1},p_i'}$ by translation along h_i, and so the same is true in the limit:

$$\varphi_{p_{i-1},p_i'} = \varphi_{p_{i-1},p_i} \circ T_{h_i}^{\mathfrak{s}(u_{i-1},u_i') - \mathfrak{s}(u_{i-1},u_i)}. \tag{45}$$

Applying axiom (SH3) for shear-shape cocycles (Definition 7.11) multiple times, we compute that

$$\mathfrak{s}(u_{i-1}, u_i') - \mathfrak{s}(u_{i-1}, u_i) = \varepsilon_2 \sum_{j=1}^m \mathfrak{s}(\alpha_j), \tag{46}$$

where $\varepsilon_2 = +1$ if p_i precedes p_i' along h_i and -1 if p_i' precedes p_i. Combining (44), (45) and (46),

$$\varphi_{p_{i-1},p_i'} \circ \varphi_{p_i',p_{i+1}} = \varphi_{p_{i-1},p_i} \circ T_{h_i}^{\varepsilon_2 \sum_{j=1}^m \mathfrak{s}(\alpha_j)} \circ T_{h_i}^{\varepsilon_1 \sum_{j=1}^m \mathfrak{s}(\alpha_j)} \circ \varphi_{p_i,p_{i+1}} = \varphi_{p_{i-1},p_i} \circ \varphi_{p_i,p_{i+1}}$$

since $\varepsilon_2 = -\varepsilon_1$. This completes the proof of the lemma in the case when (u_{i-1}, u_i) is simple.

Similarly, if p_{i-1} and p_i lie on the same isolated leaf of λ, then so must p_i'. Unpacking the definitions shows that (45) holds in this case, and Lemma 7.14 implies that (46) does as well. Therefore, in this case we also see that the desired equality holds. □

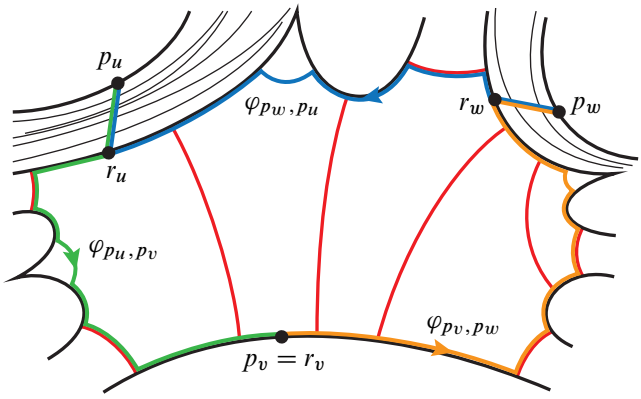


Figure 29: The cocycle relation for admissible routes.

Finally, now that we have constructed shape-shifting maps for arbitrary pairs of pointed geodesics in $\partial_\lambda \mathcal{H}$, we can prove that they piece together into an $\mathrm{Isom}^+(\widetilde{X})$–valued cocycle.

Proposition 14.26 *The map constructed from X and $\mathfrak{s}$,*

$$\varphi_{\mathfrak{s}} \colon \partial_\lambda \mathcal{H} \times \partial_\lambda \mathcal{H} \to \mathrm{Isom}^+(\widetilde{X}), \quad ((h_v, p_v), (h_w, p_w)) \mapsto \varphi_{p_v,p_w},$$

is a $\pi_1(X)$–equivariant 1–cocycle as long as $\|\mathfrak{s}\|_{\tau_{\underline{\alpha}}} < \frac{1}{2} D_\lambda(X)$.

Proof That φ is $\pi_1(X)$–equivariant means that $\varphi_{\gamma p_v, \gamma p_w} = \gamma \circ \varphi_{p_v,p_w} \circ \gamma^{-1}$ for $\gamma \in \pi_1(X)$; this follows directly from the construction.

That φ is a 1–cocycle means it satisfies the familiar cocycle condition on triples, ie

$$\varphi_{p_u,p_v} \circ \varphi_{p_v,p_w} = \varphi_{p_u,p_w}.$$

Note that, if p_v lies on some admissible route from p_u to p_w, then this is fulfilled automatically by unpacking the definitions and invoking Lemma 14.25.

One special case of the cocycle condition is when $p_u = p_w$; in this case we must show that $\varphi_{p_v,p_w} = \varphi^{-1}_{p_w,p_v}$. To demonstrate this, observe that reversing an admissible route from v to w produces an admissible route from w to v. Moreover, by Corollary 14.14 in the simple case and by definition in the other cases, each $\varphi_{p_i,p_{i+1}}$ also inverts when we flip i and $i+1$, proving that reversing v and w inverts φ_{p_v,p_w}.

Now suppose that u, v and w are all distinct; then there exists a unique subsurface Y of $\widetilde{X} \setminus \widetilde{\lambda}$ such that each component of $\widetilde{X} \setminus Y$ contains at most one of u, v or w (note that some of u, v and w may be inside Y). Choose basepoints r_u, r_v and r_w on the boundary components of Y that are closest to u, v and w (if any $\bullet \in \{u, v, w\}$ is in Y then set $r_\bullet = p_\bullet$). See Figure 29.

Choose an admissible route from p_u to p_v containing r_u and r_v, and similarly for the other two pairs. Then, by Lemma 14.25 and the observation that the cocycle condition holds along admissible routes, we

may write

$$\varphi_{p_u,p_v} = \varphi_{p_u,r_u} \circ \varphi_{r_u,r_v} \circ \varphi_{r_v,p_v}$$

and similarly for the other two pairs. Combining all three equations and applying the cocycle relation for Y (Proposition 14.24),

$$\varphi_{p_u,p_v} \circ \varphi_{p_v,p_w} = \varphi_{p_u,r_u} \circ \varphi_{r_u,r_v} \circ \varphi_{r_v,p_v} \circ \varphi_{p_v,r_v} \circ \varphi_{r_v,r_w} \circ \varphi_{r_w,p_w} = \varphi_{p_u,r_u} \circ \varphi_{r_u,r_w} \circ \varphi_{r_w,p_w} = \varphi_{p_u,p_w},$$

finishing the proof. See Figure 29 for a graphical depiction of this argument. □

15 Shear-shape coordinates are a homeomorphism

We now finish the proof of Theorem 12.1 by proving that the map $\sigma_\lambda \colon \mathcal{T}(S) \to \mathcal{SH}^+(\lambda)$ is open (Theorem 15.1) and proper and thus, by invariance of domain, a homeomorphism.

In Section 15.1, we use the shape-shifting cocycle $\varphi_{\mathfrak{s}}$, built in the previous section, to deform the representation $\rho \colon \pi_1(S) \to \mathsf{PSL}_2\,\mathbb{R}$ that induces the hyperbolic structure X. The deformed representation $\rho_{\mathfrak{s}}$ is then discrete and faithful (Lemma 15.3) and the quotient surface $X_{\mathfrak{s}}$ has the desired shear-shape cocycle (Lemma 15.6). In particular, this gives us a continuous local inverse to σ_λ, proving openness. These statements are similar in spirit to those in [Bonahon 1996], but the specifics of our proofs are different. In particular, instead of adjusting the relative placements of ideal triangles of $\widetilde{X} \setminus \widetilde{\lambda}$, we adjust the relative position of pointed geodesics in $\widetilde{\lambda}$.

We then prove properness of σ_λ in Section 15.2, concluding the proof of Theorem 12.1. Here we return to Bonahon's argument [1996, Theorem 20], but applying this strategy in our setting still requires a bit of extra care due to the polyhedral structure of $\mathcal{SH}^+(\lambda)$.

Finally, in Section 15.3, we show that the action of $\mathbb{R}_{>0}$ on $\mathcal{SH}^+(\lambda)$ by dilation produces lines in $\mathcal{T}(S)$ that can sometimes be identified with directed Thurston geodesics.

15.1 Deforming by shape-shifting

In this section, we show that any positive shear-shape cocycle close enough to $\sigma_\lambda(X)$ is actually the geometric shear-shape cocycle of a hyperbolic structure. Compare with [Bonahon 1996, Proposition 13].

Theorem 15.1 *Let $\underline{\beta}$ be a maximal arc system containing $\underline{\alpha}(X)$ and let $\tau_{\underline{\beta}}$ be a standard smoothing. Then, for any $\mathfrak{s} \in W(\tau_{\underline{\beta}})$ such that $\|\mathfrak{s}\|_{\tau_{\underline{\beta}}} < \frac{1}{2} D_\lambda(X)$ and such that $\sigma_\lambda(X) + \mathfrak{s}$ represents a positive shear-shape cocycle, there exists $X_{\mathfrak{s}} \in \mathcal{T}(S)$ close to X with*

$$\sigma_\lambda(X_{\mathfrak{s}}) = \sigma_\lambda(X) + \mathfrak{s}.$$

In particular, the image of $\sigma_\lambda(X)$ is open in $\mathcal{SH}^+(\lambda)$.

The proof of this theorem appears at the end of this subsection as the culmination of a series of structural lemmas. Our strategy is to explicitly define $X_{\mathfrak{s}}$ by using the shape-shifting cocycle constructed in Section 14 to deform the hyperbolic structure on X. Before proceeding we note the following:

Corollary 15.2 *For all $t \in \mathbb{R}$ and for all $\mu \in \Delta(\lambda)$,*

$$\sigma_\lambda(\mathrm{Eq}_{t\mu}(X)) = \sigma_\lambda(X) + t\mu.$$

Proof That the earthquake $\mathrm{Eq}_{t\mu}(X)$ is defined for all time is a consequence of countable additivity (equivalently positivity) of μ; a complete proof can be found in [Epstein and Marden 2006, Section III]. Viewing the set of measures supported on λ as a subset of $\mathcal{H}(\lambda)$, the formula is immediate from Theorem 15.1 once we note that $\mathrm{Eq}_{t\mu}(X) = X_{t\mu}$, which follows from the description of $\varphi_{t\mu}$ as a limit of simple left (or right) earthquakes; see (37) and Lemma 14.13. □

Fix $\mathfrak{s}$ as in the statement of the theorem and pick an arbitrary $v \in \mathcal{H}$ and $(h_v, p_v) \in \partial_\lambda H_v$. Identifying $\widetilde{X}$ isometrically with $\mathbb{H}^2$ and (h_v, p_v) with a pointed line picks out a representation $\rho\colon \pi_1(S) \to \mathsf{PSL}_2\,\mathbb{R}$ that induces X. Since $\|\mathfrak{s}\|_{\tau_{\underline{\alpha}}} < \frac{1}{2} D_\lambda(X)$, Proposition 14.26 allows us to construct the shape-shifting cocycle $\varphi_{\mathfrak{s}}$.

We may now deform the representation ρ by $\varphi_{\mathfrak{s}}$ by defining

$$\rho_{\mathfrak{s}}\colon \pi_1(S) \to \mathsf{PSL}_2\,\mathbb{R}, \quad \gamma \mapsto \varphi_{p_v,\gamma p_v} \circ \rho(\gamma).$$

The equivariance and cocycle properties of Proposition 14.26 ensure that $\rho_{\mathfrak{s}}$ is itself a representation. Indeed,

$$\begin{aligned}
\rho_{\mathfrak{s}}(\gamma_1\gamma_2) &= \varphi_{p_v,\gamma_1\gamma_2 p_v} \circ \rho(\gamma_1\gamma_2) \\
&= \varphi_{p_v,\gamma_1 p_v} \circ \varphi_{\gamma_1 p_v,\gamma_1\gamma_2 p_v} \circ \rho(\gamma_1) \circ \rho(\gamma_2) \\
&= \varphi_{p_v,\gamma_1 p_v} \circ \rho(\gamma_1) \circ \varphi_{p_v,\gamma_2 p_v} \circ \rho(\gamma_1)^{-1} \circ \rho(\gamma_1) \circ \rho(\gamma_2) \\
&= \rho_{\mathfrak{s}}(\gamma_1) \circ \rho_{\mathfrak{s}}(\gamma_2)
\end{aligned}$$

for all $\gamma_1, \gamma_2 \in \pi_1(S)$. Our goal in the rest of the section is then to show that $\rho_{\mathfrak{s}}$ is discrete and faithful, and that the quotient surface has the correct geometric shear-shape cocycle.

Adjusting geodesics To show that $\rho_{\mathfrak{s}}$ has the desired properties, we use $\varphi_{\mathfrak{s}}$ to adjust the position of $\widetilde{\lambda}$ in $\widetilde{X}$. Ultimately, these adjusted geodesics correspond to the realization of λ on the quotient surface $\widetilde{X}/\mathrm{im}\,\rho_{\mathfrak{s}}$.

Let $\mathcal{G}(\widetilde{X})$ be the space of geodesics in $\widetilde{X}$, and let $\partial\widetilde{\lambda} \subset \mathcal{G}(\widetilde{X})$ denote the set of boundary leaves of $\widetilde{\lambda}$. Define a map

$$\Phi_{p_v}\colon \partial\widetilde{\lambda} \to \mathcal{G}(\widetilde{X})$$

as follows: If h is a leaf of $\partial\widetilde{\lambda}$, then $h = h_u$ for some (h_u, p_u) in $\partial_\lambda\mathcal{H}$. The map Φ_{p_v} then takes (h_u, p_u) isometrically to the pointed geodesic $\varphi_{p_v,p_u}(h_u, p_u) \subset \widetilde{X}$. Note that, if $h_u = h_w$ for some other (h_w, p_w) in $\partial_\lambda\mathcal{H}$, then

$$\varphi_{p_v,p_u}^{-1} \circ \varphi_{p_v,p_w} = \varphi_{p_u,p_w}$$

by the cocycle relation (Proposition 14.26) and φ_{p_u,p_w} is by definition a translation along h. Therefore $\Phi_{p_v} h_w = \Phi_{p_v} h_u$, so Φ_{p_v} is indeed well defined.

Using the fact that S is closed, the following lemma follows directly from the fact that $\rho_{\mathfrak{s}}$ defines a representation of $\pi_1(S)$ in the same component of representations as ρ. We give a hands-on explanation that does not use this fact.

Lemma 15.3 *The representation $\rho_{\mathfrak{s}}$ constructed above is discrete and faithful.*

Proof For distinct leaves h_u and $h_w \in \partial\widetilde{\lambda}$, we claim that $\Phi_{p_v}(h_u)$ is disjoint from $\Phi_{p_v}(h_w)$. Indeed, by the cocycle relation, the position of $\Phi_{p_v}(h_w)$ relative to $\Phi_{p_v}(h_u)$ is the same as the position of $\varphi_{p_u,p_w}h_w$ relative to h_u. Every finite approximation of φ_{p_u,p_w} by compositions of elementary shape-shifting transformations preserves the property that the image of h_w is disjoint from h_u, so the same is true in the limit.

Therefore, as long as $\rho(\gamma)$ does not stabilize h_v, $\Phi_{p_v}(\rho(\gamma)h_v) = \rho_{\mathfrak{s}}(\gamma)h_v$ is different from $\Phi_{p_v}(h_v) = h_v$. If $\rho(\gamma)$ is a translation along h_v, we can find γ_0 such that $\rho(\gamma_0\gamma\gamma_0^{-1})h_v \neq h_v$, and so $\rho_{\mathfrak{s}}(\gamma)$ does not stabilize $\rho(\gamma_0\gamma\gamma_0^{-1})h_v$. In either case, this implies that $\rho_{\mathfrak{s}}(\gamma)$ acts nontrivially on the space of geodesics, so in particular $\rho_{\mathfrak{s}}(\gamma) \neq 1$, ie $\rho_{\mathfrak{s}}$ is faithful.

Since $\pi_1(S)$ is a nonelementary group and $\rho_{\mathfrak{s}}$ is faithful, $\operatorname{im}\rho_{\mathfrak{s}}$ is a nonelementary subgroup of isometries. So assume towards contradiction that $\rho_{\mathfrak{s}}$ is indiscrete. Then $\operatorname{im}\rho_{\mathfrak{s}}$ must be dense in $\mathsf{PSL}_2\,\mathbb{R}$; see eg [Sullivan 1985, Proposition, page 246]. In particular, there is an element $\gamma \in \pi_1(S)$ such that $\rho_{\mathfrak{s}}(\gamma)$ is arbitrarily close to a rotation of angle $\frac{\pi}{2}$ around p_v. Then $\rho_{\mathfrak{s}}(\gamma)h_v = \Phi_{p_v}\rho(\gamma)h_v$ meets h_v in a point, which is impossible because $\Phi_{p_v}(h_v)$ is either equal to or disjoint from $\Phi_{p_v}(\rho(\gamma)h_v)$. We conclude that $\rho_{\mathfrak{s}}$ is discrete, completing the proof of the lemma. □

By Lemma 15.3, the quotient $X_{\mathfrak{s}} = \widetilde{X}/\operatorname{im}\rho_{\mathfrak{s}}$ is a hyperbolic surface equipped with a homeomorphism $S \to X_{\mathfrak{s}}$ in the homotopy class determined by $\rho_{\mathfrak{s}}$. As such, $\rho_{\mathfrak{s}}$ induces a $(\rho, \rho_{\mathfrak{s}})$–equivariant homeomorphism $\partial\widetilde{X} \to \partial\widetilde{X}$, and hence a continuous, equivariant map on the space of geodesics.

Lemma 15.4 *The map Φ_{p_v} extends continuously to $\widetilde{\lambda}$, and $\Phi_{p_v}(\widetilde{\lambda})$ descends to the geodesic realization of λ on $X_{\mathfrak{s}}$.*

Proof By equivariance, the induced map on geodesics agrees with Φ_{p_v} on $\partial\widetilde{\lambda}$. The leaves of $\partial\widetilde{\lambda}$ are dense in $\widetilde{\lambda}$, so the closure of the image of Φ_{p_v} is the geodesic realization of $\widetilde{\lambda}$ on $\widetilde{X}_{\mathfrak{s}}$, which is invariant under the action of $\rho_{\mathfrak{s}}$. □

Since $\Phi_{p_v}(\widetilde{\lambda})$ is the lift of the realization of λ on $X_{\mathfrak{s}}$, we may now leverage our understanding of the shape-shifting cocycle to show that the complementary subsurfaces of $X_{\mathfrak{s}} \setminus \lambda$ have the desired shapes.

Lemma 15.5
$$\underline{A}(X_{\mathfrak{s}}) = \underline{A}(X) + \mathfrak{a}.$$

Proof Recall that by construction the unweighted arc systems of $X \setminus \lambda$ and $X_{\mathfrak{s}} \setminus \lambda$ are both contained in some joint maximal arc system $\underline{\beta}$, leading to an identification of hexagons of $\widetilde{X} \setminus \lambda$ with those of $\widetilde{X} \setminus \lambda_{\mathfrak{s}}$.

So let β be an arc of $\underline{\beta}$, realized orthogeodesically in $X_{\mathfrak{s}}$. Let β also denote a choice of lift, orthogonal to $\lambda_{\mathfrak{s}}$ in $\widetilde{X}$, and let u and w denote the hexagons adjacent to β. Choose either of the geodesics g of λ meeting β and let q_u and q_w be the basepoints of u and w on g. Then, by the cocycle relation (Proposition 14.26),

$$\varphi_{p_v,p_w} = \varphi_{p_v,p_u} \circ \varphi_{p_u,p_w}$$

and so, applying equivariance, the placement of q_w relative to q_u differs from the placement of p_w relative to p_u only by φ_{p_u,p_w}.

But now, since u and w are in the same subsurface, we see by definition that φ_{p_u,p_w} is translation along g by exactly $\mathfrak{s}(\beta)$. Therefore, the distance along $\varphi_{p_u,p_w} g$ between q_u and q_w is exactly the distance along g between p_u and p_w plus $\mathfrak{s}(\beta)$. Translated into arc weights,

$$\sigma_\lambda(X_{\mathfrak{s}})(\beta) = \sigma_\lambda(X)(\beta) + \mathfrak{s}(\beta),$$

completing the proof of the lemma. □

Now that we know that the "shape" part of the data of $\sigma_\lambda(X_{\mathfrak{s}})$ is what it is supposed to be, we need only check that the "shearing" data is as specified. Compare [Bonahon 1996, Lemma 19].

Lemma 15.6 *The surface $X_{\mathfrak{s}}$ has geometric shear-shape cocycle $\sigma_\lambda(X_{\mathfrak{s}}) = \sigma_\lambda(X) + \mathfrak{s}$.*

Proof Observe that, by the cocycle relation (Proposition 14.26) and the discussion in Section 13.2, it suffices to compute the change in shearing data between simple pairs.

So suppose that (v, w) is simple. For each integer r, recall that $\mathcal{H}^r_{v,w} = (u_i)_{i=1}^n$ denotes the set of hexagons such that the intersection of the geodesic from p_v to p_w with u_i has depth at most r with respect to a fixed geometric train track. Set $v = u_0$ and $w = u_{n+1}$, and let $h_i = g^v_{u_i}$, the pointed boundary geodesic of u_i closest to v. Then, by Lemma 14.13,

$$\varphi^r_{p_v,p_w} = T_{h_0}^{\mathfrak{s}(u_0,u_1)} \circ A(s_1) \circ T_{h_1}^{\mathfrak{s}(u_1,u_2)} \circ A(s_2) \circ \cdots \circ A(s_n) \circ T_{h_n}^{\mathfrak{s}(u_n,u_{n+1})}$$

is a good approximation of φ_{p_v,p_w} for large enough r.

Now, for each r, we can deform the hyperbolic structure on $\widetilde{X}$ by $\varphi^r_{p_v,p_w}$ (sacrificing equivariance) and measure the shear $\sigma_r(v, w)$ between v and w in that deformed structure. More precisely, we recall that, if h'_i denotes the other geodesic in u_i that separates v from w, the spike-shaping transformation is equal to a translation along h'_i then along h_i. We may then deform $\widetilde{X}$ by replacing each translation in the factorization of $\varphi^r_{p_v,p_w}$ with a (right) earthquake along the same geodesic; compare with our "geometric explanation" of spike-shaping in Section 14.2.

Since each translation $T_{h_i}^{\mathfrak{s}(u_i,u_{i+1})}$ appearing in $\varphi^r_{p_v,p_w}$ shears $\widetilde{X}$ along a leaf of λ, it preserves the orthogeodesic foliation in complementary components. Therefore, each such term in the deformation thus changes the shear between v and w by exactly $\mathfrak{s}(u_i, u_{i+1})$.

On the other hand, each spike-shaping transformation $A(s_i)$ is a parabolic transformation fixing the vertex of the spike and thus preserves horocycles based at that point. In particular, the distinguished basepoints of each h_i and h_i' remain on the same horocycle and hence deforming by $A(s_i)$ does not affect $\sigma_r(v,w)$.

In summary, deforming $\widetilde{X}$ by the approximation $\varphi^r_{p_v,p_w}$ changes the shear between v and w by

$$\sigma_r(v,w) - \sigma(v,w) = \sum_{i=0}^{n} \mathfrak{s}(u_i, u_{i+1}) = \mathfrak{s}(v,w),$$

where the last equality follows from finite additivity (axiom (SH2)).

Since this equality holds in each approximation and $\varphi^r_{p_v,p_w} \to \varphi_{p_v,p_w}$ as $r \to \infty$, the equality holds in the limit as well. Therefore, deforming $\widetilde{X}$ by φ_{p_v,p_w} changes the shear between v and w by exactly $\mathfrak{s}(v,w)$, which is what we needed to show. □

Proof of Theorem 15.1 As $\|\mathfrak{s}\|_{\tau_{\underline{\alpha}}} < \frac{1}{2} D_\lambda(X)$, Lemmas 14.10 and 14.13 ensure that the limits in the definition of φ_{p_v,p_w} make sense for all simple pairs (v,w). Proposition 14.26 then allows us to construct $\varphi_{\mathfrak{s}}$. By Lemma 15.3, the deformed representation $\rho_{\mathfrak{s}} = \varphi_{\mathfrak{s}} \cdot \rho$ is discrete and faithful, and, by Lemma 15.6, the quotient surface has the correct geometric shear-shape cocycle.

Finally, we observe that the values of shape-shifting cocycle $\varphi_{\mathfrak{s}}$ all converge to the identity as $\|\mathfrak{s}\|_{\tau_{\underline{\alpha}}} \to 0$, and consequently $X_{\mathfrak{s}} \to X$. □

15.2 The global structure of the shear-shape map

We have already proven in Corollary 13.14 that the image of σ_λ lies in $\mathcal{SH}^+(\lambda)$. We now show that this containment is in fact an equality, completing the proof of Theorem 12.1.

We proceed in two steps; the first is to show that:

Proposition 15.7 *The shear-shape map σ_λ is a homeomorphism onto its image.*

Proof Proposition 13.12 (injectivity of σ_λ) allows us to invert σ_λ on its image and, for each $X \in \mathcal{T}(S)$, Theorem 15.1 provides us with an open neighborhood of $\sigma_\lambda(X) \in \mathcal{SH}^+(\lambda)$ on which σ_λ^{-1} is defined and continuous. By Proposition 8.5, $\mathcal{SH}^+(\lambda) \subset \mathcal{SH}(\lambda)$ is an open cell of dimension $6g-6$. Invoking invariance of domain, σ_λ^{-1} and hence σ_λ are local homeomorphisms. An additional application of Proposition 13.12 implies that σ_λ is globally injective, so σ_λ is a homeomorphism onto its image, as claimed. □

The second step is to prove that $\sigma_\lambda : \mathcal{T}(S) \to \mathcal{SH}^+(\lambda)$ is a proper map. That is, we must show that, when X_k escapes to infinity in $\mathcal{T}(S)$, the corresponding shear-shape cocycles $\sigma_\lambda(X_k)$ must diverge in $\mathcal{SH}^+(\lambda)$. Since proper local homeomorphisms are coverings and $\mathcal{SH}^+(\lambda)$ is a cell, the map σ_λ must be a homeomorphism.

The proof we present below is essentially just that of [Bonahon 1996, Theorem 20], but we have to address the additional complications introduced by the PL structure of $\mathcal{SH}^+(\lambda)$; this manifests itself in the *stratified* real-analytic structure of the map.[23]

Proof of Theorem 12.1 We begin by recording an estimate for the geometry of surfaces near the boundary of the image of σ_λ (where "near" is measured in a train track chart).

So suppose that $X \in \mathcal{T}(S)$, set $\underline{\alpha} = \underline{\alpha}(X)$, and build a standard smoothing $\tau_{\underline{\alpha}}$ carrying λ geometrically on X. Fix $\epsilon > 0$ and suppose that there exists some $\mathfrak{s} \in W(\tau_{\underline{\alpha}})$ with $\|\mathfrak{s}\|_{\tau_{\underline{\alpha}}} < \epsilon$ such that $\sigma_\lambda(X) + \mathfrak{s} \in \mathcal{SH}^+(\lambda)$ is not in the image of σ_λ; then Theorem 15.1 implies that

$$\tfrac{1}{2} D_\lambda(X) \leq \|\mathfrak{s}\|_{\tau_{\underline{\alpha}}} < \epsilon.$$

The following claim can be extracted from the proof of [Bonahon 1996, Theorem 20]; we outline a proof for the convenience of the reader.

Claim 15.8 *There is a transverse measure $\mu \in \Delta(\lambda)$ with $1/9\chi(S) \leq \|\mu\|_{\tau_{\underline{\alpha}}} \leq 1$ and*

$$\ell_X(\mu) = \omega_{\mathcal{SH}}(\sigma_\lambda(X), \mu) < \epsilon.$$

Proof If there is a simple closed curve component of λ with length at most ϵ, then we are done. Otherwise, even though $\underline{A} + \mathfrak{a}$ defines a hyperbolic structure on each piece of $S \setminus \lambda$, the overall shear-shape cocycle $\sigma_\lambda(X) + \mathfrak{s}$ does not define a hyperbolic structure on S because the proof of Lemma 14.10 or Lemma 14.13 fails. Therefore, there is a simple pair (v, w) for which the finite products $\varphi_{\underline{\mathcal{H}}}$ (or $\varphi^r_{v,w}$) fail to converge as $\underline{\mathcal{H}}$ tends to $\mathcal{H}_{v,w}$ (or $r \to \infty$).

We claim that there exists u between v and w and a spike $s = (g, h)$ of H_u such that the following holds: for any geodesic transversal $k \subset X$ to λ meeting the spike s, the countably many points of $\tilde{k} \cap g \subset g$ (labeled by $r \in \mathbb{N}$) exiting one end of g escape at a rate strictly slower than $\epsilon(r-1)$. In other words, there are segments $d_r \subset g$ such that $\ell_X(d_r) \leq \epsilon(r-1)$ and d_r meets k exactly r times.

If this were not the case, then, as in the proof of Lemma 14.5, the "gaps" $c_r \subset k_{v,w} \setminus \lambda$ have length $\ell_X(c_r) = O(e^{-\epsilon r})$, where $c_r \cap g$ is labeled by $r \in \mathbb{N}$. This estimate on the decay of gaps implies that $\varphi_{\underline{\mathcal{H}}}$ converges as $\underline{\mathcal{H}} \to \mathcal{H}_{v,w}$ and that $\varphi^r_{v,w} \to \varphi_{v,w}$ as $r \to \infty$ (see the proof of Lemma 14.10), contradicting our assumption.

Now consider the weight system w_r on $\tau_{\underline{\alpha}}$ (not satisfying the switch conditions) defined by counting the number of times d_r travels along each branch of $\tau_{\underline{\alpha}}$, and dividing by the total number of branches n_r that d_r traverses, with multiplicity. Observe that $n_r \geq r$ by definition. Then $\|w_r\|_{\tau_{\underline{\alpha}}} \leq 1$ in the vector space $\mathbb{R}^{b(\tau_{\underline{\alpha}})}$ and w_r takes value zero on branches corresponding to arcs of $\underline{\alpha}$. Moreover, w_r is

[23] Recall that $\mathcal{H}^+(\lambda)$ is an open cone with finitely many faces in a vector space, while $\mathcal{SH}^+(\lambda)$ is an affine cone bundle over a piecewise-linear space with no obvious way of extending the smooth structure over faces of $\mathcal{B}(S \setminus \lambda)$.

nonnegative on each branch and approaches the weight space $W(\tau) \subset \mathbb{R}^{b(\tau)}$ as $r \to \infty$. Since w_r are built from leaves of λ, any limit point μ defines a transverse measure supported on λ (compare also [Penner and Harer 1992, Proposition 3.3.2]).

There are at most $9\chi(S)$ branches of $\tau_{\underline{\alpha}}$, so by the pigeonhole principal there is a branch such that each w_r has mass at least $1/9\chi(S)$, and therefore so must μ. But now, by construction,

$$\ell_X(\mu) = \lim_{r\to\infty} \frac{\ell_X(d_r)}{n_r} < \frac{(r-1)\epsilon}{n_r} < \epsilon,$$

providing the desired measure. □

Now suppose towards contradiction that $\underline{\alpha}$ is maximal and $\sigma_\lambda(X_k) \in \mathcal{SH}^+(\lambda;\underline{\alpha})$ is a sequence approaching some $\sigma \in \mathcal{SH}^+(\lambda;\underline{\alpha})$ that is not in the image of σ_λ. We may then apply the above construction to $\sigma - \sigma_\lambda(X_k)$ to extract a family of measures μ_k on λ satisfying $1/9\chi(S) \le \|\mu_k\|_{\tau_{\underline{\alpha}}} \le 1$ and $\omega_{\mathcal{SH}}(\sigma_\lambda(X), \mu_k) \to 0$. By compactness of the set measures on λ with norm bounded away from zero and infinity, there is some nonzero accumulation point μ of μ_k. Continuity of $\omega_{\mathcal{SH}}$ (Lemma 8.3) then gives

$$\omega_{\mathcal{SH}}(\sigma_\lambda(X_k), \mu_k) \to \omega_{\mathcal{SH}}(\sigma, \mu) = 0,$$

and so $\sigma \notin \mathcal{SH}^+(\lambda;\underline{\alpha})$, a contradiction. Hence, $\mathrm{im}(\sigma_\lambda) \cap \mathcal{SH}^+(\lambda;\underline{\alpha})$ is relatively closed. On the other hand, σ_λ is a local homeomorphism by Proposition 15.7, and hence $\mathrm{im}(\sigma_\lambda) \cap \mathcal{SH}^+(\lambda;\underline{\alpha})$ is relatively open.

If we knew that the projection of $\mathrm{im}(\sigma_\lambda)$ surjected onto $\mathscr{B}(S \setminus \lambda)$ (or at least met each top-dimensional face), we would be done. Since we do not a priori have this information, we instead work our way out in $\mathscr{B}(S \setminus \lambda)$ cell by cell.

To wit, we may invoke Theorem 15.1 once more to deduce that $\mathrm{im}(\sigma_\lambda) \cap \mathcal{SH}^+(\lambda;\underline{\alpha}')$ is relatively open for every filling arc system $\underline{\alpha}'$ that shares a common filling arc subsystem with $\underline{\alpha}$ (hence, $\mathcal{SH}^+(\lambda;\underline{\alpha})$ and $\mathcal{SH}^+(\lambda;\underline{\alpha}')$ intersect). Repeating the argument above for these cells, $\mathrm{im}(\sigma_\lambda) \supset \mathcal{SH}^+(\lambda;\underline{\alpha}')$ as well. Since $\mathscr{B}(S \setminus \lambda)$ is connected, iterating this procedure allows us to deduce that $\mathrm{im}(\sigma_\lambda) \supset \mathcal{SH}^+(\lambda)$. The reverse inclusion follows from Corollary 13.14, so σ_λ is a homeomorphism onto $\mathcal{SH}^+(\lambda)$.

To address the regularity of σ_λ, we note that while $\mathcal{T}(S)$ has a natural $\mathbb{R}$–analytic structure, $\mathcal{SH}(\lambda)$ does not. However, for each arc system $\underline{\alpha}$, filling or not, the open cell $\mathscr{B}^\circ(\underline{\alpha})$ has a well-defined analytic structure compatible with that of the analytic submanifold of $\mathcal{T}(S \setminus \lambda)$ that it parametrizes. The total space of the bundle $\mathcal{SH}^\circ(\lambda;\underline{\alpha}) \to \mathscr{B}^\circ(\underline{\alpha})$ also carries an analytic structure, invariant under train track coordinate–transformations (Proposition 8.5); thus $\mathcal{SH}(\lambda)$ has a stratified $\mathbb{R}$–analytic structure.

The shape-shifting cocycle $\varphi_{\mathfrak{s}}$, and hence the surface $X_{\mathfrak{s}}$, then depends real-analytically on $\mathfrak{s} \in W(\tau_{\underline{\alpha}})$ (where $\underline{\alpha}$ here is equal to the support of $\underline{A}(X)$, not a maximal completion). The reason for this is clear: all elementary shape-shifting transformations are products of small parabolic transformations (see [Thurston 1986, Section 9] or [Bonahon 1996, Theorem A]) or translations with translation distance that

are (restrictions of) real-analytic functions on (an analytic submanifold of) $\mathcal{T}(S \setminus \lambda)$. These products converge absolutely to the shape-shifting cocycle, and hence uniformly on compact sets to an analytic deformation. □

15.3 Dilation rays and Thurston geodesics

Using our coordinatization, we can define an extension of the earthquake flow to an action by the upper-triangular subgroup.

Definition 15.9 Given a measured geodesic lamination λ, a hyperbolic surface $X \in \mathcal{T}(S)$, and $t \in \mathbb{R}$, define an analytic path of surfaces $\{X^t_\lambda\}_{t\in\mathbb{R}}$ by

$$X^t_\lambda := \sigma_\lambda^{-1}(e^t \sigma_\lambda(X)),$$

called the *dilation ray*[24] based at X directed by λ.

As the earthquake flow acts by translation in coordinates (Corollary 15.2), dilation and earthquake along λ (together with scaling the measure on λ) fit together into an action by the upper-triangular subgroup $B < \mathsf{GL}_2^+\mathbb{R}$ on $\mathcal{PT}_g$. More explicitly, we can specify an action of B on $\mathcal{T}(S) \times \mathbb{R}_{>0}\lambda$ (by homeomorphisms) by setting

$$\begin{pmatrix} a & b \\ 0 & c \end{pmatrix} \cdot (X, \lambda) := \big(\sigma_\lambda^{-1}(a\sigma_\lambda(X) + b\lambda), c\lambda\big). \tag{47}$$

These B–actions assemble into a $\mathrm{Mod}(S)$–equivariant B–action on $\mathcal{PT}_g$ (observe that σ_λ depends only on the support of λ and not the actual measure). Quotienting by the mapping class group and restricting to the unit-length locus then gives a P–action on $\mathcal{P}^1\mathcal{M}_g$, and since dilation preserves the property of being regular, a P–action on each stratum $\mathcal{P}^1\mathcal{M}_g^{\mathrm{reg}}(\underline{\kappa})$. We call any such action an action by *stretchquakes*.

Using the commutativity of (2) (Theorem 13.13), we can compare (47) with the computations performed in Lemmas 11.1 and 11.2 to see that:

Proposition 15.10 *The map $\mathbb{O}$ takes the P–action of* (47) *on $\mathcal{P}^1\mathcal{M}_g$ to the standard P–action on $\mathcal{Q}^1\mathcal{M}_g$.*

While we have defined them via coordinates, it is not hard to see that dilation rays are geometrically meaningful families of surfaces. Generally, we obtain paths of surfaces along which the length of λ scales nicely, and we can identify some dilation rays as directed lines in Thurston's asymmetric metric on $\mathcal{T}(S)$.

Mirzakhani [2008, Remark, page 33] observed that, for a maximal lamination μ, the dilation ray $t \mapsto X^t_\mu$ corresponds to the stretch path directed by μ defined by Thurston [1986, Section 4]. Very roughly, stretch paths are obtained by gluing together certain expanding self-homeomorphisms of the ideal triangles that form $X \setminus \mu$ along the leaves of μ.

[24]We are abusing terminology here by declaring that the image of $\mathbb{R}$ under an analytic mapping is a ray. Our aim is to emphasize that the dilation ray should be thought of as directed toward the future, even though it can be defined for all time.

Lemma 15.11 [Thurston 1986, Proposition 2.2] *Let P_n be a regular ideal hyperbolic n–gon. For any $K \geq 1$, there is a K–Lipschitz self-homeomorphism $P_n \to P_n$ that maps each side to itself and expands arclength along the boundary by a constant factor of K.*

Proof The orthogeodesic foliation $\mathcal{O}(P_n)$ is measure equivalent to a partial foliation by horocycles centered at the spikes of P_n. The desired K–Lipschitz homeomorphism $P_n \to P_n$ is constructed by fixing the central horocyclic n–gon and mapping each horocyclic arc at distance s from the central region to the horocyclic arc at distance Ks in the same spike. □

Any partition $\underline{\kappa} = (\kappa_1, \dots, \kappa_n)$ of $4g-4$ determines a regular locus $\mathcal{PT}_g^{\mathrm{reg}}(\underline{\kappa})$ of pairs (X, λ), where the complement of λ in X is a union of regular ideal (κ_i+2)–gons. Then $\mathcal{P}^1\mathcal{M}_g^{\mathrm{reg}}(\underline{\kappa})$ is the moduli space of pairs where $\ell_X(\lambda) = 1$.

Gluing together the expanding maps of regular polygons provides an explicit model of dilation rays in $\mathcal{P}^1\mathcal{M}_g(\underline{\kappa})$ and identifies them with geodesics for the Thurston metric. A survey of some basic properties of Thurston's metric as well as similarities and differences between directed stretch rays and Teichmüller geodesics can be found in [Papadopoulos and Théret 2007]. The following proposition was inspired in part by recent work of Horbez and Tao [≥ 2024], in which they investigate the minimally displaced sets in the Thurston's metric using a similar construction.

Proposition 15.12 *For any $(X, \lambda) \in \mathcal{PT}_g^{\mathrm{reg}}(\underline{\kappa})$, the dilation ray $\{X_\lambda^t : t \in \mathbb{R}\} \subset \mathcal{PT}_g^{\mathrm{reg}}(\underline{\kappa})$ is a directed unit-speed geodesic in Thurston's asymmetric Lipschitz metric.*

Proof Since λ is regular on X, $\sigma_\lambda(X) \in \mathcal{SH}^+(\lambda)$ lies in the fiber over the empty arc system. Scaling $\sigma_\lambda(X)$ preserves this arc system, so X_λ^t is regular for all t. It suffices to prove that the optimal Lipschitz constant for a map $X \to X_\lambda^t$ in the homotopy class determined by markings is e^t for all $t \geq 0$.

Let $H_\lambda(X)$ denote the (partial) foliation of X by horocyclic arcs that is measure equivalent to $\mathcal{O}_\lambda(X)$. The maps of Lemma 15.11 assemble to an e^t–Lipschitz homeomorphism $X \setminus \lambda \to X_t^\lambda \setminus \lambda$ such that $H_\lambda(X)$ maps to $H_\lambda(X_\lambda^t) = e^t H_\lambda(X)$ on each component (as measured foliations). Now, using the fact that $\sigma_\lambda(X_\lambda^t) = e^t\sigma_\lambda(X)$, we can adapt the argument of [Bonahon 1996, Lemma 11] (as sketched in Proposition 13.12) to show that this map is locally Lipschitz and hence extends across λ to an e^t–Lipschitz homeomorphism $X \to X_\lambda^t$.

Thus e^t provides an upper bound for the optimal Lipschitz constant in the homotopy class determined by markings. On the other hand,

$$\ell_{X_\lambda^t}(\lambda) = \omega_{\mathcal{SH}}(\sigma_\lambda(X_\lambda^t), \lambda) = \omega_{\mathcal{SH}}(e^t\sigma_\lambda(X), \lambda) = e^t\ell_X(\lambda),$$

so e^t is also a lower bound for the optimal Lipschitz constant. □

Remark 15.13 As in the last line of the proof of Proposition 15.12 we always have $\ell_{X_\lambda^t}(\lambda) = e^t\ell_X(\lambda)$ for arbitrary $\lambda \in \mathcal{ML}(S)$. Thus the distance from X to X_λ^t in Thurston's metric is at least t. However, we

do not always know how to build e^t–Lipschitz proper homotopy equivalences $X \setminus \lambda \to X^t_\lambda \setminus \lambda$ (in the correct homotopy class) that expand arclength along $\partial X \setminus \lambda$ by a constant factor of e^t.

Remark 15.14 (added in proof) In recent work, Pan and Wolf [2022] build new families of geodesics for the Lipschitz metric using harmonic maps. Their work also uses our coordinates to show that certain "Hopf differential disks" in $\mathcal{T}_g$ converge to "stretch–earthquake disks". It would be interesting to know if their new geodesics coincide with the dilation rays defined here, and, by extension, if their stretch–earthquake disks are the same as the orbits of the stretchquake action defined here.

Remark 15.15 Our dilation rays are different from Thurston's stretch rays defined with respect to one of the finitely many maximal completions of λ when λ is not maximal. This follows from the fact that $\mathbb{O}_\lambda(X) \neq \mathbb{O}_{\lambda'}(X)$, where λ' is a maximal completion of λ.

The map $\mathcal{PT}^{\mathrm{reg}}_g(\underline{\kappa}) \times \mathbb{R} \to \mathcal{PT}^{\mathrm{reg}}_g(\underline{\kappa})$ defined by the rule $(X, \lambda, t) \mapsto (X^t_\lambda, e^{-t}\lambda)$ is called the *stretch flow*. The stretch flow is $\mathrm{Mod}(S)$–equivariant and

$$\ell_{X^t_\lambda}(e^{-t}\lambda) = \ell_X(\lambda),$$

and hence descends to $\mathcal{P}^1\mathcal{M}^{\mathrm{reg}}_g(\underline{\kappa})$.

Corollary 15.16 *Let ν be a P–invariant ergodic probability measure on $\mathcal{P}^1\mathcal{M}_g$.*

- *For ν–almost every (X, λ), the dilation ray $t \mapsto X^t_\lambda$ is a unit-speed geodesic in Thurston's asymmetric metric.*
- *On a set of full ν–measure, the action of the diagonal subgroup of P is identified with the stretch flow and $\mathbb{O}$ conjugates stretch flow to Teichmüller geodesic flow.*

In particular, the stretch flow is ergodic with respect to ν.

Proof By Corollary 1.2, ν–almost every point is regular (with respect to the same topological type of lamination), so the first statement of the theorem is immediate from Proposition 15.12.

The second statement is essentially a restatement of Theorem B combined with the previous statement. Alternatively, in the Gardiner–Masur parametrization of $\mathcal{QT}_g$ (Theorem 2.1), the Teichmüller geodesic flow at time t is given by $(\eta, \lambda) \mapsto (e^t\eta, e^{-t}\lambda)$, so unraveling the definitions and using commutativity of (2) (Theorem 13.13) gives the result.

For ergodicity, we apply Theorem C which asserts, in particular, that $\mathbb{O}_*\nu$ is an ergodic $\mathsf{SL}_2\,\mathbb{R}$–invariant probability measure on $\mathcal{Q}^1\mathcal{M}_g(\underline{\kappa})$. The Howe–Moore Theorem implies that any noncompact, closed subgroup of $\mathsf{SL}_2\,\mathbb{R}$ inherits ergodicity (see eg [Feres and Katok 2002, Theorem 3.3.1]); in particular, the Teichmüller geodesic flow is ergodic with respect to $\mathbb{O}_*\nu$. So $\mathbb{O}$ maps any stretch flow–invariant set B of positive ν–measure to an $\mathbb{O}_*\nu$ Teichmüller geodesic flow–invariant set of positive measure, which must have full measure by ergodicity. Thus $\nu(B) = 1$, demonstrating ergodicity of the stretch flow. □

Recently, Allessandrini and Disarlo [2022] constructed Lipschitz maps between some pairs of degenerate right-angled hexagons that stretch alternating boundary geodesics by a constant factor. Recall from Section 6 that the Teichmüller space of an ideal quadrilateral is 1–dimensional and can be described as the cone over a pair of points corresponding to the two arcs α and β that join opposite sides of Q.

Lemma 15.17 *Let Q be an ideal quadrilateral with weighted filling arc system $s\delta$, where $\delta \in \{\alpha, \beta\}$. Let Q^t be the quadrilateral with arc system $e^t s\delta$. There is an e^t–Lipschitz surjection $Q \to Q^t$ that multiplies arclength along the boundary of Q by a factor of e^t. Moreover, the projection of the compact edge of the spine of Q is mapped to the projection of the compact edge of the spine of Q^t.*

Proof Every ideal quadrilateral has an orientation-preserving isometric involution swapping opposite sides. Thus the orthogeodesic representative of δ cuts Q into two isometric pieces, each of which is a right-angled hexagon with two degenerate sides. On each piece, we can apply [Alessandrini and Disarlo 2022, Lemma 6.9] to obtain maps which glue together along δ to give a map with the desired properties. □

We immediately obtain some new geodesics for Thurston's metric.

Proposition 15.18 *If $S \setminus \lambda$ consists of ideal triangles and quadrilaterals, then, for any $X \in \mathcal{T}(S)$, $t \mapsto X^t_\lambda$ is a directed, unit-speed geodesic for Thurston's asymmetric metric.*

Proof The proof is nearly identical to the proof of Proposition 15.12, so we only provide a brief outline. Construct an e^t–Lipschitz surjective map $X \setminus \lambda \to X^\lambda_t \setminus \lambda$ from the units of Lemmas 15.11 and 15.17. For the same reason as before, this map extends continuously across the leaves of λ and provides an e^t–Lipschitz homotopy equivalence $X \to X^t_\lambda$ in the homotopy class determined by markings. Thus e^t is an upper bound for the Lipschitz constant among homotopy equivalences $X \to X^t_\lambda$ in the correct homotopy class. This is clearly an upper bound for the ratio

$$\max_{\mu \in \mathcal{ML}(S)} \frac{\ell_\mu(X^t_\lambda)}{\ell_\mu(X)}.$$

But e^t is also a lower bound for this ratio, because the length of λ is scaled by a factor of e^t.

By [Thurston 1986, Theorem 8.5], there is an e^t–Lipschitz *homeomorphism* $X \to X^t_\lambda$ homotopic to the map constructed above. □

Remark 15.19 The proof of Proposition 15.18 clearly supplies a more general statement: if λ is filling and cuts $X \in \mathcal{T}(S)$ into a regular polygons and quadrilaterals of any shape, then $t \mapsto X^t_\lambda$ is a geodesic for Thurston's metric.

There are other cases in which we can glue Lipschitz maps between degenerate right-angled hexagons that can be found in the literature (eg [Alessandrini and Disarlo 2022; Papadopoulos and Yamada 2017]). However, these other cases require additional symmetry that is not always present in our setting. We suspect that there is a different approach that would prove that dilation rays can always be identified with Thurston geodesics, so that $\mathbb{O}$ conjugates a kind of Thurston geodesic flow to Teichmüller geodesic flow.

16 Future and ongoing work

There is much more to understand about the correspondence between hyperbolic and flat geometry described in this paper. In addition to using the orthogeodesic foliation to import tools from Teichmüller dynamics into the world of hyperbolic geometry (and vice versa), the authors expect this link to provide retroactive explanations for analogous phenomena in the two settings.

We describe a number of future directions and potential applications of the correspondence below, some of which will be addressed in forthcoming sequels.

Continuity and equidistribution Theorem D states that, for a fixed lamination $\mathbb{C}_\lambda : \mathcal{T}(S) \to \mathcal{MF}(\lambda)$ is a homeomorphism, but, as Mirzakhani [2008, page 33] already observed, $\mathbb{C}$ cannot be continuous on $\mathcal{PT}_g$. Moreover, Arana-Herrera and Wright [2024] have proven that the earthquake and horocycle flow are not topologically conjugate by any map. At fault is the basic fact that the support of a measured lamination does not vary continuously in the relevant topology.

In [Calderon and Farre 2024a], we investigate the continuity properties of $\mathbb{C}$ restricted to specific families of (X, λ) with constrained geometry and topology. On these families, the support of λ is forced to vary continuously in the Hausdorff topology as the pair varies (in the usual topology on $\mathcal{PT}_g$). For example, each of the regular loci has this property. With this extra geometric control in hand, we prove that $\mathbb{C}$ restricts to a homeomorphism $\mathcal{PT}_g^{\mathrm{reg}}(\underline{\kappa}) \leftrightarrow \mathcal{QT}_g^{\mathrm{nsc}}(\underline{\kappa})$ on each regular locus.

By imposing a stronger (yet still geometrically meaningful) topology on $\mathcal{ML}(S)$, we ensure the continuity of $\mathbb{C}$ varying over all pairs: let $\mathfrak{ML}(S)$ denote the set of measured laminations with the "Hausdorff + measure" topology, so that measured laminations are close in $\mathfrak{ML}(S)$ if they are close both in measure and their supports are Hausdorff close. We prove a general phenomenon that $\mathbb{C} : \mathcal{T}(S) \times \mathfrak{ML}(S) \to \mathcal{QT}(S)$ is locally Hölder continuous with respect to a nice family of locally defined metrics in geometric train track coordinates [Calderon and Farre 2024a, Theorem 12.7].

Our continuity arguments depend on a detailed analysis of the geometric structure of small foliated train track neighborhoods of a lamination on a hyperbolic surface. This analysis is sufficiently robust to produce "enough continuity" to deduce that $\mathbb{C}$ is a Borel-measurable isomorphism, a fact which is pivotal for applications. The results of Section 1.2 then live in a more natural setting, as well.

Combined with this work, the conjugacy of Theorems A and B allows us to import techniques of flat geometry to the hyperbolic setting. In particular, while $\mathbb{C}$ is not continuous, its discontinuity is controlled enough that we can translate between equidistribution in $\mathcal{P}^1\mathcal{T}_g$ and equidistribution in $\mathcal{Q}^1\mathcal{T}_g$ [Calderon and Farre 2024b].

Symplectic structure For a maximal lamination λ, Sözen and Bonahon [2001] identified the Goldman symplectic form on the Teichmüller component of $\mathrm{Hom}(\pi_1 S, \mathsf{PSL}_2\,\mathbb{R})/\mathsf{PSL}_2\,\mathbb{R}$ (also the Weil–Petersson symplectic form) as $\sigma_\lambda^* \omega_{\mathcal{H}}$ in shear coordinates. For arbitrary $\lambda \in \mathcal{ML}(S)$ and $X \in \mathcal{T}(S)$, the shape-shifting

cocycles built in Section 14 provide an open set of deformations of the hyperbolization $[\rho\colon \pi_1 S \to \mathsf{PSL}_2\,\mathbb{R}]$ of X (Theorem 15.1). Taking derivatives (as in [Sözen and Bonahon 2001]) identifies the tangent space to $[\rho]$ with the vector space of Ad_ρ–invariant Lie algebra valued 1–cocycles, yielding a reasonably explicit formula for a vector in the tangent space at $[\rho]$. Using this formula, it is then possible to compute an expression for the Goldman symplectic form in shear-shape coordinates. It remains to understand precisely how $\mathbb{O}$ interacts with the various natural symplectic forms on $\mathcal{PM}_g$ and $\mathcal{QM}_g$ and the (degenerate) symplectic forms on strata, a question that is made technical by the lack of regularity of $\mathbb{O}$.

Measures To each $\mathsf{PSL}_2\,\mathbb{R}$–invariant ergodic probability measure on $\mathcal{Q}^1\mathcal{M}_g$, pushforward along $\mathbb{O}^{-1}$ produces a P–invariant ergodic probability measure on $\mathcal{P}^1\mathcal{M}_g$ (and vice versa). An important class of such measures on the singular flat side is furnished by the Masur–Veech measure $\mu_{\underline{\kappa}}$ on a component of a stratum $\mathcal{Q}^1\mathcal{M}_g(\underline{\kappa})$. In [Calderon and Farre 2024a], we give a geometric description of $\nu_{\underline{\kappa}} := \mathbb{O}_*^{-1}(\mu_{\underline{\kappa}})$ on the corresponding "stratum" $\mathcal{P}^1\mathcal{M}_g^{\mathrm{reg}}(\underline{\kappa})$, which parallels [Mirzakhani 2008, Theorem 1.4] that on the principal stratum, $\nu_{\underline{\kappa}}$ disintegrates into the Weil–Petersson measure on hyperbolic surfaces and Thurston measure on laminations (up to a normalization factor). We give an outline of the various ingredients required to make the analogous statement for $\nu_{\underline{\kappa}}$ with $\underline{\kappa}$ arbitrary.

As discussed in Section 9.3, the piecewise-integral-linear (PIL) structure on $\mathcal{SH}(\lambda)$ endows it with an integer lattice and distinguished measure in the class of Lebesgue. Indeed, for each filling λ, the integer lattice in $\mathcal{SH}^+(\lambda)$ restricts to an integer lattice on the fiber $\mathcal{H}^+(\lambda)$ over the empty arc system due to integrality of the equations defining the piecewise-linear structure of $\mathscr{B}(S \setminus \lambda)$. The empty arc system corresponds to the set of X on which λ is regular, and so the PIL structure induces a measure (in the class of Lebesgue) on this regular locus.

We identify the kernel of the Goldman symplectic form restricted to regular loci as tangent to certain "hyperbolic Schiffer deformations" associated to each even-gon in the complement of λ. These directions admit explicit descriptions as weight systems on a snug train track for λ [Bonahon and Wong 2017, Appendix] which can be reinterpreted as 1–forms on regular loci obtained as the differentials of coordinate functions. Using our formula for the Goldman symplectic form restricted to regular loci, we identify the pullback of the Lebesgue measure on the fiber $\mathcal{H}(\lambda)$ over the empty arc system with an analytic volume form obtained as a wedge power of the restricted symplectic form then wedged together with the distinguished 1–forms associated to the kernel.

Using snug train tracks, one can define a $\underline{\kappa}$–Thurston measure on the space $\mathcal{ML}(\underline{\kappa})$ of polygonal measured laminations of a given topological type. While this is essentially Lebesgue measure in train track coordinates for the "measure + Hausdorff" topology, it is not locally finite in the usual topology on $\mathcal{ML}(S)$. One can construct natural train track coordinate charts that give local measurable trivializations of $\mathcal{P}^1\mathcal{M}_g(\underline{\kappa})$ and exhibit $\nu_{\underline{\kappa}}$ as the product of $\underline{\kappa}$–Thurston measure and the Weil–Petersson-type volume form.

Expanding horospheres Counting problems for square-tiled surfaces/curves on hyperbolic surfaces are intricately related to the equidistribution of L–level sets for the intersection number with/hyperbolic

length of laminations as one takes $L \to \infty$. When λ is a multicurve, the equidistribution of such "expanding horospheres" to the Masur–Veech measure on the principal stratum of $\mathbb{Q}^1\mathcal{M}_g$/ the pullback by $\mathbb{O}$ of this measure on $\mathcal{P}^1\mathcal{M}_g$ (sometimes called *Mirzakhani measure*) was established in [Mirzakhani 2007; Arana-Herrera 2021; Liu 2022] using the geometry of the (symmetrized) Lipschitz metric, the nondivergence of the earthquake flow, and a no-escape-of-mass argument. On the other end of the spectrum, the equidistribution of expanding horospheres for maximal λ to $\mathbb{Q}^1\mathcal{M}_g$ can be proven using a standard "thickening plus mixing" argument from homogeneous dynamics; in the flat setting this is implicit in the work of Lindenstrauss and Mirzakhani [2008], and was recently generalized in [Forni 2021, Theorem 1.6] using different methods. Equidistribution in the hyperbolic setting then follows from the continuity results described above.

Using our extension of Mirzakhani's conjugacy (and the continuity results described above), the same "thickening plus mixing" technique can be used to prove that expanding horospheres based at any λ equidistribute to the Mirzakhani measure on $\mathcal{P}^1\mathcal{M}_g$. Moreover, an analogous result holds for strata: intersections of expanding horospheres based at λ and the regular locus should equidistribute to the pullback to $\mathcal{P}^1\mathcal{M}_g$ of the Masur–Veech measure for a component of $\mathbb{Q}^1\mathcal{M}_g(\underline{\kappa})$ [Calderon and Farre 2024b, Theorem 1.4].

Index

References

[Alessandrini and Disarlo 2022] **D Alessandrini**, **V Disarlo**, *Generalizing stretch lines for surfaces with boundary*, Int. Math. Res. Not. 2022 (2022) 18919–18991 MR Zbl

[Arana-Herrera 2021] **F Arana-Herrera**, *Equidistribution of families of expanding horospheres on moduli spaces of hyperbolic surfaces*, Geom. Dedicata 210 (2021) 65–102 MR Zbl

[Arana-Herrera and Wright 2024] **F Arana-Herrera**, **A Wright**, *The asymmetry of Thurston's earthquake flow*, Geom. Topol. 28 (2024) 2125–2144

[Birman and Series 1985] **J S Birman**, **C Series**, *Geodesics with bounded intersection number on surfaces are sparsely distributed*, Topology 24 (1985) 217–225 MR Zbl

[Bonahon 1996] **F Bonahon**, *Shearing hyperbolic surfaces, bending pleated surfaces and Thurston's symplectic form*, Ann. Fac. Sci. Toulouse Math. 5 (1996) 233–297 MR Zbl

[Bonahon 1997a] **F Bonahon**, *Geodesic laminations with transverse Hölder distributions*, Ann. Sci. École Norm. Sup. 30 (1997) 205–240 MR Zbl

[Bonahon 1997b] **F Bonahon**, *Transverse Hölder distributions for geodesic laminations*, Topology 36 (1997) 103–122 MR Zbl

[Bonahon and Dreyer 2017] **F Bonahon**, **G Dreyer**, *Hitchin characters and geodesic laminations*, Acta Math. 218 (2017) 201–295 MR Zbl

[Bonahon and Wong 2017] **F Bonahon**, **H Wong**, *Representations of the Kauffman bracket skein algebra, II: Punctured surfaces*, Algebr. Geom. Topol. 17 (2017) 3399–3434 MR Zbl

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

[Buser 1992] **P Buser**, *Geometry and spectra of compact Riemann surfaces*, Progr. Math. 106, Birkhäuser, Boston, MA (1992) MR Zbl

[Calderon and Farre 2024a] **A Calderon**, **J Farre**, *Continuity of the orthogeodesic foliation and ergodic theory of the earthquake flow*, preprint (2024) arXiv 2401.12299

[Calderon and Farre 2024b] **A Calderon**, **J Farre**, *On Mirzakhani's twist torus conjecture*, preprint (2024) arXiv 2405.12106

[Canary et al. 2006] **R D Canary**, **D B A Epstein**, **P L Green**, *Notes on notes of Thurston*, from "Fundamentals of hyperbolic geometry: selected expositions" (R D Canary, D Epstein, A Marden, editors), Lond. Math. Soc. Lect. Note Ser. 328, Cambridge Univ. Press (2006) 1–115 MR Zbl

[Casson and Bleiler 1988] **A J Casson**, **S A Bleiler**, *Automorphisms of surfaces after Nielsen and Thurston*, Lond. Math. Soc. Stud. Texts 9, Cambridge Univ. Press (1988) MR Zbl

[Chaika et al. 2020] **J Chaika**, **J Smillie**, **B Weiss**, *Tremors and horocycle dynamics on the moduli space of translation surfaces*, preprint (2020) arXiv 2004.04027

[Chen and Möller 2014] **D Chen**, **M Möller**, *Quadratic differentials in low genus: exceptional and non-varying strata*, Ann. Sci. École Norm. Sup. 47 (2014) 309–369 MR Zbl

[Do 2008] **N N V Do**, *Intersection theory on moduli spaces of curves via hyperbolic geometry*, PhD thesis, University of Melbourne (2008) Available at `https://users.monash.edu/~normd/documents/Do-Phd-Thesis.pdf`

[Dumas 2009] **D Dumas**, *Complex projective structures*, from "Handbook of Teichmüller theory, II" (A Papadopoulos, editor), IRMA Lect. Math. Theor. Phys. 13, Eur. Math. Soc., Zürich (2009) 455–508 MR Zbl

[Dumas 2015] **D Dumas**, *Skinning maps are finite-to-one*, Acta Math. 215 (2015) 55–126 MR Zbl

[Epstein and Marden 2006] **D B A Epstein**, **A Marden**, *Convex hulls in hyperbolic space, a theorem of Sullivan, and measured pleated surfaces*, from "Fundamentals of hyperbolic geometry: selected expositions" (R D Canary, D Epstein, A Marden, editors), Lond. Math. Soc. Lect. Note Ser. 328, Cambridge Univ. Press (2006) 117–266 MR Zbl

[Eskin and Mirzakhani 2018] **A Eskin**, **M Mirzakhani**, *Invariant and stationary measures for the* $\mathrm{SL}(2,\mathbb{R})$ *action on moduli space*, Publ. Math. Inst. Hautes Études Sci. 127 (2018) 95–324 MR Zbl

[Eskin et al. 2015] **A Eskin**, **M Mirzakhani**, **A Mohammadi**, *Isolation, equidistribution, and orbit closures for the* $\mathrm{SL}(2,\mathbb{R})$ *action on moduli space*, Ann. of Math. 182 (2015) 673–721 MR Zbl

[Feres and Katok 2002] **R Feres**, **A Katok**, *Ergodic theory and dynamics of G–spaces (with special emphasis on rigidity phenomena)*, from "Handbook of dynamical systems, IA" (B Hasselblatt, A Katok, editors), North-Holland, Amsterdam (2002) 665–763 MR Zbl

[Forni 2021] **G Forni**, *Limits of geodesic push-forwards of horocycle invariant measures*, Ergodic Theory Dynam. Systems 41 (2021) 2782–2804 MR Zbl

[Gardiner and Masur 1991] **F P Gardiner**, **H Masur**, *Extremal length geometry of Teichmüller space*, Complex Variables Theory Appl. 16 (1991) 209–237 MR Zbl

[Gelander 2014] **T Gelander**, *Lectures on lattices and locally symmetric spaces*, from "Geometric group theory" (M Bestvina, M Sageev, K Vogtmann, editors), IAS/Park City Math. Ser. 21, Amer. Math. Soc., Providence, RI (2014) 249–282 MR Zbl

[Gupta 2021] **S Gupta**, *Harmonic maps and wild Teichmüller spaces*, J. Topol. Anal. 13 (2021) 349–393 MR Zbl

[Han et al. 1995] **Z-C Han**, **L-F Tam**, **A Treibergs**, **T Wan**, *Harmonic maps from the complex plane into surfaces with nonpositive curvature*, Comm. Anal. Geom. 3 (1995) 85–114 MR Zbl

[Horbez and Tao $\geq$ 2024] **C Horbez**, **J Tao**, *Nielsen–Thurston classification and isometries*, in preparation

[Hubbard and Masur 1979] **J Hubbard**, **H Masur**, *Quadratic differentials and foliations*, Acta Math. 142 (1979) 221–274 MR Zbl

[Kerckhoff 1983] **S P Kerckhoff**, *The Nielsen realization problem*, Ann. of Math. 117 (1983) 235–265 MR Zbl

[Kontsevich and Zorich 2003] **M Kontsevich**, **A Zorich**, *Connected components of the moduli spaces of Abelian differentials with prescribed singularities*, Invent. Math. 153 (2003) 631–678 MR Zbl

[Lanneau 2008] **E Lanneau**, *Connected components of the strata of the moduli spaces of quadratic differentials*, Ann. Sci. École Norm. Sup. 41 (2008) 1–56 MR Zbl

[Levitt 1983] **G Levitt**, *Foliations and laminations on hyperbolic surfaces*, Topology 22 (1983) 119–135 MR Zbl

[Lindenstrauss and Mirzakhani 2008] **E Lindenstrauss**, **M Mirzakhani**, *Ergodic theory of the space of measured laminations*, Int. Math. Res. Not. 2008 (2008) art. id. rnm126 MR Zbl

[Liu 2022] **M Liu**, *Length statistics of random multicurves on closed hyperbolic surfaces*, Groups Geom. Dyn. 16 (2022) 437–459 MR Zbl

[Luo 2007] **F Luo**, *On Teichmüller spaces of surfaces with boundary*, Duke Math. J. 139 (2007) 463–482 MR Zbl

[Masur 1982] **H Masur**, *Interval exchange transformations and measured foliations*, Ann. of Math. 115 (1982) 169–200 MR Zbl

[Minsky 1992] **Y N Minsky**, *Harmonic maps, length, and energy in Teichmüller space*, J. Differential Geom. 35 (1992) 151–217 MR Zbl

[Minsky and Weiss 2002] **Y Minsky**, **B Weiss**, *Nondivergence of horocyclic flows on moduli space*, J. Reine Angew. Math. 552 (2002) 131–177 MR Zbl

[Minsky and Weiss 2014] **Y Minsky**, **B Weiss**, *Cohomology classes represented by measured foliations, and Mahler's question for interval exchanges*, Ann. Sci. École Norm. Sup. 47 (2014) 245–284 MR Zbl

[Mirzakhani 2007] **M Mirzakhani**, *Random hyperbolic surfaces and measured laminations*, from "In the tradition of Ahlfors–Bers, IV" (D Canary, J Gilman, J Heinonen, H Masur, editors), Contemp. Math. 432, Amer. Math. Soc., Providence, RI (2007) 179–198 MR Zbl

[Mirzakhani 2008] **M Mirzakhani**, *Ergodic theory of the earthquake flow*, Int. Math. Res. Not. 2008 (2008) art. id. rnm116 MR Zbl

[Mondello 2009a] **G Mondello**, *Riemann surfaces, ribbon graphs and combinatorial classes*, from "Handbook of Teichmüller theory, II" (A Papadopoulos, editor), IRMA Lect. Math. Theor. Phys. 13, Eur. Math. Soc., Zürich (2009) 151–215 MR Zbl

[Mondello 2009b] **G Mondello**, *Triangulated Riemann surfaces with boundary and the Weil–Petersson Poisson structure*, J. Differential Geom. 81 (2009) 391–436 MR Zbl

[Pan and Wolf 2022] **H Pan**, **M Wolf**, *Ray structures on Teichmüller space*, preprint (2022) arXiv 2206.01371

[Papadopoulos 1986] **A Papadopoulos**, *Geometric intersection functions and Hamiltonian flows on the space of measured foliations on a surface*, Pacific J. Math. 124 (1986) 375–402 MR Zbl

[Papadopoulos 1991] **A Papadopoulos**, *On Thurston's boundary of Teichmüller space and the extension of earthquakes*, Topology Appl. 41 (1991) 147–177 MR Zbl

[Papadopoulos and Théret 2007] **A Papadopoulos**, **G Théret**, *On Teichmüller's metric and Thurston's asymmetric metric on Teichmüller space*, from "Handbook of Teichmüller theory, I" (A Papadopoulos, editor), IRMA Lect. Math. Theor. Phys. 11, Eur. Math. Soc., Zürich (2007) 111–204 MR Zbl

[Papadopoulos and Yamada 2017] **A Papadopoulos**, **S Yamada**, *Deforming hyperbolic hexagons with applications to the arc and the Thurston metrics on Teichmüller spaces*, Monatsh. Math. 182 (2017) 913–939 MR Zbl

[Penner and Harer 1992] **R C Penner**, **J L Harer**, *Combinatorics of train tracks*, Ann. of. Math. Stud. 125, Princeton Univ. Press (1992) MR Zbl

[Sözen and Bonahon 2001] **Y Sözen**, **F Bonahon**, *The Weil–Petersson and Thurston symplectic forms*, Duke Math. J. 108 (2001) 581–597 MR Zbl

[Sullivan 1985] **D Sullivan**, *Quasiconformal homeomorphisms and dynamics, II: Structural stability implies hyperbolicity for Kleinian groups*, Acta Math. 155 (1985) 243–260 MR Zbl

[Thurston 1979] **W P Thurston**, *The geometry and topology of three-manifolds*, lecture notes, Princeton University (1979) Available at `https://url.msp.org/gt3m`

[Thurston 1986] **W P Thurston**, *Minimal stretch maps between hyperbolic surfaces*, preprint (1986) Reprinted in his "Collected works with commentary, I: Foliations, surfaces and differential geometry" (B Farb, D Gabai, S P Kerckhoff, editors), Amer. Math. Soc., Providence, RI (2022) 533–585

[Thurston 1997] **W P Thurston**, *Three-dimensional geometry and topology, I*, Princeton Math. Ser. 35, Princeton Univ. Press (1997) MR Zbl

[Ushijima 1999] **A Ushijima**, *A canonical cellular decomposition of the Teichmüller space of compact surfaces with boundary*, Comm. Math. Phys. 201 (1999) 305–326 MR Zbl

[Veech 1982] **W A Veech**, *Gauss measures for transformations on the space of interval exchange maps*, Ann. of Math. 115 (1982) 201–242 MR Zbl

[Wolpert 1983] **S Wolpert**, *On the symplectic geometry of deformations of a hyperbolic surface*, Ann. of Math. 117 (1983) 207–234 MR Zbl

[Wright 2020] **A Wright**, *A tour through Mirzakhani's work on moduli spaces of Riemann surfaces*, Bull. Amer. Math. Soc. 57 (2020) 359–408 MR Zbl

[Wright 2022] **A Wright**, *Mirzakhani's work on earthquake flow*, from "Teichmüller theory and dynamics" (P Dehornoy, E Lanneau, editors), Panor. Synth. 58, Soc. Math. France, Paris (2022) 101–134 MR Zbl

[Zhu and Bonahon 2004] **X Zhu**, **F Bonahon**, *The metric space of geodesic laminations on a surface, I*, Geom. Topol. 8 (2004) 539–564 MR Zbl

Department of Mathematics, University of Chicago
Chicago, IL, United States

Faculty of Mathematics and Computer Science, Universität Heidelberg
Heidelberg, Germany

aaroncalderon@uchicago.edu, jfarre@mathi.uni-heidelberg.de

Proposed: Mladen Bestvina
Seconded: David Fisher, Anna Wienhard

Received: 12 July 2021
Revised: 21 January 2023

msp

Geometry & Topology **28**:5 (2024) 2125–2144
DOI: 10.2140/gt.2024.28.2125
Published: 24 August 2024

The asymmetry of Thurston's earthquake flow

FRANCISCO ARANA-HERRERA
ALEX WRIGHT

We show that Thurston's earthquake flow is strongly asymmetric in the sense that its normalizer is as small as possible inside the group of orbifold automorphisms of the bundle of measured geodesic laminations over moduli space. (At the level of Teichmüller space, such automorphisms correspond to homeomorphisms that are equivariant with respect to an automorphism of the mapping class group.) It follows that the earthquake flow does not extend to an $\mathrm{SL}(2,\mathbb{R})$–action of orbifold automorphisms and does not admit continuous renormalization self-symmetries. In particular, it is not conjugate to the Teichmüller horocycle flow via an orbifold map. This contrasts with a number of previous results, most notably Mirzakhani's theorem that the earthquake and Teichmüller horocycle flows are measurably conjugate.

30F60; 32G15

1 Introduction

Context The bundle $\mathcal{P}^1\mathcal{M}_g$ of unit-length measured geodesic laminations over the moduli space $\mathcal{M}_g$ of hyperbolic or Riemann surfaces of genus g is most naturally seen as a construction of hyperbolic geometry, whereas the bundle $\mathcal{Q}^1\mathcal{M}_g$ of unit-area quadratic differentials over $\mathcal{M}_g$ is most naturally seen from the perspective of either complex analysis or flat geometry. The bundle $\mathcal{P}^1\mathcal{M}_g$ supports Thurston's rather mysterious earthquake flow, which is most concisely defined as a Hamiltonian flow using the Weil–Petersson symplectic form, whereas the bundle $\mathcal{Q}^1\mathcal{M}_g$ supports the Teichmüller horocycle flow, easily defined as part of the much-studied $\mathrm{SL}(2,\mathbb{R})$–action. Mirzakhani showed that, despite their different origins, these flows are measurably isomorphic.

Theorem 1.1 [Mirzakhani 2008] *There is a measurable conjugacy $\mathcal{P}^1\mathcal{M}_g \to \mathcal{Q}^1\mathcal{M}_g$ between the earthquake flow and the Teichmüller horocycle flow.*

In addition to being of fundamental interest as a bridge between different perspectives on the geometry of surfaces and their moduli spaces, this theorem has powered many applications concerning equidistribution, counting and the study of random surfaces [Mirzakhani 2007a; 2016; Arana-Herrera 2021; 2022; Liu 2022; Lu and Su 2022].

Mirzakhani's conjugacy is only defined on a full-measure subset of $\mathcal{P}^1\mathcal{M}_g$, and, as remarked by Mirzakhani herself [2008, Section 6], this conjugacy cannot be extended to a continuous map on all of $\mathcal{P}^1\mathcal{M}_g$. Despite

this, Calderon and Farre [2024] extended Mirzakhani's conjugacy to a bijection which, although not continuous, is geometrically natural and has exciting new applications.

One reason Theorem 1.1 is plausible is that there are many conceptual similarities between the earthquake flow and the Teichmüller horocycle flow, such as the following:

(1) Both arise from some notion of shearing.

(2) Both have been understood by analogy to unipotent flows on homogeneous spaces.

(3) Both are Hamiltonian with respect to related symplectic structures [Masur 1995; Sözen and Bonahon 2001].

(4) Both are associated to natural complex disks in Teichmüller space, namely Teichmüller discs for the Teichmüller horocycle flow and complex earthquake discs for the earthquake flow [McMullen 1998].

(5) Both have quantitative nondivergence properties [Minsky and Weiss 2002].

No continuous conjugacy In light of all these similarities and the work of Mirzakhani, Calderon and Farre, one might wonder if a result stronger than Theorem 1.1 holds: perhaps the earthquake and Teichmüller horocycle flows are isomorphic from the point of view of continuous dynamics, ie perhaps there is a different conjugacy between these flows that is also a homeomorphism. This question was advertised by Wright [2020, Problem 12.3; 2022, Remark 5.6]. Our main result on asymmetry, which we will state shortly as Theorem 1.4, implies a negative solution to this problem.

Theorem 1.2 *There does not exist an orbifold conjugacy $\mathcal{P}^1\mathcal{M}_g \to \mathcal{Q}^1\mathcal{M}_g$ between the earthquake flow and the Teichmüller horocycle flow.*

The technical restriction in Theorem 1.2 that the conjugacy respects the orbifold structure of these spaces is natural since both spaces have the same orbifold structure [Hubbard and Masur 1979].

The existence of an orbifold conjugacy $\mathcal{P}^1\mathcal{M}_g \to \mathcal{Q}^1\mathcal{M}_g$ as in Theorem 1.2 is equivalent to the existence of a topological conjugacy $\mathcal{P}^1\mathcal{T}_g \to \mathcal{Q}^1\mathcal{T}_g$ of the lifts to Teichmüller space of the earthquake and Teichmüller horocycle flows that intertwines an automorphism $\rho\colon \mathrm{Mod}_g \to \mathrm{Mod}_g$ of the mapping class group. For detailed discussions on the theory of orbifolds, see [Thurston 1979, Chapter 13; Erlandsson and Souto 2022, Section 2]. In particular, the following corollary holds:

Corollary 1.3 *There does not exist a mapping class group equivariant topological conjugacy $\mathcal{P}^1\mathcal{T}_g \to \mathcal{Q}^1\mathcal{T}_g$ between the earthquake flow and the Teichmüller horocycle flow.*

Strong asymmetry A flow $E = \{E_t : \mathcal{X} \to \mathcal{X}\}_{t\in\mathbb{R}}$ on a space $\mathcal{X}$ can be interpreted as a group homomorphism $E\colon \mathbb{R} \to \mathrm{Aut}(\mathcal{X})$ mapping $t \in \mathbb{R}$ to $E_t \in \mathrm{Aut}(\mathcal{X})$, where the automorphism group $\mathrm{Aut}(\mathcal{X})$ is defined in whatever category (smooth, continuous, measurable, etc) is under consideration.

The centralizer of the flow E is defined as

$$C(E) = \{F \in \mathrm{Aut}(\mathcal{X}) : (\forall t \in \mathbb{R})\ E_t \circ F = F \circ E_t\}.$$

The centralizer corresponds to the most narrow concept of the set of symmetries of a flow one can consider, consisting only of the automorphisms that commute with it. A slightly broader notion is the extended centralizer of a flow, defined here as

$$C_{\pm}(E) = \{F \in \mathrm{Aut}(\mathscr{X}) : (\exists \varepsilon \in \{1,-1\})(\forall t \in \mathbb{R})\ E_t \circ F = F \circ E_{\varepsilon t}\}.$$

The extended centralizer includes time-reversing symmetries of a flow.

Even more broadly, one can consider the normalizer of a flow, defined as

$$N(E) = \{F \in \mathrm{Aut}(\mathscr{X}) : (\exists \varepsilon \in \{1,-1\},\, s \in \mathbb{R})(\forall t \in \mathbb{R})\ E_t \circ F = F \circ E_{\varepsilon e^{2s} t}\}.$$

The normalizer includes symmetries that scale time, ie which conjugate the flow to a constant-speed reparametrization of itself. If $F \in N(E)$ is as above, we call F a *normalizer* of the flow, or an *s–normalizer* if we wish to specify the time dilation factor e^{2s}.

The smallest $N(E)$ can be is the flow itself, namely $N(E) = \{E_t\}_{t\in\mathbb{R}}$. When this is the case, we say that the flow E is *strongly asymmetric*. Our main result establishes this strong asymmetry property for the earthquake flow.

Theorem 1.4 *The normalizer of the earthquake flow inside the group of orbifold automorphisms of $\mathcal{P}^1\mathcal{M}_g$ is the flow itself.*

Theorem 1.2 follows immediately from Theorem 1.4, since the Teichmüller horocycle flow is normalized by the Teichmüller geodesic flow.

A few remarks Before discussing the proof of Theorem 1.4, let us make a couple of remarks.

Remark 1.5 In testing the plausibility of Theorem 1.4, it is natural to consider both Thurston's stretch map flow, defined in [Thurston 1998], and grafting, so we discuss both in turn.

The stretch map flow already makes a natural appearance in any discussion regarding Mirzakhani's conjugacy. Indeed, Mirzakhani's conjugacy shows that the earthquake flow is part of a measurable $\mathrm{SL}(2,\mathbb{R})$–action in which the diagonal subgroup acts via the stretch map flow. The stretch map flow does normalize the earthquake flow, but, since it is only defined on a full-measure subset of $\mathcal{P}^1\mathcal{M}_g$, this does not contradict Theorem 1.4.

Grafting plays a central role in the definition of complex earthquake discs. If one compares Teichmüller discs to complex earthquake discs, the Teichmüller geodesic flow corresponds to grafting. Grafting is continuous, but, since it does not normalize the earthquake flow, this does not contradict Theorem 1.4.

In the next two remarks, it is implicit that we are working in the category of topological orbifolds (so in particular all conjugacies are continuous).

Remark 1.6 Theorem 1.4 implies that the earthquake flow is not conjugate to its own inverse. (The inverse of a flow $t \mapsto E_t$ is the flow $t \mapsto E_{-t}$.)

Remark 1.7 Theorem 1.4 implies that the earthquake flow is not the restriction of any $\mathrm{SL}(2,\mathbb{R})$–action to any one-parameter subgroup. (One way to see this is to note that every noncompact one-parameter subgroup of $\mathrm{SL}(2,\mathbb{R})$ has nontrivial normalizer, since the horocycle flow is normalized by the geodesic flow and the geodesic flow is normalized by an involution.)

Outline of the proof Every normalizer can and should be considered as a conjugacy between the earthquake flow and a (possibly trivial) linear time change of itself. Given an s–normalizer $F\colon \mathcal{P}^1\mathcal{M}_g \to \mathcal{P}^1\mathcal{M}_g$, we constrain its behavior until we are eventually able to show it is an element of the flow. This involves four main steps, each occupying a different section of this paper. Throughout we assume $(X,\lambda) \in \mathcal{P}^1\mathcal{M}_g$ and $F(X,\lambda) = (Y,\mu)$.

(1) By studying minimal sets, we show in Proposition 2.1 that μ is a multicurve if and only if λ is, and, moreover, that the number of components of μ is equal to the number of components of λ. This is strongly related to [Minsky and Weiss 2002; Smillie and Weiss 2004].

(2) Leveraging the rigidity of the curve complex, we show in Proposition 3.1 that μ is a multiple of λ. This relies on [Ivanov 1997] and applies to all $(X,\lambda) \in \mathcal{P}^1\mathcal{M}_g$.

(3) By carefully analyzing the periods of specific closed orbits, we determine in Lemma 4.2 what the multiple is, showing $\mu = e^s \cdot \lambda$. We moreover show in Lemma 4.3 that, often, many curves shrink by at least a factor of e^{-s} in the passage of X to Y. This gives a contradiction unless $s = 0$, showing that the normalizer is equal to the extended centralizer, a conclusion we record as Proposition 4.1.

(4) In Proposition 5.1, we show that the extended centralizer of the earthquake flow is trivial, by showing that many and hence all orbits are preserved, and using ergodicity. We use the generalized McShane identity of [Mirzakhani 2007b] as a technical tool.

Open problems Many interesting questions related to Mirzakhani's conjugacy remain open. We highlight a few of them here.

To our knowledge, the only previously established dynamical difference between the earthquake and Teichmüller horocycle flows concerns cusp excursions in the specific case of once-punctured tori [Fu 2019]. Previous to this, it was known that certain orbits of the two flows do not stay finite distance apart in one-dimensional Teichmüller spaces [Minsky and Weiss 2002, Proposition 8.1].

Theorem 1.4 is a dynamical difference, since it relates to renormalization, but it would be illuminating to find less subtle differences.

Problem 1.8 Find a dynamical, non-group-theoretic property that is invariant under topological conjugacies and which holds for exactly one of the earthquake flow and the Teichmüller horocycle flow.

It is easy to construct topological joinings between the earthquake flow and the Teichmüller horocycle flow. For example, consider the set of pairs

$$((X,\lambda),q) \in \mathcal{P}^1\mathcal{M}_g \times \mathcal{Q}^1\mathcal{M}_g$$

such that the horizontal foliation of q is equal to λ. This construction of a topological joining admits many different variations.

Problem 1.9 Classify all the topological joinings between the earthquake flow and the Teichmüller horocycle flow.

More generally, our dynamical understanding of the earthquake flow remains incomplete, leaving questions such as the following open.

Question 1.10 Is the earthquake flow polynomially mixing?

In comparison, it is known that the Teichmüller horocycle flow is polynomially mixing [Avila et al. 2006; Avila and Resende 2012; Avila and Gouëzel 2013; Ratner 1987].

There are also interesting open questions related to strong asymmetry, including the following deliberately vague question:

Question 1.11 How common is strong asymmetry in smooth dynamics?

The most interesting setting for this question may be flows that share some properties with the earthquake flow, such as volume-preserving flows with zero entropy and having closed orbits of all periods.

Centralizers of flows (and diffeomorphisms) have been studied, for example, in [Obata 2021; Bakker and Fisher 2014; Bonomo and Varandas 2019]. Actions of Baumslag–Solitar groups and other discrete solvable groups have been studied, for example, in [Bonatti et al. 2017; Guelman and Liousse 2011; 2013; Wilkinson and Xue 2020; Burslem and Wilkinson 2004; McCarthy 2010]. Actions of solvable Lie groups have been studied, for example, in [Ghys 1985; Ghys and Verjovsky 1994]. See [Wilkinson 2010; Navas 2018] for some open questions and additional context. See [Navas 2011] for more on the one-dimensional case.

In [Frączek et al. 2014], a continuous flow on the torus is constructed that (in particular) has a measurable s–normalizer for every $s \in \mathbb{R}$ but has no continuous s–normalizers for $s \neq 0$. In light of Theorem 1.4 and the work of Mirzakhani, this is analogous to the situation for the earthquake flow. In [Frączek and Lemańczyk 2009], the symmetries of certain flows are studies in the measurable category. In [Berk et al. 2020], time-reversing translation flows are studied.

Acknowledgements Arana-Herrera is very grateful to Steve Kerckhoff for his invaluable advice, patience and encouragement. The authors are grateful to Giovanni Forni, Krzysztof Frączek, Corinna Ulcigrai and Amie Wilkinson for enlightening conversations on previous work in dynamics, and Aaron Calderon and Barak Weiss for helpful conversations regarding the appendix. This work was finished while Arana-Herrera was a member of the Institute for Advanced Study (IAS). Arana-Herrera is very grateful to the IAS for its hospitality. This material is based upon work supported by the National Science Foundation under grant DMS-1926686. During the preparation of this paper, Wright was partially supported by NSF grant DMS-1856155 and a Sloan Research Fellowship.

2 A dimension argument using minimal sets

In this section we analyze minimal sets to obtain the following:

Proposition 2.1 *Let $F\colon \mathcal{P}^1\mathcal{M}_g \to \mathcal{P}^1\mathcal{M}_g$ be a normalizer of the earthquake flow, and suppose $(X,\lambda) \in \mathcal{P}^1\mathcal{M}_g$ and $F(X,\lambda) = (Y,\mu)$. Then, for any $k \in \mathbb{N}$, λ is a simple closed multicurve with k components if and only if μ is a simple closed multicurve with k components.*

We begin by showing that every normalizer must preserve the locus of points $(X,\lambda) \in \mathcal{P}^1\mathcal{M}_g$ with λ a simple closed multicurve. We do this using the minimal sets of the earthquake flow.

A minimal set of the earthquake flow is a closed, earthquake flow–invariant subset of $\mathcal{P}^1\mathcal{M}_g$ that does not contain any proper, nonempty, closed, earthquake flow–invariant subsets.

We will be interested in compact minimal sets. Minsky and Weiss [2002] showed that all minimal sets for the earthquake flow are compact, but we will not require such a strong statement. The result we will need is the following:

Theorem 2.2 *A point $(X,\lambda) \in \mathcal{P}^1\mathcal{M}_g$ is contained in a compact minimal set if and only if λ is a simple closed multicurve.*

Smillie and Weiss [2004] prove the analogous statement for the Teichmüller horocycle flow and state that it should be possible to similarly obtain a result for the earthquake flow. However, as far as we know, even the statement of Theorem 2.2 has not previously appeared in the literature. For the convenience of the reader we sketch a proof in the appendix.

Since normalizers preserve minimal sets, we deduce the following corollary:

Corollary 2.3 *Let $F\colon \mathcal{P}^1\mathcal{M}_g \to \mathcal{P}^1\mathcal{M}_g$ be a normalizer of the earthquake flow, and suppose $(X,\lambda) \in \mathcal{P}^1\mathcal{M}_g$ and $F(X,\lambda) = (Y,\mu)$. Then λ is a simple closed multicurve if and only if μ is a simple closed multicurve.*

To get a grasp on the number of components of a simple closed multicurve, we study the local topology of the lift to $\mathcal{P}^1\mathcal{T}_g$ of the union of the compact minimal sets of the earthquake flow on $\mathcal{P}^1\mathcal{M}_g$. The following result is crucial to our approach:

Lemma 2.4 *Let $\gamma \in \mathcal{PML}_g$ be the projective class of a simple closed multicurve with $k \in \mathbb{N}$ components, $U \subseteq \mathcal{PML}_g$ be a small open neighborhood of γ in $\mathcal{PML}_g$, and W be the path-connected component containing γ of the intersection of U with the subset of $\mathcal{PML}_g$ of projective classes of simple closed multicurves. Then, if U is sufficiently small, $U \cap \overline{W}$ is locally homeomorphic to $\mathbb{R}^{6g-7-k}$.*

Proof Denote $\gamma := \sum_{i=1}^{k} a_i \gamma_i \in \mathcal{PML}_g$. Then, if U is sufficiently small, W consists of projective classes of simple closed multicurves of the form

$$\gamma' := \sum_{i=1}^{k} (a_i + \epsilon_i)\gamma_i + \sum_{j=1}^{k'} \delta_j \gamma'_j,$$

where $\epsilon := (\epsilon_i)_{i=1}^{k} \in \mathbb{R}^k$ is a small vector, $k' \geq 0$ is a nonnegative integer, $(\gamma'_j)_{j=1}^{k'}$ are pairwise nonhomotopic and nonintersecting simple closed curves that are not homotopic and do not intersect any of the components of γ, and $\delta := (\delta_j)_{j=1}^{k'} \in \mathbb{R}_+^{k'}$ is a small vector with positive entries. This fact can be readily verified using Dehn–Thurston coordinates [Penner and Harer 1992, Section 1.2]. Indeed, if U is sufficiently small, projective classes in W correspond to simple closed multicurves whose geometric intersection number with any of the components of γ is zero.

Furthermore, the closure of W in U is given by the connected component containing γ of the intersection of U with the projectivization of

$$\mathcal{Z}_g(\gamma) := \{\lambda \in \mathcal{ML}_g : i(\gamma, \lambda) = 0\}.$$

Notice that $\mathcal{Z}_g(\gamma)$ is homeomorphic to $\mathbb{R}^k \times \mathbb{R}^{6g-6-2k}$, where the first term of this product corresponds to changing the weights of the components of γ and the second term corresponds to choosing a measured geodesic lamination on S_g supported away from γ. In particular, $U \cap \overline{W}$ is locally homeomorphic to $\mathbb{R}^{6g-7-k}$. □

Suppose $(X, \gamma) \in \mathcal{P}^1\mathcal{T}_g$, where γ is a simple closed multicurve with $k \in \mathbb{N}$ components. Consider a small open neighborhood $U \subseteq \mathcal{P}^1\mathcal{T}_g$ of (X, γ). Denote by W the path-connected component containing (X, γ) of the intersection of U with the subset of points of $\mathcal{P}^1\mathcal{T}_g$ where the lamination is a simple closed multicurve. Directly from Lemma 2.4, we see that, if U is sufficiently small, $U \cap \overline{W}$ is locally homeomorphic to $\mathbb{R}^{12g-13-k}$; the $6g-6$ increase in dimension with respect to Lemma 2.4 comes from the dimension of Teichmüller space. In particular, we can recover the number of components of γ from the dimension of this intersection.

As the number of components of γ can be recovered from information depending exclusively on the minimal sets of $\mathcal{P}^1\mathcal{M}_g$, this quantity is preserved by any earthquake flow normalizer. This concludes the proof of Proposition 2.1.

3 An automorphism of the curve complex

In this section we use the rigidity of the curve complex to obtain the following:

Proposition 3.1 *Every normalizer $F : \mathcal{P}^1\mathcal{M}_g \to \mathcal{P}^1\mathcal{M}_g$ of the earthquake flow admits a Mod_g–equivariant lift $\widehat{F} : \mathcal{P}^1\mathcal{T}_g \to \mathcal{P}^1\mathcal{T}_g$ such that, for every $(X, \lambda) \in \mathcal{P}^1\mathcal{T}_g$, if $\widehat{F}(X, \lambda) = (Y, \mu)$, then μ belongs to the projective class of $\lambda \in \mathcal{ML}_g$.*

Because we assume F is an orbifold map, there exists a lift $\widehat{F}\colon \mathcal{P}^1\mathcal{T}_g \to \mathcal{P}^1\mathcal{T}_g$ that is equivariant with respect to some automorphism of Mod_g. We start with this lift and show how to modify it to get the desired lift $\widehat{F}$.

Denote by $\mathcal{S}_g$ the discrete set of free homotopy classes of unoriented simple closed curves on the marking surface S_g. By Proposition 2.1, every $X \in \mathcal{T}_g$ induces a map $\Psi_X\colon \mathcal{S}_g \to \mathcal{S}_g$ in the following way: given $\gamma \in \mathcal{S}_g$, let $\Psi_X(\gamma) \in \mathcal{S}_g$ be the free homotopy class of the simple closed curves γ' given by

$$(Y, \gamma'/\ell_{\gamma'}(Y)) := \widehat{F}(X, \gamma/\ell_\gamma(X)).$$

As $\mathcal{T}_g$ is connected and as $\mathcal{S}_g$ is discrete, the map $\Psi_X\colon \mathcal{S}_g \to \mathcal{S}_g$ is independent of $X \in \mathcal{T}_g$. From now on we denote this map simply by $\Psi\colon \mathcal{S}_g \to \mathcal{S}_g$.

We claim that Ψ induces an automorphism of the curve complex of S_g, meaning that it is bijective and that any pair of simple closed curves can be realized disjointly if and only if their images under Ψ can be realized disjointly.

Lemma 3.2 *The map $\Psi\colon \mathcal{S}_g \to \mathcal{S}_g$ defined above induces an automorphism of the curve complex of S_g.*

Proof An inverse of $\Psi\colon \mathcal{S}_g \to \mathcal{S}_g$ can be constructed using the inverse of $\widehat{F}$. It follows that Ψ is bijective.

Notice that a pair $\alpha, \beta \in \mathcal{S}_g$ of simple closed curves can be realized disjointly if and only if there exists a path

$$[0,1] \to \mathcal{P}^1\mathcal{T}_g, \quad t \mapsto (X_t, \gamma_t),$$

such that γ_t is a simple closed multicurve on S_g for every $t \in [0,1]$, $\gamma_0 = \alpha/\ell_{X_0}(\alpha)$, $\gamma_1 = \beta/\ell_{X_1}(\beta)$ and γ_t has exactly two components for every $t \in (0,1)$. It follows from Proposition 2.1 that $\widehat{F}$ preserves these types of paths. In particular, for every pair of simple closed curves $\alpha, \beta \in \mathcal{S}_g$, their images $\Psi(\alpha), \Psi(\beta) \in \mathcal{S}_g$ are nonintersecting if and only if α and β are nonintersecting. □

A well-known result of Ivanov [1997] shows that every automorphism of the curve complex of a closed, connected, oriented surface S_g of genus $g \geq 2$ is induced by the isotopy class of a diffeomorphism of S_g. Thus there exists a diffeomorphism $\psi\colon S_g \to S_g$ such that the map $\Psi\colon \mathcal{S}_g \to \mathcal{S}_g$ defined above is given by $\Psi(\gamma) = \psi(\gamma)$ for every $\gamma \in \mathcal{S}_g$. The diffeomorphism ψ acts on $\mathcal{P}^1\mathcal{T}_g$ by changing the markings even if it does not preserve the orientation of S_g. It also acts naturally on the mapping class group Mod_g by conjugation.

Since $\widehat{F}\colon \mathcal{P}^1\mathcal{T}_g \to \mathcal{P}^1\mathcal{T}_g$ is the lift of an orbifold map, there exists an automorphism $\rho\colon \mathrm{Mod}_g \to \mathrm{Mod}_g$ such that

$$\widehat{F}(\phi.(X,\lambda)) = \rho(\phi).\widehat{F}(X,\lambda)$$

for every $\phi \in \mathrm{Mod}_g$ and every $(X,\lambda) \in \mathcal{P}^1\mathcal{T}_g$. Consider the lift $\widehat{F}'\colon \mathcal{P}^1\mathcal{T}_g \to \mathcal{P}^1\mathcal{T}_g$ of F defined by

$$\widehat{F}'(X,\lambda) := \psi^{-1}.\widehat{F}(X,\lambda).$$

This lift intertwines the automorphism $\rho' : \mathrm{Mod}_g \to \mathrm{Mod}_g$ given by

$$\rho'(\phi) := \psi^{-1} \circ \rho(\phi) \circ \psi$$

for every $\phi \in \mathrm{Mod}$. Thus, by replacing $\widehat{F}$ with $\widehat{F}'$, we can assume without loss of generality that the map $\Psi : \mathcal{S}_g \to \mathcal{S}_g$ defined above is the identity.

As $\widehat{F}$ intertwines the automorphism $\rho : \mathrm{Mod}_g \to \mathrm{Mod}_g$, the map $\Psi : \mathcal{S}_g \to \mathcal{S}_g$ defined above, which we are assuming is the identity, also intertwines this automorphism. It follows that $\rho(\phi).\gamma = \phi.\gamma$ for every $\phi \in \mathrm{Mod}_g$ and every $\gamma \in \mathcal{S}_g$. As the kernels of the Mod_g–actions on $\mathcal{S}_g$ and $\mathcal{T}_g$ are equal, $\rho(\phi).X = X$ for every $\phi \in \mathrm{Mod}_g$ and every $X \in \mathcal{T}_g$. It follows that, without loss of generality, we can assume that the automorphism $\rho : \mathrm{Mod}_g \to \mathrm{Mod}_g$ is the identity.

The discussion above shows that the lift $\widehat{F}$ satisfies the following property: for every $X \in \mathcal{T}_g$ and every simple closed curve $\gamma \in \mathcal{S}_g$, if $(Y, \mu) := \widehat{F}(X, \gamma/\ell_\gamma(X)) \in \mathcal{P}^1\mathcal{T}_g$, then μ belongs to the projective class of $\gamma \in \mathcal{ML}_g$. As simple closed curves are dense in $\mathcal{PML}_g$, the same property holds for arbitrary measured geodesic laminations. This concludes the proof of Proposition 3.1.

4 Inspecting the periods of closed orbits

In this section we show that the normalizer of the earthquake flow is equal to its extended centralizer.

Proposition 4.1 $$N(E) = C_{\pm}(E).$$

In other words, given an s–normalizer F as above, we show that $s = 0$. We begin by strengthening Proposition 3.1 to control the scaling between λ and μ.

Lemma 4.2 *Let $\widehat{F}$ be the lift produced by Proposition 3.1 of an s–normalizer F. Then, for every $(X, \lambda) \in \mathcal{P}^1\mathcal{T}_g$, if $(Y, \mu) := \widehat{F}(X, \lambda)$, then $\mu = e^s \cdot \lambda$.*

Proof Since for every $(X, \lambda) \in \mathcal{P}^1\mathcal{T}_g$ the measured geodesic lamination $\mu := \mu(X, \lambda)$ given by $(Y, \mu) := \widehat{F}(X, \lambda)$ belongs to the projective class of $\lambda \in \mathcal{ML}_g$, we can consider the continuous function $c : \mathcal{P}^1\mathcal{T}_g \to \mathbb{R}^+$ which to every $(X, \lambda) \in \mathcal{P}^1\mathcal{T}_g$ assigns the unique scaling factor $c(X, \lambda) > 0$ such that

$$\mu(X, \lambda) = c(X, \lambda) \cdot \lambda. \tag{4-1}$$

Our goal is to show that $c : \mathcal{P}^1\mathcal{T}_g \to \mathbb{R}^+$ is identically equal to e^s.

Denote by $T_\gamma \in \mathrm{Mod}_g$ the Dehn twist of S_g along a simple closed curve γ. One can check that, for every $(X, a \cdot \gamma) \in \mathcal{P}^1\mathcal{T}_g$ with $a > 0$ and γ a simple closed curve on S_g, the period of the earthquake flow orbit of

$$(X, a \cdot \gamma) \in \mathcal{P}^1\mathcal{T}_g/\langle T_\gamma \rangle$$

is exactly $\ell_\gamma(X)/a$.

Now consider $\lambda = \gamma/\ell_\gamma(X)$ with γ a simple closed curve on S_g. Note that (X, λ) has period $\ell_\gamma(X)^2$ in $\mathcal{P}^1\mathcal{T}_g/\langle T_\gamma\rangle$ and $\widehat{F}(X, \lambda) = (Y, c(X, \lambda)\lambda)$ has period

$$\frac{\ell_\gamma(Y)\ell_\gamma(X)}{c(X,\lambda)} = \frac{\ell_\gamma(X)^2}{c(X,\lambda)^2},$$

where the last equality uses the fact that $c(X, \lambda)\lambda$ must have length 1 on Y. As $\widehat{F}$ is Mod_g–equivariant and as s–normalizers multiply periods by e^{-2s}, it follows that

$$\frac{\ell_\gamma(X)^2}{c(X,\lambda)^2} = e^{-2s}\ell_\gamma(X)^2.$$

Hence, $c(X, \gamma/\ell_\gamma(X)) = e^s$. As $c\colon \mathcal{P}^1\mathcal{T}_g \to \mathbb{R}^+$ is continuous and as points of the form $(X, \gamma/\ell_\gamma(X)) \in \mathcal{P}^1\mathcal{T}_g$ with γ a simple closed curve on S_g are dense in $\mathcal{P}^1\mathcal{T}_g$, this finishes the proof. □

We now prove a loop-shrinking property for lifts $\widehat{F}$ of s–normalizers of the earthquake flow. This property will play a crucial role in the proof of Theorem 1.4.

Lemma 4.3 *Let $\widehat{F}$ be the lift produced by Proposition 3.1 of an s–normalizer F of the earthquake flow. Then, for every $X \in \mathcal{T}_g$ and every simple closed curve $\alpha \in \mathcal{S}_g$, if $(Y, \mu) := \widehat{F}(X, \alpha/\ell_\alpha(X))$, then*

$$\ell_\beta(Y) \le e^{-s}\ell_\beta(X)$$

for every simple closed curve $\beta \in \mathcal{S}_g$ that can be realized disjointly from α, with equality if $\beta = \alpha$.

Proof By Lemma 4.2, $\mu = e^s \cdot \alpha/\ell_\alpha(X)$. It follows that

$$1 = \ell_\mu(Y) = e^s \cdot \ell_\alpha(X)^{-1} \cdot \ell_\alpha(Y).$$

Reorganizing the terms in this equality, we deduce

$$\ell_\alpha(Y) = e^{-s} \cdot \ell_\alpha(X).$$

Let $\beta \in \mathcal{S}_g$ be a simple closed curve that can be realized disjointly from α and is not equal to α. We average α and β with appropriate weights to obtain simple closed multicurves on S_g converging to $\alpha/\ell_\alpha(X)$, with unit length with respect to X, and whose corresponding earthquake flow orbits are periodic with explicit periods. Indeed, for every $k \in \mathbb{N}$, consider the positive weights

$$a_k = a_k(X, \alpha, \beta) := (\ell_\alpha(X) + k^{-1} \cdot \ell_\alpha(X)^{-1} \cdot \ell_\beta(X)^2)^{-1},$$
$$b_k = b_k(X, \alpha, \beta) := (\ell_\beta(X) + k \cdot \ell_\alpha(X)^2 \cdot \ell_\beta(X)^{-1})^{-1}.$$

These choices guarantee that, for every $k \in \mathbb{N}$,

$$\ell_\beta(X)/b_k = k \cdot \ell_\alpha(X)/a_k. \tag{4-2}$$

For every $k \in \mathbb{N}$, consider the simple closed multicurve on S_g given by

$$\gamma_k = \gamma_k(X, \alpha, \beta) := a_k(X, \alpha, \beta) \cdot \alpha + b_k(X, \alpha, \beta) \cdot \beta.$$

Direct computations show that $\ell_{\gamma_k}(X) = 1$ for every $k \in \mathbb{N}$. Directly from the definitions, one can also check that

$$\lim_{k\to\infty} \gamma_k = \alpha/\ell_\alpha(X).$$

For every $k \in \mathbb{N}$, consider $(Y_k, \mu_k) := \widehat{F}(X, \gamma_k)$. By Lemma 4.2, $\mu_k = e^s \cdot \gamma_k$ for every $k \in \mathbb{N}$. As $\widehat{F}$ is continuous,

$$Y = \lim_{k\to\infty} Y_k. \tag{4-3}$$

Fix $k \in \mathbb{N}$. Denote by $T_\alpha, T_\beta \in \mathrm{Mod}_g$ the Dehn twists of S_g along α and β. A direct computation using (4-2) shows that the earthquake flow orbit of the image of (X, γ_k) in $\mathcal{P}^1\mathcal{T}_g/\langle T_\alpha, T_\beta\rangle$ is periodic with period given by the least common multiple

$$\mathrm{lcm}(\ell_\alpha(X)/a_k, \ell_\beta(X)/b_k) = \ell_\beta(X)/b_k. \tag{4-4}$$

Analogously, the earthquake flow orbit of the image of (Y_k, μ_k) in $\mathcal{P}^1\mathcal{T}_g/\langle T_\alpha, T_\beta\rangle$ is periodic if and only if the following least common multiple is finite, in which case it is exactly the period of the orbit:

$$\mathrm{lcm}(\ell_\alpha(Y_k)/(e^s \cdot a_k), \ell_\beta(Y_k)/(e^s \cdot b_k)). \tag{4-5}$$

Since s–normalizers multiply periods by e^{-2s}, for the periods in (4-4) and (4-5) to agree, it is necessary that

$$\ell_\beta(Y_k) \le e^{-s} \cdot \ell_\beta(X).$$

Taking limits as $k \to \infty$ and using (4-3), we conclude

$$\ell_\beta(Y) \le e^{-s} \cdot \ell_\beta(X). \qquad \square$$

We can now conclude the proof of Proposition 4.1 as follows:

Proof of Proposition 4.1 Suppose by contradiction that $s \neq 0$. By working with the inverse of F if $s < 0$, we can assume without loss of generality that $s > 0$. Denote by $\widehat{F}$ the Mod_g–equivariant lift provided by Proposition 3.1. Let $\alpha, \beta, \gamma \in \mathcal{S}_g$ be simple closed curves such that α can be realized disjointly from β and γ, and such that β and γ have positive geometric intersection number. Fix $X \in \mathcal{T}_g$ and let

$$(X_n, \lambda_n) := \widehat{F}^n(X, \alpha/\ell_\alpha(X))$$

for every $n \in \mathbb{N}$. By Lemma 4.3, there exists $N \in \mathbb{N}$ such that $\ell_\beta(X_N)$ and $\ell_\gamma(X_N)$ are arbitrarily small, contradicting the collar lemma for hyperbolic surfaces. $\square$

5 The centralizer of the earthquake flow

In this section we show that the extended centralizer of the earthquake flow is trivial.

Proposition 5.1 $$C_\pm(E) = E.$$

We proceed in several steps, starting with the following geometric result:

Lemma 5.2 *Let X and Y be a pair of compact, connected and orientable diffeomorphic hyperbolic surfaces with at least one totally geodesic boundary component. Suppose that, for some pair of markings on X and Y, the lengths of the boundary components of X agree with those of Y, and, for every simple closed curve, the length of its geodesic representative on Y is at most the length of its geodesic representative on X. Then X and Y are isometric.*

An analogous statement for closed surfaces is well known [Thurston 1998, Theorem 3.1]. We do not know if the exact statement of Lemma 5.2 has appeared before in the literature, but, in any case, a short proof is possible from known results.

Proof The monotonicity of the summands in Mirzakhani's generalized McShane's identity [2007b, Theorem 1.3] guarantees that, if X and Y satisfy the assumptions, then, for every simple closed curve, the lengths of its geodesic representatives on X and Y are equal. As the isometry class of a marked hyperbolic structure with totally geodesic boundary components on a compact, connected, orientable surface is determined by its marked length spectrum,[1] we conclude that X and Y are isometric. □

The following result shows that centralizers of the earthquake flow map points of the form $(X, \alpha/\ell_\alpha(X)) \in \mathcal{P}^1\mathcal{M}_g$ into their own earthquake flow orbit.

Lemma 5.3 *Suppose $F \in C_\pm(E)$ and let $\widehat{F}$ be the lift provided by Proposition 3.1. Then, for every $X \in \mathcal{T}_g$ and every simple closed curve $\alpha \in \mathcal{S}_g$, there exists a unique $t \in \mathbb{R}$ satisfying*

$$\widehat{F}(X, \alpha/\ell_\alpha(X)) = E_t(X, \alpha/\ell_\alpha(X)).$$

For the proof, it is helpful to recall that an element of the extended centralizer is nothing other than an s–normalizer with $s = 0$.

Proof Let $(Y, \mu) := \widehat{F}(X, \alpha/\ell_\alpha(X)) \in \mathcal{P}^1\mathcal{T}_g$. Lemmas 4.2 and 4.3 ensure that $\mu = \alpha/\ell_\alpha(X) \in \mathcal{ML}_g$, $\ell_\alpha(Y) = \ell_\alpha(X)$, and $\ell_\beta(Y) \le \ell_\beta(X)$ for every simple closed curve $\beta \in \mathcal{S}_g$ that can be realized disjointly from α.

Cutting X and Y along the corresponding geodesic representatives of α on each surface yields a pair of (possibly disconnected) hyperbolic surfaces with totally geodesic boundary components of matching lengths. Lemma 5.2 guarantees these surfaces are isometric. As X and Y can be recovered from isometric pieces by gluing along the boundary components corresponding to α, we deduce that X and Y only differ by a Fenchel–Nielsen twist along α. In other words,

$$\widehat{F}(X, \alpha/\ell_\alpha(X)) = (Y, \mu) = E_t(X, \alpha/\ell_\alpha(X)).$$ □

The following result extends the conclusion of Lemma 5.3 to arbitrary points $(X, \lambda) \in \mathcal{P}^1\mathcal{T}_g$:

[1] A proof can be obtained by adapting the arguments in [Farb and Margalit 2012, Proof of Theorem 10.7].

Lemma 5.4 *Suppose $F \in C_{\pm}(E)$ and let $\widehat{F}$ be the lift provided by Proposition 3.1. Then there exists a continuous, Mod_g–invariant function $t\colon \mathcal{P}^1\mathcal{T}_g \to \mathbb{R}$ such that, for every $(X,\lambda) \in \mathcal{P}^1\mathcal{T}_g$, $t = t(X,\lambda)$ satisfies*

$$\widehat{F}(X,\lambda) = E_t(X,\lambda)$$

and is the unique real number satisfying this equation.

Furthermore, if $F \in C(E)$, then t is earthquake flow–invariant, and, if $F \in C_{\pm}(E) \setminus C(E)$, then T is "twisted-equivariant", in the sense that

$$t(E_s(X,\lambda)) = t(X,\lambda) - 2s.$$

Proof Fix $(X,\lambda) \in \mathcal{P}^1\mathcal{T}_g$. As weighted simple closed curves are dense in $\mathcal{ML}_g$, one can find a sequence $(\lambda_n)_{n\in\mathbb{N}}$ of length 1 weighted simple closed curves such that $\lambda_n \to \lambda$ in $\mathcal{ML}_g$ as $n \to \infty$. By Lemma 5.3, for every $n \in \mathbb{N}$, there exists $t_n \in \mathbb{R}$ such that

$$\widehat{F}(X,\lambda_n) = E_{t_n}(X,\lambda_n). \tag{5-1}$$

Claim 5.5 *The sequence $(t_n)_{n\in\mathbb{N}}$ is bounded.*

Proof Suppose by contradiction this was not the case. Assume t_n diverges to $+\infty$ along a subsequence; the case when t_n diverges to $-\infty$ along a subsequence can be treated in an analogous way. Rename this subsequence as $(t_n)_{n\in\mathbb{N}}$ and assume without loss of generality that all of its terms are positive. Let $\mu \in \mathcal{ML}_g$ be a measured geodesic lamination such that

$$\iint_X \cos\theta \, d\lambda \, d\mu > 0, \tag{5-2}$$

where θ is the angle measured counterclockwise from μ to λ at each intersection between μ and λ. The existence of such a measured geodesic lamination $\mu \in \mathcal{ML}_g$ can be argued as follows. By the infinitesimal version of Thurston's earthquake theorem (see for instance [Kerckhoff 1983, Appendix, Theorem 2]), every tangent vector at $X \in \mathcal{T}_g$ can be realized by an infinitesimal earthquake. In particular, by Kerckhoff's derivative formula [loc. cit., Corollary 3.4], the only way μ could not exist is if the function $Y \mapsto \ell_\lambda(Y) > 0$ for $Y \in \mathcal{T}_g$ had a critical point, and, by convexity of length functions [loc. cit., Section 3, Theorem 1], a minimum at X. This is not possible, as can be seen, for instance, using shear coordinates and reverse stretch lines.

By [loc. cit., Corollary 3.4], the integral in (5-2) is equal to the derivative at $t = 0$ of the convex function $t \mapsto \ell_\mu(E_t(X,\lambda))$. By continuity, there exists $c > 0$ and $N \in \mathbb{N}$ such that, for every $n \geq N$,

$$\iint_X \cos\theta \, d\lambda_n \, d\mu > c.$$

Kerckhoff's work guarantees that, for every $n \geq N$,

$$\ell_\mu(E_{t_n}(X,\lambda_n)) \geq \ell_\mu(X) + t_n \cdot c. \tag{5-3}$$

Denote by $\pi : \mathcal{P}^1\mathcal{T}_g \to \mathcal{T}_g$ the natural projection defined by $\pi(X, \lambda) = X$. By definition,

$$E_{t_n}(X, \lambda_n) = \widehat{F}(X, \lambda_n) \in \widehat{F}(\pi^{-1}(X)).$$

As $\widehat{F}$ is continuous, the set $\widehat{F}(\pi^{-1}(X)) \subseteq \mathcal{P}^1\mathcal{T}_g$ is compact. Thus, the sequence $\big(\ell_\mu(E_{t_n}(X, \lambda_n))\big)_{n\in\mathbb{N}}$ must be bounded. Taking limits as $n \to \infty$ in (5-3) yields a contradiction, concluding the proof of the claim. □

As $(t_n)_{n\in\mathbb{N}}$ is bounded, it admits a subsequence converging to some $t \in \mathbb{R}$. Taking limits in (5-1) along this subsequence, we deduce

$$\widehat{F}(X, \lambda) = E_t(X, \lambda). \tag{5-4}$$

The uniqueness of $t \in \mathbb{R}$ satisfying this condition follows directly from the fact that earthquake flow orbits in $\mathcal{P}^1\mathcal{T}_g$ are embedded. The continuity of the corresponding function $t\colon \mathcal{P}^1\mathcal{T}_g \to \mathbb{R}$ follows from (5-4) and uniqueness. The Mod_g–invariance of t can be verified using (5-4) and the fact that $\widehat{F}$ is Mod_g–equivariant. The earthquake flow–invariance or twisted-equivariance of t can be verified directly from (5-4) and the fact that $\widehat{F}$ is in the extended centralizer of the earthquake flow. □

We are now ready to conclude.

Proof of Proposition 5.1 Consider the function $t : \mathcal{P}^1\mathcal{T}_g \to \mathbb{R}$ above. Since it is Mod_g–equivariant, it induces a function $t : \mathcal{P}^1\mathcal{M}_g \to \mathbb{R}$.

If $F \in C(E)$, the function t is earthquake flow–invariant. As the earthquake flow on $\mathcal{P}^1\mathcal{M}_g$ is ergodic with respect to a measure of full support, t is equal to a constant $t_0 \in \mathbb{R}$ on a dense set of $\mathcal{P}^1\mathcal{M}_g$. Applying continuity and density, we conclude $F = E_{t_0}$, as desired.

Suppose $F \in C_\pm(E) \setminus C(E)$. There exists c such that the set $t^{-1}((c, c+2))$ has positive measure. The twisted-equivariance gives that, for all k, E_k maps $t^{-1}((c, c+2))$ into $t^{-1}((c-2k, c+2-2k))$. For different k integral, the sets $t^{-1}((c-2k, c+2-2k))$ are disjoint, and, since earthquake flow is measure-preserving, they all have the same measure. So considering all k integral contradicts the fact that the space has finite measure, showing that such an F cannot exist. □

We are now ready to prove that the earthquake flow is strongly asymmetric.

Proof of Theorem 1.4 Proposition 4.1 shows that $N(E) = C_\pm(E)$ and Proposition 5.1 shows that $C_\pm(E) = E$. □

Appendix Minimal sets

In this appendix we sketch, for the convenience of the reader, a proof of Theorem 2.2. The corresponding result in the case of the Teichmüller horocycle flow is discussed in detail by Smillie and Weiss [2004], who remark there that "an analogous result for the earthquake flow may be proved by a similar argument". Our starting point is the following observation, the details of whose proof are left to the reader:

Lemma A.1 *If $K \subset \mathcal{P}^1\mathcal{M}_g$ is a minimal set for the earthquake flow, and (X,λ) and (X',λ') are in K, then $X-\lambda$ is isometric to $X'-\lambda'$.*

Sketch of proof For any fixed $(X,\lambda) \in K$, consider the set $K' \subseteq K$ of all $(X',\lambda') \in K$ for which there exists an isometric embedding

$$X-\lambda \hookrightarrow X'-\lambda'$$

of complementary regions. Since the complementary regions are not changed by the earthquake flow, K' is invariant. A limit argument shows that K' is closed, so the definition of minimality guarantees $K' = K$.

Thus, for every $(X,\lambda), (X',\lambda') \in K$, each complementary region embeds isometrically into the other. Hence $X-\lambda = X'-\lambda'$. □

We also need the following nontrivial result:

Proposition A.2 *If λ is not a multicurve and the orbit of (X,λ) is bounded in $\mathcal{P}^1\mathcal{M}_g$, then the orbit accumulates on some (X',λ') with $X-\lambda \neq X'-\lambda'$.*

In fact, experts believe the following stronger statement is true (and accessible):

Problem A.3 Prove that, if λ is not a multicurve, then the earthquake flow orbit of (X,λ) is not bounded.

We will not consider this problem here since it is certainly harder than what we require. The analogous problem for the Teichmüller horocycle flow is item (IV) in the list of problems at the end of [Smillie and Weiss 2004] and has been considered in unpublished work of those authors.

Before addressing Proposition A.2, we note it implies Theorem 2.2.

Proof of Theorem 2.2 assuming Proposition A.2 If K is a compact minimal set and $(X,\lambda) \in K$, then Lemma A.1 implies that any (X',λ') in the orbit closure of (X,λ) has $X-\lambda = X'-\lambda'$, and so Proposition A.2 implies λ is a multicurve.

The converse implication — that if λ is a multicurve then the orbit closure of (X,λ) is a minimal set — is well known. Indeed, if $T \subseteq \mathcal{P}^1\mathcal{M}_g$ is the subset obtained by starting at (X,λ) and independently twisting at each component of λ, then T is an invariant torus and the earthquake flow is continuously conjugate to a straight-line flow on T. The converse implication follows from the fact that, for straight-line flows on tori, every orbit closure is a minimal set. □

We conclude by briefly sketching how the ideas of Smillie and Weiss apply to Proposition A.2. Most of the work is divided into two lemmas.

Lemma A.4 *Suppose λ is a measured geodesic lamination on X that is not a multicurve. Then there exists some $\delta > 0$ such that, for all $\epsilon > 0$, we can find segments γ_1 and γ_2 of leaves of λ that stay within distance 1 of each other and are such that all leaves of λ that come within δ of the starting point p_1 of γ_1 do so on the side of γ_1 containing γ_2, and all leaves that come within δ of the endpoint p_2 of γ_2 do so*

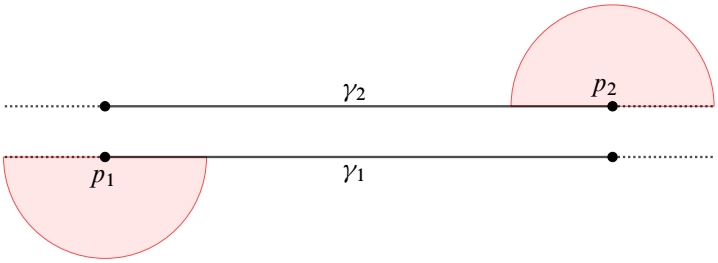


Figure 1: The γ_i. The shaded half balls of radius δ do not intersect λ.

on the side of γ_2 containing γ_1, and such that the transverse measure of a segment from γ_1 to γ_2 is less than ϵ. Moreover, γ_1 and γ_2 can be taken to lie on nonisolated leaves of λ.

In particular, it follows that both γ_i are segments of leaves of λ adjacent to regions of $X - \lambda$. See Figure 1.

The proof will use the concept of the thick part of a surface with boundary, which can be defined by embedding the surface in its double and taking the thick part there; see for example [Lipnowski and Wright 2024, Section 2.1] for details.

Sketch of proof Without loss of generality assume γ has no closed leaves. Start with p_1 on the boundary of the thick part of $X - \lambda$, on a leaf α of λ. Pick a point q that is very close to p_1 and on a leaf β of λ. Follow both leaves α and β in the same direction until they are distance $\frac{1}{10}$ apart. The region R between these segments of α and β, illustrated in Figure 2, has definite area.

The area of the thin part of $X - \lambda$ is small, so the thick part must intersect R. (Here the thick part should be defined appropriately using δ, and δ should be taken small enough.)

We then pick p_2 to be on the boundary of the thick part of $X - \lambda$ intersected with R. (One should pick p_2 so that the thick part and α are on different sides of the leaf through p_2.) We define γ_1 to be the segment of α from p_1 to the projection of p_2 onto α, and similarly define γ_2 using the leaf through p_2. □

Lemma A.5 *There exists a universal constant $C > 0$ such that the following holds. Consider any measured geodesic lamination on $\mathbb{H}$, any segments γ_1 and γ_2 of nonatomic leaves of λ that stay within distance 1 of each other, and any $p_1 \in \gamma_1$ and $p_2 \in \gamma_2$. Assume there are leaves of λ that go between p_1 and p_2. Let $\lambda_{\max}$ be a maximal geodesic lamination containing λ. Assume the p_i lie on the boundary of $\mathbb{H} - \lambda_{\max}$. Then there is a unique $t \in \mathbb{R}$ such that the image of p_1 and p_2 under the time t earthquake*

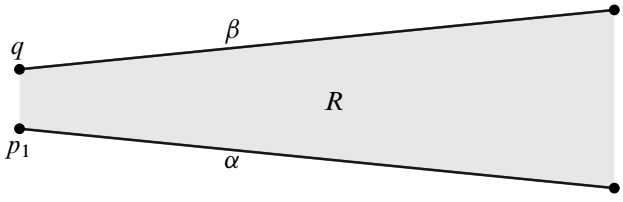


Figure 2: The region R.

of λ can be joined by a segment s of a leaf of the horocyclic foliation of $\lambda_{\max}$ and this segment has length at most C.

In applications, often λ is already maximal, so $\lambda_{\max} = \lambda$. The main conclusion here is that p_1 and p_2 become bounded distance from each other; the use of the horocyclic foliation (and $\lambda_{\max}$) is merely a convenient technical tool to obtain this.

One should of course think of $\mathbb{H}$ as the universal cover of a closed surface X; we use the universal cover only so that we do not have to specify a homotopy class for the arc s.

Sketch of proof The first claim is related to the fact that shears change linearly under earthquakes; see for example the survey [Wright 2022, Section 4].

If one considers a rectangle R bounded by γ_1 and γ_2, then λ divides this rectangle up into countably many small rectangles bounded by leaves of λ. The preimage of s on (X, λ) consists of one horocyclic arc in each small rectangle; compare to a Cantor staircase.

For each small rectangle, one can define its maximum height to be the maximum length of a horocyclic arc crossing that rectangle. A standard estimate shows that the sum of the maximum heights is at most some constant C; see [Thurston 1998, page 16]. This uses the fact that the γ_i remain within distance 1 of each other.

The length of s is the sum of the lengths of the horocyclic arcs of $E_{-t}(s)$, which is at most C. This gives the result. □

Sketch of proof of Proposition A.2 Consider a sequence $\epsilon_n \to 0$ and, for each $n \in \mathbb{N}$, let $\gamma_{1,n}$, $\gamma_{2,n}$, $p_{1,n}$ and $p_{2,n}$ be as provided by Lemma A.4 with $\epsilon = \epsilon_n$.

The output of Lemma A.5 is a sequence of points $(X_n, \lambda_n) \in \mathcal{P}^1\mathcal{M}_g$ on the earthquake flow orbit of (X, λ) such that two points on the boundary of the thick part of $X_n - \lambda_n = X - \lambda$ are joined by a path on X_n of hyperbolic length at most C and transverse measure going to 0 as $n \to \infty$. By extending these paths into the thick part and taking geodesic representatives, we obtain geodesic paths σ_n on X_n, of lengths bounded above and below, which are uniformly transverse to λ_n and which have the same transverse measure as the original paths of length at most C.

Passing to a subsequence if necessary, we can assume (X_n, λ_n) converges to some $(X_\infty, \lambda_\infty) \in \mathcal{P}^1\mathcal{M}_g$. For convenience, we can also assume that the supports of the λ_n converge to a geodesic lamination $\widehat{\lambda}_\infty$ which contains the support of λ_∞.

Since the complementary regions $X_n - \lambda_n$ are constant, it follows that $X - \lambda = X_\infty - \widehat{\lambda}_\infty$. Thus, to show that $X - \lambda \neq X_\infty - \lambda_\infty$, it suffices to show that some leaves of the geodesic lamination $\widehat{\lambda}_\infty$ are not contained in the support of λ_∞.

This is verified by considering a limit σ of the geodesic segments σ_n; the limit σ has length bounded above and below, is transverse to $\widehat{\lambda}_\infty$, and has 0 transverse measure with respect to λ_∞. □

References

[Arana-Herrera 2021] **F Arana-Herrera**, *Equidistribution of families of expanding horospheres on moduli spaces of hyperbolic surfaces*, Geom. Dedicata 210 (2021) 65–102 MR Zbl

[Arana-Herrera 2022] **F Arana-Herrera**, *Counting hyperbolic multigeodesics with respect to the lengths of individual components and asymptotics of Weil–Petersson volumes*, Geom. Topol. 26 (2022) 1291–1347

[Avila and Gouëzel 2013] **A Avila**, **S Gouëzel**, *Small eigenvalues of the Laplacian for algebraic measures in moduli space, and mixing properties of the Teichmüller flow*, Ann. of Math. 178 (2013) 385–442 MR Zbl

[Avila and Resende 2012] **A Avila**, **M J Resende**, *Exponential mixing for the Teichmüller flow in the space of quadratic differentials*, Comment. Math. Helv. 87 (2012) 589–638 MR Zbl

[Avila et al. 2006] **A Avila**, **S Gouëzel**, **J-C Yoccoz**, *Exponential mixing for the Teichmüller flow*, Publ. Math. Inst. Hautes Études Sci. 104 (2006) 143–211 MR Zbl

[Bakker and Fisher 2014] **L Bakker**, **T Fisher**, *Open sets of diffeomorphisms with trivial centralizer in the C^1 topology*, Nonlinearity 27 (2014) 2869–2885 MR Zbl

[Berk et al. 2020] **P Berk**, **K Frączek**, **T de la Rue**, *On typicality of translation flows which are disjoint with their inverse*, J. Inst. Math. Jussieu 19 (2020) 1677–1737 MR Zbl

[Bonatti et al. 2017] **C Bonatti**, **I Monteverde**, **A Navas**, **C Rivas**, *Rigidity for C^1 actions on the interval arising from hyperbolicity, I: Solvable groups*, Math. Z. 286 (2017) 919–949 MR Zbl

[Bonomo and Varandas 2019] **W Bonomo**, **P Varandas**, *A criterion for the triviality of the centralizer for vector fields and applications*, J. Differential Equations 267 (2019) 1748–1766 MR Zbl

[Burslem and Wilkinson 2004] **L Burslem**, **A Wilkinson**, *Global rigidity of solvable group actions on S^1*, Geom. Topol. 8 (2004) 877–924 MR Zbl

[Calderon and Farre 2024] **A Calderon**, **J Farre**, *Shear-shape cocycles for measured laminations and ergodic theory of the earthquake flow*, Geom. Topol. 28 (2024) 1995–2124

[Erlandsson and Souto 2022] **V Erlandsson**, **J Souto**, *Counting curves on orbifolds*, Trans. Lond. Math. Soc. 9 (2022) 56–85 MR Zbl

[Farb and Margalit 2012] **B Farb**, **D Margalit**, *A primer on mapping class groups*, Princeton Math. Ser. 49, Princeton Univ. Press (2012) MR Zbl

[Frączek and Lemańczyk 2009] **K Frączek**, **M Lemańczyk**, *On the self-similarity problem for ergodic flows*, Proc. Lond. Math. Soc. 99 (2009) 658–696 MR Zbl

[Frączek et al. 2014] **K Frączek**, **J Kułaga-Przymus**, **M Lemańczyk**, *Non-reversibility and self-joinings of higher orders for ergodic flows*, J. Anal. Math. 122 (2014) 163–227 MR Zbl

[Fu 2019] **S-W Fu**, *Cusp excursions for the earthquake flow on the once-punctured torus*, Conform. Geom. Dyn. 23 (2019) 251–261 MR Zbl

[Ghys 1985] **É Ghys**, *Actions localement libres du groupe affine*, Invent. Math. 82 (1985) 479–526 MR Zbl

[Ghys and Verjovsky 1994] **É Ghys**, **A Verjovsky**, *Locally free holomorphic actions of the complex affine group*, from "Geometric study of foliations" (T Mizutani, K Masuda, S Matsumoto, T Inaba, T Tsuboi, Y Mitsumatsu, editors), World Sci., River Edge, NJ (1994) 201–217 MR Zbl

[Guelman and Liousse 2011] **N Guelman**, **I Liousse**, *C^1–actions of Baumslag–Solitar groups on S^1*, Algebr. Geom. Topol. 11 (2011) 1701–1707 MR Zbl

[Guelman and Liousse 2013] **N Guelman**, **I Liousse**, *Actions of Baumslag–Solitar groups on surfaces*, Discrete Contin. Dyn. Syst. 33 (2013) 1945–1964 MR Zbl

[Hubbard and Masur 1979] **J Hubbard**, **H Masur**, *Quadratic differentials and foliations*, Acta Math. 142 (1979) 221–274 MR Zbl

[Ivanov 1997] **N V Ivanov**, *Automorphism of complexes of curves and of Teichmüller spaces*, Int. Math. Res. Not. 1997 (1997) 651–666 MR Zbl

[Kerckhoff 1983] **S P Kerckhoff**, *The Nielsen realization problem*, Ann. of Math. 117 (1983) 235–265 MR Zbl

[Lipnowski and Wright 2024] **M Lipnowski**, **A Wright**, *Towards optimal spectral gaps in large genus*, Ann. Probab. 52 (2024) 545–575 MR Zbl

[Liu 2022] **M Liu**, *Length statistics of random multicurves on closed hyperbolic surfaces*, Groups Geom. Dyn. 16 (2022) 437–459 MR Zbl

[Lu and Su 2022] **S Lu**, **W Su**, *Counting mapping class group orbits under shearing coordinates*, Geom. Dedicata 216 (2022) art. id. 16 MR Zbl

[Masur 1995] **H Masur**, *The Teichmüller flow is Hamiltonian*, Proc. Amer. Math. Soc. 123 (1995) 3739–3747 MR Zbl

[McCarthy 2010] **A E McCarthy**, *Rigidity of trivial actions of abelian-by-cyclic groups*, Proc. Amer. Math. Soc. 138 (2010) 1395–1403 MR Zbl

[McMullen 1998] **C T McMullen**, *Complex earthquakes and Teichmüller theory*, J. Amer. Math. Soc. 11 (1998) 283–320 MR Zbl

[Minsky and Weiss 2002] **Y Minsky**, **B Weiss**, *Nondivergence of horocyclic flows on moduli space*, J. Reine Angew. Math. 552 (2002) 131–177 MR Zbl

[Mirzakhani 2007a] **M Mirzakhani**, *Random hyperbolic surfaces and measured laminations*, from "In the tradition of Ahlfors–Bers, IV" (D Canary, J Gilman, J Heinonen, H Masur, editors), Contemp. Math. 432, Amer. Math. Soc., Providence, RI (2007) 179–198 MR Zbl

[Mirzakhani 2007b] **M Mirzakhani**, *Simple geodesics and Weil–Petersson volumes of moduli spaces of bordered Riemann surfaces*, Invent. Math. 167 (2007) 179–222 MR Zbl

[Mirzakhani 2008] **M Mirzakhani**, *Ergodic theory of the earthquake flow*, Int. Math. Res. Not. 2008 (2008) art. id. rnm116 MR Zbl

[Mirzakhani 2016] **M Mirzakhani**, *Counting mapping class group orbits on hyperbolic surfaces*, preprint (2016) arXiv 1601.03342

[Navas 2011] **A Navas**, *Groups of circle diffeomorphisms*, Univ. of Chicago Press (2011) MR Zbl

[Navas 2018] **A Navas**, *Group actions on* 1*–manifolds: a list of very concrete open questions*, from "Proceedings of the International Congress of Mathematicians, III" (B Sirakov, P N de Souza, M Viana, editors), World Sci., Hackensack, NJ (2018) 2035–2062 MR Zbl

[Obata 2021] **D Obata**, *Symmetries of vector fields: the diffeomorphism centralizer*, Discrete Contin. Dyn. Syst. 41 (2021) 4943–4957 MR Zbl

[Penner and Harer 1992] **R C Penner**, **J L Harer**, *Combinatorics of train tracks*, Ann. of Math. Stud. 125, Princeton Univ. Press (1992) MR Zbl

[Ratner 1987] **M Ratner**, *The rate of mixing for geodesic and horocycle flows*, Ergodic Theory Dynam. Systems 7 (1987) 267–288 MR Zbl

[Smillie and Weiss 2004] **J Smillie**, **B Weiss**, *Minimal sets for flows on moduli space*, Israel J. Math. 142 (2004) 249–260 MR Zbl

[Sözen and Bonahon 2001] **Y Sözen**, **F Bonahon**, *The Weil–Petersson and Thurston symplectic forms*, Duke Math. J. 108 (2001) 581–597 MR Zbl

[Thurston 1979] **W P Thurston**, *The geometry and topology of three-manifolds*, lecture notes, Princeton University (1979) Available at `https://url.msp.org/gt3m`

[Thurston 1998] **W P Thurston**, *Minimal stretch maps between hyperbolic surfaces*, preprint (1998) arXiv math/9801039 Reprinted in his "Collected works with commentary, I: Foliations, surfaces and differential geometry" (B Farb, D Gabai, S P Kerckhoff, editors), Amer. Math. Soc., Providence, RI (2022) 533–585

[Wilkinson 2010] **A Wilkinson**, *Conservative partially hyperbolic dynamics*, from "Proceedings of the International Congress of Mathematicians, III" (R Bhatia, A Pal, G Rangarajan, V Srinivas, M Vanninathan, P Gastesi, editors), Hindustan, New Delhi (2010) 1816–1836 MR Zbl

[Wilkinson and Xue 2020] **A Wilkinson**, **J Xue**, *Rigidity of some abelian-by-cyclic solvable group actions on* $\mathbb{T}^N$, Comm. Math. Phys. 376 (2020) 1223–1259 MR Zbl

[Wright 2020] **A Wright**, *A tour through Mirzakhani's work on moduli spaces of Riemann surfaces*, Bull. Amer. Math. Soc. 57 (2020) 359–408 MR Zbl

[Wright 2022] **A Wright**, *Mirzakhani's work on earthquake flow*, from "Teichmüller theory and dynamics" (P Dehornoy, E Lanneau, editors), Panor. Synth. 58, Soc. Math. France, Paris (2022) 101–134 MR Zbl

Department of Mathematics, University of Maryland
College Park, MD, United States

Department of Mathematics, University of Michigan
Ann Arbor, MI, United States

`farana@umd.edu, alexmw@umich.edu`

Proposed: Anna Wienhard
Seconded: Dmitri Burago, Mladen Bestvina

Received: 3 February 2022
Accepted: 25 February 2023

msp

Geometry & Topology **28**:5 (2024) 2145–2206
DOI: 10.2140/gt.2024.28.2145
Published: 24 August 2024

The persistence of a relative Rabinowitz–Floer complex

GEORGIOS DIMITROGLOU RIZELL
MICHAEL G SULLIVAN

We give a quantitative refinement of the invariance of the Legendrian contact homology algebra in general contact manifolds. We show that in this general case, the Lagrangian cobordism trace of a Legendrian isotopy defines a DGA stable tame isomorphism, which is similar to a bifurcation invariance proof for a contactization contact manifold. We use this result to construct a relative version of the Rabinowitz–Floer complex defined for Legendrians that also satisfies a quantitative invariance, and study its persistent homology barcodes. We apply these barcodes to prove several results, including: displacement energy bounds for Legendrian submanifolds in terms of the oscillatory norms of the contact Hamiltonians; a proof of Rosen and Zhang's nondegeneracy conjecture for the Shelukhin–Chekanov–Hofer metric on Legendrian submanifolds; and the nondisplaceability of the standard Legendrian real-projective space inside the contact real-projective space.

53D10, 53D42

1 Introduction

Let (Y^{2n+1}, ξ) be a $(2n+1)$–dimensional contact manifold with contact form α, and $\Lambda \subset Y$ be a (closed) n–dimensional Legendrian submanifold. Specific assumptions we make for Λ and (Y, ξ) vary based on our result. This list is summarized in Remark 1.3. A *Reeb chord* (or *α–Reeb chord*) of Λ is a nontrivial flow starting and ending on Λ, of the Reeb vector field $R_\alpha \in \Gamma(TY)$, which is defined by $\alpha(R_\alpha) = 1$ and $d\alpha(R_\alpha, \cdot) = 0$. We are interested in estimating the number of Reeb chords from a given Legendrian (closed submanifold) Λ to its image under a contact isotopy with compact support. If there are no such Reeb chords, we say that the contact isotopy *displaces* Λ for that given α. This is the contact analogue of a Hamiltonian isotopy displacing a Lagrangian submanifold; see Chekanov [7]. Our Main Theorem, Theorem 1.4, has more information about Reeb chords than the known analogous results for Lagrangian intersection points, because we not only give a single lower bound on how long some fixed number of chords persist, but rather, a sequence of lower bounds depending on the number of chords required to persist.

Our main tool is a filtered *Legendrian contact homology differential graded algebra* (also called the *Chekanov–Eliashberg DGA*). Let $\mathcal{A}(\Lambda)$ be the free noncommutative unital algebra over the field (or ring) $\boldsymbol{k}$ freely generated by α–Reeb chords of Λ. If the moduli spaces of these disks can be oriented in a coherent way, for example by the choice of a spin structure on the Legendrian as in Ekholm, Etnyre and Sullivan [24] and Karlsson [29], then $\boldsymbol{k}$ is $\mathbb{Z}$ or $\mathbb{Z}_p$. Otherwise $\boldsymbol{k} = \mathbb{Z}_2$.

The grading is induced by the Conley–Zehnder index of Reeb chords; see Section A.2.

The differential ∂ has degree -1 and counts J–holomorphic disks in the symplectization $(\mathbb{R}_\tau \times Y, d(e^\tau \alpha))$ with Lagrangian boundary condition $\mathbb{R}_\tau \times \Lambda$.

Each Reeb chord c has a *length* (or *action*) $\ell(c) := \int_c \alpha > 0$. For $0 < l \leq \infty$, let $\mathscr{A}^l(\Lambda)$ be the unital subalgebra generated by those generators c length bounded from above by $\int_c \alpha < l$. The action-decreasing property of the differential, which is a direct consequence of the positivity of $d\alpha$–area and Stokes' theorem, implies that $\mathscr{A}^l(\Lambda) \subset \mathscr{A}(\Lambda)$ is a unital sub-DGA.

An *augmentation* for the DGA $\mathscr{A}^l$ is a (graded) DGA–morphism $\varepsilon\colon (\mathscr{A}^l, \partial) \to (\boldsymbol{k}, \partial_{\boldsymbol{k}} := 0)$ to the ground field. Because $\ell(1) = 0$, for any Chekanov–Eliashberg DGA $\mathscr{A}$, there exists $l > 0$ such that $\mathscr{A}^l$ has an augmentation.

The *oscillation* of a contact (α–)Hamiltonian $H_t\colon Y \times \mathbb{R}_\tau \to \mathbb{R}$ is

$$\|H_t\|_{\mathrm{osc}} := \int_0^1 \left(\max_{y \in Y} H_t - \min_{y \in Y} H_t\right) dt.$$

This oscillation defines the Hofer norm of the corresponding contact Hamiltonian isotopy ϕ^t_{α, H_t}.

In order to circumvent the analytical difficulties of establishing invariance of the Legendrian contact homology algebra for general contact forms, we will make certain technical assumptions on the contractible periodic Reeb orbits of (Y, α), at least below some fixed length. To this purpose, we introduce the following definition.

Definition 1.1 Consider a contractible and nondegenerate periodic Reeb orbit γ of (Y, α). We let $|\gamma| \in \mathbb{Z} \cup \{-\infty\}$ denote the *minimum* of the expected dimensions of the moduli spaces of unparametrized pseudoholomorphic planes inside the symplectization $\mathbb{R} \times Y$ that are asymptotic to the Reeb orbit γ at the convex end, where the symplectization has been equipped with a cylindrical almost complex structure. In the case when the first Chern class vanishes, this expected dimension does not depend on the chosen plane, and $|\gamma| \in \mathbb{Z}$.

Note that, in the aforementioned moduli space, we do not identify planes that differ by a translation of the symplectization coordinate. See Section A.1 for more details.

Example 1.2 The expected dimension $|\gamma|$ of a contractible Reeb orbit is at least two for suitable nondegenerate perturbations of the round contact sphere

$$\left(S^{2n+1}, \alpha_{\mathrm{st}} = \frac{1}{2}\sum_i (x_i\, dy_i - y_i\, dx_i)\right) \quad \text{for } n \geq 1,$$

as well as for the high-dimensional "lens spaces" given as the quotients $S^{2n+1}/\mathbb{Z}_k$ for a subgroup $\mathbb{Z}_k \subset S^1$. We will here consider the case $\mathbb{R}P^{2n+1} = S^{2n+1}/\mathbb{Z}_2$; see Proposition 6.4 for the relevant index computation in the case of $\mathbb{R}P^{2n+1}$. The computations for the sphere and lens spaces are analogous.

The assumption $|\gamma| > 1$ for all contractible Reeb orbits allows us to define the Legendrian contact homology algebra without involving the contact homology algebra for the periodic Reeb orbits; see Dimitroglou Rizell [13, Section 3.3.3]. (Note that the contact homology algebra of periodic Reeb orbits has a canonical augmentation in this case.) The assumption $|\gamma| > 1$ also eliminates the need for considerations of the periodic Reeb orbits in the proof of the invariance result from Ekholm [19] for the Legendrian contact homology algebra under a Legendrian isotopy (while fixing the ambient contact form).

Remark 1.3 Before we state our results in detail, we give a quick summary, noting how general the setup is in light of the above technical discussion on the Legendrian contact homology algebra.

- Theorem 1.4 proves lower bounds for the number of Reeb chords between a Legendrian and its image under a contact Hamiltonian in terms of the oscillation norm. This is for an arbitrary Legendrian in an arbitrary contact manifold, but the bounds incorporate the technical condition on $|\gamma|$ mentioned above.
- Theorem 1.6 proves that the Shelukhin–Chekanov–Hofer metric of a Legendrian orbit space is nondegenerate. This is for an arbitrary Legendrian in an arbitrary contact manifold.
- Corollary 1.8 proves that the C^0–limit of a sequence of Legendrians is again Legendrian. Here we assume the contact manifold is geometrically bounded and there exists some lower bound on the length of Reeb chords in the sequence. But there is no assumption on the closed Reeb orbits.
- Theorem 1.9 generalizes the "interlinkedness" of an ordered pair of Legendrians; see Entov and Polterovich [26, Theorem 1.5]. We make several assumptions here: the two Legendrians are individually augmentable; $|\gamma| > 1$ holds for all closed Reeb orbits; and the resulting well-defined Rabinowitz–Floer complex is not acyclic.
- Theorem 1.10 proves the nondisplaceability of the Legendrian equator in standard contact $\mathbb{R}P^{2n+1}$ equipped with a small perturbation of the standard round S^1–invariant contact form. We will show that the $|\gamma| > 1$ assumption is satisfied because the perturbation is small.
- Proposition 1.11 proves roughly that the Legendrian contact homology algebra, below any fixed action level, of an arbitrary Legendrian undergoing a small generic isotopy in an arbitrary contact manifold is invariant under stable-tame isomorphism. Here we assume $|\gamma| > 1$ holds for those closed orbits below this action level bound.
- Usually we assume the contact manifold is closed. However, Theorems 1.4, 1.6 and 1.9 automatically carry over also to open contact manifolds, given that they arise as open subsets of closed contact manifolds, and that the contact form on the open contact manifold is taken to be the restriction of a contact form on the ambient closed manifold. In addition, Proposition 1.11 requires us to be in a setting where the DGA can be defined: aside from closed contact manifolds, cases that can be treated are contactizations and prequantizations of symplectic manifolds with a convex boundary or noncompact end.

Theorem 1.4 (Main Theorem) *Fix a generic closed Legendrian $\Lambda \subset (Y, \alpha)$ of a contact manifold with a fixed contact form. Generically, we can order the Reeb chords by action*

$$0 < \ell(c_1) < \ell(c_2) < \cdots .$$

Also, write $\hbar \in (0, +\infty]$ for the minimal length of a contractible periodic Reeb orbit γ that satisfies $|\gamma| \leq 1$. Suppose that $\mathscr{A}^l(\Lambda)$ with $0 < l \leq \infty$ admits an augmentation to the field $\boldsymbol{k}$, where $l \leq \hbar$. Fix k and consider any compactly supported contact Hamiltonian $H_t : Y \to \mathbb{R}$ such that $\|H_t\|_{\mathrm{osc}} < \min\{l, \ell(c_k)\}$. Then there exist at least

$$\sum_{i=0}^{n} \dim(H_i(\Lambda; \boldsymbol{k})) - 2(k-1)$$

many Reeb chords with one endpoint on Λ and the other endpoint on $\phi^1_{\alpha, H_t}(\Lambda)$.

Remark 1.5 The assumption on the expected dimension of the pseudoholomorphic planes inside the symplectization should not be strictly necessary, but possible to replace by a condition on the existence of an augmentation of the contact homology algebra of periodic Reeb orbits below the action level $l > 0$. However, the setup and analysis become significantly simplified under our stronger assumption. More precisely, without these assumptions, one would have to define the Chekanov–Eliashberg algebra with coefficients in the periodic orbit contact homology algebra in order to define the differential; or, in the case when (Y, α) admits an exact symplectic filling, use anchored disks as in Ekholm [21]. More importantly, our additional hypothesis allows us to avoid the technical gluing and transversality results for pseudoholomorphic planes asymptotic to periodic orbits that require the use of virtual perturbations as done by Pardon [36], Bao and Honda [1], and the polyfold theory of Hofer, Wysocki and Zehnder applied in the SFT setting by Fish and Hofer [27]. With the additional hypothesis $l \leq \hbar$, we only need SFT compactness for closed Reeb orbits (see Bourgeois, Eliashberg, Hofer, Wysocki and Zehnder [3]), and the gluing/transversality results for pseudoholomorphic discs with a single positive Reeb chord asymptotic.

Finally note that, even when replacing the assumption on the expected dimension of the pseudoholomorphic half-planes by the existence of an augmentation, the contact form remains fixed throughout our Legendrian isotopies. Hence, the difficult analytical issues that arise when proving the invariance under deformations of the contact form would not be present even in the more general case.

In [17], we studied this problem when $Y = P \times \mathbb{R}_z$ and $\alpha = dz - \theta$, where $(P, d\theta)$ is an exact symplectic manifold with a certain bounded geometry at infinity. In both cases we used a persistence homology theory (via barcodes) defined by Reeb chords. The main difference between our setup in that article and the typical setup, is that we considered complexes with a finite action filtration.

The new aspects of this article are: we translate known Floer-continuation results to new (but needed) Floer-bifurcations results; we prove an invariance for a newly defined limited-action-window Rabinowitz–Floer homology theory, which allows for chords of zero length to appear.

We now describe several new applications of the Main Theorem, Theorem 1.4.

Fix a subset $N \subset Y$ and let $\mathscr{L}(N)$ be its orbit space under the identity component of the contactomorphism group, $\mathrm{C}_0(Y,\xi)$. Given a contact isotopy ϕ^t and contact form α, let $H_t^{\alpha,\phi}$ be the contact Hamiltonian for ϕ^t. Following Rosen and Zhang's [39, Definition 1.7], define on $\mathscr{L}(N)$ the pseudometric

$$\delta_\alpha(N,N') = \inf\left\{ \int_0^1 \max|H_t|\,dt \;\middle|\; \phi^1_{\alpha,H_t}(N) = N' \right\}.$$

Shelukhin [41] shows that this defines a right-invariant nondegenerate norm on $\mathrm{C}_0(Y,\xi)$. If N is n–dimensional and somewhere not Legendrian (or more generally, if N is not contact coisotropic) then this pseudometric identically vanishes; see [39, Proposition 7.4]. We answer [39, Conjecture 1.10].

Theorem 1.6 *The Shelukhin–Chekanov–Hofer distance δ_α on a contact manifold that is assumed to be either closed, or which admits a codimension-zero contact embedding into a closed contact manifold, and where α moreover is the restriction of a contact form on the closed manifold, is nondegenerate when restricted to closed Legendrian submanifolds.*

Remark 1.7 (1) The condition on admitting an embedding into a closed contact manifold holds, for example, for contactizations and prequantizations of Liouville domains.

(2) The nondegeneracy should hold for more general open contact manifolds which are of bounded geometry. For the proof, the following features are crucial: the constant $\hbar \geq 0$ must be strictly positive for some contact form, and we must be in a setting in which the SFT compactness theorem for pseudoholomorphic curves is satisfied (we need, for example, convexity assumptions at infinity).

(3) We are grateful to Nakamura, who pointed out that it is most likely the case that the nondegeneracy vs degeneracy of the distance depends on the particular choice of contact form in the case when the contact manifold is open.

Two special cases of this were already proved: when the Legendrian is hypertight (no contractible Reeb orbits or chords), by Usher [43], and when the Legendrian is orderable (no positive loops of Legendrians), by Hedicke [28].

Consider a sequence of closed Legendrians Λ_i which C^0–converges to a smooth (not necessarily Legendrian) embedding Λ_∞. Assume that there exists a δ such that each Λ_i has no Reeb chord γ of α–length $\ell(\gamma) < \delta$. Nakamura [33, Theorem 3.4] proves that Λ_∞ is again Legendrian assuming two conditions:

(1) The Reeb vector field R_α is nowhere tangent to Λ_∞, and certain geometric boundedness conditions hold for M.

(2) $(Y,\alpha) = (P \times \mathbb{R}_z, dz - \theta)$.

The only need for the second hypothesis in Nakamura's proof is to use [17] to show the Reeb chords persist under a contact isotopy of small oscillatory norm. Since our Theorem 1.4 generalizes the persistence of Reeb chords proven in [17] to more general contact manifolds, we get an easy corollary.

Corollary 1.8 *Nakamura's convergence result holds only assuming the initial hypotheses: the Reeb chord lengths cannot approach zero, Λ_i is closed, and the ambient contact manifold (M,α) is either closed, or admits a codimension-zero contact embedding into a closed contact manifold.*

Entov and Polterovich define an ordered pair of Legendrians (Λ_0,Λ_1) in (Y,α) to be *interlinked* if for some $\mu, c>0$ and every contact Hamiltonian H satisfying $H\geq c$, there is a Hamiltonian orbit γ from Λ_0 to Λ_1 such that $\ell(\gamma)\leq \mu/c$. Using a persistence homology theory for Reeb chords, they proved if Λ_0 is the 0–section in $(Y=J^1Q,\alpha=dz-p\,dq)$, $\Lambda_1\subset J^1Q$ has an augmentation, and there exists a unique nondegenerate Reeb chord from Λ_0 to Λ_1, then (Λ_0,Λ_1) are interlinked; see Entov and Polterovich [26, Theorem 1.5]. We can generalize this.

Theorem 1.9 *Suppose $\Lambda_0,\Lambda_1\subset(Y,\alpha)$ are a generic pair of Legendrians with generic contact form α for which $|\gamma|\geq 2$ is satisfied for all contractible periodic Reeb orbits γ* (cf Remark 1.5). *Assume that Λ_0 and Λ_1 have augmentations.*

Assume that the Rabinowitz–Floer complex $\mathrm{RFC}^{[0,+\infty)}(\Lambda_0,\Lambda_1)=\mathrm{LCH}(\Lambda_0,\Lambda_1)$, *which is well defined* (*see Section 4.2*), *is not acyclic.* (This is automatically the case for instance when the chords of positive action cannot be partitioned into pairs (c,d) with $\mathfrak{a}(d)>\mathfrak{a}(c)$ and index difference $|d|-|c|=1$.) *Then* (Λ_0,Λ_1) *are interlinked.*

Some hypothesis on the mixed chords is needed. Suppose Λ_0,Λ_1 are both in two Darboux charts in $(J^1Q,dz-p\,dq)$, separated by a large z–distance. Isotope Λ_0 in the (p,q) direction such that the T^*Q–projections of Λ_0 and Λ_1 no longer intersect. During the isotopy we see (for example, via the T^*Q– or $J^0(Y)$–projection) the mixed chords canceling in pairs with strips connecting them of expected dimension 1. In this case neither (Λ_0,Λ_1) nor (Λ_1,Λ_0) are interlinked.

The standard Legendrian $(n+1)$–sphere inside the standard contact sphere is given by

$$\Lambda_0 := S^{2n+1}\cap\mathfrak{Re}\mathbb{C}^{n+1},$$

ie the intersection of the conical Lagrangian in $\mathbb{C}^{n+1}$ which is given by the real-part and the unit sphere. Quotienting by the $\mathbb{Z}_2$–antipodal map $z\mapsto -z\in\mathbb{C}^{n+1}$, $\Lambda_0/\mathbb{Z}_2$ is the standard Legendrian embedding of $\mathbb{R}P^n$ into the standard contact $\mathbb{R}P^{2n+1}=S^{2n+1}/\mathbb{Z}_2$. It is easy to show that Λ_0 can be displaced from itself in S^{2n+1}. By computing a Rabinowitz–Floer theory (Definition 4.6), we prove this is not true for this real projective plane.

Theorem 1.10 *Consider the standard contact $\mathbb{R}P^{2n+1}$ equipped with a small nondegenerate perturbation of the standard round S^1–invariant contact form as described in Proposition 6.12. The standard Legendrian $\mathbb{R}P^n=\Lambda_0/\mathbb{Z}_2$ has a Chekanov–Eliashberg algebra that admits an augmentation, and a well-defined and invariant Rabinowitz–Floer homology that is equal to*

$$\mathrm{RFH}(\mathbb{R}P^n,\mathbb{R}P^n)=\bigoplus_{i\in\mathbb{Z}}\mathbb{Z}_2[i].$$

In particular, the standard Legendrian $\mathbb{R}P^n$ cannot be displaced from itself for these contact forms.

There are two standard methods to prove the invariance of holomorphic-curve-based theories, such as the one underlying Theorem 1.4.

Given a generic Legendrian isotopy $\{\Lambda_t\}_{t\in[-,+]}$ between two generic Legendrian submanifolds Λ_- and Λ_+, there is a filtered unital DGA–morphism

$$\Phi\colon \mathscr{A}(\Lambda_+) \to \mathscr{A}(\Lambda_-)$$

given by counting certain holomorphic disks in the symplectization $\mathbb{R}\times Y$ with boundary lying on an exact Lagrangian constructed from the trace of the isotopy with Λ_+ (resp. Λ_-) at the positive (resp. negative) end. See Chantraine [4], Chantraine, Colin and Dimitroglou Rizell [5] and Ekholm, Honda and Kálmán [25] for details on these Lagrangian trace constructions. The filtration is defined by the length. The map Φ has a DGA–homotopy inverse, and so induces an isomorphism on homology; see Ekholm [19]. This invariance approach is known as the *continuation method*.

The *bifurcation method* studies how the DGA $\mathscr{A}(\Lambda_t)$ varies with $t\in[-,+]$. Generically, there are two events that occur at isolated t: a pair of Reeb chords can appear or disappear (*birth/deaths*); an isolated holomorphic curve can appear (*handleslide disk*), and via one-parameter Gromov compactness, can change the moduli spaces which define the differential. Each birth/death induces on the DGA a stabilization/destabilization combined with possible filtered tame automorphisms, while each handleslide disk induces a filtered tame automorphism; see Ekholm, Etnyre and Sullivan [23] and Chekanov [8]. So the isotopy overall induces a sequence of (filtered) *stable-tame isomorphisms*.

A stable-tame isomorphism is conjecturally stronger than a homotopically invertible DGA–morphism. For example, when making the analogous comparison for Morse chain complex invariance, the bifurcation method preserves the stable–Morse number and certain torsion-based invariances (see Damian [12] and Sullivan [42]), which the continuation method, a priori, does not. However, due to sensitive gluing analysis, for studying Chekanov–Eliashberg invariance the bifurcation method has only been proved if $(Y,\alpha) = (P\times\mathbb{R}_z, dz-\theta)$; see again [23; 8]. This (and the proof of Theorem 1.4) motivates Proposition 1.11. The statement is somewhat unwieldy because a general contact manifold may have infinite Reeb chords with actions arbitrarily close, while we can only analyze events involving finite numbers of birth/deaths and handleslides. However, it can be roughly formulated in the following (imprecise) way: the Chekanov–Eliashberg algebra in a general contact manifold is invariant under stable-tame isomorphism below any fixed action level.

Proposition 1.11 (bifurcation analysis for concordance maps) *Let $\{\Lambda_t\}_{t\in[-,+]}$ be a generic Legendrian isotopy between two generic Legendrian submanifolds Λ_- and Λ_+. Denote by $\Phi_{[a,b]}$ the unital DGA–morphism induced by the Lagrangian trace of the isotopy $\{\Lambda_t\}_{t\in[a,b]}$ with $[a,b]\subset[-,+]$. Recall that this is an exact Lagrangian cobordism diffeomorphic to a cylinder; see eg Appendix B for its construction.*

Fix some number $l>0$ and assume that all birth/deaths in the Legendrian isotopy Λ_t are generic, so that none occur precisely at action l, while there are finitely many that occur at action strictly less than l at distinct times.

For any sufficiently fine generic subdivision

$$-=t_1<t_2<\cdots<t_N=+$$

for which each (t_i,t_{i+1}) contains at most one birth/death below action l, the following holds. The restricted DGA–morphism $\Phi_{[t_i,t_{i+1}]}|_{\mathscr{A}^l(\Lambda_{t_{i+1}})}$ can be conjugated to the algebra map that is defined by mapping to zero any generator involved in a death moment and canonically identifying all remaining Reeb chord generators, where the conjugation is by filtered tame automorphisms of the domain and target. This means, in particular, that $\Phi_{[t_i,t_{i+1}]}|_{\mathscr{A}^l(\Lambda_{t_{i+1}})}$ is a stable-tame isomorphism of DGAs.

Remark 1.12 Of course the sub-DGA $\mathscr{A}^l$ is itself not invariant; the restriction $\Phi_{[t_i,t_{i+1}]}|_{\mathscr{A}^l(\Lambda_{t_{i+1}})}$ is, for example, not necessarily contained inside $\mathscr{A}^l(\Lambda_{t_i})$.

Remark 1.13 Since we do not discuss the virtual perturbation schemes for defining the contact homology algebra for periodic Reeb orbits, we again need to assume that $|\gamma|\geq 2$ holds for any contractible Reeb orbit of length at most l in the above proposition; cf Remark 1.5.

In [17], we exploited the bifurcation-type invariance in Ekholm, Etnyre and Sullivan [23; 24] to show that the barcode from persistent homology induced by the action filtration is continuous with respect to the oscillatory norm in the case of contactizations. The invariance by DG–homotopies from [19] that holds in a general contact manifold also satisfies filtration-preserving properties, but these are a priori not continuously depending on the oscillation alone; see the notion of length introduced in Sabloff and Traynor [40]. Proposition 1.11 allows us to reprove the results from [17] in the general setting.

In Section 2 we review some algebra and combinatorics of DGAs, mapping cones and barcodes. In Section 3, we prove Proposition 1.11. In Section 4, we introduce a Rabinowitz–Floer complex generated by Reeb chords between two Legendrians, and study how a certain mapping cone of this complex changes when one of the Legendrians isotopes (possibly through the other one). This version of the complex was previously defined by Legout [31] in the case of a contactization, and is also related to the Floer homology for Lagrangian cobordisms defined in [6]. In Section 5, we use the changing barcodes of these mapping cone complexes to prove Theorems 1.4, 1.6 and 1.9. In Section 6, we compute the example of Theorem 1.10; see Proposition 6.12.

Remark 1.14 We learned that Oh [34] has posted before us, by a couple of weeks, a related result: if the Hamiltonian oscillation is less than the length of the shortest Reeb chord between, then the number of Reeb chords is bounded below by the sum of the Betti numbers. This improves one of our earlier results [16], in which we require the upper bound on the oscillation to be less than the length of the shortest Reeb chord, multiplied by the conformal factor of the contact form. Whereas our prior results [16; 17] and this paper use versions of Floer theory which have already been established by others, Oh's approach is different in that he establishes the analytical framework for a new theory called Hamiltonian perturbed contact-instantons. With this new theory established, his result follows from Chekanov's original argument for Lagrangians [7].

Acknowledgements Dimitroglou Rizell is supported by the Knut and Alice Wallenberg Foundation under the grant KAW 2016.0198, and by the Swedish Research Council under the grant 2020-04426. Sullivan is supported by the Simons Foundation grant 708337. The authors thank input from Yasin Karacan, Lukas Nakamura, Stefan Nemirovski, and especially the referee.

2 Algebraic preliminaries

This section is purely algebraic, with no mention of the geometric applications.

2.1 Filtered DGAs, augmentations and stable tame isomorphisms

Let $(\mathscr{A}, \partial)$ be either a noncommutative or a graded-commutative semifree unital DGA over the ground field $\boldsymbol{k}$. Assume $(\mathscr{A}, \partial)$ has an action filtration $\ell\colon \mathscr{A} \to \{-\infty\} \cup \mathbb{R}$, where $\ell(\partial(x)) < \ell(x)$; we say that ∂ is (strictly) filtration-decreasing, or action-decreasing. See [17] for more details. Suppose $\mathscr{A}$ admits a $\mathbb{Z}_2$–graded augmentation $\varepsilon\colon \mathscr{A} \to \boldsymbol{k}$. There is an induced filtration-preserving unital algebra-automorphism $\Psi_\varepsilon\colon \mathscr{A} \to \mathscr{A}$ defined by

$$\Psi_\varepsilon(a) = a - \varepsilon(a)$$

on the generators, whose inverse is defined by

$$\Psi_\varepsilon^{-1}(a) = a + \varepsilon(a).$$

In particular, the inverse is also filtration-preserving, and these automorphisms conjugate the differential to

$$\partial^\varepsilon := \Psi_\varepsilon \circ \partial \circ \Psi_\varepsilon^{-1},$$

which preserves both word-length and action. In particular, if we write $\partial^\varepsilon(a) = \sum_{i=0}(\partial^\varepsilon(a))_i$ as a sum of monomials, then $(\partial^\varepsilon)_1$ defines a strictly filtration-decreasing differential on the graded vector space generated by the DGA generators, which is of degree-(-1).

A filtered tame automorphism $\Phi\colon (\mathscr{A}, \partial) \to (\mathscr{A}, \partial')$ of a semifree DGA $\mathscr{A}$ with preferred basis is defined on the generators by

$$\Phi(x) = kx + \delta_x^y w,$$

where $k \in \boldsymbol{k}$ is a unit, δ_x^y is the Kronecker-delta, and $w \in \mathscr{A}$ is a word such that $\ell(w) < \ell(y)$.

A canonical identification between semifree filtered DGAs with preferred bases is a DGA isomorphism induced by an identification of the generators which preserves the grading and differential, but not necessarily the filtration.

The stabilization $(\mathscr{B}, \partial_{\mathscr{B}})$ of $(\mathscr{A}, \partial_{\mathscr{A}})$ is constructed by adding to $\mathscr{A}$ two generators x, y such that $\partial_{\mathscr{B}}(y) = x$ and $\partial_{\mathscr{B}}|_{\mathscr{A}} = \partial_{\mathscr{A}}$. Note that there is a canonical DGA inclusion $\mathscr{A} \to \mathscr{B}$ and DGA quotient $\mathscr{B} \to \mathscr{A}$.

A (filtered) *stable-tame isomorphism* (STI) $\Phi\colon (\mathscr{A}, \partial_{\mathscr{A}}) \to (\mathscr{B}, \partial_{\mathscr{B}})$ is a finite composition of (filtered) tame automorphisms, (possibly nonfiltered) canonical identifications, stabilizations, and inverse stabilizations.

Let $\Phi\colon (\mathscr{A},\partial_{\mathscr{A}})\to(\mathscr{B},\partial_{\mathscr{B}})$ be a DGA–morphism. For any augmentation ε of $\mathscr{B}$ we can define the conjugation

$$\Phi^{\varepsilon} := \Psi_{\varepsilon}\circ\Phi\circ\Psi_{\varepsilon\circ\Phi}^{-1}\colon (\mathscr{A},\partial_{\mathscr{A}}^{\varepsilon\circ\Phi})\to(\mathscr{B},\partial_{\mathscr{B}}^{\varepsilon}),$$

which satisfies

$$(\Phi^{\varepsilon})_1(\partial_{\mathscr{A}}^{\varepsilon\circ\Phi})_1 = (\partial_{\mathscr{B}}^{\varepsilon})_1(\Phi^{\varepsilon})_1.$$

If Φ is a (filtered) STI, then $(\Phi^{\varepsilon})_1$ is a finite composition of (filtered) handleslides, and birth/deaths at the chain level; see Section 2.4.

2.2 Mapping cones with filtrations and invariance

Let (C_{01},d_{01}) and (C_{10},d_{10}) be filtered graded complexes with action filtrations $\ell\colon C\to\mathbb{R}\cup\{-\infty\}$, and strictly filtration-decreasing differentials. (The grading subscript is suppressed while the "01" and "10" subscripts will be justified in Section 4.) Moreover, assume that the generators of C_{01} all have actions above some fixed $\gamma\in\mathbb{R}$ while the generators of C_{10} all have actions below γ. We write

$$C^{<a} := \ell^{-1}(-\infty,a)\subset C$$

for the subcomplex consisting of chains of action less than $a\in\mathbb{R}$, and denote the quotient complex consisting of chains in the action window $[a,b)$ by $C^{[a,b)} := C^{<b}/C^{<a}$.

Let $B\colon C_{01}\to C_{10}$ be a chain map, which automatically is strictly filtration-decreasing by the above assumptions. For this reason, we get an induced action filtration on the chain complex given by the mapping cone $(\mathrm{Cone}(B),d_B)$, which we represent by

$$\left(C_{10}\oplus C_{01},\partial_{\mathrm{Cone}}=\begin{pmatrix}-\partial_{10} & B\\ 0 & \partial_{01}\end{pmatrix}\right),$$

ie ∂_{Cone} is again strictly filtration-decreasing.

Suppose $B'\colon C'_{01}\to C'_{10}$ is another chain map between filtered complexes. Again we assume that all generators of the domain have action greater than the action of the generators in the target, which means that B' is automatically strictly filtration-decreasing. A map $f\colon C\to C'$ between filtered chain complexes with filtrations ℓ,ℓ' is said to have *degree* $\epsilon\in\mathbb{R}$ if $\ell'(f(x))\le\ell(x)+\epsilon$ for all $x\in C$. Assume the following:

(A1) There exist chain maps $\phi_{01}\colon C_{01}\to C'_{01}$ and $\psi_{10}\colon C'_{10}\to C_{10}$ with homotopy inverses ψ_{01} and ϕ_{10}, with chain homotopies $h_{ij}\colon\psi_{ij}\circ\phi_{ij}\sim\mathrm{id}_{C_{ij}}$ and $k_{ij}\colon\phi_{ij}\circ\psi_{ij}\sim\mathrm{id}_{C'_{ij}}$, where all maps above are of degree ϵ.

(A2) The map B is chain homotopic to $\psi_{10}B'\phi_{01}$ via a chain homotopy $H\colon B\sim\psi_{10}B'\phi_{01}$. Note that this homotopy automatically has negative degree, ie it is strictly filtration-decreasing.

(Note in (A1) that the homotopies have degree ϵ instead of 2ϵ as is common in Hamiltonian Floer theory literature, for example [44, Definition 8.1].)

Then we have a homotopy commutative square

$$\begin{array}{ccc} C_{01} & \xrightarrow{B} & C_{10} \\ {\scriptstyle\phi_{01}}\downarrow & \swarrow^{h} & \downarrow{\scriptstyle\phi_{10}} \\ C'_{01} & \xrightarrow[B']{} & C'_{10} \end{array}$$

where $h = \phi_{10}H + k_{10}B'\phi_{01}$ is the induced chain homotopy between $\phi_{10}B$ and $B'\phi_{01}$, which thus is a map of degree ϵ. Similarly, there is a chain homotopy $h' : B\psi_{01} \sim \psi_{10}B'$ given by $h' = H\psi_{01} + \psi_{01}B'k_{01}$, which makes the square in the diagram

$$\begin{array}{ccc} C'_{01} & \xrightarrow{B'} & C'_{10} \\ {\scriptstyle\psi_{01}}\downarrow & \swarrow^{h'} & \downarrow{\scriptstyle\psi_{10}} \\ C_{01} & \xrightarrow[B]{} & C_{10} \end{array}$$

commute up to homotopy. Both h and h' are maps of degree ϵ.

It follows that there are induced chain maps of the cones

$$\begin{array}{ccccc} C_{10} & \longrightarrow & \mathrm{Cone}(B) & \longrightarrow & C_{01} \\ \downarrow & & {\scriptstyle f}\downarrow & & \downarrow \\ C'_{10} & \longrightarrow & \mathrm{Cone}(B') & \longrightarrow & C'_{01} \\ \downarrow & & {\scriptstyle g}\downarrow & & \downarrow \\ C_{10} & \longrightarrow & \mathrm{Cone}(B) & \longrightarrow & C_{01} \end{array}$$

given by

$$f = \begin{pmatrix} \phi_{10} & h \\ 0 & \phi_{01} \end{pmatrix} \quad \text{and} \quad g = \begin{pmatrix} \psi_{10} & h' \\ 0 & \psi_{01} \end{pmatrix}$$

of degree ϵ. By the five lemma, the induced homology maps are isomorphisms.

Lemma 2.1 *The maps $f \circ g$ and $g \circ f$ are each chain homotopic to automorphisms of filtered chain complexes* (ie degree-zero chain maps with degree-zero inverses), *via chain homotopies that are of degree ϵ.*

Proof We show the statement for $g \circ f$; the argument for $f \circ g$ is completely analogous.

There is a chain homotopy

$$\begin{pmatrix} h_{10} & 0 \\ 0 & h_{01} \end{pmatrix}$$

of degree ϵ from $f \circ g$ to a chain automorphism of the form

$$\begin{pmatrix} \mathrm{id}_{C_{10}} & K \\ 0 & \mathrm{id}_{C_{01}} \end{pmatrix}.$$

Since the entry $K\colon C_{01} \to C_{10}$ is of nonpositive degree, the matrix is of degree zero. The matrix is a chain map since it is chain homotopic to $f \circ g$. This chain map property translates to the fact that K is a chain map (that performs a grading shift by $+1$). Note that the inverse map is

$$\begin{pmatrix} \mathrm{id}_{C_{10}} & -K \\ 0 & \mathrm{id}_{C_{01}} \end{pmatrix},$$

which hence is also of degree zero. □

2.3 Simple equivalences from small degree homotopy equivalences

Next we relate the homotopy equivalences of small degree as above with the invariance of bifurcation-type that our previous work [17] was based on.

Recall that a basis $\{\boldsymbol{e}_i\}$ of the filtered complex C is *compatible* with the filtration if there is an action function $\ell(\boldsymbol{e}_i) \in \mathbb{R}$ defined on the basis so that $c \in C^{<a}$ if and only if c can be written as a sum of basis elements of action less than a; see [17, Section 2.1]. We say that a complex is *δ–gapped* if two different basis elements in a compatible basis have action values that either coincide, or differ by at least $\delta > 0$.

Lemma 2.2 *Consider two filtered complexes C and C', where C is δ–gapped and satisfies the property that each action level has at most finitely many generators, and for which*

(1) *C and C' admit bases compatible with the filtration whose elements are in a bijection $x \mapsto x'$, under which the action satisfies $\ell(x) - \ell'(x') \leq \epsilon$; and*

(2) *there are chain maps $\phi\colon C \to C'$ and $\psi\colon C' \to C$, which both are of degree $\epsilon > 0$, where $\psi\phi$ and $\phi\psi$ are homotopic to automorphisms of filtered chain complexes via chain homotopies of degree ϵ.*

If $\delta > 4\epsilon > 0$, then ϕ is a isomorphism with inverse ψ.

If we endow the complexes with bases that are compatible with the filtrations, ordered in decreasing action, with the additional assumption that two different elements have distinct action values, then ϕ is upper triangular with units on the diagonal.

Proof By the assumptions, we can write

$$\psi\phi = \Phi + \partial K + K\partial,$$

where Φ is an automorphism of filtered chain complexes. By the assumption that C is δ–gapped, and since K is of degree $\epsilon < \frac{1}{4}\delta$, we conclude that K is in fact filtration preserving. Hence $\partial K + K\partial$ is strictly filtration-decreasing. It follows by a standard fact that $\psi\phi$ itself is an automorphism of filtered chain complexes.

The map ϕ is injective by the above. It thus suffices to show that ϕ is surjective since, in that case, $\psi = \phi^{-1}$.

Note that, for any basis element $x \in C$ in a basis compatible with the filtration, the map induced by quotient and restriction

$$\phi\colon C^{[\ell(x)-3\epsilon,\ell(x)+3\epsilon)} \to C'^{[\ell(x)-3\epsilon,\ell(x)+3\epsilon)}$$

is a map between equidimensional vector spaces by the assumption on the action spectrum of the involved complexes. Since $\psi\phi$ is an automorphism of filtered complexes, together with the assumption that C is δ–gapped, we deduce that the above map on the quotient is injective as well, and thus surjective. Consequently the map ϕ itself is surjective, as sought. □

Lemma 2.3 (characterization of a birth/death) *Consider two filtered complexes C and C', and assume that $C^{[a+\delta,a+3\delta)} = C^{[a,a+4\delta)}$ are both two-dimensional, while $C'^{[a,a+4\delta)} = 0$. Assume that there exist chain maps $\phi\colon C \to C'$ and $\psi : C' \to C$ which both are of degree $\epsilon > 0$, where $\psi\phi$ and $\phi\psi$ are homotopic to automorphisms of filtered chains complexes via chain homotopies of degree ϵ. If $\delta > \epsilon > 0$, then $C^{[a+\delta,a+3\delta)}$ is a complex generated by x and y, with $\partial x = ky$ for some unit k.*

Proof Since the map $\phi\colon C^{[a,a+4\delta)} \to C'^{[a,a+4\delta)} = 0$ induced by quotient and surjection is a homotopy equivalence, $C^{[a,a+4\delta)}$ is acyclic. The only two-dimensional acyclic complexes are the ones described above. □

2.4 Barcodes

We sketch without details the modified barcode theory done in [17, Section 2].

For $t \in \mathbb{R}$, let $C(t)_{a_t}^{b_t}$ be a filtered complex with filtration action ℓ taking values in $[a_t, b_t) \subset \mathbb{R}$ and $-\infty$. A *piecewise continuous* (PWC) *family* of such filtered complexes, parametrized by t, is characterized by the following properties:

- The endpoints of the action window $[a_t, b_t)$ vary continuously with t.
- There exists a discrete set of $t_1 < \cdots < t_N$ such that during any component $I \subset \mathbb{R}\setminus\{t_1, \dots, t_N\}$, there are canonically identified (for different $t \in I$) generators of the complexes which are compatible with the action.
- The action of each such generator is continuous and almost everywhere differentiable with respect to $t \in I$.
- For each $t \in I$, the differential strictly decreases the action.
- For each $t \in \{t_1, \dots, t_N\}$, the chain complex undergoes at most one of the following possible "simple bifurcations":
  - The algebraic equivalent of a Morse handleslide.
  - The algebraic equivalent of a Morse birth/death.
  - An entrance (resp. exit) of one generator into (resp. from) either the top or bottom of the action window.

We continue to use [17, Section 2] in defining barcodes, but for the equivalent definition based on normal forms, see [37, Section 2.1] and [44, Definition 6.2]. A barcode is a finite collection of "bars" $[s,e)$, where the endpoint e might be $+\infty$. Let $\phi_{c_0,c_1}: H(C(t)_{a_t}^{c_0}) \to H(C(t)_{a_t}^{c_1})$ be induced from the inclusion $C(t)_{a_t}^{c_0} \hookrightarrow C_*(t)_{a_t}^{c_1}$, where $a_t \le c_0 \le c_1 \le b_t$. The *barcode of the complex* $(C(t)_{a_t}^{b_t}, \partial_t)$ is the barcode uniquely characterized by the following properties:

- The number of bars with starting point s is equal to the dimension of the quotient
$$\operatorname{coker}(\phi_{s,s+\epsilon}) = H(C(t)_{a_t}^{s+\epsilon}, \partial_t)/\operatorname{im}\phi_{s,s+\epsilon},$$
where $\epsilon > 0$ is any sufficiently small number.
- The number of bars with starting point s that persist at action level $l \ge s$ is equal to the dimension of the subspace
$$[\phi_{s+\epsilon,l+\epsilon}](\operatorname{coker}(\phi_{s,s+\epsilon})) \subset H(C(t)_{a_t}^{l+\epsilon}, \partial_t)/\operatorname{im}\phi_{s,l+\epsilon},$$
where $\epsilon > 0$ is any sufficiently small number and where the map
$$[\phi_{s+\epsilon,l+\epsilon}]\colon \operatorname{coker}(\phi_{s,s+\epsilon}) \to H(C(t)_{a_t}^{l+\epsilon}, \partial_t)/\operatorname{im}\phi_{s,l+\epsilon}$$
is induced by descending $\phi_{s+\epsilon,l+\epsilon}$ to the quotients.

Proposition 2.4 [17, Proposition 2.7] *Consider a PWC family $C(t)_{a_t}^{b_t}$ of filtered complexes with action window. When the complex undergoes no such bifurcation, the barcode undergoes a continuous change of action levels for its starting and endpoints. At the bifurcations the barcode undergoes the following corresponding changes*:

- ***Handle-slide*** *The barcode is unaffected.*
- ***Birth/death*** *When two generators x, y undergo a birth/death, then a bar connecting $\ell(x)$ to $\ell(y)$ is added to/removed from the barcode. (The bar is not present at the exact time of the birth/death, but immediately after/before it is visible and of arbitrarily short length.)*
- ***Exit below*** *A generator slides below the action level a_t at time t. If the uniquely determined bar which starts at the action level of that generator is of finite length, then that bar gets replaced with a bar of infinite length whose starting point is located at the same action level as the endpoint of the original bar. If the bar has infinite length, then it simply disappears from the barcode.*
- ***Entry below*** *This is the same as an exit below, but in backwards time.*
- ***Exit above*** *A generator slides beyond the action level b_t at time t. There is a uniquely determined bar which either ends or starts at the action level of that generator. In the first case, the bar gets replaced with one that has the same starting point but which is of infinite length. In the second case, when the bar necessarily is infinite, then that bar simply disappears from the barcode.*
- ***Entry above*** *This is the same as an exit above, but in backwards time.*

2.5 A piecewise continuous family of complexes from small-degree homotopy equivalences

In order to investigate the continuous dependence of the barcodes in relation to the invariance properties established in Section 2.2, we need to relate invariance under small-degree homotopy equivalence as in Lemma 2.1 to the bifurcation-type invariance that Proposition 2.4 above is based upon.

Let $C(t)$, $t \in [0,1]$, be a family of finite-dimensional filtered complexes with choices of compatible basis elements. We assume that all basis elements vary smoothly with t except that there are finitely many times $t_1 < \cdots < t_N$ at which there is a unique birth/death moment. Roughly speaking, at these moments precisely two basis elements of the same action either appear or disappear. The next paragraph gives the precise characterization of a birth/death.

Since we require the differential to be strictly filtration-decreasing, the two basis elements that undergo a birth/death at t_i are necessarily missing from the complex $C(t_i)$. However, in the case of a birth (resp. death) two generators for $t > t_i$ (resp. $t < t_i$) are assumed to have actions that extend continuously to $t = t_i$, such that the extensions moreover attain the same action value at $t = t_i$. (Note that, in particular, the action of any basis element is bounded in the family, and there is a global bound on the dimension of the complexes $C(t)$ in the family.) For simplicity we make the additional assumptions that, at each birth/death moment $t = t_i$, all elements in the compatible basis have distinct action values and, moreover, their action values differ from the action of the (continuous extension of the) birth/death pair. In addition, we assume that there is a finite set of times when the action values of a compatible basis are not distinct.

In order to simplify the notation, we will now assume that the finitely many times when the action spectrum is not injective, as well as the birth/death moments, all occur at rational times $t \in \mathbb{Q} \cap [0,1]$.

It is worth stressing that, at this moment, we have not yet made any assumptions on how the differential of the complexes $C(t)$ varies; we are simply prescribing how their compatible bases depend on t. Under the further assumptions of the next result, Proposition 2.5, we establish an invariance result for this family of complexes; this is in fact one of the main goals of this section.

Proposition 2.5 *Let $C(t)$ be a family of complexes as above that satisfies the following additional requirement. For all $\epsilon > 0$, all sufficiently large $N \gg 0$ and all $i \in \{0, \dots, N-1\}$, there are chain maps*

$$\phi_i : C(i/N) \to C((i+1)/N) \text{ and } \psi_i : C((i+1)/N) \to C(i/N)$$

of degree $\epsilon > 0$, where $\psi_i \phi_i$ and $\phi_i \psi_i$ are homotopic to automorphisms of filtered chain complexes via chain homotopies of degree ϵ. Then, for $N \gg 0$ is sufficiently large, there exists a piecewise continuous family of complexes $D(t)$ that admit isomorphisms $C(t) \cong D(t)$ of filtered vector spaces, that moreover are chain maps for all $t = i/N$, where $i = 0, 1 \dots, N$.

Proof We start by prescribing

$$D(i/N) := C(i/N)$$

for all $i = 0, 1 \dots, N$.

First we consider the special case when $C(t)$ has no birth/deaths, and the action values of the basis elements are all distinct for all times. We then construct the PWC family as follows. The complexes $D(t)$ for $t \in [i/N, (i+1)/N)$ are constructed by setting $D(t) := D(i/N)$ as a complex, and then simply letting the action values of a compatible basis vary accordingly with $t \in [i/N, (i+1)/N)$. (In particular, the differential remains unchanged.) It is immediate that $D(t) \cong C(t)$ holds on the level of *filtered vector spaces*.

The family of complexes $D(t)$ for $t \in [i/N, (i+1)/N)$ extends by continuity to also $t = (i+1)/N$; denote the limit filtered complex by $\widetilde{D}((i+1)/N)$. What remains is to construct an isomorphism of filtered complexes $\widetilde{D}((i+1)/N) \cong D((i+1)/N)$.

The assumptions of Lemma 2.2 are satisfied for all maps ϕ_i and ψ_i whenever $N \gg 0$. In particular, all complexes $D(i/N)$ can be assumed to be δ–gapped for some fixed $\delta > 0$. Hence $\phi_i \colon D(i/N) \to D((i+1)/N)$ is a chain isomorphism of degree ϵ. Since $\widetilde{D}((i+1)/N)$ is canonically identified with $D(i/N)$ (only the action values of the compatible basis have changed slightly), it follows that the induced chain isomorphism $\phi_i \colon \widetilde{D}((i+1)/N) \to D((i+1)/N)$ is of degree zero. This implies that we have a PWC family, as sought.

In the case when the family $C(t)$ has birth/deaths or compatible basis elements of the same action value, then we need to take care at those moments separately. This we do in the subsequent paragraph. After having constructed the PWC family in a small neighborhood of these points in time, the family for the remaining times can be constructed as above.

In the case when two action values for a compatible basis coincide at some $t = i/N$, then we can again use Lemma 2.2 as above to construct the family $D(t)$ for $t \in [i/N, (i+1)/N]$ and $[(i-1)/N, i/N]$. In the case when there is a birth/death at $t = i/N$, the same can be done by alluding to Lemma 2.3. Once we have taken care of the construction of $D(t)$ for these times, we simply invoke the construction in the first simple case, ie the case when action values are distinct, and when there are no birth/death moments. □

3 Proof of Proposition 1.11

Consider, in the symplectization $(\mathbb{R}_\tau \times Y, d(e^\tau \alpha))$, the *Lagrangian trace* of a Legendrian isotopy Λ_t with t ranging from $-$ to $+$. This (exact embedded) Lagrangian concordance coincides with the cylinder $\mathbb{R} \times \Lambda_\pm$ for $\pm\tau \geq k$ for some $k \gg 0$, and after reparametrizing the Legendrian isotopy, the τ–level set of the trace is close to $\{\tau\} \times \Lambda_\tau$ For example, see [4, Theorem 1.2], the proof of [19, Lemma A.1], the proof of [25, Lemma 6.1], or [35, Definition 2.10] on constructing this trace. We recall the version of the construction from Chantraine [4, Theorem 1.2] in Appendix B.

The *length* $\delta \in \mathbb{R}_{\geq 0}$ of the cobordism, as defined by Sabloff and Traynor in [40], is the shortest δ such that $\{\tau \in [\tau_0, \tau_0 + \delta]\} \subset \mathbb{R} \times Y$ contains the noncylindrical portions of the cobordism and almost complex structure. Proposition B.1 shows how the length of the trace cobordism constructed by [4] depends on

the conformal factor of the contact isotopy; we also consider the length of the "inverse cobordism". The length is crucial for analyzing the filtered invariance result, since the chain maps produced by the trace cobordism have filtration properties that depend on this.

We denote by x^t a continuous family of chords of Λ_t. Recall that the family of Legendrians is assumed to be generic, which means that a chord $x^\pm$ of $\Lambda_\pm$ that does not undergo a death-type bifurcation in the family corresponds to a unique chord $x^\mp$ of $\Lambda_\mp$.

The Lagrangian together with an appropriately compatible almost complex structure J induce a DGA–morphism

$$\Phi\colon (\mathscr{A}_+, \partial_+) \to (\mathscr{A}_-, \partial_-)$$

between the DGAs $(\mathscr{A}_\pm, \partial_\pm)$ of $\Lambda_\pm$; see [19]. More precisely, we require as in [19] that the almost complex structure is *adjusted*, which means that it is compatible with the symplectic form and *cylindrical* outside of a compact subset; the latter means that the almost complex structure preserves the contact planes lifted to $\mathbb{R}_\tau \times Y$, is invariant under the $\mathbb{R}_\tau$–translation, and sends ∂_τ to the lifted Reeb flow.

Suppose the Lagrangian trace of an isotopy Λ_t induces the map Φ and the inverse trace induces the map Ψ. Construct a generic 1–family of Lagrangians connecting the trivial cobordism $\Lambda_+ \times \mathbb{R}$, with induced DGA map $\mathrm{id}\colon (\mathscr{A}_+, \partial_+) \to (\mathscr{A}_+, \partial_+)$, to the trace concatenated with its inverse. Let G count index -1 punctured pseudoholomorphic curves that occur at isolated moments in this family of cobordism, as defined in [19]. Then

$$\mathrm{id} = \Psi\Phi - (\partial_+ G - G\partial_+). \tag{3-1}$$

The result below follows from the filtration-preserving properties of the DGA–morphisms and chain homotopy involved.

Lemma 3.1 (1) *Consider the Legendrian isotopy Λ_t generated by a time-dependent contact Hamiltonian H_t. For any $\delta > 0$ the Legendrian isotopies*

$$\{\Lambda_t\}_{t\in[i/N,(i+1)/N]} \quad \text{and} \quad \{\Lambda_{-t}\}_{t\in[-(i+1)/N,-i/N]}$$

may all be assumed to have a trace cobordism of length less than δ whenever $N \gg 0$. In addition, both concatenations of these two trace cobordisms may be assumed to be compactly supported Hamiltonian isotopic to the trivial cylinder through cobordisms of length at most 2δ.

(2) *Let Φ and Ψ above be induced by Lagrangian cobordisms of length $\delta > 0$, where the concatenation of the cobordisms are Hamiltonian isotopic to trivial cobordism through cobordisms of length at most 2δ.* (For instance the trace cobordisms from part (1).) *Then*

$$\ell(F(x^+)) < \ell(x^+)e^{2\delta}$$

holds for $F \in \{\Psi, \Phi\}$, while any word $\boldsymbol{x}^+$ that consists of letters that satisfy $\ell(x^+) \le a$ has the property that $G(\boldsymbol{x}^+)$ consists of words of letters that all satisfy $\ell(x^-) \le e^{2\delta}a$. (Notation as in equation (3-1).)

Proof (1) We start by fixing a global contact isotopy that generates the Legendrian isotopy Λ_t. For $N \gg 0$, we may assume by continuity that the restriction of the contact isotopy that generates $\{\Lambda_t\}_{t\in[i/N,(i+1)/N]}$ has a conformal factor that is bounded by $\delta > 0$. The construction of the trace cobordisms with the sought properties can now be deduced from Proposition B.1.

(2) The statement is clear for any map F that is defined by a count of finite-energy pseudoholomorphic disks in $\mathbb{R}_\tau \times Y$ with boundary on a Lagrangian cobordism of length at most 2δ and a single positive puncture, when the almost complex structure is adjusted and, moreover, cylindrical where the cobordism is cylindrical (eg outside of $[-\delta,\delta]\times Y$). Namely, [6, Lemma 3.4 and Proposition 3.5(9)] explicitly bound the Hofer energy of such curves used to define $F(x^+)$ from below by 0 and from above by the quantity $\ell(x^+)e^{\delta} - \ell(F(x^+))e^{-\delta}$. To match with their convention, we ignore their distinction of pure and mixed chords, and we center the concordance around $\tau_0 = 0$ (from our definition of length above).

Since the chain maps $F = \Phi, \Psi$ are defined by pseudoholomorphic disks of the type mentioned above, the statement now directly follows in these cases.

The chain homotopy G has a more complicated construction, which was carried out in [19, Appendix B]. Each term in $G(x_1^+ \cdots x_k^+)$ corresponds to a count of disconnected pseudoholomorphic *buildings* [3], where each component of the building has the topological type of a broken disk with a single positive puncture at x_i^+ for $i = 1,\dots,k$. In addition, each component satisfies the following:

- There is a single level consisting of a number of pseudoholomorphic disks of index -1 and 0 inside $\mathbb{R}\times Y$, each with boundary on one of the Lagrangian cobordisms in the family that interpolates between the concatenation and the trivial cylinder (these are all of length at most 2δ), and which all have a single positive puncture. As in the first case, we can again assume that the almost complex structure is cylindrical in the subset where the family of cobordisms are cylindrical (eg outside of $[-\delta,\delta]\times Y$).
- All other levels consist of punctured disks of index 1 and trivial strips of index 0 inside $\mathbb{R}\times Y$, with boundary on either $\mathbb{R}\times\Lambda_{i/N}$ or $\mathbb{R}\times\Lambda_{(i+1)/N}$ and a single positive puncture, which are pseudoholomorphic for a cylindrical almost complex structure.

If one considers these terms as a composition of operations, the fact that the disks of index 1 in the second bullet point define the differential (which is strictly filtration-decreasing), the statement finally follows by an energy estimate similarly to the first case. □

The remainder of the proof of Proposition 1.11 is similar to the proof of Proposition 2.5; roughly, we decompose the isotopy into small steps that then are shown to induce homotopy equivalences of small degree. Lemma 3.1(1) implies that, for a sufficiently fine decomposition of $[-,+]$, each map $\Phi_{[-_i,+_i]}$ has an arbitrarily small cobordism length. To prove Proposition 1.11, we thus restrict to a single interval of a sufficiently fine generic decomposition (this single small interval we continue to label $[-,+]$) and show that it is a finite composition of (de)stabilizations and tame automorphisms.

Note that the Reeb chord lengths vary continuously with the parameter t. For a very fine decomposition we may thus assume that these lengths are almost constant in the interval $[-,+]$. Together with the proposition's hypothesis of genericity, and since $\delta > 0$ can be assumed to be arbitrarily small, we get three cases listed below. For all cases, we assume that no chord has action less than $e^{2\delta}$. Recall that we only consider the Chekanov–Eliashberg algebra $\mathscr{A}^l$ generated by chords with actions less than l. Below we thus disregard all chords of action greater than l' for some suitable action level $l' \gg l$. Moreover, after further shrinking $\delta > 0$, we can assume that no Reeb chords on the family Λ_t of Legendrians has length contained in the interval $[e^{-\delta}l', e^{\delta}l']$.

Case 1 There are no births/deaths in $[-,+]$. Further, any x^+ satisfies

$$\ell(x^{\mp}) \in [e^{-\delta}\ell(x^{\pm}), e^{\delta}\ell(x^{\pm})],$$

while any y^+ different from x^+ satisfies

$$[e^{-\delta}\ell(y^-), e^{\delta}\ell(y^-)] \cap [e^{-\delta}\ell(x^+), e^{\delta}\ell(x^+)] = \varnothing.$$

(In particular, any two Reeb chords have distinct lengths.)

Case 2 The chords whose lengths are contained in $[e^{-2\delta}\ell_0, e^{2\delta}\ell_0]$ are precisely two, and undergo a birth/death at $0 \in [-,+]$; ie there are precisely two chords x^+, y^+ of lengths

$$\ell(x^t), \ell(y^t) \in [e^{-\delta}\ell_0, e^{\delta}\ell_0]$$

for $t > 0$ (resp. $t < 0$), while there are no such chords for $t < 0$ (resp. $t > 0$). Furthermore, $\ell(x^+) > \ell(y^+)$ (resp. $\ell(x^-) > \ell(y^-)$). The chords z^t of length

$$\ell(z^t) \notin [e^{-2\delta}\ell_0, e^{2\delta}\ell_0]$$

satisfy the assumptions of Case 1.

Case 3 There are no births/deaths in $[-,+]$. There are precisely two chords whose lengths are contained in $[e^{-2\delta}\ell_0, e^{2\delta}\ell_0]$, which moreover have lengths contained inside $[e^{-\delta}\ell_0, e^{\delta}\ell_0]$, and satisfy $\pm\ell(x^{\pm}) > \pm\ell(y^{\pm})$ while $\ell(x^0) = \ell(y^0)$. The chords z^t of length

$$\ell(z^t) \notin [e^{-2\delta}\ell_0, e^{2\delta}\ell_0]$$

satisfy the assumptions of Case 1.

In the case when there are no births/deaths, the invariance under DG–homotopy together with Lemma 3.1(2) now implies that

$$\begin{aligned} x^+ = (\Psi\Phi - (\partial_+ G - G\partial_+))x^+ &= k_\Psi k_\Phi x^+ + \sum_j v_j^+ - (\partial_+ G - G\partial_+)x^+ \\ &= k_\Psi k_\Phi x^+ + \sum_j v_j^+ + \sum_k u_k^+ \end{aligned} \tag{3-2}$$

is satisfied, where $k_\Phi \in \boldsymbol{k}$ (resp. $k_\Psi \in \boldsymbol{k}$) are the coefficients $\langle \Phi(x^+), x^- \rangle$ and $\langle \Psi(x^-), x^+ \rangle$, and v_j^+, u_k^+ are monomials of chords of Λ_+ that satisfy

$$e^{2\delta}\ell(x^+) \geq \ell(v_j^+), \ell(u_k^+).$$

Case 1 Looking at the last two terms in equation (3-2), Lemma 3.1(2) implies two things. First, if $v_j^+ \in (\boldsymbol{k} \setminus \{0\})x^+$, then v_j^+ is in the image of Ψ of an element of action at most $e^\delta \ell(x^+)$, which by definition is not contained in $\boldsymbol{k}x^-$. Second, if $u_k^+ \in (\boldsymbol{k} \setminus \{0\})x^+$, then either x^+ appears as a term in the differential of a word $G(x^+)$ all of whose letters are of action at most $e^{2\delta}\ell(x^+)$, or as a term in $G(w^+)$ for a word w^+ of action $\ell(w^+) < \ell(x^+)$. Moreover, in the latter case, all letters in $G(w^+)$, and thus in particular x^+ itself, have action bounded by $e^{2\delta}\ell(w^+)$. The hypotheses in Case 1 imply that $\ell(v_j^+), \ell(u_k^+) < \ell(x^+)$. Thus $k_\Phi = k_\Psi^{-1}$ in (3-2) is a unit.

Cases 2 and 3 Case 1 handles all chords without other chords of approximately the same action. So we have reduced to studying the maps at x and y which have approximately the same action. Consider the 2×2 matrix

$$[F] := \begin{pmatrix} F_{xx} & F_{xy} \\ F_{yx} & F_{yy} \end{pmatrix}$$

in the x, y basis, for the map $F \in \{\Phi, \Psi, G, \partial_+, \mathrm{id}\}$. The bound from below of $\ell(z)$ implies there is no additional (nonlinear) term involving x or y in either $F(x)$ or $F(y)$. Since $\ell(x^+) > \ell(y^+)$,

$$[\partial_+] = \begin{pmatrix} 0 & 0 \\ (\partial_+)_{yx} & 0 \end{pmatrix}$$

for some $(\partial_+)_{yx} \in \boldsymbol{k}$. We also consider the corresponding 2×2 matrix version of (3-2).

Case 2 Since (x^-, y^-) do not exist,

$$[\Psi] = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} = [\Phi].$$

So (3-2) implies $1 = \mathrm{id}_{xx} = G_{xy}(\partial_+)_{yx}$. In particular, after the tame automorphism of scaling y by $((\partial_+)_{yx})^{-1} = G_{xy} \in \boldsymbol{k}$, we get $\partial_+(x^+) = y^+ + \sum_i w_i$ with $\ell(w_i) < \ell(y^+)$. The result now follows from Chekanov's algebraic treatment of birth/deaths [8, Sections 8.4–8.5].

Case 3 Assume at $t = 0$ that J is generic so that the DGA differential ∂_0 is well defined. Let $\Phi^0 \colon (\mathscr{A}, \partial_+) \to (\mathscr{A}, \partial_0)$ and $\Psi^0 \colon (\mathscr{A}, \partial_0) \to (\mathscr{A}, \partial_+)$ be the DGA morphisms induced by the trace and its inverse over the subinterval $[0, +] \subset [-, +]$. Let G^0 be the homotopy relating $\Psi^0\Phi^0$ and id_0. As above, define the 2×2 matrix

$$\begin{pmatrix} F_{xx} & F_{xy} \\ F_{yx} & F_{yy} \end{pmatrix}$$

in the x^0, y^0 basis for the map $F \in \{\Phi^0, \Psi^0, G^0, \partial_0, \mathrm{id}_0\}$. Stokes' theorem implies that $[\partial_0] = 0$ as a 2×2 matrix. From the 2×2 matrix equations

$$[\mathrm{id}_0] = [G^0\partial_0] + [\partial_0 G^0] + [\Psi^0\Phi^0] = [\Psi^0\Phi^0], \quad [\partial_+\Phi^0] = [\Phi^0\partial_0] = 0, \quad [\Psi^0\partial_+] = [\partial_0\Psi^0] = 0,$$

we get

$$1 = (\mathrm{id}_0)_{xx} = \Psi^0_{xx}\Phi^0_{xx} + \Psi^0_{xy}\Phi^0_{yx}, \quad (\partial_+)_{yx}\Phi^0_{xx} = 0, \quad \Psi^0_{xy}(\partial_+)_{yx} = 0,$$

which imply $(\partial_+)_{yx} = 0$.

Since $[\partial_+]=0$ as a 2×2 matrix, equation (3-2) implies that $[\Psi],[\Phi]\in \mathrm{GL}(2,\mathbb{Z})$. [11, Corollary 2.6] proves that $\mathrm{SL}(2,\mathbb{Z})$, which is an index 2 subgroup of $\mathrm{GL}(2,\mathbb{Z})$, is generated by the two tame automorphisms

$$\begin{pmatrix}1 & 1\\ 0 & 1\end{pmatrix} \quad \text{and} \quad \begin{pmatrix}1 & 0\\ 1 & 1\end{pmatrix}.$$

But we also allow the map

$$\begin{pmatrix}-1 & 0\\ 0 & 1\end{pmatrix} \in \mathrm{GL}(2,\mathbb{Z}) \setminus \mathrm{SL}(2,\mathbb{Z}).$$

Thus Ψ and Φ are compositions of our allowable tame automorphisms.

4 A Rabinowitz–Floer theory for Legendrians

Rabinowitz–Floer homology in the case of a contact type hypersurface was originally defined by Cieliebak, Frauenfelder and Oancea [9]. We present a version of the theory in the relative case, RFH, which previously has been considered in the Hamiltonian setting by Merry [32] and Cieliebak and Oancea [10], and in the SFT setting by Legout [31]. Our construction of the complex is the direct generalization of the construction from [31] to the case of an arbitrary contact manifold, while allowing augmentations that are only defined under some action level.

In Section 4.1, we compactify the moduli spaces used in Sections 4.2 and 4.3. In Section 4.2, we introduce a Rabinowitz–Floer complex (RFC) as a mapping cone complex generated by Reeb chords. In Section 4.3, we prove the invariance of this mapping cone complex. Compared to the invariance result from [31], we here need to take extra care to control the filtration-preserving properties, in order to establish invariance by a PWC family of complexes.

4.1 Compactification of certain moduli spaces

Let $\Lambda_0^+,\Lambda_1^+\subset Y$ be two Legendrians isotopic to Λ_0^-,Λ_1^-. Assume $\overline{\Lambda}^{\pm}:=\Lambda_0^{\pm}\cup\Lambda_1^{\pm}$ is embedded. Let $-\le t\le +$ parametrize this isotopy. (We use $\pm$ instead ± 1 to avoid notational overuse of the symbol 1.) Further, let $L_0,L_1\subset(\mathbb{R}_\tau\times Y, d(e^\tau\alpha))$ be the exact Lagrangian concordance arising from the trace of the isotopy, with $L_i\cap\{\tau\}\times Y=\Lambda_i^{\pm}$ for $\pm\tau\gg 1$. Assume that $L:=L_0\cup L_1$ has at most one transverse double point q.

There exist primitives of $e^\tau\alpha|_{TL_i}$ by exactness. Since this primitive is necessarily locally constant wherever L_i is cylindrical, we can fix a unique primitive that vanishes on the negative ends of L_i. After a small perturbation of L_i, we may assume that there is a nonzero difference of primitives at the unique intersection point $\{q\}=L_0\cap L_1$.

We consider several types of asymptotic behaviors for our holomorphic disks.

- **Mixed $\alpha^{\pm}$–chords** Such a chord starts on $\Lambda_0^{\pm}$ (resp. $\Lambda_1^{\pm}$) and ends on $\Lambda_1^{\pm}$ (resp. $\Lambda_0^{\pm}$).
- **Pure $\alpha^{\pm}$–chords** Such a chord both starts and ends on $\Lambda_0^{\pm}$ (resp. $\Lambda_1^{\pm}$).
- **Lagrangian intersection point q**

Let Γ be a nonempty cyclically ordered set of the above, each endowed with a sign. Repetition is allowed. Let $\mathrm{Bd} \in \{L, \overline{\Lambda}^{\pm}\}$. Let $\mathcal{M}^d(\Gamma; \mathrm{Bd})$ denote the moduli space of J–holomorphic disks $u\colon D \to \mathbb{R} \times Y$, with boundary marked points, satisfying the following conditions:

- The boundary of the disk maps to L if $\mathrm{Bd} = L$ and to $\mathbb{R} \times \overline{\Lambda}^{\pm}$ if $\mathrm{Bd} = \overline{\Lambda}^{\pm}$.
- The (formal) dimension of the component is d.
- Each marked point maps to an element of Γ. The cyclic ordering of marked points induced by the boundary orientation matches the cyclic ordering of the chords/double points in Γ.
- If a marked point maps to the double point (ie when $\mathrm{Bd} = L$), then the puncture is *positive* (resp. *negative*) if the primitive of $e^{\tau}\alpha|_L$ evaluated along the boundary of the disk makes a jump to a lower (resp. higher) value at the puncture, while traversing the boundary in the positive direction. If a marked point maps to a chord, the endowed sign $\pm$ indicates that it is an asymptotic limit at the $\pm\infty$ end of the symplectization boundary.

Remark 4.1 Consider the Legendrian lift of $L_0 \cup L_1$ to the contactization $(\mathbb{R}_\tau \times Y \times \mathbb{R}_Z, dZ + e^{\tau}\alpha)$ that is uniquely determined by the requirement that its Z–coordinate vanishes at $\tau = -\infty$. The sign of the puncture at a double point has the following direct reformulation in terms of the Reeb chord on this Legendrian. A disk has a positive (resp. negative) puncture at q if and only if the value of the Z–coordinate along the boundary of the disk, as specified by the Legendrian lift, jumps to a higher (resp. lower) value at the puncture when traversing the boundary according to the orientation. This will be important later, when we describe the cobordism by the front projection of its Legendrian lift.

Note that $\mathcal{M}^d(\Gamma; \overline{\Lambda}^{\pm})$ has an $\mathbb{R}$–translation in the range. We denote the quotient of this translation by $\widehat{\mathcal{M}}^{d-1}(\Gamma; \overline{\Lambda}^{\pm}) = \mathcal{M}^d(\Gamma; \overline{\Lambda}^{\pm})/\mathbb{R}$. We next list different types of boundary conditions for the moduli spaces. In all cases below, Γ may have some additional negative pure chords.

($1_\pm$) $\mathrm{Bd} = \overline{\Lambda}^{\pm}$; Γ contains two positive mixed chords.

($2_\pm$) $\mathrm{Bd} = \overline{\Lambda}^{\pm}$; Γ contains one positive and one negative mixed chord.

($3_\pm$) $\mathrm{Bd} = \overline{\Lambda}^{\pm}$; Γ contains one positive pure chord and no mixed chords.

(4) $\mathrm{Bd} = L$; Γ contains two mixed chords (both positive).

(5) $\mathrm{Bd} = L$; Γ contains two mixed chords (one positive, one negative).

(6) $\mathrm{Bd} = L$; Γ contains one mixed chord (positive) and q (positive or negative).

(7) $\mathrm{Bd} = L$; Γ contains one mixed chord (negative) and q (positive or negative).

Note that the sign of the puncture at q in cases (6) and (7) can be recovered by the following data: the component of the starting point (or endpoint) of the mixed Reeb chord asymptotic of the disk, together with the two action values of the potentials at q for each of the two sheets of L.

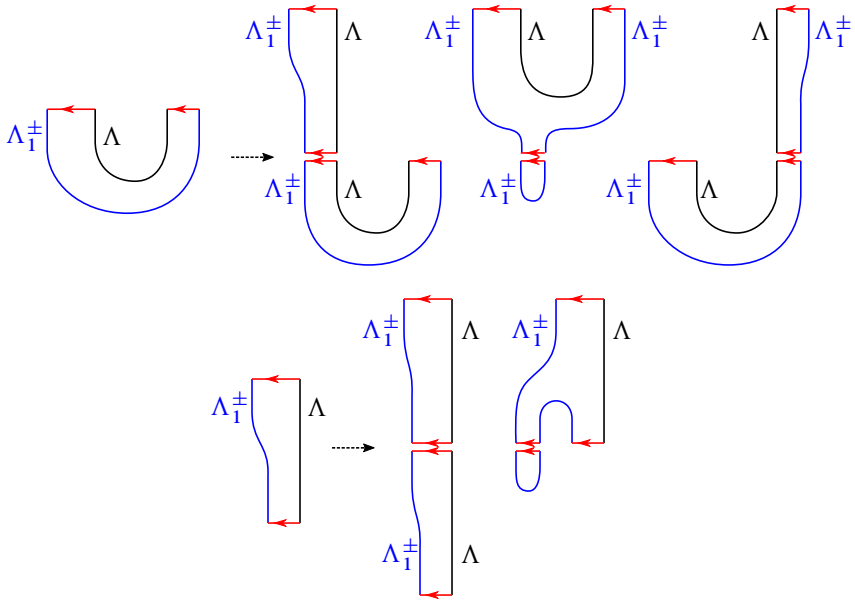


Figure 1: Top: Boundary points of the compactification of the moduli space $(1_\pm)$ of punctured disks with boundary on the cylinders over $\Lambda \cup \Lambda_1^{\pm}$ with two positive mixed Reeb chord asymptotics. Note that we have omitted any trivial strip inside the symplectization level from the figure. The broken configurations on the left and right belong to $(2_\pm \times 1_\pm)$, while the middle configuration belongs to $(1_\pm \times 3_\pm)$. Bottom: Boundary points of the compactification of the moduli space $(2_\pm)$ of punctured disks with boundary on the cylinders over $\Lambda \cup \Lambda_1^{\pm}$. Note that we have omitted any trivial strip inside the symplectization level from the figure. The broken configuration on the left belongs to $(2_\pm \times 2_\pm)$ while the one on the right belongs to $(2_\pm \times 3_\pm)$. There are similar breakings for the disks whose mixed Reeb chords start on $\Lambda_1^{\pm}$ and end on Λ.

We now describe the boundary ∂ in the sense of the SFT compactification [3], also called Gromov–Hofer compactification, of certain one-dimensional moduli spaces, modding out by the $\mathbb{R}$–translation when one can. We illustrate the notation with some examples. The broken curve $(2_+ \times 1_+)$ has boundaries in two copies of $(\mathbb{R} \times Y, \mathbb{R} \times \Lambda^+)$. In the lower copy there is a curve of index 1 (rigid after $\mathbb{R}$–translation) of type (1_+). In the upper copy there is one curve of index 1 (rigid after $\mathbb{R}$–translation) of type (2_+) and one "trivial strip" of index 0, which is a curve of the form $(\mathbb{R} \times \text{chord})$. (We omit listing the trivial strips.) The broken curve (6×6) has two index 0 curves of type (6) in the same copy of $(\mathbb{R} \times Y, L)$, one with a positive puncture at q and the other with a negative puncture at q.

Figures 1 through 3 depict the broken configurations corresponding to the boundary strata of the moduli spaces of the corresponding type. Note that, in these figures, any trivial strip inside a cylindrical level has been omitted. For every broken configuration in which there is a nonempty level with boundary on $\overline{\Lambda}^+$, as well as a middle level with boundary on L, there might be noncylindrical components in the middle level with only pure punctures. Such components are not exhibited in the aforementioned figures; see Figure 6 for an example where levels of this type arise in the boundary of the moduli space of type (5).

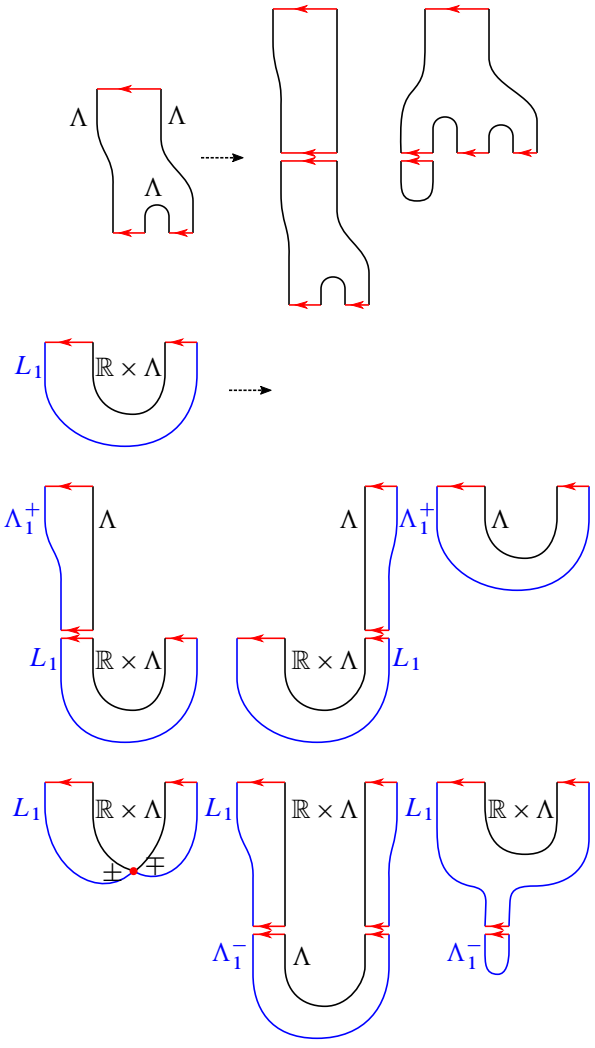


Figure 2: Top: Boundary points of the compactification of the moduli space $(3_\pm)$ of punctured disks with all boundary components on either the cylinder over Λ, Λ_1^-, or Λ_1^+. In other words, these are the curves that are used in the definition of the Chekanov–Eliashberg algebra of either Legendrian. All broken curves belong to $(3_\pm \times 3_\pm)$. Note that we have omitted any trivial strip inside the symplectization level from the figure. Bottom: Boundary points of the compactification of the moduli space (4) of punctured disks with boundary on $L = L_0 \cup L_1$ with two positive mixed Reeb chords. Note that we have omitted any trivial strip inside the symplectization level from the figure. The broken configurations shown are as follows: The top row the left and middle configurations are in $(2_+ \times 4)$, while the one on the right is in (1_+). On the bottom row from left to right, the configurations shown are in (6×6), $(5\times 5\times 1_-)$, and $(4\times 3_-)$.

The figures depict the most general case that we will need, ie the special case when $L_0 = \mathbb{R}\times\Lambda$ is a trivial cylinder, which in particular means that $\Lambda_0^\pm = \Lambda$, while L_1 is a Lagrangian cylinder from Λ_1^- to Λ_1^+.

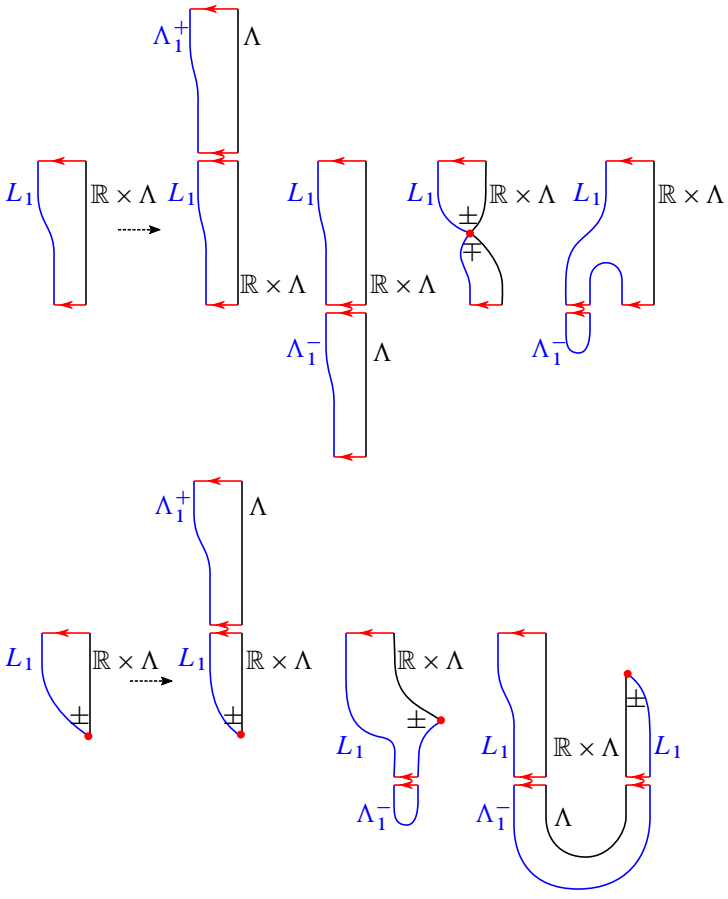


Figure 3: Top: Boundary points of the compactification of the moduli space (5) of punctured disks with boundary on L with one positive and one negative mixed Reeb chord puncture. Note that we have omitted any trivial strip inside the symplectization level from the figure. The broken configurations shown are in $(5\times 2_-)$, $(5\times 2_+)$, (6×7), and $(5\times 3_-)$, going from left to right. There are analogous configurations when the mixed Reeb chords start on L_1 and end on L_0. Bottom: Boundary points of the compactification of the moduli space (6) of pseudoholomorphic curves with one puncture asymptotic to a positive mixed Reeb chord from $\mathbb{R}\times\Lambda$ to L_1, and one positive puncture at the unique intersection point. The broken configurations shown are in $(2_+\times 6)$, $(6\times 3_-)$, and $(2\times 7\times 1_-)$, going from left to right. There are analogous configurations in the case when the positive mixed Reeb chord starts on L_1 and ends on $\mathbb{R}\times\Lambda$.

Proposition 4.2 *For a generic almost complex structure, the boundary of a one-dimensional moduli space is made of the following configurations of rigid moduli spaces*:

- $(1_\pm)$ $(1_\pm\times 3_\pm)$, $(2_\pm\times 1_\pm)$,
- $(2_\pm)$ $(2_\pm\times 2_\pm)$, $(2_\pm\times 3_\pm)$,
- $(3_\pm)$ $(3_\pm\times 3_\pm)$,
- (4) $(2_+\times 4)$, $(5\times 5\times 1_-)$, (1_+), (6×6), $(4\times 3_-)$,

(5) $(2_+ \times 5), (5 \times 2_-), (6 \times 7), (5 \times 3_-)$,

(6) $(2_+ \times 6), (5 \times 7 \times 1_-), (6 \times 3_-)$.

The Lagrangian cobordisms we study in Section 4 will not have any one-dimensional moduli space of type (7).

Proof The mixed chords (or double points) each appears exactly once in the word of chords defining a given moduli space. If a disk has only pure chords, then Stokes' theorem implies it has a unique positive pure chord. Therefore, transversality for these spaces can be achieved by either perturbing J near this distinguished positive chord as in [13, Proposition 3.13], or by perturbing the Lagrangian boundary condition near p as in [23].

These configurations are the only ones that can appear because

(a) a dimension argument implies exactly one (nontrivial) moduli space in a symplectization;

(b) each moduli space has at least one positive puncture. □

4.2 The Rabinowitz complex as a mapping cone

In the following we assume that $\Lambda_0 = \Lambda$ is fixed, and that $\Lambda_1^t = \phi^t_{\alpha,H_t}(\Lambda_1)$ is a Legendrian isotopy of $\Lambda_1^- = \Lambda_1$ for $- \leq t \leq +$. We write $\overline{\Lambda}^t := \Lambda_0 \cup \Lambda_1^t$, while $\overline{\Lambda} = \Lambda_0 \cup \Lambda_1$ as before.

Fix $t \in [-, +]$ as in Section 4.1. Choose a cylindrical almost complex structure that is regular for the moduli spaces involved (ie that consist of disks for which there is a distinguished asymptotic that only appears once, and where the moduli space is of expected dimension at most two before taking the quotient by the action of translation). Recall that this is possible by [13, Proposition 3.13]. Assume there exist augmentations ε^i defined for $\mathscr{A}^l(\Lambda_i)$, $i = 0, 1$, which can be identified with an augmentation ε_0 of the DGA $\mathscr{A}^l_{\text{pure}}(\overline{\Lambda}^0)$ that is generated by the pure chords. In [17, Lemma 3.4], we describe how to define an augmentation ε_t for $\mathscr{A}^{l-l(t)}_{\text{pure}}(\overline{\Lambda}^t)$, where

$$l(t) = \int_0^t \big(\max_{y \in Y} H_\tau(y) - \min_{y \in Y} H_\tau(y)\big)\, d\tau.$$

The setting for [17] was limited to $(Y, \alpha) = (P \times \mathbb{R}_z, dz - \theta)$. However, [17, Lemma 3.4] used only that the DGA underwent a stable-tame isomorphism. Thus, Proposition 1.11 implies that [17, Lemma 3.4] applies to our more current general setting.

Since the stable tame isomorphism given by Proposition 1.11 are the DGA–quasi-isomorphism induced by the Lagrangian trace cobordisms, we get

Lemma 4.3 *For a sufficiently fine subdivision $t_1 < \cdots < t_N$, the augmentation ε_{t_i} can, moreover, be assumed to coincide with the pullback of the augmentation $\varepsilon_{t_{i-1}}$ under the DGA–morphism induced by the exact Lagrangian trace of the isotopy Λ_1^t for $t \in [t_{i-1}, t_i]$.*

In the following, let x_{01} (resp. x_{10}) indicate a mixed chord starting on Λ (resp. Λ_1^t) and ending on Λ_1^t (resp. Λ). All other chords z_i below are pure. We define the actions of x_{01} and of x_{10} to be, respectively,

$$\mathfrak{a}(x_{01}) := +\int_{x_{01}} \alpha > 0 \quad \text{and} \quad \mathfrak{a}(x_{10}) := -\int_{x_{10}} \alpha < 0.$$

Fix $a_t, b_t \in \mathbb{R}$ such that

$$0 < b_t - a_t < l - l(t). \tag{4-1}$$

Denote by

$$C_*([a_t, b_t)) = C_*(\mathbb{R}_{\geq 0} \cap [a_t, b_t)) \quad \text{and} \quad C^*([a_t, b_t)) = C^*(\mathbb{R}_{\leq 0} \cap [a_t, b_t))$$

the "linearized Legendrian contact homology complex" and the "dual cocomplex", respectively, where the former is generated by the mixed chords x_{01} while the latter is generated by the mixed Reeb chords x_{10}. In both cases we assume that the action $\mathfrak{a}$ of the generators is contained inside the interval $[a_t, b_t)$.

As prescribed below, the differential restricted to $C_*([a_t, b_t))$ (resp. $C^*([a_t, b_t))$) is the usual differential of the Legendrian contact homology (co)complex linearized by the augmentation ε_t that counts trips with one positive (resp. negative) mixed input Reeb chord and one negative (resp. positive) mixed output Reeb chord; see below for the precise formula.

Remark 4.4 For the cocomplex, the differential increases the Reeb chord length ℓ, hence decreases the above action $\mathfrak{a}$. This is why when taking the quotient complex, we consider $C^*([a_t, b_t))$ and not $C^*((a_t, b_t])$.

Recall that the linearized Legendrian contact (co)homology complex of a pair of Legendrians can be endowed with a grading that depends on additional choices of the two Legendrians involved; see [6] as well as Section A.3. For simplicity we restrict ourselves to the case when the first Chern class of (Y, α) vanishes, which means that we can choose a symplectic trivialization of the (square of the) determinant $\mathbb{C}$–line bundle $\det \xi \to Y$. Given the choice of a homotopy class of such a symplectic trivialization, together with choices of Maslov potentials of Λ_i as described in Section A.3, we get a canonically induced $\mathbb{Z}$–grading for which the linearized differential (resp. codifferential) is of degree -1 (resp. 1). In addition, the choice of a Maslov potential can be naturally extended over a Legendrian isotopy. Note that a loop of Legendrians can induce a nontrivial action on its set of Maslov potentials; see Lemma 6.11 for an example.

For a pair of mixed chords x, y in either complex, and ordered (possibly empty) sets of pure chords $\boldsymbol{z}$ and $\boldsymbol{z}'$, define

$$\boldsymbol{m}_{\overline{\Lambda}^t}(x^+, y^\pm) = \sum_{\boldsymbol{z}, \boldsymbol{z}'} \sharp \widehat{\mathcal{M}}^0(x^+ \boldsymbol{z}^- y^\pm \boldsymbol{z}'^-; \overline{\Lambda}^t) \varepsilon_t(\boldsymbol{z}\boldsymbol{z}').$$

Note that the action of the individual pure chord z in $\boldsymbol{z}$ or $\boldsymbol{z}'$ is less than $b_t - a_t$, and so $\varepsilon_t(z)$ is defined. We suppress the subscript notation when we define the maps

$$d_{01}\colon C_*([a_t,b_t)) \to C_*([a_t,b_t)), \quad d_{10}\colon C^*([a_t,b_t)) \to C^*([a_t,b_t)), \quad B\colon C_*([a_t,b_t)) \to C^*([a_t,b_t]),$$

where

$$\begin{aligned} d_{01}(x_{01}) &= \sum_{\{y_{01} \mid \mathfrak{a}(y_{01}) \in [a_t,b_t) \cap \mathbb{R}_{>0}\}} \boldsymbol{m}_{\overline{\Lambda}^t}(x_{01}^+, y_{01}^-) y_{01}, \\ d_{10}(x_{10}) &= \sum_{\{y_{10} \mid \mathfrak{a}(y_{10}) \in [a_t,b_t) \cap \mathbb{R}_{<0}\}} \boldsymbol{m}_{\overline{\Lambda}^t}(y_{10}^+, x_{10}^-) y_{10}, \\ B(x_{01}) &= \sum_{\{y_{10} \mid \mathfrak{a}(y_{10}) \in [a_t,b_t) \cap \mathbb{R}_{<0}\}} \boldsymbol{m}_{\overline{\Lambda}^t}(x_{01}^+, y_{10}^+) y_{10}. \end{aligned}$$

Note that d_{01} and d_{10} are the linearized Legendrian contact homology differential and codifferential, respectively.

Proposition 4.5 *The matrix*

$$d_B := \begin{pmatrix} d_{01} & 0 \\ B & d_{10} \end{pmatrix}$$

is a filtered mapping cone differential for the filtered chain map B.

Proof By Proposition 4.2($1_\pm$) and 4.2($2_\pm$), the matrix squares to zero. Consider $x_{01} \in C_*([a_t,b_t))$ and $z_{10} \in C^*([a_t,b_t))$, for example. Then

$$\langle d_B^2 x_{01}, z_{10}\rangle = \sum_{y_{01}} \boldsymbol{m}_{\overline{\Lambda}^t}(x_{01}^+, y_{01}^-)\boldsymbol{m}_{\overline{\Lambda}^t}(y_{01}^+, z_{10}^+) + \sum_{y_{10}} \boldsymbol{m}_{\overline{\Lambda}^t}(x_{01}^+, y_{10}^+)\boldsymbol{m}_{\overline{\Lambda}^t}(y_{10}^-, z_{10}^+),$$

which vanishes by Proposition 4.2($1_\pm$). □

Definition 4.6 Let $(a_t, b_t, \varepsilon_t, l - l(t), \overline{\Lambda}^t)$ denote the auxiliary information used to define this cone complex. We denote this *mapping cone complex* as

$$\mathrm{RFC}_*^{[a_t,b_t)}(\Lambda_0, \Lambda_1^t; \varepsilon_t) = \big(C_*(t)[a_t,b_t) \oplus C^{n-*-2}(t)[a_t,b_t), d_B\big),$$

which naturally is a filtered chain complex with action window $[a_t, b_t)$ and filtration induced by $\mathfrak{a}$.

Remark 4.7 (1) The sign change and shift of grading of the second summand is needed in order for the summand B of the differential d_B to be of degree -1. The relevant index for the disks with two positive punctures whose count defines B was computed in [22, Lemma 2.5].

Further, with this convention, the degree of a generator is continuous under deformations through transverse chords, even when the length of the chord at some point becomes zero, so that the component of the starting and end points become interchanged; this readily follows from the definition of the grading in Section A.3.

The mapping cone complex also can be defined without a grading. While the grading is necessary (and warranted) to prove Theorem 1.10, the other results in Section 1 need only the ungraded complex.

(2) We suppress the upper bound notation $l - l(t)$ in an attempt to reduce the overbearing set of decorations. To be consistent with similar concepts in other literature, we sometimes call the mapping cone complex the *Rabinowitz–Floer complex*, and when the Legendrian is augmentable (so that we can choose $b_t = l = -a_t = \infty$) we denote it by $\mathrm{RFC}_*(\Lambda, \Lambda_t)$.

(3) The Legendrian contact cohomology complex $C^*((-\infty, a])$ is the degree-wise dual of a possibly infinite dimensional complex. In such a case the chords x_{10} do not form a basis of $\mathrm{RFC}_*(\Lambda, \Lambda_t)$, since one must also allow formal infinite sums of such chords. In other words, in each fixed degree $i \in \mathbb{Z}$, the filtered vector space $\mathrm{RFC}_i^{(-\infty,+\infty)}(\Lambda, \Lambda_t)$ is the inverse limit of the directed system

$$\mathrm{RFC}_i^{[0,+\infty)} \leftarrow \mathrm{RFC}_i^{[-1,+\infty)} \leftarrow \mathrm{RFC}_i^{[-2,+\infty)} \leftarrow \cdots$$

of filtered vector spaces and canonical quotient maps induced by the filtration.

Theorem 4.8 *Assume that the following conditions are satisfied*:

- $l - l(-) > 0$ *is smaller than the length of any contractible periodic Reeb orbit* γ *of degree* $|\gamma| \leq 1$;
- ε_- *is an augmentation of the sub-DGA* $\mathcal{A}_{\mathrm{pure}}^{l-l(-)}(\Lambda_0 \cup \Lambda_1)$ *generated by pure Reeb chords*; *and*
- Λ_1^t *is generated by a Legendrian isotopy of oscillation*

$$l(t) = \int_-^t \big(\max_{y \in Y} H_\tau(y) - \min_{y \in Y} H_\tau(y)\big)\, d\tau.$$

Then, as long as $l - l(t) > b_t - a_t > 0$ *and* $[a_t, b_t)$ *is a finite interval, there exists a sequence of augmentations* ε_t *of the DGAs* $\mathcal{A}_{\mathrm{pure}}^{l-l(t)}(\Lambda_0 \cup \Lambda_1^t)$ *that makes* $\mathrm{RFC}_*^{[a_t,b_t)}(\Lambda_0, \Lambda_1^t; \varepsilon_t)$ *into a well-defined PWC family of complexes with action-window* $[a_t, b_t)$. *In particular, the complexes undergo the deformations specified by the barcode proposition (Proposition 2.4) as* t *varies.*

In the case when $l = +\infty$, $a_t = -\infty$ *and* $b_t = +\infty$ *holds for all* t, *then the homology of the entire complex is invariant under Legendrian isotopy. In addition, in any smooth family of finite action windows, we may assume that we have a PWC family of complexes.*

The analogous invariance also holds when the first Legendrian is deformed by a Legendrian isotopy, while the second copy is being fixed. The proof is completely analogous and left to the reader.

Remark 4.9 • Theorem 4.8 holds if we replace $l(t)$ with the oscillation underlying the Shelukhin–Chekanov–Hofer norm or the Usher norm

$$l_1(t) = \int_-^t \max_{y \in Y} H_\tau(y)\, d\tau \quad \text{or} \quad l_2(t) = l(t) + \max_{y \in Y} |g(y)|.$$

Here $g(y)$ is the conformal factor for the time-t contactomorphism defined by H; see [39, Definition 10.1]. This follows because $l(t) \leq l_1(t), l_2(t)$.

• If H defines a contact-form-preserving contactomorphism, ie if the conformal factor is 0, then the length of the pure chords are preserved. It then follows from Proposition 1.11 that the stable-tame isomorphism class of the Chekanov–Eliashberg algebra $\mathscr{A}^l$ does not depend on t. In particular, $\mathscr{A}^l$ has an augmentation for all t, and we can improve Theorem 4.8 by dropping the condition $l - l(t) > 0$.

To prove Theorem 4.8, we write the isotopy as a concatenation of short isotopies. Below any action level upper bound, we can assume there is at most one of the following singular moments: a mixed chord that enters or leaves the action window $[a_t, b_t)$; the birth/death of a pair of chords occurs; the actions of exactly two mixed chords coincide; the action of a pure chord equals $l - l(t)$; and the action of a mixed chord vanishes. All of these moments, except the last one, are considered in [17, Proposition 3.5], which is the equivalent of Theorem 4.8 in the special case $(Y, \alpha) = (P \times R_z, dz - \theta)$. The case when the action of a mixed chord vanishes corresponds to when the Legendrians Λ_0 and Λ_1^t intersect. By genericity we can always assume that there are only finitely many such moments in the family. The invariance at these moments is taken care of in Section 4.3.1 below.

To recycle the reasoning of [17, Proposition 3.5], we will apply the algebraic machinery from Section 2, in particular Propositions 2.4 and 2.5, to our current geometric setup. (In particular, the algebraic interpretation of bifurcation analysis done in [17] is replaced by Proposition 1.11.) And to apply the Section 2 results, we need to establish the hypotheses (A1) and (A2) for Lemma 2.1.

Finally, in the case when $l = +\infty$, $a_t = -\infty$ and $b_t = +\infty$, the proof of the invariance of the complex is simpler, since we do not need to care about whether the maps consist of standard bifurcations, ie handleslides or birth/deaths. (In other words, we do not need Lemma 4.13.) Instead, the weaker property of quasi-isomorphism follows by standard invariance properties of the linearized Legendrian contact homology as in [20], combined with the treatment of the double point and hypothesis (A2) in Section 4.3.1 below.

4.3 Mapping cone complex invariance during a short Legendrian isotopy

We now consider a varying $t \in [-, +]$. Assume that the interval $[-, +]$ is small. As in Section 4.1, L denotes the Lagrangian isotopy-trace concordance of $\overline{\Lambda}^t$. Since the oscillatory norm $\epsilon = l(+) - l(-)$ isotopy from $\overline{\Lambda}_-$ to $\overline{\Lambda}_+$ is arbitrarily small, we assume that there are no pure chords whose lengths are in the interval $(l - l(+), l - l(-))$. Thus we can use the same action window $[a_-, b_-) = [a_+, b_+) = [a, b)$ for both complexes. Moreover, we assume that no mixed chords enter or leave this action window, and that there are no birth/death pairs of chords when $t \in [-, +]$.

4.3.1 One double point Suppose $\Lambda \cap \Lambda_1^t = \varnothing$ when $t \in [-, +] \setminus \{0\}$ and that there exists a unique intersection point $\Lambda \cap \Lambda_1^0 = \{q\}$ that is transverse in a one-parameter family. In particular, the double point arises as a transverse family of mixed Reeb chords c_t, whose action $\mathfrak{a}(c_t)$ changes sign at $t = 0$.

Remark 4.10 For the remainder of this discussion, we assume that c_- runs from Λ_1^- to Λ, and hence $\mathfrak{a}(c_-) < 0 < \mathfrak{a}(c_+)$. The case when c_- runs from Λ to Λ_1^- follows from this case by considering the reverse-time isotopy.

For some sufficiently small interval $[-,+]$, after a possible C^0–small perturbation, we can choose a contact-form-preserving Darboux neighborhood $U \subset Y$ of q such that the following hold:

- There is a contact-form-preserving identification
$$U \cong (B_1)_x \times (B_1)_y \times [-\epsilon,\epsilon]_z \subset (B_1)_x \times \mathbb{R}^n_y \times \mathbb{R}_z = (J^1(B_1), \eta(dz - p_x\,dx))$$
for some $\eta > 0$ that we can make arbitrarily small, and where $B_r \subset \mathbb{R}^n$ is the disk centered at $\mathbf{0}$ of radius r.
- $\Lambda \cap U$ is the $\mathbf{0}$–section $j^1 0$.
- $\Lambda_1^t \cap J^1(B_1) = j^1(f_t)$ for some $f_t : B_1 \to \mathbb{R}$ that smoothly varies with t.
- For all $x \in B_1 \setminus B_{2/3}$, $f_t(x)$ is independent of t.
- For all $x \in B_{1/3}$, $f_t(x) = \|x\|^2 + t$.
- For all $x \in B_1$, $(df_t)^{-1}(0) = \mathbf{0}$.
- The chord $c_t \subset \mathbf{0} \times \mathbf{0} \times \mathbb{R}_z$ is the unique mixed chord of Λ and Λ_1^t in U.

In order to describe the trace cobordism it is useful to utilize the exact symplectomorphism

$$\big(\mathbb{R}_\tau \times \mathbb{R}^n_x \times \mathbb{R}^n_y \times \mathbb{R}_z, d(e^\tau(dz - y\,dx))\big) \xrightarrow{\cong} (\mathbb{R}_{>0})_q \times \mathbb{R}^n_x \times \mathbb{R}^n_{p_x} \times \mathbb{R}_{p_q}, \quad (\tau, x, y, z) \mapsto (e^\tau, x, e^\tau y, z),$$

where the target is equipped with the symplectic form $d(q\,dp_q - p_x\,dx)$. For the choice of primitive $d(-p_q\,dq - p_x\,dx)$, one can describe exact Lagrangians via their front projection to the contactization

$$\big(((\mathbb{R}_{>0})_q \times \mathbb{R}^n_x \times \mathbb{R}^n_{p_x} \times \mathbb{R}_{p_q}) \times \mathbb{R}_Z, dZ - p_q\,dq - p_x\,dx\big).$$

The noncylindrical part of the immersed Lagrangian trace cobordism inside $\mathbb{R} \times U$ is constructed in these coordinates via the front projection shown in Figures 4 and 5. We can assume that:

- The Lagrangian trace is cylindrical away from the intersection point, ie
$$L \cap (\mathbb{R}_\tau \times (Y \setminus U)) = \mathbb{R}_\tau \times ((\Lambda_1^+ \cup \Lambda) \setminus U) = \mathbb{R}_\tau \times ((\Lambda_1^- \cup \Lambda) \setminus U)$$
holds for some small neighborhood U of q.
- There exists $A < 0 < B$ so that $L \cap \{A \le \tau \le B\}$ contains the unique intersection point q in the level $\{\tau = 0\}$.
- In the above identification of $\mathbb{R} \times U$ with a subset of $\mathbb{R} \times J^1 B_1$, the noncylindrical part of L is identified with the exact Lagrangian immersion whose front is as shown in Figures 4 and 5. (The figures depict L in the coordinates $(q = e^\tau, x, p, p_x, Z)$ on the contactization of the symplectization $\mathbb{R} \times U$ described above.)

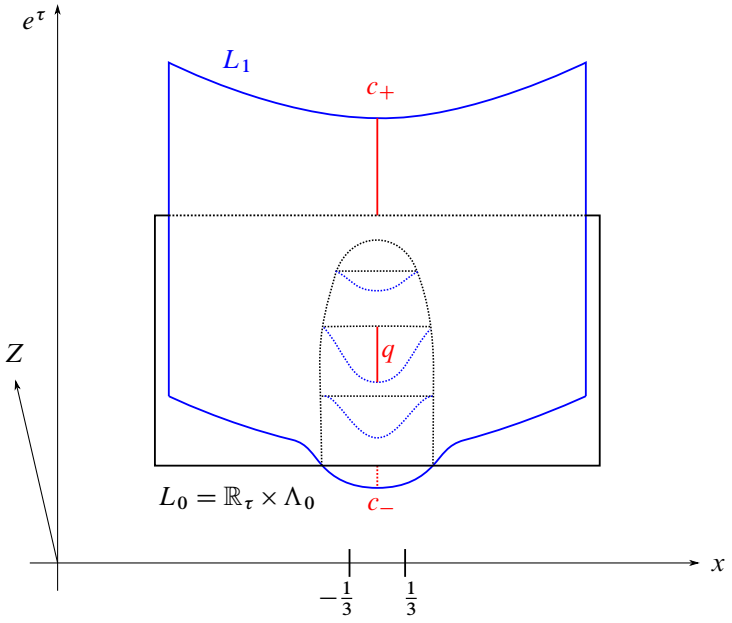


Figure 4: The front projection of the Legendrian lift of the exact Lagrangian immersion $L = L_0 \cup L_1$ inside $\mathbb{R}_\tau \times U$ to the contactization $(\mathbb{R}_\tau \times U) \times \mathbb{R}_Z$ of the symplectization. The Reeb chord q on the Legendrian lift corresponds to the unique double point $\{q\} = L_0 \cap L_1$. The cobordism is cylindrical with a vanishing primitive of $e^\tau\alpha$ outside of a compact subset if all sheets of the front are of the form $e^\tau f(x)$ outside of a compact subset. In order for (4-2) to hold inside the subset $\{\|x\| \leq \frac{1}{3}\}$, it suffices to consider a front which is the graph of a function of the form $e^\tau(\widetilde{f}(x) + g(e^\tau))$ above $B^n_{1/3}$.

- The pullback of the Liouville form $e^\tau\alpha$ to L has a primitive that vanishes outside of a compact subset; this is equivalent to the front projection in Figure 4 to consist of sheets that are of the form $e^\tau f(x)$ outside of a compact subset.
- In the subset of $\mathbb{R} \times U \hookrightarrow \mathbb{R} \times J^1B_1$ that lives inside $\mathbb{R} \times J^1B_{1/3}$, the immersed trace L is, moreover, of the form

$$\mathbb{R} \times \Lambda \cup \{(\tau, \phi_R^{\gamma(\tau)}(x)); x \in \Lambda_1^-\} \tag{4-2}$$

for some smooth $\gamma(\tau) \geq 0$ (but not monotonously increasing) function that vanishes when $\tau \leq A$, and which is constant when $\tau \geq B$. Recall that ϕ_R^t is the Reeb flow of α.

Remark 4.11 Equality (4-2) holds because Λ_1^t is assumed to induced by the Reeb flow inside the neighborhood $J^1B_{1/3}$ of the double point.There is one subtle point here: the function $\gamma(\tau)$ must be carefully chosen in order for $e^\tau\alpha$ to admit a compactly supported primitive when pulled back to L, as postulated by the fourth bullet point above. This condition on the primitive is easier to describe in the coordinates of the front projection: the front should be the graph of a suitable function of the form $e^\tau(\widetilde{f}(x) + g(e^\tau))$ above $B_{1/3}$, where $g(e^\tau)$ is constant for all $\pm\tau \gg 0$ sufficiently large.

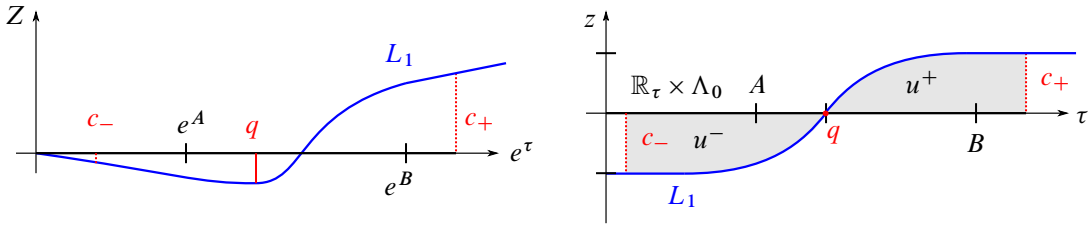


Figure 5: Left: the front projection the Legendrian lift of the exact immersed Lagrangian cobordism $L = L_1 \cup (\mathbb{R} \times \Lambda_0)$ to the contactization $(\mathbb{R}_\tau \times B_{1/3}) \times \mathbb{R}_Z$ of the symplectization. Right: the image of small disks $u^\pm$ with one puncture at the Reeb chord $c_\pm$ and a positivepuncture at the double point q which is contained inside a slice consisting of the image of the line $\mathbb{R}_\tau \times \{\mathbf{0}\}$ under the Reeb flow.

Lemma 4.12 *For $[-,+]$ sufficiently small, the following holds. For a suitable compatible almost complex structure on $\mathbb{R} \times Y$ which is cylindrical outside of a compact subset, there exists a* ***unique*** *pseudoholomorphic disk u^+ (resp. u^-) with positive (resp. negative) asymptotic puncture at c_+ (resp. c_-), a positive puncture at q, and no other punctures. There exists a neighborhood of adjusted almost complex structures* (see Section 3) *in which a generic choice J makes $u^\pm$ rigid and transversely cut out. Furthermore, there are no disks with a negative puncture at q and with all other punctures negative Reeb chord asymptotics, of which precisely one is mixed (and thus necessarily going from Λ to Λ_1^-).*

Proof From the above assumptions (including formula (4-2)), there exists a local projection

$$\pi : \mathbb{R} \times U \subset \mathbb{R} \times J^1(B_1) \to T^*(B_1)$$

such that

$$\pi(L_1 \cap (\mathbb{R} \times U)) \cap T^* B_{1/3} = \{(x, 2x)\} \cap T^* B_{1/3} \quad \text{and} \quad \pi(\mathbb{R} \times (\Lambda \cap B_{1/3})) = \{(x, 0)\}.$$

Let $J_{T^*(B_1)}$ be any almost complex structure compatible with the standard symplectic structure ω_0 on $T^*(B_1)$. This lifts to a unique cylindrical almost complex structure J_U on $\mathbb{R} \times U$ for which π is $(J_U, J_{T^*(B_1)})$–holomorphic.

If $J_{T^*(B_1)}$ is integrable in a neighborhood of the double point $\mathbf{0} \in T^*(B_1)$, then [14, Lemma 8.3(1)] proves the transversality result for the strip $u_\pm \subset \pi^{-1}(\mathbf{0})$ whose image is the trace of the isotopy of the Reeb chord c_t between q and $c_\pm$. This strip is depicted in Figure 5.

Extend J_U to an adjusted almost complex structure J for $\mathbb{R} \times Y$. Since transversality is an open condition, we can perturb J generically to assume all rigid holomorphic disks are transversely cut out.

We will show uniqueness of $u = u^+$ first using a monotonicity result [13, Lemma 5.1] for Lagrangian surgery cobordisms (the current setup is similar), then using a monotonicity result [30, Proposition 4.7.2] for compact Lagrangians. The u^- case is similar.

Using the notation of [13], let $-=A$ and $+=B$ so that L is cylindrical outside of $\{A \leq \tau \leq B\} \subset \mathbb{R}_\tau \times Y$. In [13, Section 3.3.1] the *total energy* is defined by

$$E_{[A,B]}(u) = e^{-A} \int_u d(e^{\varphi(\tau)}\alpha) + \sup_{\rho(\tau)} \int_u \rho(\tau)\, d\tau \wedge \alpha.$$

This total energy is the sum of the $d(e^{\varphi(\tau)}\alpha)$–energy of u, where

$$\varphi(\tau) = \begin{cases} A & \text{if } \tau \leq A, \\ \tau & \text{if } \tau \in [A, B], \\ B & \text{if } \tau \geq B, \end{cases}$$

and the $(d\tau \wedge \alpha)$–energy of $u \cap \{\tau \notin [A, B]\}$, and where the $\rho(\tau)$ are nonnegative bump functions that have compact support contained in precisely one of the subsets

- $(-\infty, A]$, in which case $\int_{\mathbb{R}} \rho(\tau)\, dt = 1$, or in
- $[B, +\infty)$, in which case $\int_{\mathbb{R}} \rho(\tau)\, dt = e^{B-A}$.

Let $h(q_+) > h(q_-) = 0$ be the primitives of $e^\tau \alpha|_{TL}$ at the two sheets of the double point of L, whose size is controlled by the size of $\ell(c_\pm)$; see the difference of Z–coordinates in Figure 4. Applying [13, Proposition 3.11(2) and (3)] to the first and second term of the total energy, we get

$$e^{-A} \int_u d(e^{\varphi(\tau)}\alpha) \leq e^{B-A}\ell(c_+) + e^{-A}(h(q_+) - h(q_-)), \quad \sup_{\rho(\tau)} \int_u \rho(\tau) d\tau \wedge \alpha \leq e^{B-A}\ell(c_+),$$

and thus

$$E_{[A,B]}(u') \leq e^{B-A}\ell(c_+) + e^{-A}(h(q_+) - h(q_-)) + e^{B-A}\ell(c_+),$$

for any J–holomorphic strip u' with its unique positive Reeb chord puncture at c_+, one positive puncture at q, and possibly additional negative Reeb chord punctures. For the computation of the above bound on the energy, we have used that the primitive of $e^\tau \alpha$ pulled back to L can be taken to have compact support.

In view of the above energy bound, and dependence of $(h(q_+) - h(q_-))$ on $\ell(c_\pm)$, we get the following crucial property: total energy of u' can be assumed to be *arbitrarily small*, after shrinking the interval of the isotopy used in the construction of the cobordism (in order for $\ell(c_\pm)$ to become arbitrarily small).

Following the proof of [13, Lemma 5.1] based upon the standard monotonicity property for pseudoholomorphic curves in symplectic manifolds (see [30, Proposition 4.3.1]), there is a constant E_0 such that if u' intersects the subset $\{z = \pm 1\} \subset J^1 B_1$ (which is disjoint from L), then

$$E_{[A,B]}(u') \geq E_0.$$

Under the assumption that the cobordism has been constructed so that $\ell(c_\pm) > 0$ is sufficiently small (while keeping the above coordinates, almost complex structure, and projection of L to $T^*B_{1/3}$ fixed), we can hence conclude that u' is disjoint from some fixed neighborhood of $\{|z| = 1\}$. Since

$$\partial U = \{|z| = 1\} \cup \partial(B_1 \times B_1) \times (-1, 1),$$

the projected curve $v = \pi \circ u'|_{u'^{-1}(U)}$ has boundary $v|_{u'^{-1}(\partial U)}$ contained in $\partial(B_1 \times B_1) \subset T^*B_1$.

In view of the above bound on the $|z|$–coordinate of u', we conclude the following. If v intersects the boundary $\partial(B_{1/3}\times B_1)$, then its symplectic area can be bounded from below by Sikorav's original monotonicity result [30, Proposition 4.7.2]. More precisely, we apply this monotonicity to $J_{T^*B_1}$–holomorphic curves inside in $(B_{1/3}\times B_1,\omega_0)$ with boundary on the transversally intersecting Lagrangian planes

$$\pi(L\cap(B_{1/3}\times B_1\times[-1,1]))\subset B_{1/3}\times B_1.$$

(The Lagrangian planar property of the projection of L is a consequence of formula (4-2).) We conclude that any such v has ω_0–area bounded from below by some constant $C>0$.

The holomorphicity of the projection π, together with the fact that the two-forms $d(e^{\varphi(\tau)}\alpha)$ and $\rho(\tau)\,d\tau\wedge\alpha$ pull back to nonnegative two-forms on any pseudoholomorphic curve for a cylindrical almost complex structure (recall that $\varphi'(\tau),\rho(\tau)\geq 0$), implies the inequalities

$$0\leq\int_v e^A\,d\alpha\leq E_{[A,B]}(u).$$

In particular, we conclude that $E_{[A,B]}(u)$ has a lower bound in terms of the symplectic area of v which, in turn, is bounded from below by $C>0$ in view of the aforementioned monotonicity argument. For $\ell(c_\pm)>0$ sufficiently small, the upper bound on the left-hand side implies that the image of u must be contained entirely inside $B_{1/3}\times B_1\times[-1,1]$.

Since the projected boundary condition $\pi(L\cap(B_{1/3}\times B_1\times[-1,1]))$ consists of two transversely intersecting Lagrangian planes, Stokes' theorem can be applied to show that v must have vanishing ω_0–area. Hence the projection v is constantly equal to $0\in B_{1/3}\times B_1$. This means that u' is contained inside the Reeb orbit that projects to the origin in the above coordinates. The sought claim concerning the uniqueness of $u^\pm$ follows from this.

The claim about the nonexistence of discs with only negative asymptotics is a consequence of Stokes' theorem. □

For a pair of mixed chords x,y and ordered (possibly empty) sets of pure chords $\boldsymbol{z},\boldsymbol{z}'$ define

$$\boldsymbol{m}_L(x^+,y^\pm)=\sum_{\boldsymbol{z},\boldsymbol{z}'}\sharp\mathcal{M}^0(x^+\boldsymbol{z}y^\pm\boldsymbol{z}';L)\varepsilon_-(\boldsymbol{z}\boldsymbol{z}').$$

Lemma 4.12 implies (up to a $q\mapsto\pm q$ basis change)

$$\boldsymbol{m}_L((c_+)^+,q^+)=1=\boldsymbol{m}_L(q^+,(c_-)^-). \tag{4-3}$$

Since a positive (resp. negative) puncture at q means that the boundary of the disk makes a jump from (resp. to) the component $\mathbb{R}\times\Lambda$ at the puncture, we immediately get the vanishing result

$$\boldsymbol{m}_L(q^-,(c_-)^-)=0=\boldsymbol{m}_L((c_+)^+,q^-).$$

Choose δ such that

$$\max_{x_{10}\neq c_-}\ell(x_{10})<\delta<\ell(c_-)<0<\min_{x_{01}}\ell(x_{01}).$$

Let $\mathrm{RFC}_*^{[\delta,b)}(\pm)$ and $\mathrm{RFC}_*^{[a,\delta)}(\pm)$ be the Rabinowitz–Floer complexes for $\overline{\Lambda}_\pm$ with maps $d_{01}^\pm, d_{10}^\pm, B^\pm$ from Definition 4.6.

Define the linear maps

$$\phi_{01}\colon \mathrm{RFC}_*^{[\delta,b)}(+) \to \mathrm{RFC}_*^{[\delta,b)}(-) \quad \text{and} \quad \phi_{10}\colon \mathrm{RFC}_*^{[a,\delta)}(-) \to \mathrm{RFC}_*^{[a,\delta)}(+)$$

via the generators, as follows:

$$\phi_{01}(x_{01}) = \sum_{y_{01}} \boldsymbol{m}_L(x_{01}^+, y_{01}^-)y_{01} + \boldsymbol{m}_L(x_{01}^+, q^+)c_-, \quad \phi_{10}(x_{10}) = \sum_{y_{10}} \boldsymbol{m}_L(x_{10}^-, y_{10}^+)y_{10}. \tag{4-4}$$

Note that $\phi_{10}(c_-) = c_-$, and c_- is not in the domain of either map.

Fix a small $\epsilon > 0$. By Lemma 3.1, we can assume that the time interval of the isotopy, $[-,+]$, is small enough that the following holds. For any pair of distinct chords $\{x, y\} \neq \{c_-, c_+\}$ and any chord $z \notin \{c_-, c_+\}$,

$$|\ell(x^-) - \ell(y^+)| > e^\epsilon\epsilon, \quad |\ell(z^-) - \ell(z^+)| < e^{-\epsilon}\epsilon, \quad |\ell(c_-) - \ell(c_+)| < e^{-\epsilon}\epsilon.$$

Lemma 4.13 *For ϵ sufficiently small and for all $x_{01}, x_{10} \notin \{c_-, c_+\}$,*

$$\langle \phi_{01}x_{01}, x_{01}\rangle = k_{01} \quad \textit{and} \quad \langle \phi_{10}x_{10}, x_{10}\rangle = k_{10}$$

are units.

Proof The exact Lagrangian concordance L_1 shown in Figure 4 has an inverse cobordism L_2 constructed in the same manner, so that the concatenation $L_2 \odot L_1$ is (compactly supported) Hamiltonian isotopic to $\mathbb{R}\times\Lambda^+$. (One can either construct it by the front projection, or use the fact that the noncylindrical component of L_1 can be realized as a Lagrangian trace cobordism as constructed by Proposition B.1.) Let $L^0 = \mathbb{R}\times\overline{\Lambda}^+$ and let $L^1 = L_2 \odot L_1$, such that $L^1 \cap \{|\tau| > \tau_0\}$ agrees with L^0.

For $0 \le s \le 1$, let L^s be a smooth family of exact Lagrangian concordances with only double points of arbitrarily small action, such that $L^s \cap \{|\tau| > \tau_0\}$ agrees with L^0. As in the last statement of Lemma 4.12 (proven by an application of Stokes' theorem and the positivity of the total energy) we conclude that $\mathcal{M}^d(\Gamma, L^s) = \varnothing$ whenever Γ has a double point of L^s (thus, of very small action), no positive Reeb chords, and no negative Reeb at c_+.

$\mathcal{M}^d(\Gamma, L^s) = \varnothing$ implies that if we ignore disks with punctures at double points of the concordance, the trace concordance and inverse-trace concordance induce maps Φ, Ψ which satisfy (3-2). In particular, if we set $x = x_{01}$ (resp. x_{10}) then we get that k_Φ is a unit, arguing as in Case 1 in the proof of Proposition 1.11 in Section 3. But k_Φ is also a count of the disks contributing to $\langle \phi_{01}x_{01}, x_{01}\rangle$ (resp. $\langle \phi_{10}x_{10}, x_{10}\rangle$). □

Proposition 4.14 *The maps ϕ_{01} and ϕ_{10} satisfy axiom* (A1) *with degree ϵ.*

Proof We first verify that $\phi_{01}d_{01}^+ = d_{01}^-\phi_{01}$:

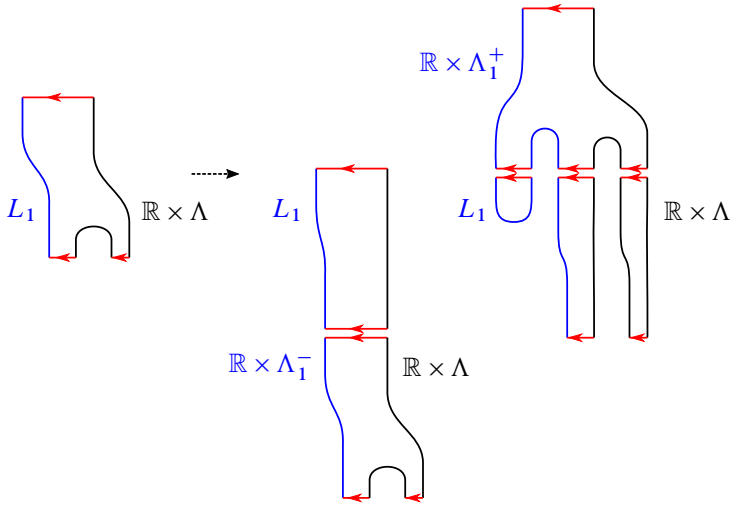


Figure 6: Further examples of boundary points of type $(5\times 2_\pm)$ for the compactification of the one-dimensional moduli space of pseudoholomorphic disks with precisely one positive and precisely one negative mixed Reeb chord asymptotic, and boundary on $L_1\cup(\mathbb{R}\times\Lambda)$. See Figure 3. The components in the middle level that have only pure Reeb chord asymptotics are responsible for pulling back the augmentations under the DGA–morphism induced by the concordance.

$$\begin{aligned}\langle(\phi_{01}d_{01}^+ - d_{01}^-\phi_{01})x_{01}, c_-\rangle &\\ &= \sum_{y_{01}}\langle d_{01}^+x_{01}, y_{01}\rangle\langle\phi_{01}y_{01}, c_-\rangle - \langle\phi_{01}x_{01}, y_{01}\rangle\langle d_{01}^-y_{01}, c_-\rangle\\ &= \sum_{y_{01}}\boldsymbol{m}_{\overline{\Lambda}_+}(x_{01}^+, y_{01}^-)\boldsymbol{m}_L(y_{01}^+, q^+) - \boldsymbol{m}_L(x_{01}^+, y_{01}^-)\boldsymbol{m}_L(q^+, (c_-)^-)\boldsymbol{m}_{\overline{\Lambda}^-}(y_{01}^+, c_-^+).\end{aligned}$$

The term $\boldsymbol{m}_L(q^+,(c_-)^-)=1$ is added to illustrate how the first and second sums are of type $(2_+\times 6)$ and $(2\times 7\times 1_-)$, respectively.

The breaking of curves that involve only pure chords can be disregarded when the strips are counted with augmentations; see the analysis of the so-called "δ–breakings" in [6]. This is due to two different mechanisms. First, a broken configuration such as $(5\times 3_-)$ shown in Figure 3, ie with a disk with only pure punctures in the bottom level (these discs define the differential of the Chekanov–Eliashberg algebra of $\overline{\Lambda}^-$), are canceled algebraically when the count is weighted by the value of the augmentation. (Recall that augmentations vanish on boundaries of the Chekanov–Eliashberg differential by definition.) Second, a disk with only pure punctures in the middle level, shown for instance in Figure 6, plays the role of pulling back the augmentation ε_- under the DGA–morphism induced by the concordance L. In view of Lemma 4.3, this pullback is equal to ε_+, as sought. For more details, see [6, page 416, I and II].

Since $(6\times 3_-)$ corresponds to such an augmentation-related breaking, Proposition 4.2(6) then implies the right-hand side is 0. The relevant boundary of the moduli space is shown in Figure 3.

For $z_{01} \neq c_-$,

$$\langle(\phi_{01}d_{01}^+ - d_{01}^-\phi_{01})x_{01}, z_{01}\rangle = \sum_{y_{01}} \boldsymbol{m}_{\overline{\Lambda}^+}(x_{01}^+, y_{01}^-)\boldsymbol{m}_L(y_{01}^+, z_{01}^-) - \boldsymbol{m}_L(x_{01}^+, y_{01}^-)\boldsymbol{m}_{\overline{\Lambda}^-}(y_{01}^+, z_{01}^-),$$

which are the $(2_+ \times 5)$ and $(5 \times 2_-)$ terms in Proposition 4.2(5). See Figures 3 and 6 for illustrations. There is no (6×7) term since the component with a negative mixed Reeb chord asymptotic to z_{01}^- must have a negative puncture at q, while $\boldsymbol{m}_L(q^-, z_{01}^-) = 0$ holds by Lemma 4.12. The $(5 \times 3_-)$ terms are the augmented terms. Finally,

$$(\phi_{01}d_{01}^+ - d_{01}^-\phi_{01})c_+ = \phi_{01}(0) - d_{01}^-c_- = 0 - 0.$$

Verifying $\phi_{10}d_{10}^- = d_{10}^+\phi_{10}$ has only one computation,

$$\langle(\phi_{10}d_{10}^- - d_{10}^+\phi_{10})x_{10}, z_{10}\rangle = \sum_{y_{10}} \boldsymbol{m}_L(y_{10}^+, x_{10}^-)\boldsymbol{m}_{\overline{\Lambda}^-}(z_{10}^+, y_{10}^-) - \boldsymbol{m}_{\overline{\Lambda}^+}(y_{10}^+, x_{10}^-)\boldsymbol{m}_L(z_{10}^+, y_{10}^-),$$

which are all the terms in Proposition 4.2(5). Again the boundary is depicted in Figure 3, and we use Lemma 4.12 to conclude that $\boldsymbol{m}_L(q^+, x_{10}^-) = 0$ since $x_{10}^- \neq c_-$ holds by assumption.

For $x_{01} \neq c_+$, Lemma 4.13 implies $\langle\phi_{01}x_{01}, x_{01}\rangle = k_{01}$ and $\langle\phi_{10}x_{10}, x_{10}\rangle = k_{10}$ are units. So to construct (strict, not just homotopy) inverses ψ_{01}, ψ_{10} for ϕ_{01}, ϕ_{10}, it suffices to set

$$\begin{aligned}
\psi_{01}(x_{01}) &= k_{01}^{-1}x_{01} - \sum_{y_{01}\neq x_{01}} \langle\phi_{01}(x_{01}), y_{01}\rangle y_{01} - \langle\phi_{01}(x_{01}), c_-\rangle c_+,\\
\psi_{01}(c_-) &= c_+,\\
\psi_{10}(x_{10}) &= k_{10}^{-1}x_{10} - \sum_{y_{10}\neq x_{10}} \langle\phi_{10}(x_{10}), y_{10}\rangle y_{10}.
\end{aligned}$$

It is easy to check that these are chain maps.

Stokes' theorem finally bounds all the maps' degrees by $\epsilon > 0$. □

Proposition 4.15 *The maps B^- and B^+ are chain maps which satisfy axiom* (A2)*. That is, $\phi_{10}B^+\phi_{01}$ is homotopic to B^-, where $B^\pm$ are of degree ϵ.*

Proof To prove axiom (A2), define (on generators, then extend linearly) the map

$$H : \mathrm{RFC}_*^{[\delta,b)}(+) \to \mathrm{RFC}_{*+1}^{[a,\delta)}(+), \quad \text{defined by } H := H_\alpha + H_\beta,$$

where

$$\langle H_\alpha w_{01}, v_{10}\rangle := \boldsymbol{m}_L(w_{01}^+, v_{10}^+) \quad \text{and} \quad \langle H_\beta w_{01}, v_{10}\rangle := \boldsymbol{m}_L(v_{10}^+, (c_-)^-)\langle\phi_{01}w_{01}, c_-\rangle. \tag{4-5}$$

The degree of H is tautologically bounded from above by 0. It suffices to show that

$$\langle(\phi_{01}B^-\phi_{01} - B^+)x_{01}, y_{10}\rangle = \langle(d_{10}^+H + Hd_{01}^+)x_{01}, y_{10}\rangle. \tag{4-6}$$

Apply Proposition 4.2(4) to analyze $\partial\mathcal{M}^1(x_{01}^+ \boldsymbol{z}\, y_{10}^+ \boldsymbol{z}'; L)$ for possibly empty words of pure cords $\boldsymbol{z}, \boldsymbol{z}'$. The broken configurations that constitutes this boundary are shown in Figure 2. If the count these boundary strata weighted by augmentations, then the strata of type $(4 \times 3_-)$ all cancel; for this reason, these configurations can be ignored.

The right-hand side of (4-6) corresponds to the boundary strata of the following types:

- $(2_+ \times 4)$ These boundary points correspond to all terms in $d_{10}^+ \circ H_\alpha - H_\alpha \circ d_{01}^+$.
- $(5 \times 5 \times 1_-)$ With the additional requirement that the component in (1_-) has one positive mixed Reeb chord asymptotic to c_-, these boundary points correspond to the terms in $H_\beta \circ d_{01}^+$. (Here we have used the chain map property of ϕ_{01} established in Proposition 4.14 together with equation (4-5).)

However, we also have:

- (†) The broken configurations that correspond to $d_{10}^+ \circ H_\beta$, which are of type $(2_+ \times 5)$ (which, thus, are not boundary points of (4)).

We continue by analyzing the left-hand side of equation (4-6):

- (1_+) These components equal the $\langle B^+ x_{01}, y_{10}\rangle$ term on the left-hand side.

By unraveling the definitions, the remaining term on the left-hand side is computed to be

$$\begin{aligned}
\phi_{10} B^- \phi_{01}(x_{01}) &= \phi_{10} B^- \Big(\boldsymbol{m}_L(x_{01}^+, q^+) c_- + \sum_{y_{01}} \boldsymbol{m}_L(x_{01}^+, y_{01}^-) y_{01} \Big) \\
&= \phi_{01} \Big(\sum_{z_{10}} \Big[\boldsymbol{m}_{\overline{\Lambda}^-}(z_{10}^+, (c_-)^-) \boldsymbol{m}_L(x_{01}^+, q^+) + \sum_{y_{01}} \boldsymbol{m}_{\overline{\Lambda}^-}(y_{01}^+, z_{10}^+) \boldsymbol{m}_L(x_{01}^+, y_{01}^-) \Big] z_{10} \Big) \\
&= \sum_{z_{10}, w_{10}} \Big[\boldsymbol{m}_L(w_{10}^+, z_{10}^-) \boldsymbol{m}_{\overline{\Lambda}^-}(z_{10}^+, (c_-)^-) \boldsymbol{m}_L(x_{01}^+, q^+) \\
&\qquad + \sum_{y_{01}} \boldsymbol{m}_L(w_{10}^+, z_{10}^-) \boldsymbol{m}_{\overline{\Lambda}^-}(y_{01}^+, z_{10}^+) \boldsymbol{m}_L(x_{01}^+, y_{01}^-) \Big] w_{10}.
\end{aligned}$$

Here we find the remaining types of broken configurations that arise in the boundary $\partial\mathcal{M}^1(x_{01}^+ \boldsymbol{z}\, y_{10}^+ \boldsymbol{z}'; L)$:

- $(5 \times 5 \times 1_-)$ With the additional constraint that no Reeb chord asymptotic of the component 1_- is asymptotic to c_-, these types correspond to the second term on the right-hand side.

What remains is the term (6×6). For this we need to analyze the term

$$\boldsymbol{m}_L(w_{10}^+, z_{10}^-) \boldsymbol{m}_{\overline{\Lambda}^-}(z_{10}^+, (c_-)^-) \boldsymbol{m}_L(x_{01}^+, q^+)$$

in the expression $\phi_{10} B^- \phi_{01}(x_{01})$ further. We start by using Proposition 4.2(5) to rewrite this as

$$(\boldsymbol{m}_{\overline{\Lambda}^+}(w_{10}^+, z_{10}^-) \boldsymbol{m}_L(z_{10}^+, (c_-)^-) + \boldsymbol{m}_L(w_{10}^+, q^-) \boldsymbol{m}_L(q^+, (c_-)^-) \boldsymbol{m}_L(x_{01}^+, q^+).$$

These broken strata are shown in Figure 3. Since $\boldsymbol{m}_L(q^+,(c_-)^-)=1$ holds by Lemma 4.12, the previous expression can be simplified even further. We have thus found:

(6×6) This corresponds to the second term in the latter expression, ie the term $\boldsymbol{m}_L(w_{10}^+,q^-)\boldsymbol{m}_L(x_{01}^+,q^+)$.

The first term $\boldsymbol{m}_{\overline{\Lambda}^+}(w_{10}^+,z_{10}^-)\boldsymbol{m}_L(z_{10}^+,(c_-)^-)\boldsymbol{m}_L(x_{01}^+,q^+)$ of the latter expression remains to be taken care of, since it is not inside boundary $\partial\mathcal{M}^1(x_{01}^+\boldsymbol{z}y_{10}^+\boldsymbol{z}';L)$. However, these terms cancel with the remaining contribution on the right-hand side of equation (4-6), since:

(†) The first term

$$\boldsymbol{m}_{\overline{\Lambda}^+}(w_{10}^+,z_{10}^-)\boldsymbol{m}_L(z_{10}^+,(c_-)^-)\boldsymbol{m}_L(x_{01}^+,q^+)$$

is equal to $d_{10}^+\circ H_\beta$ by formula (4-5) (it is not a part of the boundary of the moduli space (4)).

We have thus established equation (4-6), as sought. □

4.3.2 No double points Consider the summands $C_*(\pm)=\mathrm{RFC}_*^{[0,b)}(\pm)$ and $C^*(\pm)=\mathrm{RFC}_*^{(a,0]}(\pm)$. We define ϕ_{01} and ϕ_{10} as in equation (4-4), noting that $\boldsymbol{m}_L(x_{01}^+,q^-)=0$ since there is no double point q. The rest of the arguments are essentially simplifications of the ones above, since curves of type (6) and (7) do not exist. Note that bifurcations can occur that were not considered in Section 4.3.1. Notably, a pair of mixed chords may be born or die, making one of the above pair of RFC–complexes nonisomorphic. But even in the event of a birth (death is a reverse-time birth), the statement and proof of Lemma 4.13 still apply. To recap our argument: since there are no double points, we can apply Proposition 1.11 to replace the DGA–morphism induced by the Lagrangian concordance with a stable-tame isomorphism (STI) of DGAs. Lemma 4.3 equates how the pullback of the augmentation induced by the concordance is the same as the change in augmentation induced by the STI in [17, Lemma 3.4]. Thus, as outlined shortly after stating Theorem 4.8, we can apply the techniques of [17, Section 3], to both complexes $C_*(\pm)$ and $C^*(\pm)$, to prove Theorem 4.8.

5 Proofs of Theorems 1.4, 1.6 and 1.9

We need the following variation of [17, Lemma 3.1], which allows us to estimate the oscillation of a contact Hamiltonian based on the change of lengths of a pair of Reeb chords.

Lemma 5.1 *Consider a smooth one-parameter family $c(t)\subset(Y,\alpha)$ of Reeb chords with boundary on $a(t)\in\Lambda_0, b(t)\in\Lambda_1(t)$, where $\Lambda_0\subset Y$ is a fixed Legendrian submanifold and $\Lambda_1(t)\subset Y$ is a Legendrian isotopy. (Here a is either the endpoint or the starting point, and $c(t)$ is allowed to be of zero length, ie a double point $\Lambda_0\cap\Lambda_1(t)$.) Then*

$$\frac{d}{dt}\ell(c(t))=\alpha(X_{b(t)}(t))=H_t(b(t)), \tag{5-1}$$

where $X_{b(t)} \in TY$ is the contact vector field that generates $\Lambda_1(t)$. In particular, if $c(t)$ and $d(t)$ are two Reeb chords as above, then

$$\big|\big(\ell(c(0)) - \ell(d(0))\big) - \big(\ell(c(1)) - \ell(d(1))\big)\big| \leq \|H_t\|_{\mathrm{osc}}$$

is satisfied.

Consider a smooth one-parameter family $c_1(t)$ of pure Reeb chords with initial and terminal endpoints at $a_1(t), b_1(t) \in \Lambda_1(t)$. Then

$$\frac{d}{dt}\ell(c_1(t)) = \alpha\big(X_t(b_1(t))\big) - \alpha\big(X_t(a_1(t))\big) = H_t(b_1(t)) - H_t(a_1(t)).$$

Hence, the inequality $|\ell(c_1(1)) - \ell(c_1(0))| \leq \|H_t\|_{\mathrm{osc}}$ holds.

Proof The calculation for the pure chords c_1 was proven in [17, Lemma 3.1] for general contact manifolds, so we only need to verify the computation for c.

When $c(t)$ has positive length, we may perform the computation for a contact vector field X that can be taken to vanish along all of Λ_0 after cutting off the contact Hamiltonian in some neighborhood. The computation of $\frac{d}{dt}\ell(c(t))$ is then a direct application of [17, Lemma 3.1].

In the case when $c(t)$ is of length zero, we replace Λ_0 by its image $\phi^{\epsilon}_{R_\alpha}(\Lambda_0)$ under the time-ϵ Reeb flow, and instead compute $\frac{d}{dt}\ell(\widetilde{c}(t))$ for the induced chord between $\phi^{\epsilon}_{R_\alpha}(\Lambda_0)$ and $\Lambda_1(t)$. Here $\widetilde{c}(t)$ corresponds to $c(t)$ under the natural bijective correspondence of Reeb chords between, on one hand, Λ_0 and $\Lambda_1(t)$ and, on the other hand, $\phi^{\epsilon}_{R_\alpha}(\Lambda_0)$ and $\Lambda_1(t)$. Under this bijection, gotten by "removing" the first time-ϵ portion, the chord lengths differ by the constant ϵ.

The inequality involving the difference of lengths and oscillation of the contact Hamiltonian is finally proven by the following computation:

$$\begin{aligned}\big|\big(\ell(c(0)) - \ell(d(0))\big) - \big(\ell(c(1)) - \ell(d(1))\big)\big| &\leq \int_0^1 \Big|\frac{d}{dt}\ell(c(t)) - \frac{d}{dt}\ell(d(t))\Big|\,dt \\ &= \int_0^1 |H_t(c^+(t)) - H_t(d^+(t))|\,dt \\ &\leq \int \big(\max_{y\in Y} H_t - \min_{y\in Y} H_t\big)\,dt = \|H_t\|_{\mathrm{osc}},\end{aligned}$$

where $c^+(t)$ and $d^+(t)$ are the endpoints of the corresponding chords that lie on the component $\Lambda_1(t)$. □

5.1 Proof of the Main Theorem

Recall the notation from Theorem 1.4. A Legendrian $\Lambda \subset (Y, \alpha)$ is moved by a contact isotopy $\phi^t_{\alpha,H}$, with $0 \leq t \leq 1$. The Hamiltonian $H_t\colon Y \to \mathbb{R}$ satisfies $\|H_t\|_{\mathrm{osc}} < \min\{l, \ell(c_k)\}$. Here $0 < l \leq \infty$ is chosen such that there exists an augmentation $\varepsilon\colon \mathscr{A}^l(\Lambda) \to \boldsymbol{k}$, and c_k is the Reeb chord with the k^{th} shortest length.

Let $1 \gg \epsilon > 0$ be a constant to be determined. Let $\Lambda_0 = \Lambda_0^t = \Lambda$ and $\Lambda_1^t = \phi_{\alpha,H}^t(\Lambda_0')$, where Λ_0' is a perturbation of a push-off (of flow ϵ in the positive Reeb direction) of Λ_0. Then let $\overline{\Lambda}^t = \Lambda_0^t \cup \Lambda_1^t$.

First set $-a_0 = 2\epsilon = b_0$ and $l(t) = \int_0^t \max H_\tau - \min H_\tau \, d\tau$. We claim that $\mathrm{RFC}_*^{[a_t, b_t)}(\Lambda_0, \Lambda_1^0; \varepsilon_0)$ (notation as in Definition 4.6) is quasi-isomorphic to the Morse complex of Λ_0. This can be seen since $C^*(a_0, 0] = 0$ and d_{01}^0 counts only holomorphic strips which approximate the gradient flows of the function which represents Λ_1^0 graphically in a small $J^1\Lambda_0^0$ neighborhood of Λ_0^0 for a suitable cylindrical almost complex structure; see eg [14].

We wish to adapt our situation in order to replicate [17, Section 3.4] as much as possible, which proves the special case of Main Theorem, Theorem 1.4, when $(Y = P \times \mathbb{R}_z, \alpha = dz - \theta)$. In that case, due to the global ∂_z Reeb flow, we could choose a large $N \gg 1$ push-off instead of a small $\epsilon \ll 1$ one. The complex considered in the previous setup is somewhat simpler since there never are any x_{10} chords throughout the isotopy (or chords of zero length). The chords at $t = 0$ in the action window $[N - 2\epsilon, N + 2\epsilon]$ form a subcomplex which is quasi-isomorphic to the Morse complex of Λ_0.

The action of chords ($\approx N$ versus ≈ 0) which generate a Morse complex, and the generators of the complex (type x_{01} versus type x_{01}, x_{10}) are the only distinctions between the existing argument for the Main Theorem, Theorem 1.4, when $(Y = P \times \mathbb{R}_z, \alpha = dz - \theta)$ presented in [17, Section 3.4], and the needed argument for the Main Theorem when (Y, α) is arbitrary.

We briefly sketch this argument, deferring full details to [17, Section 3.4]. Label the Morse generators $x_1, \dots, x_m$. For each family $x_i(t)$, consider families of two barcodes which are approximately based on action windows $[\ell(x_i(t)) - (l - l(t)) + \epsilon, \ell(x_i(t)) + \epsilon]$ and $[\ell(x_i(t)) - \epsilon, \ell(x_i(t)) + (l - l(t)) - \epsilon]$ as specified in [17, equations (3.6) and (3.7)]. Then [17, Lemma 3.7] implies that in each action window we see an infinite bar that starts at $\ell(x_i(t))$. Theorem 4.8 implies that we can use the barcode proposition (Proposition 2.4) for this isotopy. By looking at both action windows simultaneously during this isotopy, we verify that the bar can only disappear in the amount of time it takes $\ell(x_i(t))$ to coincide with one of the values of some other mixed chord that was not part of the original Morse complex. The rate of change of the difference of these action values is controlled by $\frac{d}{dt} l(t)$ as calculated in Lemma 5.1, which brings in the $\|H_t\|_{\mathrm{osc}}$ term as stated in the Main Theorem (Theorem 1.4). This sketch glosses over the difference between the endpoints of the bars in a barcode, and their representations as Reeb chords. However, this important distinction has been addressed in case (ii) in [17, Section 3.4], and the argument there is independent of the contact manifold and form.

5.2 Proof of Theorem 1.6

Rosen and Zhang prove that, for any submanifold $N \subset Y$ of $\dim(N) = n$, the distance δ_α is either nondegenerate or $\delta_\alpha \equiv 0$; see [39, Theorem 1.9]. So it suffices to find Λ_1 and Λ_2 which are the images of contact isotopies of Λ, and for which $\delta_\alpha(\Lambda_1, \Lambda_2) > 0$.

Consider any closed Legendrian Λ and fix a contact form α on Y. Let Λ_i, $i=1,2$, be given by $j^1(\epsilon^2\cdot i\cdot f)$ in a standard contact-form-preserving jet-neighborhood of the form

$$D_{\leq r}T^*\Lambda\times[-5\epsilon,5\epsilon]_z\subset(J^1\Lambda, dz-p\,dq)$$

inside Y which identifies Λ with $j^1 0$, where $f:\Lambda\to[0,1]$ is a Morse function, and $0<\epsilon\ll 1$. Let

$$0<\min(f)=m_1<\cdots<m_k=\max(f)$$

be an enumeration of the critical values of f that we, moreover, assume correspond to distinct critical points. After postcomposing the Morse function with a suitable change of coordinates, we may assume

$$2\min(f)>\max(f)>0. \tag{5-2}$$

Using the notation of Definition 4.6 for the Rabinowitz–Floer complex with action window

$$\mathrm{RFC}_*^{[a_t,b_t)}(\Lambda,\Lambda_t;\varepsilon_t)$$

we choose the following: the action window is constantly equal to $[a_t,b_t)=[-2\epsilon,+2\epsilon)$; the initial action threshold, $l=6\epsilon$, is sufficiently smaller than the length of any contractible periodic Reeb orbit γ of degree $|\gamma|\leq 1$; by the existence of the above standard neighborhood, all Reeb chords of Λ and Λ_i that start and end on the same component have length at least 7ϵ, so ε_0 is necessarily trivial; and Λ_t is an isotopy from Λ_1 to Λ_2 with oscillation $l(1)\leq\epsilon$.

If we had studied a Legendrian isotopy from Λ_1 to Λ_2 generated by a Hamiltonian H_t such that $l(1)$ no longer satisfies $l-l(1)>b_1-a_1$ (so that our technology breaks down), we automatically get a desired lower bound

$$\int_0^1\max|H_t|\,dt\geq\tfrac{1}{2}\|H_t\|_{\mathrm{osc}}>\tfrac{1}{2}l(1)\geq\tfrac{1}{2}(l-(b_1-a_1))=\tfrac{1}{2}(6\epsilon-4\epsilon)=\epsilon.$$

Assume we can apply Theorem 4.8; thus the isotopy deforms the barcode of $\mathrm{RFC}_*^{[-2\epsilon,+2\epsilon)}(\Lambda,\Lambda_1;\varepsilon_1)$ to that of $\mathrm{RFC}_*^{[-2\epsilon,+2\epsilon)}(\Lambda,\Lambda_2;\varepsilon_2)$ via the bifurcations specified by the barcode proposition, Proposition 2.4.

During this isotopy, one of the following scenarios occur:

Case 1 Some starting/end point of a bar in the barcode of $\mathrm{RFC}_*^{[-2\epsilon,+2\epsilon)}(\Lambda,\Lambda_1;\varepsilon_1)$ survives, and by inequality (5-2), moves at least a distance $\epsilon^2(2\min(f)-\max(f))>0$.

Case 2 Some bar in the barcode dies or escapes the action window. Note that the concerned bars all are of length at least $\min\{\epsilon^2(m_2-m_1),\dots,\epsilon^2(m_k-m_{k-1})\}$ by the assumption made on distinct critical values. (In fact, $\mathrm{RFC}_*^{[-2\epsilon,+2\epsilon)}(\Lambda,\Lambda_i;\varepsilon_i)$ for $i=1,2$ are the Morse complexes that compute the singular homology $H_*(\Lambda;\mathbb{Z}_2)$. So even more can be said about the barcode, but we do not use this.) Theorem 4.8 together with the barcode proposition, Proposition 2.4, then implies that

$$\int_0^1\max|H_t|\geq C$$

holds for this Hamiltonian that generates a Legendrian isotopy from Λ_1 to Λ_2, where

$$C := \min\{\epsilon^2(2\min(f) - \max(f)), \epsilon^2(m_2 - m_1), \dots, \epsilon^2(m_k - m_{k-1})\} > 0$$

is a constant that only depends on the contact form α and the Legendrians $\Lambda, \Lambda_1, \Lambda_2$. Here we use Lemma 5.1 to relate change in Reeb chord length and the value of the contact Hamiltonian H_t at the endpoint of the Reeb chord of Λ_t that corresponds to the moving bar.

We conclude that $\delta_\alpha(\Lambda_1, \Lambda_2) > C$ holds, as sought. □

5.3 Proof of Theorem 1.9

Let $\Lambda_0^t = \phi^t_{\alpha, H_t}(\Lambda_0)$, with $H_t \geq c > 0$, $\Lambda_1^t = \Lambda_1$ and $\overline{\Lambda}^t = \Lambda_0^t \cup \Lambda_1^t$. Recall the assumptions that Λ_0, Λ_1 are augmented and that no pseudoholomorphic plane appears in the SFT–compactifications from [3]; see Remarks 1.5 and 1.12. Thus we can set $l = +\infty$ and consider the sequence of complexes $\mathrm{RFC}_*(\Lambda_0^t, \Lambda_1)$, whose barcode in any finite action window is well defined for all $t \geq 0$ and varies continuously as in Theorem 4.8. Since each finite bar is a pairing of two mixed chords of relative grading 1, the hypothesis implies the barcode at $t = 0$ either has an infinite bar with a starting point of positive action, or a finite bar with starting point of negative action, and endpoint of positive action. The interlinkedness property then follows by the continuity of the barcode, together with Lemma 5.1. Namely, according to the latter, all endpoints of bars moves in the direction of decreasing action, with a speed bounded from below by $c > 0$. (Here we use a large enough action window for the given Hamiltonian.)

6 Computations for the standard Legendrian $\mathbb{R}P^n$ (proof of Theorem 1.10)

This section concerns computations of the Rabinowitz–Floer complex for the standard Legendrian $\mathbb{R}P^n \subset \mathbb{R}P^{2n+1}$, with the goal of proving Theorem 1.10.

Recall the definition

$$\Lambda_0 := S^{2n+1} \cap \mathfrak{Re}\, \mathbb{C}^{n+1} \subset (S^{2n+1}, \xi_{\mathrm{st}})$$

of the standard Legendrian sphere inside the standard contact sphere, as defined in Section 1. By taking the quotient under the antipodal map we obtain the standard Legendrian embedding

$$\widetilde{\Lambda}_0 := \Lambda_0/\mathbb{Z}_2 \subset S^{2n+1}/\mathbb{Z}_2 = \mathbb{R}P^{2n+1}$$

of $\mathbb{R}P^n$ into the standard contact projective space.

Remark 6.1 It is possible to also pass to further quotients $S^{2n+1}/\mathbb{Z}_k$ with $k \geq 2$ in order to obtain Legendrian embeddings in the standard contact (higher-dimensional analogues of) "Lens spaces". Most of the analysis carried out in this section should be possible to apply with only minor modifications in order to derive similar results also for general even-dimensional lens spaces, ie $k = 2m \geq 2$. However, we do not pursue this direction further.

Similarly, any linear Lagrangian subspace $V^{n+1} \subset \mathbb{C}^{n+1}$ gives rise to a Legendrian embedding

$$\Lambda_V := S^{2n+1} \cap V \subset (S^{2n+1}, \xi_{\text{st}})$$

of S^n, and hence a corresponding Legendrian embedding

$$\widetilde{\Lambda}_V := \Lambda_V/\mathbb{Z}_2 \subset (\mathbb{R}P^{2n+1}, \xi_{\text{st}})$$

of $\mathbb{R}P^n$. (Note that $\widetilde{\Lambda}_0 = \widetilde{\Lambda}_{\mathfrak{Re}\,\mathbb{C}^{n+1}}$.) All of these different Legendrian embeddings are clearly Legendrian isotopic via an ambient contact isotopy that preserves the round contact form. In particular, we obtain a C^∞–small Legendrian push-off of $\widetilde{\Lambda}_0$ by considering $\widetilde{\Lambda}_V$ for a Lagrangian plane $V \subset \mathbb{C}^{n+1}$ which is sufficiently close to $\mathfrak{Re}\,\mathbb{C}^{n+1}$.

Recall that the round contact S^{2n+1} is equipped with the contact form $\alpha_{\text{st}} = \frac{1}{2}\sum_i (x_i\, dy_i - y_i\, dx_i)$, and a time-$t$ Reeb flow given by complex scalar multiplication by e^{i2t}. Hence S^{2n+1} is foliated by simple closed Reeb orbits, all of whose periods are equal to π. It follows that the round $\mathbb{R}P^{2n+1}$ is foliated by simple closed Reeb orbits of period $\frac{1}{2}k\pi$ for each $k = 1, 2, 3, \dots$. The orbits of period $\frac{1}{2}k\pi$ form a manifold $\Gamma_k \cong \mathbb{C}P^n$ of Reeb chords which are nondegenerate in the Bott sense; see the work [2] by Bourgeois. Note that these Reeb orbits are contractible if and only if $k = 2m$.

The Reeb chords on $\widetilde{\Lambda}_V$ all come in connected families $\mathcal{Q}(\widetilde{\Lambda}_V)_k^{\text{Bott}} \cong \mathbb{R}P^n$ labeled by the Reeb chord lengths $\frac{1}{2}k\pi$, $k = 1, 2, 3, \dots$. These families are also smooth manifolds which are nondegenerate in the Bott sense.

In Sections 6.1 and 6.2 below we will compute the Conley–Zehnder indices of these orbits and chords. In addition, we compute the Conley–Zehnder indices of the chords and orbits after a generic perturbation by a Morse function on the Bott manifolds as described in [2], which makes the contact form nondegenerate. The conclusion of Proposition 6.4 in Section 6.1 is that the Chekanov–Eliashberg DGAs of the standard Legendrian $\mathbb{R}P^n \subset \mathbb{R}P^{2n+1}$ is well defined for the aforementioned nondegenerate perturbations of the round contact form. In Proposition 6.8 from Section 6.2 the degree of the Reeb chords on $\mathbb{R}P^n$ are computed. This is useful for showing that $\mathbb{R}P^n$ admits augmentations, which of course is crucial for defining the Rabinowitz–Floer complex in Proposition 6.12.

6.1 Conley–Zehnder index of a periodic Reeb orbit

Here we perform Conley–Zehnder index computations for the periodic Reeb orbits on the round $\mathbb{R}P^{2n+1}$, as well as its nondegenerate perturbations. See Section A.1 for the definition of the Conley–Zehnder index.

Recall that precisely the even covers of the simple orbits are contractible. In addition, since the universal cover of $\mathbb{R}P^{2n+1}$ is S^{2n+1}, it follows that $\pi_2(\mathbb{R}P^{2n+1}) = 0$, and any two planes with the same asymptotic Reeb orbit are homotopic through planes of the same kind. Denote by $\pi\colon \mathbb{R}P^{2n+1} \to \mathbb{C}P^n$ the prequantum bundle projection, for which $D\pi|_\xi$ is injective. For any contractible Reeb orbit $\gamma \in \Gamma_{2m}$,

take the trivialization of ξ along the Reeb orbit γ that is induced by pulling back a symplectic frame at the point $\pi(\gamma) \in \mathbb{C}P^n$ under the bundle projection π.

We will apply the index formula (A-1) to prove the following.

Lemma 6.2 *Consider a plane $u\colon \mathbb{C} \to \mathbb{R} \times \mathbb{R}P^{2n+1}$ which is asymptotic to $\gamma \in \Gamma_{2m}$. The relative first Chern class with respect to the above choice of trivialization satisfies*

$$c^{\xi}_{1,\mathrm{rel}}[u] = c_1^{\mathbb{C}P^n}[\pi \circ u] = m(n+1) \quad \text{for } m \geq 1,$$

where $c_1^{\mathbb{C}P^n}$ is the usual first Chern class and the detailed definition of $c^{\xi}_{1,\mathrm{rel}}$ is in Section A.1.

Here we identify $\pi \circ u$ with a sphere that is homologous to $m \cdot L$, where $L \in H_2(\mathbb{C}P^n)$ denotes the homology class of a line.

Proof By Stokes' theorem the symplectic area of the chain $\pi \circ u$ in $\mathbb{C}P^n$ is equal to the length of the periodic Reeb orbit. Here we have endowed $\mathbb{C}P^n = \mathbb{R}P^{2n+1}/S^1$ with the symplectic form induced by the curvature $d\alpha$ of the prequantization form α on $\mathbb{R}P^{2n+1}$ via symplectic reduction. (With these conventions, the area of a line in $\mathbb{C}P^n$ is equal to π.) Since $\mathbb{C}P^n$ is monotone, ie the area and index are proportional, the second equality follows. The first equality then follows since

$$D\pi|_{\xi}\colon \xi \to T\mathbb{C}P^n$$

is a symplectic bundle morphism which is an isomorphism on the fibers. □

Lemma 6.3 *The Conley–Zehnder index of $\gamma \in \Gamma_{2m}$ with respect to the above choice of trivialization is equal to*

$$\mu_{\mathrm{CZ}}(\boldsymbol{A}_{\gamma} - \delta \cdot \mathrm{id}) = n, \tag{6-1}$$

independently of the multiplicity $2m$.

After a perturbation by a Morse function on the Bott manifold as in [2], *the nondegenerate orbit that corresponds to a critical point has Conley–Zehnder index*

$$\mu_{\mathrm{CZ}}(\boldsymbol{A}_{\gamma} - \delta \cdot \mathrm{id}) + i - 2n = i - n,$$

where $i \in \{0, 1, \dots, \dim \Gamma_{2m} = 2n\}$ is the Morse index of the critical point.

Proof The computation in the Bott setting was performed in [45, (3.10)]. Alternatively, one can argue as follows. The linearized Reeb flow in this trivialization is constantly equal to the identity map. Hence, the aforementioned perturbation of the Reeb flow by a small positive rotation $s \mapsto e^{i\delta s}\,\mathrm{id}$ in the contact planes perturbs the nondegenerate return map to a nondegenerate one. We then compute the classical Conley–Zehnder index of this path, as defined in [38], which gives $\frac{1}{2}(2n) = n$, as sought.

The Conley–Zehnder index after the perturbation by a Morse function on the Bott manifold follows from formula (A-2). Recall that $\dim \Gamma_{2m} = 2n$ is independent of m. □

In conclusion, we have shown the following.

Proposition 6.4 *For a plane $u\colon \mathbb{C} \to \mathbb{R}\times\mathbb{R}P^{2n+1}$ which is asymptotic to a Reeb orbit in the family Γ_{2m} for some $m \geq 1$, for the round contact form we have the identity*

$$\operatorname{index}(u) = ((n+1)-3) + n + 2m(n+1) = (2m+2)(n+1) - 4$$

for the expected dimension of the moduli space of unparametrized pseudoholomorphic planes of the same type (with asymptotics that are free to vary in the Bott family Γ_{2m}).

Moreover, after a small perturbation of the contact form by a Morse function defined on the Bott manifolds, as constructed in [2], *all periodic Reeb orbits may be assumed to be nondegenerate, and to satisfy the bound*

$$|\gamma| \geq 4(n+1) - 4 - 2n = 2n$$

on their degrees.

Proof The result follows from a direct application of the index formula (A-1) combined with the above computations of the relative first Chern class and Conley–Zehnder indices. □

6.2 Conley–Zehnder index of a pure Reeb chord

The next step is to compute the Conley–Zehnder indices for the pure Reeb chords $\mathcal{Q}(\widetilde{\Lambda}_V)_k$ on Λ_V, both for the round contact form and after a nondegenerate perturbation. The definition of the Conley–Zehnder index for Reeb chords will be recalled in Section A.2.

First, note that the Reeb chords on $\mathbb{R}P^n \subset \mathbb{R}P^{2n+1}$ are all null-homotopic when considered as elements in $\pi_1(\mathbb{R}P^{2n+1}, \mathbb{R}P^n)$ when $n \geq 1$. This follows since any Reeb chord lifts to a Reeb chord on the standard Legendrian sphere under the universal cover

$$(S^{2n+1}, S^{2n+1} \cap \mathfrak{Re}\mathbb{C}^{n+1}) \to (\mathbb{R}P^{2n+1}, \mathbb{R}P^n),$$

where both the map and the restriction $S^{2n+1} \cap \mathfrak{Re}\mathbb{C}^{n+1} \cong S^n \to \mathbb{R}P^n$ are universal covers. Moreover, any element in $\pi_2(\mathbb{R}P^{2n+1}, \mathbb{R}P^n)$ lifts to an element in $\pi_2(S^{2n+1}, S^n)$ under this map, where the latter group vanishes whenever $n > 1$. Hence the Maslov class automatically vanishes on $\pi_2(\mathbb{R}P^{2n+1}, \mathbb{R}P^n)$ when $n \geq 2$. This is also the case for $n = 1$, by a standard calculation.

Contractibility of a Reeb chord is equivalent to the existence of a continuous half-plane

$$u\colon (\boldsymbol{H}, \partial\boldsymbol{H}) \to (\mathbb{R}\times\mathbb{R}P^{2n+1}, \mathbb{R}\times\mathbb{R}P^n)$$

with boundary condition on the cylinder over the Legendrian $\mathbb{R}P^n$ and puncture asymptotic to the Reeb chord. It follows similarly to the case of planes discussed above that two half-planes that share the same asymptotics are homotopic through half-planes of the same type when $n \geq 2$; indeed, $\pi_2(S^{2n+1}, S^n) \cong \pi_1(S^n) = 0$ whenever $n \geq 2$. In the case $n = 1$, $\pi_2(S^3, S^1) \cong \pi_1(S^1) = \mathbb{Z}$ and there are infinitely many homotopy classes of planes asymptotic to any given family of Reeb orbits. The implication is the following.

Lemma 6.5 *Any Reeb chord c is the asymptotic of a half-plane, and for any two pseudoholomorphic half-planes u_1 and u_2 asymptotic to c, we have* $\operatorname{index}(u_1) = \operatorname{index}(u_2)$. *Here,* $\operatorname{index}(u)$ *denotes the Fredholm index of a linearization of the $\overline{\partial}_J$ at u, viewed as an element of a standard Banach space of maps.*

The Reeb chords on the Legendrians $\widetilde{\Lambda}_V$ for the round contact form come in Bott families $\mathcal{Q}(\widetilde{\Lambda}_V)_k^{\mathrm{Bott}} \cong \mathbb{R}P^n$, where the images of these chords coincide with k–fold multiples of the simply covered periodic Reeb orbits for $k = 1, 2, 3, \ldots$. The index formula for a pseudoholomorphic half-plane inside $\mathbb{R}\times\mathbb{R}P^{2n+1}$ with boundary on $\widetilde{\Lambda}_V$ and asymptotics to the chord $c \in \mathcal{Q}(\widetilde{\Lambda}_V)_k^{\mathrm{Bott}}$ (without a constraint at a fixed asymptotic orbit) is given by formula (A-3) in Section A.2. Namely,

$$\operatorname{index}(u) = (\mathrm{CZ}(c) - 1) + \mu_{\mathbb{R}\times\mathbb{R}P^n}[u],$$

where $\mathrm{CZ}(c)$ is the Conley–Zehnder index of the path of Legendrian planes along c; see Section 6.2.

In order to compute the Conley–Zehnder index and the relative Maslov class, we need to make the choice of a capping path (up to homotopy) as described in Section A.2. Since the Reeb flow is totally periodic, we will simply choose the constant capping path.

Lemma 6.6 *For the constant capping path, we have the identity*

$$\mu_{\mathbb{R}\times\mathbb{R}P^n}[u] = \mu_{\mathbb{R}P^n}(\pi\circ u) = k(n+1) \quad \text{for some } k \geq 1,$$

where $\mu_{\mathbb{R}P^n}$ is the classical Maslov index for a disc with boundary on the Lagrangian $\mathbb{R}P^n \subset \mathbb{C}P^n$.

Here we identify $\pi\circ u$ with a disk in $(\mathbb{C}P^n, \mathbb{R}P^n)$ that is homologous to kD, where $D \in H_2(\mathbb{C}P^n, \mathbb{R}P^n)$ is the homology class represented by either of the two hemispheres in a complex line $\mathbb{C}P^1 \hookrightarrow \mathbb{C}P^n$ that intersects $\mathbb{R}P^n$ in an equator.

Proof The claim about the homology class of $\pi\circ u$ follows an area consideration similar to the proof of Lemma 6.2.

The first equality is immediate from the choice of capping path, together with the fact that $D\pi|_\xi : \xi \to T\mathbb{C}P^N$ is a bundle morphism that is a symplectic isomorphism on the fibers.

The Maslov index computation

$$\mu_{\mathbb{R}P^n}(\pi\circ u) = k(n+1)$$

is well known. (Note that the symplectic area of the projected disc in $(\mathbb{C}P^n, \mathbb{R}P^n)$ is equal to $\frac{1}{2}k\pi$, where π is the symplectic area of a line.) □

Lemma 6.7 *With the above choice of constant capping paths, the Conley–Zehnder index satisfies*

$$\mathrm{CZ}(c) = n \tag{6-2}$$

for any $c \in \mathcal{Q}(\widetilde{\Lambda}_V)_k^{\mathrm{Bott}}$, independently of $k = 1, 2, 3, \ldots$.

After a perturbation by a Morse function on the Bott manifold as in [2], *the nondegenerate Reeb chord that corresponds to a critical point has Conley–Zehnder index*

$$\mathrm{CZ}(c) + i - n = i,$$

where $i \in \{0, 1, \dots, \dim \mathcal{Q}(\widetilde{\Lambda}_V)_k^{\mathrm{Bott}} = n\}$ *is the Morse index of the critical point.*

Proof This is similar to the computation of equation (6-1) in the periodic case. More precisely, the Reeb flow on $\mathbb{R}P^{2n+1}$ is totally periodic, and the starting point and end point of all Reeb chords on $\mathbb{R}P^n$ coincide. The return map of the Reeb flow is the identity, and we make it nondegenerate by performing a small positive rotation $s \mapsto e^{i\delta s}\,\mathrm{id}$ in the contact planes. Finally, the result is obtained by computing the standard Conley–Zehnder index of this nondegenerate path. □

Proposition 6.8 *For a half-plane* u *as above with boundary on* $\mathbb{R}\times\widetilde{\Lambda}_V$, *and boundary puncture asymptotic to a Reeb orbit in the family* $\mathcal{Q}(\widetilde{\Lambda}_V)_k^{\mathrm{Bott}}$ *for some* $k \geq 1$, *for the round contact form we have the identity*

$$\mathrm{index}(u) = (n-1) + k(n+1) = (1+k)(n+1) - 2$$

for the expected dimension of the moduli space of unparametrized pseudoholomorphic half-planes of the same type (*with asymptotics that are free to vary in the Bott family* $Q(\widetilde{\Lambda}_V)_k^{\mathrm{Bott}}$).

Moreover, *after a small perturbation of the contact form by a Morse function defined on the Bott manifolds as in* [2], *all Reeb chords with boundary on* $\widetilde{\Lambda}_V$ *may be assumed to be nondegenerate*, *and to satisfy the bound*

$$|c| \geq 2(n+1) - 2 - n = n$$

on their degrees.

Proof It suffices to use the index formula (A-3) with the computations from the above lemmas. □

6.3 Computing the DGA and the Rabinowitz complex

First we show that the Chekanov–Eliashberg algebra of $\mathbb{R}P^n \subset \mathbb{R}P^{2n+1}$ has a (canonically defined) augmentation. Of course, here it is important that the Chekanov–Eliashberg algebra is well defined in the first place, ie there is no Gromov-bubbling which is not captured in the algebra. This nonbubbling follows from Proposition 6.4 above, which bounds the degree of a contractible periodic Reeb orbits from below by $2n$.

Proposition 6.9 *Consider a small Morse perturbation of the round contact form to a nondegenerate one as described above. The DGA of* $\widetilde{\Lambda}_V$ *has an augmentation in* $\mathbb{Z}_2$ *that sends all Reeb chord generators to* 0.

Proof In the case $n > 1$ this follows directly from the degree computation in Proposition 6.8; since all chords have degree strictly greater than one, the differential of the DGA has no constant terms.

In the case $n = 1$ we cannot argue merely by considerations of the degree. In this case there is a single Reeb chord c in degree $|c| = 1$, while all other chords have degree at least 2. There are thus two possibilities when coefficients in $\mathbb{Z}_2$ are used: either $\partial c = 1$, and there are no augmentations, or $\partial c = 0$, and there is a unique "trivial" augmentation that sends all chords to zero. We will see that the latter case holds.

Since $\mathbb{R}P^3 = UT^*S^2 = \{(q, p) \in T^*S^2 \mid \|p\| = 1\}$ and $\widetilde{\Lambda}_V$ is Legendrian isotopic to the conormal lift $\{(q_0, p) \mid \|p\| = 1\}$ of a point $q_0 \in S^2$, the Legendrian $\widetilde{\Lambda}_V$ admits an exact Lagrangian filling isotopic to $\{(q_0, p)\} \subset T^*S^2$ inside the exact symplectic filling T^*S^2 of the contact manifold $UT^*S^2 = \mathbb{R}P^{2n+1}$ when $n = 1$. Hence, in this case, the Chekanov–Eliashberg algebra admits augmentations that are geometrically induced by the fillings by the functorial properties of SFT proven in [19]. (Recall that by functoriality, a Lagrangian filling induces a DGA–morphism from the Chekanov–Eliashberg algebra of the Legendrian at the positive end to the empty-set Legendrian, whose DGA is the ground ring with trivial differential. Such a morphism is by definition an augmentation.) In view of the above degree consideration, the existence of the augmentation implies that ∂ has no constant terms. Hence, there is a (trivial) augmentation. □

The mixed Reeb chords that start on $\widetilde{\Lambda}_0$ and end on $\widetilde{\Lambda}_V$ and which are of length $t \geq 0$ can be parametrized by their starting points

$$(\Lambda_0 \cap e^{-i2t} \cdot V)/\mathbb{Z}_2 \subset \widetilde{\Lambda}_0.$$

The mixed chords from $\widetilde{\Lambda}_0$ to $\widetilde{\Lambda}_V$ are nondegenerate if and only if all Kähler angles between the two subspaces are pairwise distinct. In any case, as for the pure chords, any mixed chord γ can be concatenated by any closed Reeb orbit of period $\frac{1}{2}k\pi$, in order to form a new mixed Reeb chord of length $\ell(\gamma) + \frac{1}{2}k\pi$.

We continue to direct our attention to the following particular family of Lagrangian subspaces

$$V_s := \langle e^{is\pi/(n+1)-s\epsilon} \cdot \partial_{x_1}, e^{is2\pi/(n+1)-s\epsilon} \cdot \partial_{x_2}, \dots, e^{is(n+1)\pi/(n+1)-s\epsilon} \cdot \partial_{x_{n+1}} \rangle_{\mathbb{R}} \subset \mathbb{C}^{n+1}$$

for $\epsilon > 0$ sufficiently small, and the induced one-parameter family $\widetilde{\Lambda}_s := \widetilde{\Lambda}_{V_s} \subset \mathbb{R}P^{2n+1}$ of Legendrians.

Lemma 6.10 *For a suitable choice of Maslov potentials, the complex* $\mathrm{RFC}_*(\widetilde{\Lambda}_0, \widetilde{\Lambda}_1)$ *is generated by the mixed chords* c_j^k *with* $k \in \mathbb{Z}$ *and* $j \in \{1, 2, \dots, n, n+1\}$ *that all are nondegenerate and whose gradings are given by* $|c_j^k| = j + k(n+1) - 1$. *In addition*:

- *The chords* c_j^k *with* $k = 0, 1, 2, \dots$ *start on* $\widetilde{\Lambda}_0$, *end on* $\widetilde{\Lambda}_1$, *and are of the form*

$$c_j^k(t) := e^{i2t} \cdot \partial_{x_j}$$

  for $t \in \big[0, \frac{1}{2}\big(\pi(j/(n+1) + k) - \epsilon\big)\big]$, $j = 1, \dots, n+1$ *and* $k = 0, 1, 2, 3, \dots$.

- *The chords* c_j^k *with* $k = -1, -2, -3, \dots$ *start on* $\widetilde{\Lambda}_1$, *end on* $\widetilde{\Lambda}_0$, *and are of the form*

$$c_j^k(t) := e^{i2t} \cdot e^{i(j\pi/(n+1)-\epsilon)}\partial_{x_j}$$

  for $t \in \big[0, \frac{1}{2}\big(\pi(-j/(n+1) - k) + \epsilon\big)\big]$, $j = 1, \dots, n+1$ *and* $k = -1, -2, -3, \dots$.

- *Their actions satisfy*
$$\mathfrak{a}(c_{j_1}^{k_1}) < \mathfrak{a}(c_{j_2}^{k_2})$$
if and only if $(k_1, j_1) < (k_2, j_2)$ *with respect to the lexicographic order.*

Proof Consider the family $\widetilde{\Lambda}_s$ of Legendrian push-offs. It can be seen that for $\delta > 0$ sufficiently small, $\widetilde{\Lambda}_\delta$ is obtained by the perturbation of the image of $\widetilde{\Lambda}_0$ under a small positive Reeb flow by a perfect Morse function $\mathbb{R}P^n \to \mathbb{R}$. Recall that such a Morse function has precisely $n+1$ nondegenerate critical points, one in each degree $0, 1, 2, \dots, n$.

We proceed to perform the computation of the grading of the mixed chords, as defined in Section A.3. First, note that the chords c_j^k correspond to the critical point of the above Morse function with Morse index $j-1$. For a suitable choice of Maslov potentials, a standard computation thus gives us $|c_j^0| = j-1$; see eg [15, Lemma 2.9(2)].

The chords c_j^k with $k > 0$ have the same start and endpoints as c_j^0. By Section A.3 the difference in Maslov potential can be computed by the relative Chern number $c_{1,\mathrm{rel}}^{\xi}[u]$ of a plane $u\colon \mathbb{C} \to \mathbb{R} \times \mathbb{R}P^{2n+1}$ whose asymptotic is the k–fold cover of the simple periodic Reeb orbit that contains c_j^0, relative the trivialization that is constant under the periodic Reeb flow. We compute this index to be
$$c_{1,\mathrm{rel}}^{\xi}[u] = c_1^{\mathbb{C}P^n}[\pi \circ u] = k(n+1),$$
where the right-hand side is the first Chern class in $\mathbb{C}P^n$ of the k–fold multiple of the generator of $H_1(\mathbb{C}P^n)$. The computation of $|c_j^k|$ for all $k \geq 0$ follows from this.

We leave the computation of the degree of the chords c_j^k with $k < 0$ to the reader, since it follows by analogous computations (although the order of the Legendrians have been switched).

To deduce the statement for $\widetilde{\Lambda}_1$ from the statement for $\widetilde{\Lambda}_\delta$, it suffices to use the continuity of the degrees of the chords. This, in turn, holds since all chords remain transverse as the parameter $s \in [\delta, 1]$ varies. □

Lemma 6.11 *There is a* ***contact-form-preserving*** *isotopy ϕ^t of the round projective space $(\mathbb{R}P^{2n+1}, \alpha_{\mathrm{st}})$ such that*

- *ϕ^t acts on $\widetilde{\Lambda}_0$ by the Reeb flow and a reparametrization, more precisely, $\phi^t(\widetilde{\Lambda}_0) = e^{t\pi/(n+1)}\widetilde{\Lambda}_0$ for all $t \in [0,1]$;*
- *in addition, ϕ^1 fixes $\widetilde{\Lambda}_1$ setwise, ie $\phi^1(\widetilde{\Lambda}_1) = \widetilde{\Lambda}_1$;*
- *the mixed Reeb chords with one endpoint on $\widetilde{\Lambda}_0$ and one endpoint on $\phi^t(\widetilde{\Lambda}_1)$ remain nondegenerate for all $t \in [0,1]$, in particular there are no birth/deaths of Reeb chords during this isotopy (note that mixed chords can shrink to a zero-length chord and change direction in the path, which corresponds to the moments when $\widetilde{\Lambda}_0 \cap \widetilde{\Lambda}_1 \neq \varnothing$); and*
- *the path of nondegenerate mixed Reeb chords parametrized by $t \in [0,1]$ that is induced by the above isotopy connects the chord c_j^k (at $t=0$) with the chord c_{j+1}^k if $j < n+1$, and with c_1^{k+1} if $j = n+1$ (at $t=1$).*

Proof It suffices to consider the action on $\mathbb{C}^n$ of a path of real matrices in $\mathrm{SO}(n+1) \subset \mathrm{SU}(n+1) \subset \mathrm{GL}(n+1,\mathbb{C})$ that starts with the identity matrix, and ends at a matrix in $\mathrm{SO}(n+1)$ that represents a linear map of the form

$$\partial_{x_j} \mapsto \pm\partial_{x_{j+1}} \quad \text{for } j = 1, 2, \dots, n+1,$$

on the standard basis of $\mathbb{C}^{n+1}$ (here we write $\partial_{x_{n+2}} = \partial_{x_1}$). In other words, the latter matrix performs as a permutation of the real coordinate lines. The new Legendrian now corresponds to the subspace

$$\mathbb{C}^{n+1} \supset V' := \langle \pm e^{i(n+1)\pi/(n+1)-s\epsilon} \cdot \partial_{x_1}, \pm e^{i\pi/(n+1)-s\epsilon} \cdot \partial_{x_2}, \dots, \pm e^{in\pi/(n+1)-s\epsilon} \cdot \partial_{x_{n+1}} \rangle_{\mathbb{R}}.$$

Note that this contact isotopy fixes $\widetilde{\Lambda}_0$ set-wise for all $t \in [0,1]$, but that the time-1 map does not fix $\widetilde{\Lambda}_1$. It is finally a simple matter to apply the time-$\pi/(2(n+1))$ Reeb flow in order for the image of $\widetilde{\Lambda}_1$ to become fixed set-wise under the time-1 map.

This proves the first two bullet points, while the third and fourth follow from the first two. □

Proposition 6.12 *There exist arbitrarily small perturbations of the round contact form on $\mathbb{R}P^{2n+1}$, $n \geq 1$, to a nondegenerate contact form, for which the minimal degree of a contractible Reeb orbit is $|\gamma| \geq 2n$. Moreover,*

(1) *the Chekanov–Eliashberg algebra the standard Legendrian $\mathbb{R}P^n$, which is well defined and invariant under Legendrian isotopy by the above, admits an augmentation; and*

(2) *the Rabinowitz–Floer complex $\mathrm{RFC}_*(\widetilde{\Lambda}_0, \widetilde{\Lambda}_1)$ for the above pair of standard $\mathbb{R}P^n$'s, which thus also is well defined and invariant under Legendrian isotopy, can be assumed to have underlying graded vector space of the form*

$$\mathrm{RFC}_*(\widetilde{\Lambda}_0, \widetilde{\Lambda}_1) = \bigoplus_{i \in \mathbb{Z}} \mathbb{Z}_2 \cdot c_i = \bigoplus_{i \in \mathbb{Z}} \mathbb{Z}_2[i],$$

where $|c_i| = i \in \mathbb{Z}$ and $\mathfrak{a}(c_i) < \mathfrak{a}(c_{i+1})$, and with a differential that vanishes.

Proof We perturb the contact form by a Morse function defined on the Bott manifold of Reeb orbits as in [2]. Proposition 6.4 implies the sought bound on the degree of the periodic Reeb orbits. The well-definedness claimed in part (1) is a consequence of this bound. The augmentation of the Chekanov–Eliashberg algebra then follows from Proposition 6.9.

For part (2), the form of the graded vector space that underlies the complex can be seen by considering Lemma 6.10. Recall that the mixed chords described in that lemma already are transversely cut out, and may be assumed to be unaffected by the small perturbation of the round contact form by the Morse functions. The same is also true for the path of mixed Reeb chords in the isotopy produced by Lemma 6.11.

Consider the barcode of the filtered complex

$$\mathrm{RFC}_*(\widetilde{\Lambda}_0, \widetilde{\Lambda}_1)$$

with coefficients in $\mathbb{Z}_2$. The claim that the differential vanishes is equivalent to the claim that there are no finite bars in this barcode for any finite action window.

We argue by contradiction and assume that there exists a finite bar. By degree considerations, this finite bar must start at some Reeb chord c_i end at c_{i+1} for some $i \in \mathbb{Z}$.

Consider the family of barcodes that corresponds to the family of filtered complexes

$$\mathrm{RFC}_*(\widetilde{\Lambda}_0, \phi^t(\widetilde{\Lambda}_1)), \quad \text{with } t \in [0,1],$$

obtained from the isotopy produced by Lemma 6.11. Here we can impose a finite action window that includes $\ell(c_i)$ and $\ell(c_{i+1})$, and is much larger than the oscillation of the Hamiltonian which generates ϕ^t. This enables us to apply Theorem 4.8 to get a PWC. For this family of pairs of Legendrians, no births/deaths of mixed Reeb occur; the Reeb chords remain transverse, even if their lengths can shrink to zero and change direction. Since the barcode varies continuously, we deduce that no bar can (dis)appear during the deformation.

Lemma 6.11 moreover implies that the generators of $\mathrm{RFC}_*(\widetilde{\Lambda}_0, \phi^t(\widetilde{\Lambda}_1))$ vary continuously for $t \in [0,1]$, and connect c_i to c_{i+1}. In particular, since c_i is the start of a finite bar, we conclude that there must be a finite bar in the barcode of

$$\mathrm{RFC}_*(\widetilde{\Lambda}_0, \phi^1(\widetilde{\Lambda}_1)) = \mathrm{RFC}_*(\widetilde{\Lambda}_0, \widetilde{\Lambda}_1)$$

that starts at c_{i+1} as well. It is however impossible for a single chord to correspond to both an endpoint and a starting point of a bar, which is the sought contradiction. □

Appendix A Gradings and indices for Reeb chords and orbits

Here we recall some generalities about index formulas of pseudoholomorphic curves and Conley–Zehnder indices for a $(2n+1)$–dimensional contact manifold (Y, α) with a choice of contact form which is nondegenerate in the Bott sense. We consider both cases of periodic Reeb orbits and Reeb chords on Legendrians. We also consider Maslov potentials for pairs of Legendrians and induced gradings of mixed Reeb chords. None of these results are new, but since they can be difficult to extract from the literature, we give a brief but systematic treatment of relevant definitions and basic results.

For simplicity we assume that the first Chern class of the contact distribution $\xi \to Y$ vanishes. The Reeb chords and orbits of α are assumed to be nondegenerate in the Bott sense; see [2]. (In particular, this also includes the case when the Reeb chords and orbits are nondegenerate in the usual sense.)

A.1 Grading and Conley–Zehnder index for periodic Reeb orbits

The grading $|\gamma|$ of a contractible period Reeb orbit $\gamma \in C^\infty(S^1, Y)$ that lives in a Bott manifold Γ is defined as the expected dimension

$$|\gamma| = \mathrm{vdim}\, \mathcal{M}(\Gamma)$$

of the moduli space that consists of unparametrized finite-energy pseudoholomorphic planes

$$u\colon \mathbb{C} \to \mathbb{R}_\tau \times Y$$

that are

- asymptotic to some Reeb orbit in the (possibly zero-dimensional) Bott family $\Gamma \ni \gamma$ (which is free to vary); and
- pseudoholomorphic for an almost complex structure J that is compatible with $d(e^\tau\alpha)$ and cylindrical outside of a compact subset.

Here we do not identify two solutions that differ by a translation of the symplectization coordinate $\tau \in \mathbb{R}$.

Remark A.1 Both the contractibility of γ and the vanishing of the first Chern class on spherical classes are crucial for the well-definedness of the grading.

One can compute the expected dimension from topological data of the map u. Namely,

$$|\gamma| := \operatorname{index}(u).$$

This Fredholm index $\operatorname{index}(u)$ has been computed; see [2], [45, Proposition 3.7] or [18]. Here we use the formulation from [45, Proposition 3.7] applied to the special case of a pseudoholomorphic plane.

$$\operatorname{index}(u) = ((n+1)-3) + \mu_{\mathrm{CZ}}(\boldsymbol{A}_\gamma - \delta\cdot\mathrm{id}) + 2c^{\xi}_{1,\mathrm{rel}}[u], \tag{A-1}$$

where $\delta > 0$ is a sufficiently small number. Strictly speaking, we will not define the individual terms in the expression $\boldsymbol{A}_\gamma - \delta\cdot\mathrm{id}$, although see Remark A.3.

Remark A.2 The quantity $n+1$ in our formula corresponds to n in [45]. In addition, unlike the formulation from [45, Proposition 3.7], we consider the moduli space of *unparametrized* planes; one thus has to add $\dim_{\mathbb{R}} \operatorname{Aut}(\mathbb{C}) = 4$ to (A-1) in order to get the cited formula.

What remains is now to explain the quantities in the formula. First we need to choose a symplectic trivialization of the contact planes ξ along γ (up to homotopy).

- *The relative first Chern class* $c^{\xi}_{1,\mathrm{rel}}[u]$ is the algebraic number of zeros of a section of the determinant line $u^*\det\xi$ line that is constant in the trivialization of $\gamma^*\det\xi \to S^1$ chosen along the Reeb orbit.
- *The Conley–Zehnder index* $\mu_{\mathrm{CZ}}(\boldsymbol{A}_\gamma - \delta\cdot\mathrm{id})$ *in the nondegenerate case* can be computed as a Maslov index of a suitable (nonclosed) path of Lagrangians, as shown in [38, Remark 5.4]. The path of Lagrangians is obtained by taking the graphs of the symplectic path induced by the linearized Reeb flow (in the chosen trivialization of $\gamma^*\xi \to S^1$); this is a path of symplectic matrices that start with the identity matrix and ends with a symplectic matrix whose eigenvalues are all different from one (by the nondegeneracy assumption).
- *The Conley–Zehnder index* $\mu_{\mathrm{CZ}}(\boldsymbol{A}_\gamma - \delta\cdot\mathrm{id})$ *in the degenerate case*, ie when the linearized time-one map of the Reeb flow along γ has an eigenvalue equal to one, is computed as follows. Deform the path of symplectic matrices induced by the linearized Reeb flow by adding a small positive rotation

$s \mapsto e^{i\delta s}$ id, for some $\delta > 0$, in the contact plane. Then compute the above Conley–Zehnder index for this deformed path.

Remark A.3 The notation $\mu_{\mathrm{CZ}}(\boldsymbol{A}_\gamma - \delta \cdot \mathrm{id})$ is motivated by the nondegenerate case (second bullet point above), where the Conley–Zehnder index also can be computed by first perturbing the asymptotic operator $\boldsymbol{A}_\gamma$ for the linearized $\overline{\partial}_J$–equation by adding a small negative multiple of the identity, and then computing the corresponding classical Conley–Zehnder index. We refer to [45, Section 3.2] for a description of the asymptotic operator $\boldsymbol{A}_\gamma$, which has a discrete spectrum, and which is injective if and only if the periodic Reeb orbit γ is nondegenerate.

As described in [2], for a contact manifold (Y, α) with a Reeb flow that is nondegenerate in the Bott sense, one can use a small Morse function $f : \Gamma \to \mathbb{R}$ defined on the Bott manifolds in order to perturb the contact form to one whose Reeb flow is nondegenerate, while still keeping control of the periodic orbits. More precisely, the Reeb orbits obtained by such a perturbation coincides with the critical points of the Morse function f, where the Reeb orbit corresponding to the critical point p has Conley–Zehnder index

$$\mu_{\mathrm{CZ}}(\boldsymbol{A}_\gamma - \delta \cdot \mathrm{id}) + \mathrm{index}(p) - \dim \Gamma, \tag{A-2}$$

where $\mathrm{index}(p)$ denotes the Morse index of the critical point p of f.

Remark A.4 In particular, the Conley–Zehnder in the degenerate case coincides with the Conley–Zehnder index of the nondegenerate orbit at the maximum of the Morse function that arises from the aforementioned perturbation.

A.2 Grading and Conley–Zehnder index for pure Reeb chords which bound planes

In addition to the vanishing of the first Chern class of $\xi \to Y$, we now also assume the Maslov class of $\mathbb{R} \times \Lambda$ vanishes. In this case, the grading $|c|$ of a contractible pure Reeb chord $c \in C^\infty\big(([0, \ell], \{0, \ell\}), (Y, \Lambda)\big)$ that lives in a Bott manifold $\mathcal{Q}$ can be defined as the expected dimension

$$|c| = \mathrm{vdim}\, \mathcal{M}(c).$$

Here $\mathcal{M}(c)$ denotes the moduli space that consists of unparametrized finite-energy pseudoholomorphic half-planes

$$u : (\mathbb{C} \cap \{y \geq 0\}, \{y = 0\}) \to (\mathbb{R}_\tau \times Y, \mathbb{R} \times \Lambda)$$

that are

- asymptotic to some Reeb chord in the (possibly zero-dimensional) Bott family $\mathcal{Q} \ni c$ (which is free to vary); and
- pseudoholomorphic for an almost complex structure J that is compatible with $d(e^\tau \alpha)$ and cylindrical outside of a compact subset.

Here we do not identify two solutions that differ by a translation of the symplectization coordinate $\tau \in \mathbb{R}$.

Remark A.5 Again, for the well-definedness of the above grading, both the contractibility of γ as well as the vanishing of the Maslov class for disks are crucial properties.

As before we can compute the expected dimension from topological data of the map u by

$$|c| := \operatorname{index}(u),$$

where the Fredholm index can be expressed as

$$\operatorname{index}(u) = (\mathrm{CZ}(c) - 1) + \mu_{\mathbb{R}\times\Lambda}[u]. \tag{A-3}$$

We proceed to describe the quantities in the above formula. First we need to choose a continuous capping path of Lagrangian tangent planes along $c(t)$ that connects $T_{c(0)}\Lambda \subset \xi_{c(0)}$ to $T_{c(\ell)}\Lambda \subset \xi_{c(\ell)}$ (up to homotopy).

- *The relative Maslov class* $\mu_{\mathbb{R}\times\Lambda}[u]$ of the half-plane u which is defined by concatenating the path of Lagrangian tangent planes along $u|_{\{y=0\}}$ with the capping path at the puncture to obtain a closed path of Lagrangian tangent planes, and then computing the usual Maslov index for this loop of Lagrangian tangent planes in the trivialization induced by u.
- *The Conley–Zehnder index* $\mathrm{CZ}(c)$ *in the nondegenerate case* is defined as follows. First, we use the Reeb flow to identify the contact planes along the Reeb chord. Construct a closed loop of Lagrangian tangent planes by rotating $T_{c(0)}\Lambda$ to $T_{c(1)}\Lambda$ in the contact plane by using the minimal positive Kähler angles (these are nonzero by the nondegeneracy assumption). Then concatenate this path with the capping path to obtain a loop of Lagrangian tangent planes. The Maslov index of this loop is the Conley–Zehnder index.
- *The Conley–Zehnder index* $\mathrm{CZ}(c)$ *in the degenerate case* is computed as above, but where we first perturb the Lagrangian tangent plane $T_{c(0)}\Lambda$ at the starting point by a small positive rotation $e^{i\delta}$ id in the contact plane.

Similarly to the perturbation of Bott manifolds of periodic Reeb orbits by Morse functions as constructed in [2], one can also perform a perturbation of the Bott manifolds of Reeb chords. We again obtain an analogous formula

$$\mathrm{CZ}(c) + \operatorname{index}(p) - \dim \mathcal{Q} \tag{A-4}$$

for the Conley–Zehnder index of the nondegenerate Reeb orbit that corresponds to the critical point p of the Morse function $f : \mathcal{Q} \to \mathbb{R}$. Here $\operatorname{index}(p)$ denotes the Morse index of the critical point p of f.

Remark A.6 As in the periodic orbit case, the Conley–Zehnder in the degenerate case coincides with the Conley–Zehnder index of the nondegenerate orbit at the maximum after such a perturbation.

A.3 Grading and Conley–Zehnder index for mixed Reeb chords

In this subsection we assume that the mixed Reeb chords are all nondegenerate in the strong sense.

The grading of a mixed Reeb chord with endpoints on two different Legendrians Λ_0 and Λ_1 depends on several additional choices. First, we need to choose a symplectic trivialization of the square of the determinant $\mathbb{C}$–line bundle $(\det_{\mathbb{C}} \xi)^{\otimes_{\mathbb{C}} 2} \to Y$ (up to homotopy). This is possible since the first Chern class vanishes (in fact one only needs it to be two-torsion). Note that, since $T\Lambda_i \subset \xi$ is Lagrangian, its image $[T_{c(0)}\Lambda_i] \subset (\det_{\mathbb{C}} \xi_{c(0)})^{\otimes_{\mathbb{C}} 2} \cong \mathbb{C}$ inside the determinant line is a one-dimensional real subspace. Second, we need to make continuous choices of lifts to $\mathbb{R}$ of the angular phase in $\mathbb{R}/\pi\mathbb{Z}$, ie the argument, of the images

$$[T\Lambda_i] := (\det_{\mathbb{R}} \Lambda_i)^{\otimes_{\mathbb{R}} 2} \subset (\det_{\mathbb{C}} \xi)^{\otimes_{\mathbb{C}} 2} \to Y$$

of these real subspaces in the latter $\mathbb{C}$–line bundle. (Passing to the square means that this operation is well defined on unoriented Legendrian tangent spaces.) The choice of such a lift is called a *Maslov potential*, and it exists if and only if the Maslov classes of Λ_i vanish. See [15, Section 2.5] or [6] for more details about Maslov potentials in the Legendrian setting.

We can define the grading of a mixed Reeb chord $c\colon [0,\ell] \to Y$ from Λ_i to Λ_j as follows. Let $\bar{\phi}_i \in \mathbb{R}$ be the lifts of phases of $[T\Lambda_i] \subset (\det_{\mathbb{C}} \xi)^{\otimes_{\mathbb{C}} 2}$ at the endpoints of the Reeb chord c as defined by the choice of Maslov potential. Use the Reeb flow $\phi_R^t\colon (Y,\alpha) \to (Y,\alpha)$ to identify $T_{c(0)}\Lambda_i$ with a Lagrangian tangent plane $\phi_R^\ell(T_{c(0)}\Lambda_i) \subset \xi_{c(\ell)}$. By continuity (of course using the triviality of $(\det_{\mathbb{C}} \xi_{c(1)})^{\otimes_{\mathbb{C}} 2} \to Y$) we obtain a lift $\bar{\phi}_0'$ of the phase of $[\phi_R^\ell(T_{c(0)}\Lambda_i)] \subset (\det_{\mathbb{C}} \xi_{c(\ell)})^{\otimes_{\mathbb{C}} 2}$. Perform the smallest positive rotation $e^{i\theta_0}$, for some $\theta_0 \in (0,\pi)$, that makes the real line

$$[\phi_R^\ell(T_{c(0)}\Lambda_i)] \subset (\det_{\mathbb{C}} \xi_{c(\ell)})^{\otimes_{\mathbb{C}} 2} \cong \mathbb{C}$$

coincide with $[\phi_R^\ell(T_{c(\ell)}\Lambda_i)]$. (The nondegeneracy implies that this angle is nonzero.) The Conley–Zehnder index is then the difference

$$\mathrm{CZ}(c) = \frac{1}{\pi}(\bar{\phi}_0' + \theta_0 - \bar{\phi}_1) \in \mathbb{Z}$$

of lifts of phases, and we define the grading via

$$|c| := \mathrm{CZ}(c) - 1 \in \mathbb{Z}.$$

This grading depends on the chosen symplectic trivialization of the $\mathbb{C}$–line bundle $(\det_{\mathbb{C}} \xi)^{\otimes_{\mathbb{C}} 2} \to Y$ as well as the choice of Maslov potentials of the Legendrians involved.

A feature of this grading is that an unparametrized pseudoholomorphic strip in $\mathbb{R}_\tau \times Y$ that has a positive asymptotic to the mixed Reeb chord c^+, and a negative Reeb chord asymptotic to the mixed Reeb chord c^-, lives in a moduli space of expected dimension

$$\mathrm{vdim}\, \mathcal{M}(c^+, c^-) = |c^+| - |c^-|.$$

Note that, for this dimension, we again do not identify solutions that differ by a translation of the τ–coordinate.

We end by noting that Legendrian isotopies naturally induce continuous extension of the Maslov potential. In the case of a loop Λ^t of Legendrians, the effect on the Maslov potential at a point $p \in \Lambda^0 = \Lambda^1$

can be seen to be as follows. Take a smooth path $\gamma(t) \in \Lambda^t$ so that $\gamma(0) = p = \gamma(1)$, and consider the trivialization of $\gamma^*(\det_{\mathbb{C}} \xi)^{\otimes_{\mathbb{C}} 2}$ induced by the real lines $[T_{\gamma(t)}\Lambda^t]$. The action on this isotopy of the Maslov potential at p can then be seen to be equal to

$$c^{\xi}_{1,\mathrm{rel}}[u] \in \mathbb{Z}.$$

Here the relative first Chern class computes the algebraic number of zeros of a smooth extension of the section of $u^*(\det_{\mathbb{C}} \xi)^{\otimes_{\mathbb{C}} 2}$ along a smooth orientable compact surface $u\colon \Sigma \to Y$ whose boundary parametrizes $\gamma(t)$, where we require the section to be nonzero and constant with respect to the aforementioned trivialization along the boundary $\gamma(t)$.

Appendix B Length of trace cobordisms and conformal factors

It is well known that the trace of a Legendrian isotopy can be deformed to an exact Lagrangian concordance in the symplectization; see [4; 5; 25] for different versions of this construction. Here we revisit the version from [4, Theorem 1.2] and show that it fits our purposes as far as control on the length is concerned. The length of a Lagrangian cobordism was defined in [40] by Sabloff and Traynor; see Section 3.

Let $\Lambda \subset (Y, \alpha)$ be a Legendrian submanifold of a contact manifold with a choice of contact form α. Let $\phi^t_{\alpha,H}\colon (Y, \ker\alpha) \to (Y, \ker\alpha)$ be a contact isotopy with $\phi^0_{\alpha,H} = \mathrm{id}$, which thus is generated by a contact Hamiltonian $H_t\colon Y \to \mathbb{R}$ defined by $H_t \circ \phi^t_{\alpha,H} = \alpha(\dot{\phi}^t_{\alpha,H})$. Furthermore, let $f_t\colon Y \to \mathbb{R}$ be the smooth function for which $(\phi^t)^*_{\alpha,H}\alpha = e^{f_t}\alpha$. The function e^{f_t} is called the *conformal factor* of the contact isotopy and was introduced in Section 1. In particular, $(\tau, y) \mapsto (\tau - f_t(y), \phi^t_{\alpha,H}(y))$ is a Hamiltonian isotopy of the symplectization that is generated by the t–dependent Hamiltonian $e^{\tau}H_t\colon \mathbb{R}_\tau \times Y \to \mathbb{R}$. Note that this symplectic isotopy preserves the primitive $e^{\tau}\alpha$ of the symplectic form.

Proposition B.1 *For a contact isotopy as above, the following holds for any arbitrary choice of $\epsilon > 0$:*

(1) *There exists a Lagrangian trace cobordism $L_{01} \subset (\mathbb{R}_\tau \times Y, e^{\tau}\alpha)$ from Λ to $\phi^1_{\alpha,H}(\Lambda)$ of length equal to $-e^{1+\epsilon}\min_{x\in Y, t\in[0,1]} f_t(x) \geq 0$.*

(2) *There exists a Lagrangian trace cobordism $L_{10} \subset (\mathbb{R}_\tau \times Y, e^{\tau}\alpha)$ from $\phi^1_{\alpha,H}(\Lambda)$ to Λ of length equal to $e^{1+\epsilon}\max_{x\in Y, t\in[0,1]} f_t(x) \geq 0$.*

(3) *One can assume that the two concatenations $L_{01} \odot L_{10}$ and $L_{10} \odot L_{01}$ of traces, which are thus Lagrangian cobordisms from Λ to itself and $\phi^1_{\alpha,H}(\Lambda)$ to itself, respectively, are of length*

$$c_0 := e^{1+\epsilon}\Big(\max_{x\in Y, t\in[0,1]} f_t(x) - \min_{x\in Y, t\in[0,1]} f_t(x)\Big).$$

Moreover, these concatenations are Hamiltonian isotopic to the trivial Lagrangian cylinders $\mathbb{R}\times\Lambda$ and $\mathbb{R}\times\phi^1_{\alpha,H}(\Lambda)$, respectively, by isotopies supported in a subset of the form $[0, c_0]\times Y$ (after a suitable translation of the τ–coordinate).

(4) *All Lagrangian cobordisms constructed above have primitives of the pullback of $e^{\tau}\alpha$ that can be taken to be globally constant on each cylindrical end $\pm\tau \gg 0$.*

Proof (1) It suffices to consider the image of the trivial Lagrangian cylinder $\mathbb{R}\times\Lambda$ under the time-one map of the Hamiltonian isotopy defined by a Hamiltonian of the form $\rho(\tau)e^{\tau}H_t$. Recall that τ is the symplectization coordinate here, while t is the time-coordinate. We take $\rho(\tau)\colon \mathbb{R}_{\tau}\to 0$ to be a smooth function that vanishes on the subset $(-\infty,0]$ and is equal to one on the subset $[\delta,+\infty)$ for some $\delta>0$.

So far we have merely repeated the argument from the proof of [4, Theorem 1.2]. To achieve the bound on the length, it suffices to choose $\delta>0$ sufficiently small, so that the inequality

$$e^{1+\epsilon}\big(-\min_{x\in Y,t\in[0,1]} f_t(x)\big)\geq \delta-\min_{x\in Y,t\in[0,1]} f_t(x)$$

is satisfied. To show the result follows from this inequality, we use the fact that $e^{\tau}H_t$ generates a Hamiltonian isotopy $(\tau,y)\mapsto(\tau-f_t(y),\phi^t_{\alpha,H}(y))$ of the symplectization, which means that $\max_{y\in Y} f_t(y)$ is the maximal translation in the negative symplectization direction at time t.

(2) This is similar to (1), using the fact that $(\phi^t_{\alpha,H})^{-1}$ again is a contact isotopy (thus it is generated by a contact Hamiltonian) whose conformal factor is equal to $-f_t$. We then apply the construction from (1) to the cylinder $\mathbb{R}\times\phi^1_{\alpha,H}(\Lambda)$ instead of $\mathbb{R}\times\Lambda$, while using the contact isotopy $(\phi^t_{\alpha,H})^{-1}$ instead of $\phi^t_{\alpha,H}$.

(3) The constructions in (1) and (2) can be taken to depend smoothly on the family of contact isotopies $t\mapsto(\phi^t_{\alpha,H})_r:=\phi^{rt}_{\alpha,H}$ for $t\in[0,1]$, where the family is parametrized by $r\in[0,1]$. Writing $e^{f_{r,t}}$ for the conformal factor of $(\phi^t_{\alpha,H})_r$ we immediately note that

$$\max_{x\in Y,t\in[0,1]} f_t(x)\geq\max_{x\in Y,t\in[0,1]} f_{r,t}(x)\quad\text{and}\quad\min_{x\in Y,t\in[0,1]} f_{r,t}(x)\geq\min_{x\in Y,t\in[0,1]} f_t(x)$$

holds for any $r\in[0,1]$, from which the sought length properties follow.

We thus produce families of Lagrangian cobordisms whose concatenations $L^r_{01}\odot L^r_{10}$ (resp. $L^r_{10}\odot L^r_{01}$) interpolate between $\mathbb{R}\times\Lambda$ (resp. $\mathbb{R}\times\phi^1_{\alpha,H}(\Lambda)$) at $r=0$ and $L_{01}\odot L_{10}$ (resp. $L_{10}\odot L_{01}$) at $r=1$. The corresponding isotopy may be assumed to be supported inside $[0,c_0]\times Y$. Since this is an isotopy through exact Lagrangians, a standard fact implies that it is generated by a global Hamiltonian isotopy.

(4) As follows by Cartan's formula, a Hamiltonian isotopy $\phi^t_{\rho(\tau)e^{\tau}H_t}\colon\mathbb{R}\times Y\to\mathbb{R}\times Y$, with $\rho'(\tau)$ being of compact support, pulls back the primitive of the symplectic form $e^{\tau}\alpha$ to a one form $e^{\tau}\alpha+dG$. Since $\phi^t_{\rho(\tau)e^{\tau}H_t}$ preserves the primitive $e^{\tau}\alpha$ outside of a compact subset, we conclude that $G\colon Y\to\mathbb{R}$ is locally constant outside of a compact subset or, equivalently, dG is compactly supported. The sought statement follows from this. □

References

[1] **E Bao**, **K Honda**, *Semi-global Kuranishi charts and the definition of contact homology*, Adv. Math. 414 (2023) art. id. 108864 MR Zbl

[2] **F Bourgeois**, *A Morse–Bott approach to contact homology*, from "Symplectic and contact topology: interactions and perspectives" (Y Eliashberg, B Khesin, F Lalonde, editors), Fields Inst. Commun. 35, Amer. Math. Soc., Providence, RI (2003) 55–77 MR Zbl

[3] **F Bourgeois**, **Y Eliashberg**, **H Hofer**, **K Wysocki**, **E Zehnder**, *Compactness results in symplectic field theory*, Geom. Topol. 7 (2003) 799–888 MR Zbl

[4] **B Chantraine**, *Lagrangian concordance of Legendrian knots*, Algebr. Geom. Topol. 10 (2010) 63–85 MR Zbl

[5] **B Chantraine**, **V Colin**, **G Dimitroglou Rizell**, *Positive Legendrian isotopies and Floer theory*, Ann. Inst. Fourier (Grenoble) 69 (2019) 1679–1737 MR Zbl

[6] **B Chantraine**, **G Dimitroglou Rizell**, **P Ghiggini**, **R Golovko**, *Floer theory for Lagrangian cobordisms*, J. Differential Geom. 114 (2020) 393–465 MR Zbl

[7] **Y V Chekanov**, *Lagrangian intersections, symplectic energy, and areas of holomorphic curves*, Duke Math. J. 95 (1998) 213–226 MR Zbl

[8] **Y Chekanov**, *Differential algebra of Legendrian links*, Invent. Math. 150 (2002) 441–483 MR Zbl

[9] **K Cieliebak**, **U Frauenfelder**, **A Oancea**, *Rabinowitz Floer homology and symplectic homology*, Ann. Sci. École Norm. Sup. 43 (2010) 957–1015 MR Zbl

[10] **K Cieliebak**, **A Oancea**, *Symplectic homology and the Eilenberg–Steenrod axioms*, Algebr. Geom. Topol. 18 (2018) 1953–2130 MR Zbl

[11] **K Conrad**, $\mathrm{SL}_2(\mathbb{Z})$, preprint (2022) Available at `https://kconrad.math.uconn.edu/blurbs/grouptheory/SL(2,Z).pdf`

[12] **M Damian**, *On the stable Morse number of a closed manifold*, Bull. Lond. Math. Soc. 34 (2002) 420–430 MR Zbl

[13] **G Dimitroglou Rizell**, *Legendrian ambient surgery and Legendrian contact homology*, J. Symplectic Geom. 14 (2016) 811–901 MR Zbl

[14] **G Dimitroglou Rizell**, *Lifting pseudo-holomorphic polygons to the symplectisation of $P \times \mathbb{R}$ and applications*, Quantum Topol. 7 (2016) 29–105 MR Zbl

[15] **G Dimitroglou Rizell**, *Families of Legendrians and Lagrangians with unbounded spectral norm*, J. Fixed Point Theory Appl. 24 (2022) art. id. 43 MR Zbl

[16] **G Dimitroglou Rizell**, **M G Sullivan**, *An energy-capacity inequality for Legendrian submanifolds*, J. Topol. Anal. 12 (2020) 547–623 MR Zbl

[17] **G Dimitroglou Rizell**, **M G Sullivan**, *The persistence of the Chekanov–Eliashberg algebra*, Selecta Math. 26 (2020) art. id. 69 MR Zbl

[18] **L Diogo**, **S T Lisi**, *Morse–Bott split symplectic homology*, J. Fixed Point Theory Appl. 21 (2019) art. id. 77 MR Zbl

[19] **T Ekholm**, *Rational symplectic field theory over $\mathbb{Z}_2$ for exact Lagrangian cobordisms*, J. Eur. Math. Soc. 10 (2008) 641–704 MR Zbl

[20] **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, Birkhäuser, New York (2012) 109–145 MR Zbl

[21] **T Ekholm**, *Holomorphic curves for Legendrian surgery*, preprint (2019) arXiv 1906.07228

[22] **T Ekholm**, **J B Etnyre**, **J M Sabloff**, *A duality exact sequence for Legendrian contact homology*, Duke Math. J. 150 (2009) 1–75 MR Zbl

[23] **T Ekholm**, **J Etnyre**, **M Sullivan**, *The contact homology of Legendrian submanifolds in* $\mathbb{R}^{2n+1}$, J. Differential Geom. 71 (2005) 177–305 MR Zbl

[24] **T Ekholm**, **J Etnyre**, **M Sullivan**, *Orientations in Legendrian contact homology and exact Lagrangian immersions*, Int. J. Math. 16 (2005) 453–532 MR Zbl

[25] **T Ekholm**, **K Honda**, **T Kálmán**, *Legendrian knots and exact Lagrangian cobordisms*, J. Eur. Math. Soc. 18 (2016) 2627–2689 MR Zbl

[26] **M Entov**, **L Polterovich**, *Legendrian persistence modules and dynamics*, J. Fixed Point Theory Appl. 24 (2022) art. id. 30 MR Zbl

[27] **J W Fish**, **H Hofer**, *Lectures on polyfolds and symplectic field theory*, preprint (2018) arXiv 1808.07147

[28] **J Hedicke**, *Lorentzian distance functions in contact geometry*, J. Topol. Anal. 16 (2024) 205–225 MR Zbl

[29] **C Karlsson**, *A note on coherent orientations for exact Lagrangian cobordisms*, Quantum Topol. 11 (2020) 1–54 MR Zbl

[30] **J Lafontaine**, **M Audin**, *Introduction: applications of pseudo-holomorphic curves to symplectic topology*, from "Holomorphic curves in symplectic geometry" (M Audin, J Lafontaine, editors), Progr. Math. 117, Birkhäuser, Basel (1994) 1–14 MR Zbl

[31] **N Legout**, *A_∞–category of Lagrangian cobordisms in the symplectization of $P \times \mathbb{R}$*, Quantum Topol. 14 (2023) 101–200 MR Zbl

[32] **W J Merry**, *Lagrangian Rabinowitz Floer homology and twisted cotangent bundles*, Geom. Dedicata 171 (2014) 345–386 MR Zbl

[33] **L Nakamura**, *C^0–limits of Legendrian submanifolds*, preprint (2020) arXiv 2008.00924

[34] **Y-G Oh**, *Geometry and analysis of contact instantons and entanglement of Legendrian links*, *I*, preprint (2021) arXiv 2111.02597

[35] **Y Pan**, **D Rutherford**, *Augmentations and immersed Lagrangian fillings*, J. Topol. 16 (2023) 368–429 MR Zbl

[36] **J Pardon**, *Contact homology and virtual fundamental cycles*, J. Amer. Math. Soc. 32 (2019) 825–919 MR Zbl

[37] **L Polterovich**, **D Rosen**, **K Samvelyan**, **J Zhang**, *Topological persistence in geometry and analysis*, Univ. Lect. Ser. 74, Amer. Math. Soc., Providence, RI (2020) MR Zbl

[38] **J Robbin**, **D Salamon**, *The Maslov index for paths*, Topology 32 (1993) 827–844 MR Zbl

[39] **D Rosen**, **J Zhang**, *Chekanov's dichotomy in contact topology*, Math. Res. Lett. 27 (2020) 1165–1193 MR Zbl

[40] **J M Sabloff**, **L Traynor**, *The minimal length of a Lagrangian cobordism between Legendrians*, Selecta Math. 23 (2017) 1419–1448 MR Zbl

[41] **E Shelukhin**, *The Hofer norm of a contactomorphism*, J. Symplectic Geom. 15 (2017) 1173–1208 MR Zbl

[42] **M G Sullivan**, *K–theoretic invariants for Floer homology*, Geom. Funct. Anal. 12 (2002) 810–872 MR Zbl

[43] **M Usher**, *Local rigidity, contact homeomorphisms, and conformal factors*, Math. Res. Lett. 28 (2021) 1875–1939 MR Zbl

[44] **M Usher**, **J Zhang**, *Persistent homology and Floer–Novikov theory*, Geom. Topol. 20 (2016) 3333–3430 MR Zbl

[45] **C Wendl**, *Automatic transversality and orbifolds of punctured holomorphic curves in dimension four*, Comment. Math. Helv. 85 (2010) 347–407 MR Zbl

Department of Mathematics, Uppsala University
Uppsala, Sweden

Department of Mathematics and Statistics, University of Massachusetts
Amherst, MA, United States

georgios.dimitroglou@math.uu.se, mikesullivan@umass.edu

Proposed: Leonid Polterovich Received: 4 February 2022
Seconded: Yakov Eliashberg, András I Stipsicz Revised: 28 December 2022

msp

Geometry & Topology **28**:5 (2024) 2207–2257
DOI: 10.2140/gt.2024.28.2207
Published: 24 August 2024

Packing Lagrangian tori

RICHARD HIND
ELY KERMAN

We consider the problem of packing a symplectic manifold with integral Lagrangian tori, that is, Lagrangian tori whose area homomorphisms take only integer values. We prove that the Clifford torus in $S^2 \times S^2$ is a maximal integral packing, in the sense that any other integral Lagrangian torus must intersect it. In the other direction, we show that in any symplectic polydisk $P(a, b)$ with $a, b > 2$, there is at least one integral Lagrangian torus in the complement of the collection of standard product integral Lagrangian tori.

53D12, 53D35

1 Introduction

In this paper we consider packings of symplectic manifolds by Lagrangian tori. Since every symplectic manifold contains infinitely many disjoint Lagrangian tori, we must set a scale in order to pose meaningful questions. We therefore restrict our attention to Lagrangian tori whose area homomorphism takes only integer values. These will be referred to as *integral Lagrangian tori*.[1] The fundamental packing question, in this setting, is the following:

> *What is the maximum number of disjoint integral Lagrangian tori contained in a given (pre)compact symplectic manifold?*

A more approachable version of this question is to consider a specific collection of disjoint integral Lagrangian tori in a symplectic manifold (M, ω), and to ask if it is a *maximal integral packing* in the sense that any other integral Lagrangian torus in M must intersect at least one torus in the collection. In this paper, we study this question in the simplest nontrivial setting.

1.1 Results

Equip the sphere S^2 with its standard symplectic form ω scaled so that $\int_{S^2} \omega = 2$. Let $L_{1,1}$ be the monotone Clifford torus (product of equators) in $(S^2 \times S^2, \pi_1^*\omega + \pi_2^*\omega)$. Our first result is the following.

Theorem 1.1 *The Clifford torus $L_{1,1}$ is a maximal integral packing of $(S^2 \times S^2, \pi_1^*\omega + \pi_2^*\omega)$.*

[1] These are also sometimes called Bohr–Sommerfeld Lagrangians.

Consider $\mathbb{R}^4$ equipped with its standard symplectic structure ω_4. For real numbers $a, b > 0$, consider the symplectic polydisk

$$P(a,b) = \{(z_1, z_2) \in \mathbb{C}^2 \mid \pi|z_1|^2 < a, \pi|z_2|^2 < b\} \subset \mathbb{R}^4.$$

Identifying $L_{1,1}$ with the standard Clifford torus in $\mathbb{R}^4$, Theorem 1.1 implies that $L_{1,1}$ is a maximal integral packing of each $P(a,b)$ with $1 < a, b < 2$.

If a and b are both greater then 2, then a natural candidate for a maximal integral packing of $P(a,b)$ is the collection of integral Lagrangian tori

$$\{L_{k,l} \mid k, l \in \mathbb{N}, k \leq \lfloor a \rfloor, l \leq \lfloor b \rfloor\},$$

where $L_{k,l}$ is the product of the circle about the origin bounding area k in the z_1–plane with the circle about the origin bounding area l in the z_2–plane. The analogous packing in dimension two is always maximal. Our second result shows that, in dimension four, this candidate always fails.

Theorem 1.2 *If $\min(a,b) > 2$, then $\{L_{k,l} \mid k, l \in \mathbb{N}, k \leq \lfloor a \rfloor, l \leq \lfloor b \rfloor\}$ is not a maximal integral packing of $P(a,b)$. For every $\epsilon > 0$, there is an integral Lagrangian torus L^+ in*

$$P(2+\epsilon, 2+\epsilon) \setminus \{L_{k,l} \mid k, l \in \{1,2\}\}.$$

1.2 Overview

The first step in our proof of Theorem 1.1 is to show that any integral Lagrangian torus contained in $(S^2 \times S^2, \pi_1^*\omega + \pi_2^*\omega)$ is actually monotone. This follows from the work of Hind and Opshtein [9], and is proved in Proposition 3.2 below. Arguing by contradiction, we then assume there is a monotone Lagrangian torus $\mathbb{L}$ in $(S^2 \times S^2, \pi_1^*\omega + \pi_2^*\omega)$ that is disjoint from the Clifford torus $L_{1,1}$. The work of Ivrii [11] and Dimitroglou-Rizell, Goodman and Ivrii [5] implies that there is a finite-energy holomorphic foliation $\mathcal{F}$ of $S^2 \times S^2 \setminus (\mathbb{L} \cup L_{1,1})$ which has a normal form near $\mathbb{L}$ and $L_{1,1}$; see Section 3.5. We use $\mathcal{F}$ to establish the existence of two symplectic spheres, F and G, in $(S^2 \times S^2, \pi_1^*\omega + \pi_2^*\omega)$. These are obtained from the compactifications of the pseudoholomorphic buildings obtained in Section 3.7; see Propositions 3.20 and 3.22. Both F and G represent a homology class of the form $(1,d) \in H_2(S^2 \times S^2; \mathbb{Z}) = \mathbb{Z}^2$ for some large d. They also have special intersection properties with the leaves of $\mathcal{F}$ and with each other; see Proposition 3.24. Using the spheres F and G, together with the operations of blow-up, inflation and blow-down, we then alter the ambient symplectic manifold away from $\mathbb{L} \cup L_{1,1}$ to obtain a new monotone symplectic manifold, (X, Ω). This new manifold is symplectomorphic to $(S^2 \times S^2, (d+1)(\pi_1^*\omega + \pi_2^*\omega))$, and $\mathbb{L}$ and $L_{1,1}$ remain disjoint and monotone therein. However, the images (transforms) of the spheres F and G in (X, Ω) are now in the class $(1,0)$ and their existence implies, by the work of Cieliebak and Schwingenheuer in [4], that $\mathbb{L}$ and $L_{1,1}$ must both be Hamiltonian isotopic to the Clifford torus in (X, Ω). It then follows from standard monotone Lagrangian Floer theory (as in Oh [17]) that it is not possible for $\mathbb{L}$ and $L_{1,1}$ to be disjoint. This contradiction completes the proof of Theorem 1.1.

To prove Theorem 1.2 we construct, for every $\epsilon > 0$, an explicit embedding of the closure of $P(1,1)$ into $P(2+\epsilon, 2+\epsilon) \setminus \{L_{k,l} \mid k,l \in \{1,2\}\}$, using a time-dependent Hamiltonian flow. The desired Lagrangian, L^+, is the one on the boundary of the image.

1.3 Commentary and further questions

Given that Theorem 1.1 is reduced to the problem of detecting intersection points of two monotone Lagrangian tori, using Hind and Opshtein [9], it is natural to ask whether Lagrangian Floer theory (rigid holomorphic curves) can also be used to prove Theorem 1.1 directly. To the knowledge of the authors this is not yet possible. The following result seems to be as close to a proof of Theorem 1.1 as one can currently get using Lagrangian Floer theory.

Theorem 1.3 *Suppose that L is a monotone Lagrangian torus in $(S^2 \times S^2, \pi_1^*\omega + \pi_2^*\omega)$. If the Lagrangian Floer homology of L, with respect to some $\mathbb{C}^*$–local system, is nontrivial, then L must intersect $L_{1,1}$.*

This follows from the work of Ritter and Smith in [19].[2] In particular, Corollary 1.5 of [19] implies that the Clifford torus $L_{1,1}$ split-generates the monotone Fukaya category of $(S^2 \times S^2, \pi_1^*\omega + \pi_2^*\omega)$. It is not known whether there exist monotone Lagrangian tori in $(S^2 \times S^2, \pi_1^*\omega + \pi_2^*\omega)$ whose Lagrangian Floer homology is trivial for every choice of $\mathbb{C}^*$–local system. In [20], Vianna constructs a countably infinite collection of monotone Lagrangian tori in $(S^2 \times S^2, \pi_1^*\omega + \pi_2^*\omega)$, no two of which are Hamiltonian isotopic. Each of the tori in Vianna's collection satisfies the hypothesis of Theorem 1.3.

The following question, in the spirit of Theorem 1.1, remains unresolved.

Question 1.4 *Does every pair of monotone Lagrangian tori in $(S^2 \times S^2, \pi_1^*\omega + \pi_2^*\omega)$ intersect?*

Progress on other aspects of the study of disjoint Lagrangian tori has also recently been made in two related works by Mak and Smith [13], and by Polterovich and Shelukhin [18]. Let $\{\gamma_i\}$ be a collection of disjoint circles bounding disks of the same area, and let E be the equator in the sphere S^2. In [13] and [18] it is shown that, with respect to certain nonmonotone symplectic forms on $S^2 \times S^2$, packings of the form $\mathcal{L} = \bigsqcup \gamma_i \times E$ are maximal in the sense that any Lagrangian torus Hamiltonian isotopic to $\gamma_1 \times E$ must intersect $\mathcal{L}$. In comparison, the maximal packing given by Theorem 1.1 only includes a single torus, $L_{1,1}$, but we do not assume any other tori are Hamiltonian isotopic to it. Theorem 1.2 shows that analogous packings of the form $\bigsqcup \gamma_i \times \gamma_j$ are no longer maximal.

Below are a few of the questions suggested by Theorem 1.2, which also remain unresolved.

Question 1.5 *Is every integral Lagrangian torus in $P(2+\epsilon, 2+\epsilon) \setminus \{L_{k,l} \mid k,l \in \{1,2\}\}$ Hamiltonian isotopic to $L_{1,1}$?*

[2] We are grateful to the referee for pointing out this reference.

Question 1.6 *Suppose $2 < a, b < 3$. Are there six disjoint integral Lagrangian tori in $P(a, b)$?*

Question 1.7 *Suppose $2 < b < 3$. Are there three disjoint integral Lagrangian tori in $P(2, b)$?*

Question 1.5 has recently been answered negatively, and Questions 1.6 and 1.7 positively, by Hicks and Mak in the preprint [7]. The question of whether these domains might actually contain infinitely many disjoint integral Lagrangians remains completely open.

Acknowledgements The authors would like to thank the referee of this paper for their careful analysis and many valuable comments. We also thank Karim Boustany for helpful comments and for pointing out a few mistakes.

The authors are supported by Simons Foundation grants 633715 (Hind) and 429872 (Kerman).

2 Conventions, labels and notation

Every copy of the two-dimensional sphere S^2 will implicitly be identified with the unit sphere in $\mathbb{R}^3$ and we will label the north and south poles by ∞ and 0, respectively. In $(S^2 \times S^2, \pi_1^*\omega + \pi_2^*\omega)$, we use these points to define the four symplectic spheres $S_0 = S^2 \times \{0\}$, $S_\infty = S^2 \times \{\infty\}$, $T_0 = \{0\} \times S^2$ and $T_\infty = \{\infty\} \times S^2$. The ordered basis $\{[S_0], [T_0]\}$ of $H_2(S^2 \times S^2; \mathbb{Z})$ is used to identify it with $\mathbb{Z}^2$.

Let $L \subset (M, \Omega)$ be a Lagrangian torus in a four-dimensional symplectic manifold. A diffeomorphism ψ from $\mathbb{T}^2 = S^1 \times S^1$ to L will be referred to as a parametrization of L. It specifies a basis of $H_1(L; \mathbb{Z})$ and thus an isomorphism from $H_1(L; \mathbb{Z})$ to $\mathbb{Z}^2$. We will denote this copy of $\mathbb{Z}^2$ by $H_1^\psi(L; \mathbb{Z})$. The parametrization ψ can also be extended to a symplectomorphism Ψ from a neighborhood of the zero section in $T^*\mathbb{T}^2$ to a Weinstein neighborhood $\mathcal{U}(L)$ of L in M. We will denote the corresponding coordinates in the neighborhood $\mathcal{U}(L)$ of L by (p_1, p_2, q_1, q_2) and, for simplicity, we will assume that

$$\mathcal{U}(L) = \{|p_1| < \epsilon, |p_2| < \epsilon\} \quad \text{for some } \epsilon > 0.$$

3 Proof of Theorem 1.1

Arguing by contradiction, we begin with the following.

Assumption 1 There is an integral Lagrangian torus $\mathbb{L}$ in $(S^2 \times S^2, \pi_1^*\omega + \pi_2^*\omega)$ which is disjoint from the Clifford torus $L_{1,1}$.

We will show that Assumption 1 can be refined in three ways.

3.1 Refinement 1: we may assume that $\mathbb{L}$ is monotone

A symplectic manifold (M, Ω) is *monotone* if the Chern and area homomorphisms,

$$c_1 : \pi_2(M) \subset H_2(M, \mathbb{Z}) \to \mathbb{Z} \quad \text{and} \quad \Omega : \pi_2(M) \to \mathbb{R},$$

are positively proportional. Recall that a Lagrangian submanifold $L \subset (M, \Omega)$ is *monotone* if its Maslov and area homomorphisms,

$$\mu\colon \pi_2(M, L) \to \mathbb{Z} \quad \text{and} \quad \Omega\colon \pi_2(M, L) \to \mathbb{R},$$

are positively proportional. We will denote the constant of proportionality of L by λ.

If L is a Lagrangian torus, one can verify monotonicity by checking it for a collection of disks whose boundaries generate $H_1(L; \mathbb{Z})$.

Lemma 3.1 *Suppose that (M, Ω) is a symplectic 4–manifold which is monotone with constant $\frac{1}{2}\lambda$. A Lagrangian torus L in (M, Ω) is monotone with constant λ if there are two smooth maps $v_1, v_2\colon (D^2, S^1) \to (M, L)$ such that the boundary maps $v_1|_{S^1}$ and $v_2|_{S^1}$ determine an integral basis of $H_1(L; \mathbb{Z})$ and $\mu([v_i]) = \lambda\Omega([v_i])$ for $i = 1, 2$.*

Refinement 1 is validated by the following result.

Proposition 3.2 *Every integral Lagrangian torus L in $(S^2 \times S^2, \pi_1^*\omega + \pi_2^*\omega)$ is monotone.*

Proof By Theorem C of [5] there is a Hamiltonian diffeomorphism which displaces L from the pair of spheres $S_\infty \cup T_\infty$. Hence, L can be identified with an integral Lagrangian torus $\boldsymbol{L}$ inside the polydisk $P(2-\epsilon, 2-\epsilon) \subset (\mathbb{R}^4, \omega_4)$ for some sufficiently small $\epsilon > 0$. By Lemma 3.1, it suffices to find two smooth maps $v_1, v_2\colon (D^2, S^1) \to (\mathbb{R}^4, \boldsymbol{L})$ such that the boundary maps $v_1|_{S^1}$ and $v_2|_{S^1}$ determine an integral basis of $H_1(\boldsymbol{L}; \mathbb{Z})$ and $\mu([v_i]) = 2\omega_4([v_i])$ for $i = 1, 2$. Simplifying further, we note that, for $\mathbb{R}^4$, the maps μ and ω_4 can be recast as homomorphisms

$$\mu\colon H_1(\boldsymbol{L}; \mathbb{Z}) \to \mathbb{Z} \quad \text{and} \quad \omega_4\colon H_1(\boldsymbol{L}; \mathbb{Z}) \to \mathbb{R}$$

and it suffices to find an integral basis $\{e_1, e_2\}$ of $H_1(\boldsymbol{L}; \mathbb{Z})$ such that $\mu(e_i) = 2\omega_4(e_i)$ for $i = 1, 2$.

Since $\boldsymbol{L}$ is contained in $P(2-\epsilon, 2-\epsilon)$, it follows from [3] that there is a smooth map $f\colon (D, S^1) \to (\mathbb{R}^4, \boldsymbol{L})$ of Maslov index 2 whose symplectic area is 1. To see this we include the polydisk $P(2-\epsilon, 2-\epsilon)$ into $B^4(4-2\epsilon)$, the ball of capacity $4-2\epsilon$, and then compactify this ball to $\mathbb{C}P^2$ equipped with the Fubini–Study form rescaled by $(4-2\epsilon)/\pi$. In this setting, the proofs of Theorems 1.1 and 1.2 of [3] imply that there are three discs mapping to $\mathbb{C}P^2$, with boundary on $\boldsymbol{L}$, that each have Maslov index equal to 2 and positive symplectic areas whose sum is at most $(4-2\epsilon)$. These discs are holomorphic away from $\boldsymbol{L}$ and are obtained from a limit of spheres in the class $[\mathbb{C}P^1]$. Hence, by positivity of intersection, exactly one of the three discs intersects the line at infinity, and the other two discs, f and g, can be viewed as maps to $B^4(4-2\epsilon) \subset \mathbb{R}^4$. Since $\boldsymbol{L}$ is integral, the total symplectic area of f and g is either 2 or 3. In either case, one of them, say f, has symplectic area equal to 1. If e_1 is the element of $H_1(\boldsymbol{L}; \mathbb{Z})$ represented by $f|_{S^1}$, we then have $\mu(e_1) = 2$ and $\omega_4(e_1) = 1$.

Let c be a class in $H_1(\boldsymbol{L};\mathbb{Z})$ such that $\{e_1, c\}$ is an integral basis. Since $\mu(c)$ is even, by adding integer multiples of e_1 to c, if necessary, we may assume that $\mu(c) = 2$. It remains to show that $\omega_4(c) = 1$.

Arguing by contradiction, assume that $\omega_4(c) \neq 1$. Set

$$\widehat{c} = \begin{cases} c & \text{if } \omega_4(c) > 1, \\ c + 2(e_1 - c) & \text{if } \omega_4(c) < 1. \end{cases}$$

Then $\{e_1, \widehat{c}\}$ is an integer basis of $H_1(\boldsymbol{L};\mathbb{Z})$ that satisfies

$$\omega_4(e_1) = 1, \quad \omega_4(\widehat{c}) \geq 2 \quad \text{and} \quad \mu(e_1) = \mu(\widehat{c}) = 2.$$

In [9], Hind and Opshtein prove that if a Lagrangian torus in $P(a,b)$ admits such a basis, then either $a > 2$ or $b > 2$. This contradicts the assumption that $\boldsymbol{L}$ lies in $P(2-\epsilon, 2-\epsilon)$ and we are done. □

3.2 Refinement 2: we may assume that $\mathbb{L}$ lies in the complement of $S_0 \cup S_\infty \cup T_0 \cup T_\infty$

To verify this, we utilize the relative finite-energy foliations from [5], which we now recall.

3.2.1 Foliations of $(S^2 \times S^2) \setminus \boldsymbol{L}$ In [6], Gromov proves that if J is a smooth almost complex structure J on $S^2 \times S^2$ that is tamed by the symplectic form $\pi_1^*\omega + \pi_2^*\omega$, then there is a foliation of $S^2 \times S^2$ by J–holomorphic spheres in the class $(0,1)$, and another with fibers in the class $(1,0)$. For any monotone Lagrangian torus $L \subset (S^2 \times S^2, \pi_1^*\omega + \pi_2^*\omega)$, there is an analogous relative foliation theory, developed first by Ivrii [11] and then completed by Dimitroglou-Rizell, Goodman and Ivrii [5], with input from Wendl [21] and Hind and Lisi [8]. By stretching certain Gromov foliations along L and smoothing the compactifications of the limiting buildings with more than one level, they obtain symplectic S^2–foliations of $S^2 \times S^2$ that are *compatible with* L. A version of this is described below. As in [8], we focus on the curves which, after stretching, map to $S^2 \times S^2 \setminus L$.

Input Let L be a monotone Lagrangian torus in $(S^2 \times S^2, \pi_1^*\omega + \pi_2^*\omega)$. Fix a parametrization ψ of L and the corresponding Weinstein neighborhood

$$\mathcal{U}(L) = \{|p_1| < \epsilon, |p_2| < \epsilon\}.$$

Definition 3.3 A tame almost complex structure J on $(S^2 \times S^2 \setminus L, \pi_1^*\omega + \pi_2^*\omega)$ is said to be *adapted to the parametrization* ψ if, in $\mathcal{U}(L)$, we have

$$J\frac{\partial}{\partial q_i} = -\sqrt{p_1^2 + p_2^2}\,\frac{\partial}{\partial p_i}.$$

For such a J, each negative end of a finite-energy J–holomorphic curve u mapping to $S^2 \times S^2 \setminus L$ is asymptotic to a closed Reeb orbit on a copy of the flat unit cotangent bundle $S^*_L\mathbb{T}^2$ of $\mathbb{T}^2$, corresponding to L. This Reeb orbit covers a closed geodesic γ of the flat metric on $\mathbb{T}^2$. In this case, we simply say that the end of u is asymptotic to L along γ.

Output As described in Section 2.5 of [5], each J adapted to the parametrization ψ of L is part of the limit, as $\tau \to \infty$, of a standard family almost complex structures $J_{\tau \geq 0}$ on $S^2 \times S^2$ that are tame with respect to $\pi_1^* \omega + \pi_2^* \omega$. Taking the limit of the Gromov foliations for the J_τ as $\tau \to \infty$, it follows from Theorem D and Propositions 5.3 and 5.16 of [5], and the fact that L is monotone, that one obtains a foliation $\mathcal{F} = \mathcal{F}(L, \psi, J)$ of $S^2 \times S^2 \setminus L$ with the following properties.

- The foliation $\mathcal{F}$ has two kind of leaves: unbroken ones consisting of a single closed J–holomorphic sphere in $S^2 \times S^2 \setminus L$ of class $(0, 1)$, and broken leaves consisting of a pair of finite-energy J–holomorphic planes in $S^2 \times S^2 \setminus L$.
- Each leaf of $\mathcal{F}$ intersects S_∞ in exactly one point. For a broken leaf this means that exactly one of its planes intersects S_∞.
- The ends of two planes of a broken leaf are asymptotic to the same geodesic, but with opposite orientations. This geodesic is embedded. We denote its homology class, equipped with the orientation determined by the plane which intersects S_∞, by $\beta \in H_1(L; \mathbb{Z})$. This class is the same for all broken leaves of $\mathcal{F}$ and is referred to as the foliation class of $\mathcal{F}$.
- Limits of the Gromov spheres in the completion of a neighborhood of L, which is a copy of $T^* \mathbb{T}^2$, are cylinders asymptotic to geodesics in the classes $\pm\beta$.
- Each point $z \in S^2 \times S^2 \setminus L$ lies in a unique leaf of $\mathcal{F}$, and each point of L lies on a unique geodesic in the foliation class β that corresponds to a unique plane of a broken leaf of $\mathcal{F}$ that intersects S_∞.
- If L is disjoint from a configuration of symplectic spheres, then we may assume these spheres are complex with respect to all J_τ. In particular, if L has been displaced from $S_0 \cup S_\infty \cup T_0 \cup T_\infty$, then we may assume this configuration of symplectic spheres is J–complex.
- Suppose L is disjoint from S_∞ and we therefore take S_∞ to be complex. Then, by positivity of intersection, there is a well-defined map $p\colon S^2 \times S^2 \to S_\infty$ which takes $z \in S^2 \times S^2 \setminus L$ to the unique intersection of its leaf with S_∞, and takes $z \in L$ to the intersection with S_∞ of the broken leaf asymptotic to the unique geodesic through z representing the foliation class. The image $p(L)$ is an embedded closed curve in S_∞. Moreover, if L is homotopic to $L_{1,1}$ in the complement of $T_0 \cup T_\infty$, then $p(T_0)$ and $p(T_\infty)$ — which are points, since T_0 and T_∞ are complex — lie on opposite sides of the closed curve $p(L)$.

Lemma 3.4 (straightening) *For all sufficiently small $\epsilon > 0$ we may assume, by perturbing J outside of $\mathcal{U}(L)$, that the unbroken leaves of $\mathcal{F}$ that intersect $\mathcal{U}(L)$ do so along the annuli*

$$\{p_1 = \delta, q_1 = \theta, -\epsilon < p_2 < \epsilon\}$$

for some $\theta \in S^1$ and nonzero $\delta \in (-\epsilon, \epsilon)$.

Proof The statement for broken leaves was established in Proposition 5.16 of [5]; see the first bullet point of the proof. This means the parts of the broken leaves lying outside of $\mathcal{U}(L)$ form two S^1–families of

holomorphic disks, with boundaries $\{p_1 = 0, q_1 = \theta, p_2 = \pm\epsilon\}$. We may smoothly identify a neighborhood of such an S^1–family of holomorphic disks in the complement of $\mathcal{U}(L)$ with $(-\epsilon_0, \epsilon_0) \times S^1 \times D^2$, where the disks correspond to subsets $\{0\} \times \{\theta\} \times D^2$, and the circles $\{p_1 = \delta, q_1 = \theta, p_2 = \pm\epsilon\}$ in $\partial\mathcal{U}(L)$ match with the circles $\{\delta\} \times \{\theta\} \times \partial D^2$. Hence our S^1–families of holomorphic disks can be extended to smooth families of disjoint smoothly embedded disks $D_{\delta,\theta} = \{\delta\} \times \{\theta\} \times D^2$ with $|\delta| < \epsilon_0$ and $\theta \in S^1$. We may assume these disks extend smoothly into $\mathcal{U}(L)$ along the surfaces $\{p_1 = \delta, q_1 = \theta, -\epsilon < p_2 < \epsilon\}$. For a sufficiently small $\epsilon_0 > 0$, we may also assume that the disks $D_{\delta,\theta}$ are symplectic, since they are C^∞–close to the holomorphic disks corresponding to $\delta = 0$. Hence, we may choose a tame almost complex structure, J_0, which agrees with J inside $\mathcal{U}(L)$ but is chosen outside of $\mathcal{U}(L)$ so that the disks $D_{\delta,\theta}$ are J_0–holomorphic. With this, we replace the foliation $\mathcal{F} = \mathcal{F}(L, \psi, J)$ with the foliation $\mathcal{F}_0 = \mathcal{F}(L, \psi, J_0)$ and the neighborhood $\mathcal{U}(L)$ with

$$\mathcal{U}_0(L) = \{|p_1| < \epsilon_0, |p_2| < \epsilon\}.$$

By construction, for each annulus $\{p_1 = \delta, q_1 = \theta, -\epsilon < p_2 < \epsilon\}$ with $|\delta| < \epsilon_0$, there is a pair of J_0–holomorphic disks which join smoothly with the boundary components to form J_0–holomorphic spheres in the class $(0, 1)$. These are unbroken leaves of $\mathcal{F}_0$ for $\delta \neq 0$, and broken leaves for $\delta = 0$. Moreover, by positivity of intersection, these are the only leaves of $\mathcal{F}_0$ intersecting $\mathcal{U}_0(L)$. □

Example 3.5 (solid tori bounded by $L_{1,1}$) For the Clifford torus $L_{1,1} \subset S^2 \times S^2$ and a J adapted to the standard parametrization $\psi_{1,1}$ of $L_{1,1}$, we get a foliation $\mathcal{F}_{1,1}$ of $S^2 \times S^2 \setminus L_{1,1}$ with leaves in the class $(0, 1)$. As $L_{1,1}$ is disjoint from $S_0 \cup S_\infty \cup T_0 \cup T_\infty$ we may assume that these four spheres are J–complex. The broken leaves of $\mathcal{F}_{1,1}$ comprise two families of J–holomorphic planes with boundary on $L_{1,1}$, which can be labeled as follows: $\mathfrak{s}_0$, which consists of the planes intersecting S_0, and $\mathfrak{s}_\infty$, which consists of the planes intersecting S_∞. The families of holomorphic planes $\mathfrak{s}_0$ and $\mathfrak{s}_\infty$ can be seen directly for a model almost complex structure, but in fact exist for all J adapted to a parametrization of $L_{1,1}$. We will write $\mathfrak{s}_0(J)$ and $\mathfrak{s}_\infty(J)$ when we want to highlight the dependence of these families on J. Modulo reparametrization, $\mathfrak{s}_0$ and $\mathfrak{s}_\infty$ form compact moduli spaces, as they represent classes of minimal positive area in $H_2(S^2 \times S^2, L_{1,1})$. These moduli spaces are automatically regular by [21, Theorem 1].

For each J as above, there is an analogous foliation of $S^2 \times S^2 \setminus L_{1,1}$ with leaves in the class $(1, 0)$. The broken leaves in this case yield two families of J–holomorphic planes, $\mathfrak{t}_0$ and $\mathfrak{t}_\infty$, which consist of the planes intersecting T_0 and T_∞, respectively.

The following result establishes Refinement 2. The proof is based on that of Corollary E in [5].

Proposition 3.6 *Suppose that $\mathbb{L}$ is a monotone Lagrangian torus in $(S^2 \times S^2, \pi_1^*\omega + \pi_2^*\omega)$ that is disjoint from $L_{1,1}$. Then there is a Hamiltonian diffeomorphism ϕ of $S^2 \times S^2$ which displaces $\mathbb{L}$ from $S_0 \cup S_\infty \cup T_0 \cup T_\infty$ and is supported away from $L_{1,1}$. Moreover, $\phi(\mathbb{L})$ is homotopic to $L_{1,1}$ in the complement of $T_0 \cup T_\infty$ and also in the complement of $S_0 \cup S_\infty$.*

Proof We first displace $\mathbb{L}$ from S_∞ in the complement of $L_{1,1}$, or equivalently S_∞ from $\mathbb{L}$. Let J_0 be an almost complex structure on $S^2 \times S^2 \setminus L_{1,1}$ that is adapted to the standard parametrization $\psi_{1,1}$ of $L_{1,1}$ and such that S_∞ is J_0–complex. We deform J_0 through almost complex structures J_τ to an almost complex structure J_∞ which is also adapted to a parametrization of $\mathbb{L}$. For each τ we have a finite-energy J_τ–holomorphic foliation in the class $(1,0)$ including the broken leaves $\mathfrak{t}_0$ and $\mathfrak{t}_\infty$ as in Example 3.5. We can find a smooth family of holomorphic spheres H_τ in the class $(1,0)$, that is, unbroken leaves of the corresponding foliations, with $H_0 = S_\infty$.

In the limit $\tau \to \infty$, the moduli spaces of J_τ–holomorphic disks $\mathfrak{t}_0(J_\tau)$ and $\mathfrak{t}_\infty(J_\tau)$ both converge. Indeed, their limits $\mathfrak{t}_0(J_\infty)$ and $\mathfrak{t}_\infty(J_\infty)$ still represent classes of minimal positive area in $H_2(S^2 \times S^2, L_{1,1} \cup \mathbb{L})$, where all classes still have integral area; see also Lemma 3.23 for this. Moreover, as the union of broken spheres with respect to J_∞ still has codimension 1, we may assume the H_τ converge to an unbroken sphere, H_∞. As the H_τ are all disjoint from $L_{1,1}$ and H_∞ is disjoint from $\mathbb{L}$, we can find a Hamiltonian isotopy supported away from $L_{1,1}$ displacing S_∞ from $\mathbb{L}$, as required.

The remainder of the argument follows similar lines. We may assume that each of the spheres S_0, S_∞, T_0 and T_∞ are J_0–complex. Let $\mathcal{F}_0$ be the corresponding J_0–holomorphic foliation of $S^2 \times S^2 \setminus L_{1,1}$ in the class $(0,1)$. Let $p_0 \colon S^2 \times S^2 \to S_\infty$ be the projection map from the (sixth bullet point of the) description of $\mathcal{F}$ above. We may assume that the points $p_0(T_0)$ and $p_0(T_\infty)$ lie in different components of $S_\infty \setminus p_0(L_{1,1})$.

Fix a parametrization of $\mathbb{L}$ and let (P_1, P_2, Q_1, Q_2) be the corresponding local coordinates on the Weinstein neighborhood $\mathcal{U}(\mathbb{L})$ of $\mathbb{L}$. We may also assume that $\mathcal{U}(\mathbb{L})$ is disjoint from the Weinstein neighborhood $\mathcal{U}(L_{1,1})$ corresponding to $\psi_{1,1}$. Consider a smooth family $J_{t\in[0,1]}$ of almost complex structures on $S^2 \times S^2 \setminus L_{1,1}$ such that each J_t is equal to J_0 in $\mathcal{U}(L_{1,1})$, and in $\mathcal{U}(\mathbb{L})$ we have

$$J_1 \frac{\partial}{\partial Q_i} = -\frac{\partial}{\partial P_i}.$$

We can then smoothly extend the family J_t to $t > 1$ to stretch (to length t) along a small sphere bundle in $\mathcal{U}(\mathbb{L})$, as in [1]. This yields a family of foliations $\mathcal{F}_t$ of $S^2 \times S^2 \setminus L_{1,1}$. Since the planes of the broken leaves of $\mathcal{F}_0$ have minimal area they persist under the deformation to yield the planes of the broken leaves of $\mathcal{F}_t$. This yields a family of maps $p_t \colon S^2 \times S^2 \to S_\infty$.

Lemma 3.7 *The sets $p_t(\mathbb{L})$ in S_∞ converge in the Hausdorff topology to a subset of a circle $C_\infty \in S_\infty$ as $t \to \infty$.*

Proof Let J_∞ be the limiting almost complex structure on $S^2 \times S^2 \setminus (\mathbb{L} \cup L_{1,1})$. The circle C_∞ is the intersection with S_∞ of the broken leaves of the J_∞ foliation which are asymptotic to $\mathbb{L}$. Now, $p_t(\mathbb{L})$ consists of the intersection with S_∞ of J_t–holomorphic spheres which intersect $\mathbb{L}$. Hence a sequence of points $z_t \in p_t(\mathbb{L})$ corresponds to a sequence of J_t–holomorphic curves in the class $(0,1)$ which all intersect $\mathbb{L}$. Up to taking a subsequence, this sequence of curves converges to a broken curve asymptotic to $\mathbb{L}$ and hence the z_t converge to a point in C_∞. □

Lemma 3.8 *If we denote the projection with respect to the fully stretched almost complex structure by p_∞, then $C_\infty = p_\infty(\mathbb{L})$ is disjoint from $p_\infty(L_{1,1})$.*

Proof This follows from the fact that the original planes of the broken leaves have area 1 and so cannot degenerate further. Indeed, since $\mathbb{L}$ is monotone, any holomorphic curve asymptotic to $\mathbb{L}$ must have integral area, and in particular curves in the class $(0,1)$ cannot converge to buildings with more than two top level curves. □

The results above imply that there is an $N > 0$ such that $p_t(L_{1,1})$ is disjoint from C_∞ for all $t \geq N$. With this we can choose two continuous curves $\gamma_0, \gamma_\infty : [0,\infty) \to S_\infty$ with the following properties:

- $\gamma_0(0) = p_0(T_0)$ and $\gamma_\infty(0) = p_0(T_\infty)$,
- $\gamma_0(t)$ and $\gamma_\infty(t)$ are disjoint from $p_t(L_{1,1})$ for all $t \in [0,\infty)$,
- for some $N > 0$, both $\gamma_0(t)$ and $\gamma_\infty(t)$ are disjoint from C_∞, and C_∞ is disjoint from $p_t(L_{1,1})$ for all $t \geq N$,
- C_∞ separates $\gamma_0(N)$ and $\gamma_\infty(N)$ in S_∞.

For each $t \in [0,\infty)$, both $p_t^{-1}(\gamma_0(t))$ and $p_t^{-1}(\gamma_\infty(t))$ are J_t–holomorphic spheres in the class $(0,1)$ disjoint from $L_{1,1}$. The family of spheres

$$\{p_t^{-1}(\gamma_0(t))\}_{t\in[0,N]}$$

forms a symplectic isotopy, which displaces T_0 from $\mathbb{L}$ in the complement of $L_{1,1}$. Similarly, the family of spheres

$$\{p_t^{-1}(\gamma_\infty(t))\}_{t\in[0,N]}$$

forms a symplectic isotopy which displaces T_∞ from $\mathbb{L}$ in the complement of $L_{1,1}$. Moreover, these isotopies can be generated by a single Hamiltonian flow on $S^2 \times S^2$ that fixes $L_{1,1}$. The inverse flow displaces $\mathbb{L}$ from $T_0 \cup T_\infty$. The final separation condition is enough to guarantee the homotopy condition in the theorem.

By considering also the J_t–holomorphic foliation in the class $(1,0)$ (see Example 3.5), we can displace $\mathbb{L}$ from $S_0 \cup S_\infty$ too. After adjusting the isotopy of $S_0 \cup S_\infty$, we may assume that it fixes $T_0 \cup T_\infty$; see Corollary 3.7 of [5]. Hence the inverse flow will not reintroduce intersections with T_0 or T_∞. □

3.3 Refinement 3: we may assume that $\mathbb{L}$ is homologically trivial in $(S^2 \times S^2) \setminus (S_0 \cup S_\infty \cup T_0 \cup T_\infty)$

To see this, note that $(S^2 \times S^2) \setminus (S_0 \cup S_\infty \cup T_0 \cup T_\infty)$ can be identified with a subset of the cotangent bundle of $\mathbb{T}^2$ in which $L_{1,1}$ is identified with the zero section. In this setting we can invoke the following.

Theorem 3.9 [5, Theorem 7.1] *A homologically nontrivial Lagrangian torus L in $(T^*\mathbb{T}^2, d\lambda)$ is Hamiltonian isotopic to a constant section. In particular, if L is exact then it is Hamiltonian isotopic to the zero section.*

If our monotone Lagrangian $\mathbb{L}$ was homologically nontrivial in $(S^2\times S^2)\setminus(S_0\cup S_\infty\cup T_0\cup T_\infty)$ it would then follow from Theorem 3.9 and Section 2.3.B$''_4$ of [6] that $\mathbb{L}\cap L_{1,1}\neq\varnothing$, which would contradict our original assumption.

3.4 A path to the proof of Theorem 1.1

By the three refinements established above, it suffices to show that the following assumption is false.

Assumption 2 There is a monotone Lagrangian torus $\mathbb{L}$ in the set

$$Y=(S^2\times S^2)\setminus(S_0\cup S_\infty\cup T_0\cup T_\infty)$$

which is disjoint from the Clifford torus $L_{1,1}$ and is homologically trivial in Y.

A path to a contradiction To obtain a contradiction to Assumption 2, we will show, using a sequence of blow-ups, inflations and blow-downs, that it implies the existence of two disjoint monotone Lagrangian tori in a new (monotone) copy of $S^2\times S^2$, which are both Hamiltonian isotopic to the Clifford torus therein, and hence cannot be disjoint.

To perform the necessary sequence of blow-ups, inflations and blow-downs, we must first establish the existence of a special collection of symplectic spheres and disks in our current model; see Proposition 3.24. These spheres and discs must be well placed with respect to a holomorphic foliation of $S^2\times S^2\setminus(\mathbb{L}\cup L_{1,1})$ which we introduce below in Section 3.5. They are obtained from special holomorphic buildings, whose existence we establish in Section 3.7. These existence results rely on the analysis of a general stretching scenario that is contained in Section 3.6.

Remark 3.10 To falsify Assumption 2, we must use it to build and analyze a complicated set of secondary objects in order to derive a contradiction. The reader is asked to bear in mind that many of the results established in the remainder of this section hold in a setting which will later be shown to be impossible.

3.5 Straightened holomorphic foliations of $S^2\times S^2\setminus(\mathbb{L}\cup L_{1,1})$, under Assumption 2

Let $\mathbb{L}$ be a Lagrangian torus as in Assumption 2. Here we describe the holomorphic foliations of $S^2\times S^2\setminus(\mathbb{L}\cup L_{1,1})$ that are implied by the existence of $\mathbb{L}$.

Let ψ be a parametrization of $\mathbb{L}$ and $\psi_{1,1}$ be the standard parametrization of $L_{1,1}$. Consider a tame almost complex structure J on $(S^2\times S^2\setminus(\mathbb{L}\cup L_{1,1}),\pi_1^*\omega+\pi_2^*\omega)$ which is adapted to both ψ and $\psi_{1,1}$. We will always make the following assumption.

(A1) J is equal to the standard product complex structure near S_0, S_∞, T_0 and T_∞. In particular, T_0 and T_∞ are unbroken leaves of the foliation.

Let J_τ be the family of almost complex structures on $S^2\times S^2$ that are determined by J as in [5, Section 2.5]. Taking the limit of the Gromov foliations in the class $(0,1)$ with respect to the J_τ as $\tau\to\infty$, and arguing as in [5], we get a J–holomorphic foliation

$$\mathscr{F} = \mathscr{F}(\mathbb{L}, L_{1,1}, \psi, \psi_{1,1}, J)$$

of $S^2\times S^2\setminus(\mathbb{L}\cup L_{1,1})$. The features of this foliation are described below and are illustrated in Figure 1.

Each leaf of $\mathscr{F}$ still intersects S_∞ in exactly one point, but there are now three types of leaves. The first are unbroken leaves consisting of a single closed J–holomorphic sphere in $S^2\times S^2\setminus(\mathbb{L}\cup L_{1,1})$ of class $(0,1)$. The second type of leaves are broken and consist of a pair of finite-energy J–holomorphic planes in $S^2\times S^2\setminus(\mathbb{L}\cup L_{1,1})$ that are asymptotic to $L_{1,1}$ along the same embedded geodesic with opposite orientations. As in Example 3.5, the collection of planes like this which intersect S_∞ comprise a one-dimensional family, $\mathfrak{s}_\infty$, and their companion planes comprise a family $\mathfrak{s}_0$. The third class of leaves are also broken, but consist of pairs of finite-energy J–holomorphic planes in $S^2\times S^2\setminus(\mathbb{L}\cup L_{1,1})$ that are asymptotic to $\mathbb{L}$. These pairs also have matching ends. We denote by $\mathfrak{r}_\infty$ the set of all the J–holomorphic planes of broken leaves that are asymptotic to $\mathbb{L}$ and intersect S_∞. The collection of their companion planes will be denoted by $\mathfrak{r}_0$. As established below in Lemma 3.11, the planes of $\mathfrak{r}_\infty$ intersect both S_∞ and S_0 while the planes of $\mathfrak{r}_0$ intersect neither of these spheres. Since curves in the class $(0,1)$ have area 2, and by monotonicity planes asymptotic to our Lagrangians have integral area, no more complicated degenerations are possible.

Note that there are now two foliation classes, $\beta_{\mathbb{L}}$ and $\beta_{L_{1,1}}$, determined by each of the two classes of broken leaves.

The foliation $\mathscr{F}$ also defines a projection map

$$p\colon S^2\times S^2\to S_\infty.$$

In this setting, the images $p(L_{1,1})$ and $p(\mathbb{L})$ are disjoint embedded circles in S_∞ which, by Proposition 3.6, are disjoint from $T_0\cup T_\infty$ and are homotopic in the complement. Therefore, without loss of generality, there are disjoint closed disks $A_0\subset S_\infty$ with boundary $p(\mathbb{L})$ and $A_\infty\subset S_\infty$ with boundary $p(L_{1,1})$, such that $p(T_0)\in A_0$ and $p(T_\infty)\in A_\infty$. Denote the closed annulus defined by the closure of $S_\infty\setminus(A_0\cup A_\infty)$ by B. These features of $\mathscr{F}$ are all represented in Figure 1.

Lemma 3.11 *The planes in $\mathfrak{r}_\infty$ intersect both S_0 and S_∞. Equivalently, the planes in $\mathfrak{r}_0$ are disjoint from $S_0\cup S_\infty$.*

Proof We define a relative homology class $\Sigma\in H_2(S^2\times S^2, (S_0\cup S_\infty\cup T_0\cup T_\infty)$ by first choosing an embedded path $\gamma\colon[0,1]\to S_\infty$ with $\gamma(0)=T_0\cap S_\infty$ and $\gamma(1)=T_\infty\cap S_\infty$. Then choose a family of embedded paths σ_t in each $p^{-1}(\gamma(t))$ going from S_∞ to S_0. The union of the σ_t define Σ. By Proposition 3.6 we may assume that γ intersects $p(\mathbb{L})$ in a single point $\gamma(t_0)$. The intersection

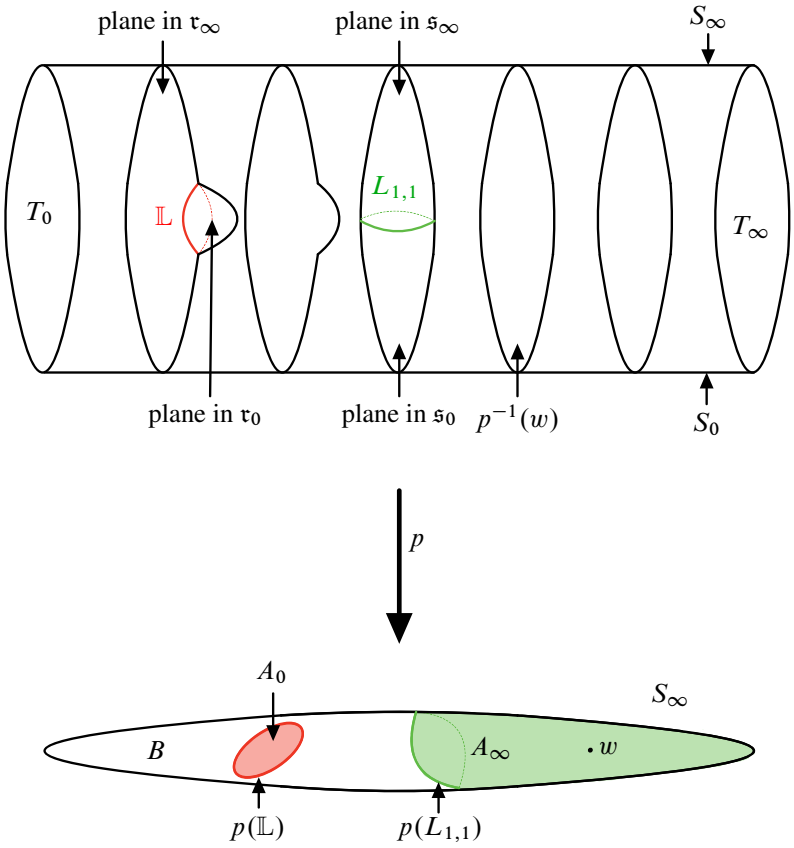


Figure 1: The foliation $\mathcal{F}$ of $S^2 \times S^2 \setminus (\mathbb{L} \cup L_{1,1})$.

$\mathbb{L} \cap p^{-1}(\gamma(t_0))$ is an embedded circle, bounding disks from $\mathfrak{r}_0$ and $\mathfrak{r}_\infty$. The disks in $\mathfrak{r}_\infty$ intersect S_∞ by definition, so arguing by contradiction, if $\mathfrak{r}_0$ happens to intersect S_0 then our circle $\mathbb{L} \cap p^{-1}(\gamma(t_0))$ separates S_0 and S_∞, and thus must intersect the path σ_{t_0}. This is the only intersection between Σ and $\mathbb{L}$, and so we would conclude that $\Sigma \bullet \mathbb{L}$ is nontrivial, contradicting Refinement 3. □

Straightening Let (P_1, P_2, Q_1, Q_2) be coordinates in the neighborhood $\mathcal{U}(\mathbb{L})$ of $\mathbb{L}$ determined by ψ, and let (p_1, p_2, q_1, q_2) be coordinates in the neighborhood $\mathcal{U}(L_{1,1})$ of $L_{1,1}$ determined by $\psi_{1,1}$. As in Lemma 3.4, where we had only one Lagrangian torus, we may assume that the leaves of $\mathcal{F}$ are straight in both $\mathcal{U}(\mathbb{L})$ and $\mathcal{U}(L_{1,1})$. In particular, we may assume that the unbroken leaves of $\mathcal{F}$ that intersect $\mathcal{U}(\mathbb{L})$ do so along the annuli $\{P_1 = \delta \neq 0, Q_1 = \theta, |P_2| < \epsilon\}$, the planes of $\mathfrak{r}_\infty$ intersect $\mathcal{U}(\mathbb{L})$ along the annuli $\{P_1 = 0, Q_1 = \theta, 0 < P_2 < \epsilon\}$, and the planes of $\mathfrak{r}_0$ intersect $\mathcal{U}(\mathbb{L})$ along the annuli $\{P_1 = 0, Q_1 = \theta, -\epsilon < P_2 < 0\}$. Similarly, we may assume that the unbroken leaves of $\mathcal{F}$ that intersect $\mathcal{U}(L_{1,1})$ do so along the annuli $\{p_1 = \delta \neq 0, q_1 = \theta, |p_2| < \epsilon\}$, the planes of $\mathfrak{s}_\infty$ intersect $\mathcal{U}(L_{1,1})$ along the annuli $\{p_1 = 0, q_1 = \theta, 0 < p_2 < \epsilon\}$, and the planes of $\mathfrak{s}_0$ intersect $\mathcal{U}(L_{1,1})$ along the annuli $\{p_1 = 0, q_1 = \theta, -\epsilon < p_2 < 0\}$.

The map p can also be described simply in these Weinstein neighborhoods. In $\mathcal{U}(\mathbb{L})$, we may assume that the region $\{P_1 < 0\} \subset \mathcal{U}(\mathbb{L})$ is mapped by p into the interior of A_0, and $\{P_1 > 0\} \subset \mathcal{U}(\mathbb{L})$ is mapped by p into the interior of B. Similarly, we may assume that in $\mathcal{U}(L_{1,1})$ the region $\{p_1 > 0\} \subset \mathcal{U}(L_{1,1})$ is mapped by p into the interior of A_∞ and $\{p_1 < 0\} \subset \mathcal{U}(L_{1,1})$ is mapped by p into the interior of B.

Using some of the freedoms available in the choice of ψ and $\psi_{1,1}$, we can add the following additional assumption:

(A2) The foliation class $\beta_{\mathbb{L}}$ is equal to $(0,-1) \in H_1^{\psi}(\mathbb{L};\mathbb{Z})$, and the foliation class $\beta_{L_{1,1}}$ is equal to $(0,-1) \in H_1^{\psi_{1,1}}(L_{1,1};\mathbb{Z})$.

3.6 Stretching scenario for class $(1, d)$, under Assumption 2

Recall that for each nonnegative integer d and a generic tame almost complex structure J on $S^2 \times S^2$ there exists a smooth J–holomorphic sphere $u\colon S^2 \to S^2 \times S^2$ representing the class $(1, d)$. Moreover, this curve is unique, up to reparametrization, if we impose $2d + 1$ constraint points. To see this, note that for the integrable product complex structure such a curve can be written explicitly as the graph of a degree d rational map, and this implies that the Gromov–Witten invariant associated to the homology class and point constraints is 1. Hence, nodal curves will exist for all tame almost complex structures and away from a codimension 2 subset of almost complex structures we will have smooth curves. The uniqueness in the assertion above follows because spheres in the class $(1, d)$ have self-intersection number $2d$, so distinct spheres cannot satisfy the same $2d + 1$ point constraints.

Let J_τ, for $\tau \geq 0$, be the family of almost complex structures on $S^2 \times S^2$ used in Section 3.5 to obtain the foliation $\mathcal{F}$. For a sequence $\tau_k \to \infty$, let $u_{k,d}\colon S^2 \to S^2 \times S^2$ be a sequence of J_{τ_k}–holomorphic curves in the class $(1, d)$ that converges to a holomorphic building $\boldsymbol{F}_d$ as in [1]. The limit $\boldsymbol{F}_d$ consists of genus zero holomorphic curves in three levels. The *top level* curves map to $S^2 \times S^2 \setminus (\mathbb{L} \cup L_{1,1})$ and are J–holomorphic. The *middle level* curves map to one of two copies of $\mathbb{R} \times S^*\mathbb{T}^2$, the symplectization of the unit cotangent bundle of the flat torus. These copies correspond to $\mathbb{L}$ and $L_{1,1}$ and the identifications are defined by the parametrizations ψ and $\psi_{1,1}$. It follows from the definition of the family J_τ that these middle level curves are all J_{cyl}–holomorphic where J_{cyl} is a fixed cylindrical almost complex structure. Similarly, the *bottom level* curves of the limiting building map to one of two copies of $T^*\mathbb{T}^2$ and are J_{std}–holomorphic, where J_{std} is a standard complex structure.

Each top level curve of $\boldsymbol{F}_d$ can be compactified to yield a map from a surface of genus zero with boundary to $(S^2 \times S^2, \mathbb{L} \cup L_{1,1})$. The components of the boundary correspond to the negative punctures of the curve. They are mapped to the closed geodesics on $\mathbb{L}$ or $L_{1,1}$ underlying the Reeb orbits to which the corresponding puncture is asymptotic. The middle and bottom level curves can be compactified to yield maps to either $\mathbb{L}$ or $L_{1,1}$ with the same type of boundary conditions. These compactified maps can all be glued together to form a map $\bar{\boldsymbol{F}}_d\colon S^2 \to S^2 \times S^2$ in the class $(1, d)$.

Definition 3.12 A J–holomorphic curve u in $S^2 \times S^2 \setminus (\mathbb{L} \cup L_{1,1})$ is said to be *essential* (*with respect to the foliation* $\mathcal{F}$) if the map $p \circ u$ is injective.

Definition 3.13 Let u be a J–holomorphic curve in $S^2 \times S^2 \setminus (\mathbb{L} \cup L_{1,1})$. A puncture of u is said to be of *foliation type with respect to* $\mathbb{L}$ ($L_{1,1}$) if it is asymptotic to a closed Reeb orbit which lies on the copy of $S^*\mathbb{T}^2$ that corresponds to $\mathbb{L}$ ($L_{1,1}$) and covers a closed geodesic in an integer multiple of the foliation class $\beta_{\mathbb{L}}$ ($\beta_{L_{1,1}}$). The puncture is of *positive* (*negative*) *foliation type* if this integer is positive (negative).

Lemma 3.14 *Let u be a J–holomorphic curve in $S^2 \times S^2 \setminus (\mathbb{L} \cup L_{1,1})$ with a puncture. Let $\{c_l\}$ be a sequence of circles in the domain of u which lie in a standard neighborhood of the puncture, wind once around it, and converge to it in the Hausdorff topology. If the puncture is of foliation type with respect to $\mathbb{L}$ ($L_{1,1}$), then the sets $p(u(c_l))$ converge to a point on $p(\mathbb{L})$ ($p(L_{1,1})$). Moreover each $p(u(c_l))$ either maps into the point (in which case u covers a plane in a broken leaf) or it winds nontrivially around the point. If the puncture is not of foliation type then the sets $p(u(c_l))$ converge to $p(\mathbb{L})$ ($p(L_{1,1})$).*

Proof This follows from the exponential convergence theorem from [10]. □

Corollary 3.15 *If u is an essential J–holomorphic curve in $S^2 \times S^2 \setminus (\mathbb{L} \cup L_{1,1})$, then its punctures on $\mathbb{L}$ are either all of foliation type or none of them are, and similarly for the punctures on $L_{1,1}$. If u has no punctures of foliation type, then it is either a J–holomorphic plane or cylinder.*

Assume u has no punctures of foliation type. If u is a plane, then the closure of the image of $p \circ u$ is A_0 or A_∞ or the closure of their complements in S_∞. If u is a cylinder, then the closure of the image of $p \circ u$ is B.

Proof Lemma 3.14 implies that if u has punctures of both foliation type and not of foliation type on $\mathbb{L}$ or $L_{1,1}$, then $p \circ u$ will not be injective. Indeed, the projection of a small circle around a puncture of foliation type on $\mathbb{L}$ (resp. $L_{1,1}$) will intersect any circle sufficiently close to $p(\mathbb{L})$ (resp. $P(L_{1,1})$), including the projections of small circles around punctures not of foliation type.

Essential curves with all punctures not of foliation type project onto connected subsets of S_∞ with boundary components equal to $\mathbb{L}$ or $L_{1,1}$. Checking possibilities, the second half of the statement follows. □

The following result can be proved in the same way as Lemma 6.2 in [8].

Lemma 3.16 *Let u be an essential curve whose punctures on $\mathbb{L}$ are all of foliation type. Then these punctures are either all positive or all negative* (see Definition 3.13).

Similarly, let v be an essential curve whose punctures on $L_{1,1}$ are all of foliation type. Then the punctures on $L_{1,1}$ are either all positive or all negative.

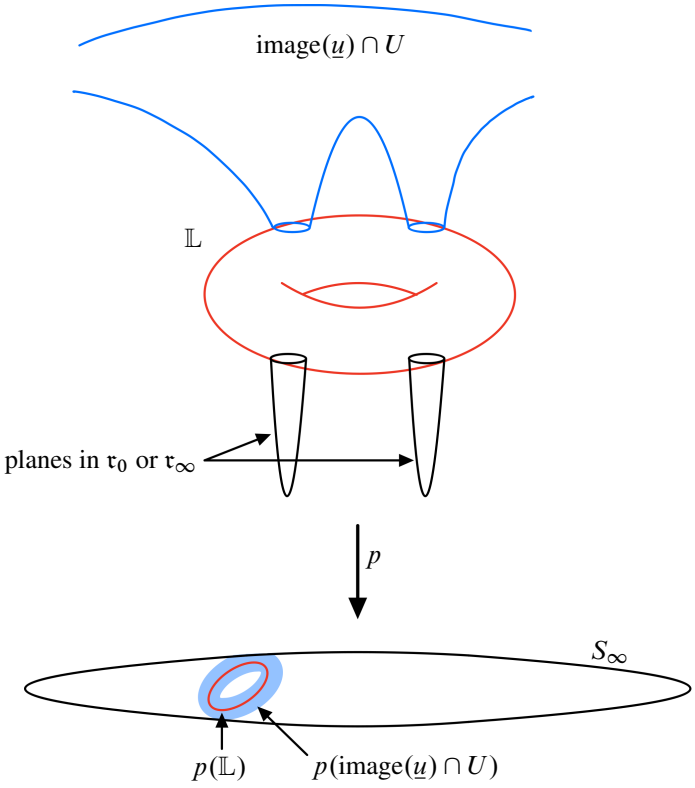


Figure 2: Images of curves of a limit $\boldsymbol{F}_d$ of Type 2b in a neighborhood U of $p^{-1}(p(\mathbb{L}))$. If one replaces $\underline{u}$ with u_d, then this picture also works for limits $\boldsymbol{F}_d$ of Type 1.

Let $u_{k,d}$ be a sequence converging to $\boldsymbol{F}_d$ as in the *stretching scenario for class* $(1,d)$. Positivity of intersection implies that the curves $u_{k,d}$ must intersect each leaf of $\mathcal{F}$ exactly once. This fact imposes several important restrictions on $\boldsymbol{F}_d$ in relation to the foliation $\mathcal{F}$, allowing us to identify a handful of possible limit types.

Proposition 3.17 *Let $\boldsymbol{F}_d$ be a limit as in the **stretching scenario for class $(1,d)$**. Then the building $\boldsymbol{F}_d$ is of one of the following types.*

- ***Type 0*** *$\boldsymbol{F}_d$ is a (possibly nodal) J–holomorphic sphere in $S^2\times S^2\setminus(\mathbb{L}\cup L_{1,1})$ in the class $(1,d)$, where one (essential) sphere lies in the class $(1,j)$ for some $1\le j\le d$, and any remaining top level curves are either spheres covering unbroken leaves of the foliation, or pairs of planes covering broken leaves of the foliation. Any middle and bottom level curves are cylinders asymptotic to Reeb orbits in multiples of the foliation class.*
- ***Type 1*** *$\boldsymbol{F}_d$ has a unique essential curve u_d. The punctures of u_d are all of foliation type, and along $\mathbb{L}$, and also $L_{1,1}$, are either all positive or all negative. The image of $p\circ u_d$ is S_∞ minus finitely many points on $p(\mathbb{L})\cup p(L_{1,1})$. The other top level curves of $\boldsymbol{F}_d$ either cover unbroken*

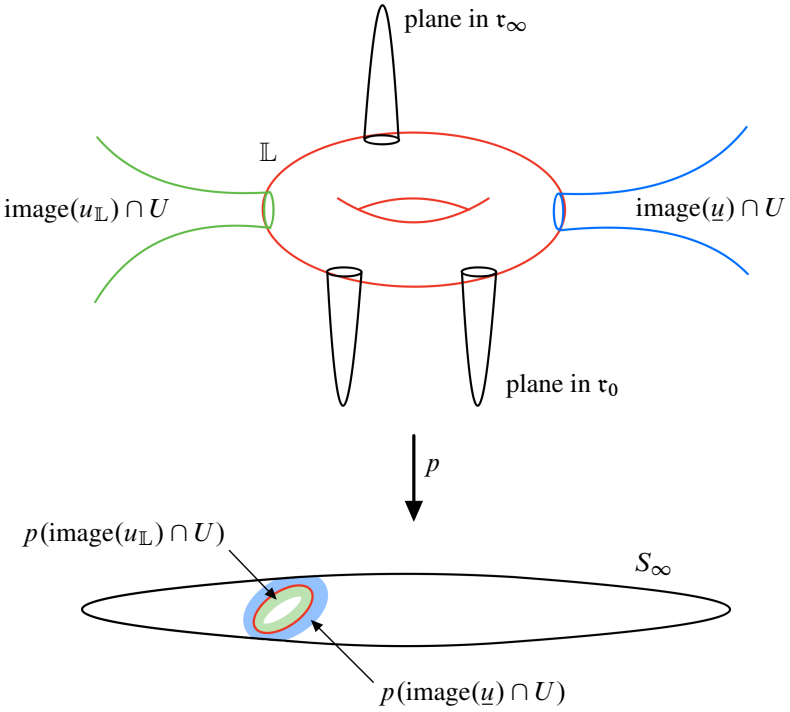


Figure 3: Images of curves of a limit $\boldsymbol{F}_d$ of Type 2a or 3 in a neighborhood U of $p^{-1}(p(\mathbb{L}))$.

leaves of the foliation, or they are J–holomorphic planes covering one of the planes of a broken leaf of the foliation. Any middle and bottom level curves cover cylinders asymptotic to Reeb orbits in multiples of the foliation class.

- **Type 2a** *$\boldsymbol{F}_d$ has exactly two essential curves, $u_{\mathbb{L}}$ and $\underline{u}$. The closures of the images of the maps $p \circ u_{\mathbb{L}}$ and $p \circ \underline{u}$ are A_0 and $B \cup A_\infty$, respectively. Any punctures of $\underline{u}$ on $L_{1,1}$ are all of foliation type and are either all positive or all negative. The other top level curves of $\boldsymbol{F}_d$ cover (broken or unbroken) leaves of $\mathcal{F}$. Any middle and bottom level curves in the copy of $T^*\mathbb{T}^2$ corresponding to $L_{1,1}$ cover cylinders asymptotic to Reeb orbits in multiples of the foliation class.*
- **Type 2b** *$\boldsymbol{F}_d$ has exactly two essential curves, $\underline{u}$ and $u_{L_{1,1}}$. The closures of the images of the maps $p \circ \underline{u}$ and $p \circ u_{L_{1,1}}$ are $A_0 \cup B$ and A_∞, respectively. Any punctures of $\underline{u}$ on $\mathbb{L}$ are all of foliation type and are either all positive or all negative. The other top level curves of $\boldsymbol{F}_d$ cover (broken or unbroken) leaves of $\mathcal{F}$. Any middle and bottom level curves in the copy of $T^*\mathbb{T}^2$ corresponding to $\mathbb{L}$ cover cylinders asymptotic to Reeb orbits in multiples of the foliation class.*
- **Type 3** *$\boldsymbol{F}_d$ has exactly three essential curves, $u_{\mathbb{L}}$, $\underline{u}$ and $u_{L_{1,1}}$. The closures of the images of the maps $u_{\mathbb{L}}$, $\underline{u}$ and $u_{L_{1,1}}$ are A_0, B and A_∞, respectively. The other top level curves of $\boldsymbol{F}_d$ again cover (broken or unbroken) leaves of $\mathcal{F}$.*

Limits of Type 2b and Type 3 are partially illustrated in Figures 2 and 3, respectively.

Proof of Proposition 3.17 We begin with the following result, which allows us to use essential curves to sort the limit structures.

Lemma 3.18 *Let $\boldsymbol{F}_d$ be a limit as in the **stretching scenario for class $(1, d)$**. If u is a top level curve of $\boldsymbol{F}_d$, then it is either essential or else the image of $p \circ u$ is a point. The essential curves have disjoint images under p, which are open sets, and these images include the complement of $p(\mathbb{L}) \cup p(L_{1,1})$.*

Proof The curves of $\boldsymbol{F}_d$ can be compactified and glued together to form a map $\bar{\boldsymbol{F}}_d \colon S^2 \to S^2 \times S^2$ in the class $(1, d)$. Let T be an unbroken leaf of the foliation. Since $(1, d) \bullet T = (1, d) \bullet (0, 1) = 1$, we see that T can only intersect one top level curve with $p \circ u$ nonconstant. If u is a top level curve such that the map $p \circ u$ is constant, then u covers part of a (possibly broken) leaf of our foliation and contributes intersection number 0 with all unbroken leaves.

Assume then that u is a top level curve such that $p \circ u$ is nonconstant. By the discussion above, u intersects any unbroken leaf T either once or not at all, and therefore if $p \circ u$ has any double points they must lie in $p(\mathbb{L}) \cup p(L_{1,1})$. But positivity of intersection again implies that the nonconstant map $p \circ u$ is an open mapping and this implies that the double points of $p \circ u$ form an open set. We conclude that there are no double points and u is essential. To see that the essential curves have disjoint images under p we can apply the same argument to a union $u \cup v$. The intersection number also implies that all unbroken fibers intersect at least one essential curve. □

Lemma 3.18 implies that there is an essential curve u of $\boldsymbol{F}_d$ that intersects T_0. The closure of the image of $p \circ u$ must contain A_0. By Corollary 3.15 the following cases are exhaustive.

Case 1 (u has no punctures) In this case, $p \circ u$ must be a bijection onto S_∞. Hence, u is a J–holomorphic sphere in a class of the form $(1, j)$ for j in $[0, d]$. By Lemma 3.18 all the other top level curves of $\boldsymbol{F}_d$ must cover leaves of the foliation. This also implies that middle and lower level curves cover cylinders asymptotic to multiples of the foliation class.

The top level curves of $\boldsymbol{F}_d$ which cover fibers fit together to form a possibly disconnected curve in the class $(0, d - j)$. If $j = d$ then $\boldsymbol{F}_d$ consists only of the curve u. Either way, the building is of Type 0.

Case 2 (u has punctures and they are all of foliation type) In this case we claim that $\boldsymbol{F}_d$ is of Type 1. By Lemma 3.14, the image of the map $p \circ u$ includes points in each component of the complement of $p(\mathbb{L}) \cup p(L_{1,1})$, and so by Lemma 3.18 we have that $p \circ u$ is a bijection onto S_∞ minus a finite set of points on $p(\mathbb{L}) \cup p(L_{1,1})$. The other top level curves of $\boldsymbol{F}_d$ must either cover unbroken leaves of $\mathcal{F}$ or they are J–holomorphic planes covering one of the planes of a broken leaf of $\mathcal{F}$. The statement about positivity or negativity of punctures is Lemma 3.16.

Case 3 (u has at least one puncture not of foliation type) Since u intersects the leaf T_0, the closure of the image of $p \circ u$ is either A_0 or $A_0 \cup B$. In either case, u has exactly one puncture not of foliation type and does not intersect T_∞.

Suppose that the closure of the image of $p \circ u$ is A_0. By Lemma 3.18, there is an essential curve v of $\boldsymbol{F}_d$ that intersects T_∞, and the images of $p \circ u$ and $p \circ v$ cannot intersect. Hence the closure of the image of $p \circ v$ is either A_∞ or $B \cup A_\infty$. In the first case, $\boldsymbol{F}_d$ is of Type 3 with $u_{\mathbb{L}} = u$ and $u_{L_{1,1}} = v$, where the third curve, $\underline{u}$, exists by Lemma 3.18. In the second case, $\boldsymbol{F}_d$ is of Type 2a with $u_{\mathbb{L}} = u$ and $\underline{u} = v$.

If, instead, the closure of the image of $p \circ u$ is $A_0 \cup B$, a similar argument implies that $\boldsymbol{F}_d$ is of Type 2b.

This completes the proof of Proposition 3.17. □

3.7 The existence of special buildings, under Assumption 2

In this section we will establish the existence of two special limits of the *stretching scenario for class* $(1, d)$ when d is sufficiently large. The following result will be used to exploit the large d limit.

Lemma 3.19 *There exists an $\epsilon > 0$ such that*

$$\operatorname{area}(u) \geq \epsilon u \bullet (S_0 \cup S_\infty)$$

for all J–holomorphic curves u in $S^2 \times S^2 \setminus (\mathbb{L} \cup L_{1,1})$.

Proof Fix an open neighborhood of S_∞ of the form $\mathcal{N}_\epsilon = S_\infty \times D^2(\epsilon)$, where $D^2(\epsilon)$ is the open disc of area ϵ. We may assume that the closure of $\mathcal{N}_\epsilon$ is disjoint from $\mathbb{L} \cup L_{1,1}$ and, by (A1), we may assume that J restricts to $\mathcal{N}_\epsilon$ as the standard split complex structure. Let $\pi_2 \colon S_\infty \times D^2(\epsilon) \to D^2(\epsilon)$ be projection and set

$$u_{\epsilon,\infty} = u|_{u^{-1}(\mathcal{N}_\epsilon)}.$$

By perturbing ϵ if needed we may assume that $u^{-1}(\mathcal{N}_\epsilon)$ is a smooth manifold. We have

$$\operatorname{degree}(\pi_2 \circ u_{\epsilon,d}) = u \bullet S_\infty.$$

This implies

$$\operatorname{area}(u_{\epsilon,\infty}) \geq \int_{u^{-1}(\mathcal{N}_\epsilon)} u_{\epsilon,\infty}^*(\omega \oplus \omega) \geq \int_{(\pi_2 \circ u_{\epsilon,\infty})^{-1}(D^2(\epsilon))} (\pi_2 \circ u_{\epsilon,\infty})^* \omega = \left(\int_{D^2(\epsilon)} \omega \right) u \bullet S_\infty = \epsilon u \bullet S_\infty.$$

A similar calculation for S_0 gives the result. □

Proposition 3.20 *For all sufficiently large d, there exists a limiting building $\boldsymbol{F}$ as in the **stretching scenario for class** $(1, d)$ such that $\boldsymbol{F}$ is of Type 3. The building consists of its three essential top level curves, $u_{\mathbb{L}}$, $\underline{u}$ and $u_{L_{1,1}}$, together with $d - 1$ planes in $\mathfrak{r}_0 \cup \mathfrak{r}_\infty$ and d planes in $\mathfrak{s}_0 \cup \mathfrak{s}_\infty$.*

Proof Fix $d + 1$ points on $L_{1,1}$ and d points on $\mathbb{L}$. For $\tau \geq 0$, let J_τ be the family of almost complex structures on $S^2 \times S^2$ from Section 3.5. It follows from the discussion in Section 3.6 and the compactness result from [1] that for a sequence $\tau_k \to \infty$ there is a sequence $u_k \colon S^2 \to S^2 \times S^2$ of J_{τ_k}–holomorphic curves in the class $(1, d)$ that pass through the $2d + 1$ constraint points and converge in the sense of [1]. Their limit, $\boldsymbol{F}$, is the desired building.

To see this we first note that the point constraints already preclude the possibility that $\boldsymbol{F}$ is of Type 0. Indeed, top level essential curves are disjoint from the point constraints, so these must be satisfied by curves of $\boldsymbol{F}$ inside copies of $T^*\mathbb{T}^2$ (corresponding to neighborhoods of $\mathbb{L}$ or $L_{1,1}$). In the Type 0 case, the nonessential curves fit together to form a union of spheres in the class $(0, d-j)$ for some $0 \leq j \leq d$. These intersect $\mathbb{L} \cup L_{1,1}$ in a finite set of geodesics, and any middle or lower level curves in our $T^*\mathbb{T}^2$ cover cylinders asymptotic to these geodesics. As there are at most d such cylinders they cannot satisfy the $2d+1$ point constraints. (The holomorphic cylinders in $T^*\mathbb{T}^2$ are described explicitly by Lemma 4.2 in [5].)

If $\boldsymbol{F}$ was of Type 1, then punctures of its essential curve on $L_{1,1}$ would all be of the foliation type. The remaining top level curves of $\boldsymbol{F}$ asymptotic to $L_{1,1}$ would cover broken planes, and all the curves of $\boldsymbol{F}$ mapping to the copy of $T^*\mathbb{T}^2$ corresponding to $L_{1,1}$ would cover cylinders over geodesics in the foliation class $\beta_{1,1}$. To satisfy the $d+1$ point constraints on $L_{1,1}$ the essential curve of $\boldsymbol{F}$ must have at least $d+1$ punctures on $L_{1,1}$, matching with at least $d+1$ cylinders in the copy of $T^*\mathbb{T}^2$. But then $\boldsymbol{F}$ would have $d+1$ curves covering planes in $\mathfrak{s}_0 \cup \mathfrak{s}_\infty$. By Lemma 3.16, the punctures of the essential curve on $L_{1,1}$ are either all positive or all negative. Hence these $d+1$ curves either all lie in $\mathfrak{s}_0$ or all lie in $\mathfrak{s}_\infty$. This contradicts the fact that (the compactification of) $\boldsymbol{F}$ has intersection number d with both S_0 and S_∞. The same argument precludes the possibility that $\boldsymbol{F}$ has Type 2a.

It remains to show that $\boldsymbol{F}$ does not have Type 2b. Assuming that $\boldsymbol{F}$ has Type 2b, we will show that it must include a collection of curves of total area equal to two, that intersect $S_0 \cup S_\infty$ d times. If d is sufficiently large, this contradicts Lemma 3.19 above.

Claim 1 *If $\boldsymbol{F}$ has Type 2b, then it includes at least d planes in $\mathfrak{s}_0 \cup \mathfrak{s}_\infty$.*

To see this, consider the subbuilding $\boldsymbol{F}_{1,1}$ of $\boldsymbol{F}$ consisting of its middle and bottom level curves mapping to the copies of $\mathbb{R} \times S^*\mathbb{T}^2$ and $T^*\mathbb{T}^2$ that correspond to $L_{1,1}$. Since it is connected and has genus zero, it follows from Proposition 3.3 of [8] that

$$\operatorname{index}(\boldsymbol{F}_{1,1}) = 2(s-1),$$

where s is the number of positive ends of $\boldsymbol{F}_{1,1}$. Since $\boldsymbol{F}_{1,1}$ passes through the $d+1$ generic point constraints on $L_{1,1}$, and the Fredholm index in these manifolds is nondecreasing under multiple covers, we must also have

$$\operatorname{index}(\boldsymbol{F}_{1,1}) \geq 2(d+1).$$

Hence, $\boldsymbol{F}_{1,1}$ has at least $d+2$ positive ends. Under the assumption that $\boldsymbol{F}$ has Type 2b, two of these positive ends match with the two essential top level curves of $\boldsymbol{F}$. This leaves at least d positive ends of $\boldsymbol{F}_{1,1}$ that match with top level curves of $\boldsymbol{F}$ that cover planes in $\mathfrak{s}_0 \cup \mathfrak{s}_\infty$.

Remark 3.21 The same argument implies that if $\boldsymbol{F}$ has Type 3, then again it must include at least d planes in $\mathfrak{s}_0 \cup \mathfrak{s}_\infty$.

Claim 2 *If $\boldsymbol{F}$ has Type 2b, then it includes d planes in $\mathfrak{r}_0$ and none in $\mathfrak{r}_\infty$.*

The d constraint points on $\mathbb{L}$ imply that, if $\boldsymbol{F}$ is of Type 2b, it must contain d planes covering broken planes asymptotic to $\mathbb{L}$. These planes match, via cylinders in $T^*\mathbb{T}^2$, with asymptotic ends of an essential curve, and by Lemma 3.16 these ends are either all positive or all negative. Hence we have d planes either all in $\mathfrak{r}_0$ or all in $\mathfrak{r}_\infty$. To show that these planes cannot be in $\mathfrak{r}_\infty$, we consider intersections with $S_0 \cup S_\infty$. Overall, the top level curves of $\boldsymbol{F}$ must intersect $S_0 \cup S_\infty$ exactly $2d$ times. The planes of $\boldsymbol{F}$ asymptotic to $L_{1,1}$ from Claim 1 account for at least d of these intersections.

Since $\mathbb{L}$ is homologically trivial in Y, by Lemma 3.11 each plane of $\mathfrak{r}_\infty$ must intersect both S_0 and S_∞, while the planes in $\mathfrak{r}_0$ intersect neither of these spheres. If the d planes of $\boldsymbol{F}$ asymptotic to $\mathbb{L}$ were in $\mathfrak{r}_\infty$ then they would contribute another $2d$ intersections with $S_0 \cup S_\infty$. By positivity of intersection, this cannot happen, so these planes must belong to $\mathfrak{r}_0$ as claimed.

To complete the argument, we now balance areas. The total area of all the curves in $\boldsymbol{F}$ is $2(d+1)$. If $\boldsymbol{F}$ has Type 2b, then the planes from Claim 1 and Claim 2 have total area at least $2d$. Its essential curves must then have total area equal to 2. Also, they must contribute the remaining d intersections with $S_0 \cup S_\infty$. It follows from Lemma 3.19, that this is impossible for all d sufficiently large. Hence $\boldsymbol{F}$ cannot be of Type 2b, and must instead be of Type 3. Arguing as above, it follows that in addition to its three essential top level curves, $\boldsymbol{F}$ must then have d planes in $\mathfrak{s}_0 \cup \mathfrak{s}_\infty$ and $d-1$ planes in $\mathfrak{r}_0 \cup \mathfrak{r}_\infty$. □

Proposition 3.22 *For all sufficiently large d, there exists a limiting building $\boldsymbol{G}$ as in the **stretching scenario for class** $(1,d)$ such that $\boldsymbol{G}$ is of Type 3. In addition to its three essential curves it consists of d planes in $\mathfrak{r}_0 \cup \mathfrak{r}_\infty$ and $d-1$ planes in $\mathfrak{s}_0 \cup \mathfrak{s}_\infty$.*

Proof Here we fix d points on $L_{1,1}$ and $d+1$ points on $\mathbb{L}$, and for J_τ as in Proposition 3.20 consider the limit, $\boldsymbol{G}$, of a convergent sequence of J_{τ_k}–holomorphic spheres, for $\tau_k \to \infty$, that represent the class $(1,d)$ and pass through the $2d+1$ constraint points. The point constraints imply that $\boldsymbol{G}$ is not of Type 0.

If $\boldsymbol{G}$ was of Type 1, the point constraints would imply that $\boldsymbol{G}$ includes at least d planes, which by Lemma 3.16 either all lie in $\mathfrak{s}_0$ or all lie in $\mathfrak{s}_\infty$, and at least $d+1$ planes either all in $\mathfrak{r}_0$ or all in $\mathfrak{r}_\infty$. From this it follows that the essential curve of $\boldsymbol{G}$ would have area 1. Recalling Lemma 3.11, since L is homologically trivial, the planes of $\mathfrak{r}_\infty$ each intersect $S_0 \cup S_\infty$ twice. Arguing as in Claim 2 from the proof of Proposition 3.20, if the planes asymptotic to $\mathbb{L}$ lie in $\mathfrak{r}_\infty$ then the broken planes will contribute a total of $d + 2(d+1)$ intersections with $S_0 \cup S_\infty$, a contradiction as there are only $2d$ such intersections. On the other hand, if these planes all lie in $\mathfrak{r}_0$ then the essential curve must contribute d intersections with $S_0 \cup S_\infty$. As this essential curve has area 1, this contradicts Lemma 3.19 when d is sufficiently large. Hence, $\boldsymbol{G}$ is not of Type 1.

Next we show that $\boldsymbol{G}$ cannot be of Type 2b. Assume that it is. Then $\boldsymbol{G}$ includes $d+1$ planes either all in $\mathfrak{r}_0$ or all in $\mathfrak{r}_\infty$. Counting intersections as above, $\boldsymbol{G}$ must have $d+1$ planes in $\mathfrak{r}_0$.

Arguing as in Claim 1 above, we consider the subbuilding $\boldsymbol{G}_{1,1}$ of $\boldsymbol{G}$ consisting of its middle and bottom level curves that map to the copies of $\mathbb{R}\times S^*\mathbb{T}^2$ and $T^*\mathbb{T}^2$ that correspond to $L_{1,1}$. Since $\boldsymbol{G}_{1,1}$ is connected and has genus zero, we have

$$\operatorname{index}(\boldsymbol{G}_{1,1}) = 2(s-1),$$

where s is the number of positive ends of $\boldsymbol{G}_{1,1}$. Since $\boldsymbol{G}_{1,1}$ passes through the d generic point constraints on $\mathbb{L}$, we also have

$$\operatorname{index}(\boldsymbol{G}_{1,1}) \geq 2d.$$

Hence, $\boldsymbol{G}_{1,1}$ has at least $d+1$ positive ends. Two of these positive ends match with negative ends of the two essential curves of $\boldsymbol{G}_{1,1}$. It follows that $\boldsymbol{G}$ must have at least $d-1$ planes in $\mathfrak{s}_0\cup\mathfrak{s}_\infty$. This means the planes covering broken leaves then have area at least $2d$. As the limiting building has total area $2d+2$ and also includes two essential curves, we see that the essential curves each have area 1 and there are exactly $d-1$ planes in $\mathfrak{s}_0\cup\mathfrak{s}_\infty$. As the planes in $\mathfrak{r}_0$ are disjoint from $S_0\cup S_\infty$, the essential curves of $\boldsymbol{G}$ must have $d+1$ intersections with $S_0\cup S_\infty$. Lemma 3.19 again implies that this is impossible for all sufficiently large d.

Finally we show that $\boldsymbol{G}$ cannot be of Type 2a. In this case $\boldsymbol{G}$ includes d planes in $\mathfrak{r}_0\cup\mathfrak{r}_\infty$ and d planes all in either $\mathfrak{s}_0$ or all in $\mathfrak{s}_\infty$. The planes asymptotic to $L_{1,1}$ thus account for all intersections with either S_0 or S_∞ and so the planes asymptotic to $\mathbb{L}$ therefore all lie in $\mathfrak{r}_0$. The essential curves have total area 2 and must together account for all intersections with either S_0 or S_∞. This contradicts Lemma 3.19 as before. □

Lemma 3.23 *All curves in the limiting buildings $\boldsymbol{F}$ and $\boldsymbol{G}$ that map to $S^2\times S^2\setminus(\mathbb{L}\cup L_{1,1})$ have area 1, and in particular are simply covered.*

Proof To see this, first observe that classes in $H_2(S^2\times S^2,\mathbb{L}\cup L_{1,1})$ all have integral area. Indeed, adding classes which lie only in $H_2(S^2\times S^2,\mathbb{L})$ or $H_2(S^2\times S^2,L_{1,1})$, which have integral area by monotonicity, any relative class can be completed to an integral area absolute homology class.

Note that since $\boldsymbol{F}$ is of Type 3, it has its three essential curves together with $2d-1$ other top level curves that cover leaves of the foliation. Since $\boldsymbol{F}$ has total area $2d+2$ and all curves have integral area, the result for $\boldsymbol{F}$ follows.

The same argument applies to $\boldsymbol{G}$. □

3.8 A collection of symplectic spheres and disks, under Assumption 2

Let J be a tame almost complex structure on $S^2\times S^2\setminus(\mathbb{L}\cup L_{1,1})$ that is adapted to parametrizations ψ and $\psi_{1,1}$ of $\mathbb{L}$ and $L_{1,1}$, respectively. Recall that for the projection $p: S^2\times S^2\to S_\infty$, defined by the foliation $\mathcal{F}$ corresponding to J, the images $p(\mathbb{L})$ and $p(L_{1,1})$ are disjoint circles. There are also disjoint disks $A_0\subset S_\infty$ with boundary $p(\mathbb{L})$ and $A_\infty\subset S_\infty$ with boundary $p(L_{1,1})$ such that $p(T_0)\in A_0$ and $p(T_\infty)\in A_\infty$; see Figure 1. In this section we will prove the following result.

Proposition 3.24 *For sufficiently large d there exist embedded symplectic spheres*

$$F, G\colon S^2 \to S^2 \times S^2 \setminus (\mathbb{L} \cup L_{1,1})$$

in the class $(1, d)$, and embedded symplectic disks

$$\mathbb{E}\colon (D^2, S^1) \to (S^2 \times S^2, \mathbb{L}) \quad \text{and} \quad E_{1,1}\colon (D^2, S^1) \to (S^2 \times S^2, L_{1,1})$$

of Maslov index 2, such that:

(1) *F, G, $\mathbb{E}$ and $E_{1,1}$ are all J–holomorphic away from arbitrarily small neighborhoods of a collection of Lagrangian tori whose elements are near to, and Lagrangian isotopic to, either $\mathbb{L}$ or $L_{1,1}$.*

(2) *The class of $\mathbb{E}|_{S^1}$ and the foliation class $\beta_{\mathbb{L}}$ form an integral basis of $H_1(\mathbb{L} : \mathbb{Z})$.*

(3) *The class of $E_{1,1}|_{S^1}$ and the foliation class $\beta_{L_{1,1}}$ form an integral basis of $H_1(L_{1,1} : \mathbb{Z})$.*

(4) *Exactly one of F and G intersects the planes of $\mathfrak{r}_0$ and the other intersects the planes of $\mathfrak{r}_\infty$.*

(5) *Exactly one of F and G intersects the planes of $\mathfrak{s}_0$ and the other intersects the planes of $\mathfrak{s}_\infty$.*

(6) $F \bullet \mathbb{E} + G \bullet \mathbb{E} = d$.

(7) $F \bullet E_{1,1} + G \bullet E_{1,1} = d$.

(8) $F \bullet G = 2d$.

(9) *The set $p(F \cap G)$ consists of d points in A_0 and d points in A_∞.*

Remark 3.25 This proposition is the key to our result. Following Theorem 1.1, the spheres F and G will eventually be transformed to form axes of a new copy of $S^2 \times S^2$ in which $\mathbb{L}$ and $L_{1,1}$ must intersect. A natural approach to finding spheres in the complement of the Lagrangians may have been to fix constraint points in the complement of $\mathbb{L} \cup L_{1,1}$ and then take a limit of holomorphic spheres through these points as we stretch along $\mathbb{L} \cup L_{1,1}$. Indeed, generically this does give holomorphic spheres in the complement, but it seems difficult to obtain in this way a pair of spheres where one intersects the family $\mathfrak{r}_0$ and the other the family $\mathfrak{r}_\infty$.

Our alternative approach is to start with the Type 3 curves given by Propositions 3.20 and 3.22, and this is why we need to assume d is large. The Type 3 buildings intersect our Lagrangians in such a way that they can be deformed, by a diffeomorphism supported near the Lagrangians, into cycles which are disjoint from $\mathbb{L}$ and $L_{1,1}$ and have the required intersections. The process is illustrated in Figure 4. In the figure, the blue curves correspond to curves in the building $\boldsymbol{F}$ and the black curves correspond to the deformed cycle contained in the complement of $\mathbb{L}$. The curves running vertically correspond to broken leaves of the foliation, and those running horizontally to essential curves.

More precisely, we will deform the building $\boldsymbol{F}$ constructed in Proposition 3.20 to a building $\boldsymbol{F}(\{\boldsymbol{v}_1, \boldsymbol{w}_1\})$ containing curves asymptotic to Lagrangians $\mathbb{L}(\boldsymbol{v}_1)$ and $L_{1,1}(\boldsymbol{w}_1)$. In local coordinates $\mathbb{L}(\boldsymbol{v}_1)$ will be a translation of $\mathbb{L}$ and $L_{1,1}(\boldsymbol{w}_1)$ a translation of $L_{1,1}$. Similarly, the building $\boldsymbol{G}$ constructed in Proposition 3.22 is deformed to a building $\boldsymbol{G}(\{\boldsymbol{v}_2, \boldsymbol{w}_2\})$ containing curves asymptotic to Lagrangians $\mathbb{L}(\boldsymbol{v}_2)$

and $L_{1,1}(\boldsymbol{w}_2)$, where again, in local coordinates, $\mathbb{L}(\boldsymbol{v}_2)$ will be a translation of $\mathbb{L}$ and $L_{1,1}(\boldsymbol{w}_2)$ a translation of $L_{1,1}$.

Our proof proceeds by describing these deformations carefully and then remarking that the images of the buildings can be smoothed to form our symplectic spheres. This smoothing occurs only near $\mathbb{L}(\boldsymbol{v}_1)$ and $L_{1,1}(\boldsymbol{w}_1)$ for $\boldsymbol{F}(\{\boldsymbol{v}_1, \boldsymbol{w}_1\})$, and only near $\mathbb{L}(\boldsymbol{v}_2)$ and $L_{1,1}(\boldsymbol{w}_2)$ for $\boldsymbol{G}(\{\boldsymbol{v}_2, \boldsymbol{w}_2\})$. As the six Lagrangian tori $\mathbb{L}, \mathbb{L}(\boldsymbol{v}_1), \mathbb{L}(\boldsymbol{v}_2), L_{1,1}, L_{1,1}(\boldsymbol{w}_1)$ and $L_{1,1}(\boldsymbol{w}_2)$ are disjoint, and in fact disjoint from any intersections between different buildings, this smoothing does not affect our intersection pattern calculation.

Proof of Proposition 3.24 In what follows, $\boldsymbol{F}$ and $\boldsymbol{G}$ will be limiting J–holomorphic buildings of Type 3 as established in Proposition 3.20 and Proposition 3.22, respectively. We assume that they are in the same class $(1, d)$ for some large d.

The top level curves of $\boldsymbol{F}$ are

$$\{u_{\mathbb{L}}, \underline{u}, u_{L_{1,1}}, u_1, \dots, u_{d-1}, \mathfrak{u}_1, \dots, \mathfrak{u}_d\}.$$

Here, $u_{\mathbb{L}}$, $\underline{u}$ and $u_{L_{1,1}}$ are the essential curves of $\boldsymbol{F}$. For some nonnegative integer $\alpha_0 \leq d-1$, the curves $u_1, \dots, u_{\alpha_0}$ belong to $\mathfrak{r}_0$ and the curves $u_{\alpha_0+1}, \dots, u_{d-1}$ belong to $\mathfrak{r}_\infty$. For some nonnegative integer $\beta_0 \leq d$, the curves $\mathfrak{u}_1, \dots, \mathfrak{u}_{\beta_0}$ belong to $\mathfrak{s}_0$ and the curves $\mathfrak{u}_{\beta_0+1}, \dots, \mathfrak{u}_d$ belong to $\mathfrak{s}_\infty$.

Similarly, the top level curves of $\boldsymbol{G}$ are

$$\{v_{\mathbb{L}}, \underline{v}, v_{L_{1,1}}, v_1, \dots, v_d, \mathfrak{v}_1, \dots, \mathfrak{v}_{d-1}\},$$

where for some nonnegative integer $\gamma_0 \leq d$ the curves $v_1, \dots, v_{\gamma_0}$ belong to $\mathfrak{r}_0$ and $v_{\gamma_0+1}, \dots, v_d$ belong to $\mathfrak{r}_\infty$, and for some nonnegative integer $\delta_0 \leq d-1$, the curves $\mathfrak{v}_1, \dots, \mathfrak{v}_{\delta_0}$ belong to $\mathfrak{s}_0$ and the curves $\mathfrak{v}_{\delta_0+1}, \dots, \mathfrak{v}_{d-1}$ belong to $\mathfrak{s}_\infty$.

3.8.1 Deformations near $\mathbb{L}$ Consider the coordinates (P_1, Q_1, P_2, Q_2) in the Weinstein neighborhood of $\mathbb{L}$,

$$\mathcal{U}(\mathbb{L}) = \{|P_1| < \epsilon, |P_2| < \epsilon\}.$$

For each *translation vector* $\boldsymbol{v} = (a, b) \in (-\epsilon, \epsilon) \times (-\epsilon, \epsilon)$, there is a corresponding nearby Lagrangian torus

$$\mathbb{L}(\boldsymbol{v}) = \mathbb{L}(a, b) = \{P_1 = a,\ P_2 = b\} \subset \mathcal{U}(\mathbb{L}).$$

Note that the parametrization ψ of $\mathbb{L}$ determines an obvious parametrization, $\psi(\boldsymbol{v}) = \psi(a, b)$ of $\mathbb{L}(a, b)$, and a canonical isomorphism from $H_1^{\psi}(L; \mathbb{Z})$ to $H_1^{\psi(a,b)}(L(a, b); \mathbb{Z})$.

Given a finite, nonempty collection of translation vectors,

$$\boldsymbol{V} = \{\boldsymbol{v}_1, \dots, \boldsymbol{v}_k\} = \{(a_1, b_1), \dots, (a_k, b_k)\},$$

let $J_{\boldsymbol{V}}$ be an almost complex structure on the complement of the collection of Lagrangians

$$\mathbb{L}(\boldsymbol{V}) = \bigcup_{i=1}^{k} \mathbb{L}(\boldsymbol{v}_i),$$

which coincides with J outside $\mathcal{U}(\mathbb{L})$ and inside has the form

$$J_{\boldsymbol{V}} \frac{\partial}{\partial Q_i} = -\rho_{\boldsymbol{V}} \frac{\partial}{\partial P_i}, \tag{3-1}$$

where $\rho_{\boldsymbol{V}}$ is a positive function away from $\mathbb{L}(\boldsymbol{V})$ and in a neighborhood of each $\mathbb{L}(\boldsymbol{v}_i)$ has the form

$$\rho_{\boldsymbol{V}} = \sqrt{(P_1 - a_i)^2 + (P_2 - b_i)^2}.$$

In this case, we say that $J_{\boldsymbol{V}}$ is *adapted to* $\mathbb{L}(\boldsymbol{V})$ *with respect to* ψ. The set of all such almost complex structures adapted to some nontrivial collection $\mathbb{L}(\boldsymbol{V}) \subset \mathcal{U}(\mathbb{L})$ will be denoted by $\mathcal{J}_{\mathcal{U}(\mathbb{L})}$.

Following Section 2.5 of [5], for each $J_{\boldsymbol{V}}$ in $\mathcal{J}_{\mathcal{U}(\mathbb{L})}$ one can construct, for $\tau \geq 0$, a standard family of almost complex structures $J_{\boldsymbol{V},\tau}$ on $S^2 \times S^2$ that are tame with respect to $\pi_1^*\omega + \pi_2^*\omega$, such that the limit $\tau \to \infty$ corresponds to the process of stretching the neck along small sphere bundles surrounding each of the components of $\mathbb{L}(\boldsymbol{V})$; see [1]. The structure $J_{\boldsymbol{V}}$ is the part of the limit of the $J_{\boldsymbol{V},\tau}$ corresponding to $S^2 \times S^2 \setminus (\mathbb{L}(\boldsymbol{V}) \cup L_{1,1})$. The limit of the Gromov foliations for the $J_{\boldsymbol{V},\tau}$, in class $(0,1)$, yields a foliation $\mathcal{F}(\boldsymbol{V})$ of $S^2 \times S^2 \setminus (\mathbb{L}(\boldsymbol{V}) \cup L_{1,1})$. For example, for $\boldsymbol{V} = \{(0,0)\}$ we have $J_{\boldsymbol{V}} = J$ and $\mathcal{F}(\boldsymbol{V}) = \mathcal{F}$.

Lemma 3.26 *Leaves of the foliation $\mathcal{F}(\boldsymbol{V})$ intersect $\mathcal{U}(\mathbb{L})$ along the annuli $\{P_1 = \delta, Q_1 = \theta, |P_2| < \epsilon\}$. A leaf of $\mathcal{F}(\boldsymbol{V})$ that intersects $\mathcal{U}(\mathbb{L})$ along the annulus $\{P_1 = \delta, Q_1 = \theta, |P_2| < \epsilon\}$ is broken if and only if the collection $\boldsymbol{V}$ contains an element of the form (δ, b_i).*

Proof It follows from equation (3-1) that these annuli are $J_{\boldsymbol{V}}$–holomorphic. By assuming J satisfies the conclusions of Lemma 3.4, they also extend to $J_{\boldsymbol{V}}$–holomorphic spheres in the class $(0,1)$. By positivity of intersection, these spheres, and indeed any holomorphic sphere in the class $(0,1)$, are leaves of the foliation $\mathcal{F}(\boldsymbol{V})$. □

First deformation process Our first deformation process allows us to deform a regular J–holomorphic curve so that its ends on $\mathbb{L}$ become ends on a nearby Lagrangian $\mathbb{L}(\boldsymbol{v})$.

Lemma 3.27 (Fukaya's trick) *Let u be a regular J–holomorphic curve with $k \geq 0$ ends on $\mathbb{L}$ and $l \geq 0$ ends on $L_{1,1}$. For all $\boldsymbol{v} = (a,b)$ with $\|\boldsymbol{v}\|^2 = a^2 + b^2$ sufficiently small, there is a regular $J_{\boldsymbol{v}}$–holomorphic curve $u(\boldsymbol{v})$ with k ends on $\mathbb{L}(\boldsymbol{v})$ and l ends on $L_{1,1}$. Moreover, the ends of $u(\boldsymbol{v})$ on $\mathbb{L}(\boldsymbol{v})$ represent the identical classes in $H_1^{\psi(\boldsymbol{v})}(L,\mathbb{R})$, as do those of u in $H_1^{\psi}(L,\mathbb{R})$. The classes corresponding to the ends of $u(\boldsymbol{v})$ on $L_{1,1}$ are also identical to those of u.*

Proof For $\|\boldsymbol{v}\|$ sufficiently small, the Lagrangian isotopy $t \mapsto \mathbb{L}(t\boldsymbol{v})$ for $0 \leq t \leq 1$ is contained in $\mathcal{U}(\mathbb{L})$. Let $f_{t,\boldsymbol{v}}$ be a family of diffeomorphisms of $S^2 \times S^2$ such that

- $f_{0,\boldsymbol{v}}$ is the identity map,
- $f_{t,\boldsymbol{v}}(\mathbb{L}) = \mathbb{L}(t\boldsymbol{v})$ for all $t \in [0,1]$,
- each $f_{t,\boldsymbol{v}}$ is equal to the identity map outside of $\mathcal{U}(\mathbb{L})$, and
- $\|f_{t,\boldsymbol{v}}\|_{C^1}$ is of order 1 in $\|\boldsymbol{v}\|$.

Let $J_{t\boldsymbol{v}}$ be a family of tame almost complex structures in $\mathcal{J}_{\mathcal{U}(\mathbb{L})}$ such that each $J_{t\boldsymbol{v}}$ is adapted to $\mathbb{L}(t\boldsymbol{v})$ with respect to ψ. Set

$$\widetilde{J}_{t\boldsymbol{v}} = (f_{t,\boldsymbol{v}}^{-1})_* J_{t\boldsymbol{v}}.$$

For $\|\boldsymbol{v}\|$ sufficiently small, $\widetilde{J}_{t\boldsymbol{v}}$ is a tame almost complex structure on $S^2\times S^2\setminus(\mathbb{L}\cup L_{1,1})$ for all $t\in[0,1]$. Since u is regular, for sufficiently small $\|\boldsymbol{v}\|$ the curve u persists to yield a regular $\widetilde{J}_{\boldsymbol{v}}$–holomorphic curve $\widetilde{u}(\boldsymbol{v})$ with the same asymptotic behavior as u. By our choice of $\widetilde{J}_{t\boldsymbol{v}}$, the curve

$$u(\boldsymbol{v}) = f_{1,\boldsymbol{v}}\circ\widetilde{u}(\boldsymbol{v})$$

is then regular, $J_{\boldsymbol{v}}$–holomorphic and has k ends on $\mathbb{L}(\boldsymbol{v})$ instead of $\mathbb{L}$. □

Applying Lemma 3.27 to $\boldsymbol{F}$ and $\boldsymbol{G}$ To apply Lemma 3.27 to the top level curves of $\boldsymbol{F}$ and $\boldsymbol{G}$ we need these curves to be regular. Lemma 3.23 implies that the top level curves of the buildings $\boldsymbol{F}$ and $\boldsymbol{G}$ are somewhere injective. Since they are the limits of embedded curves, they are actually embedded and hence regular for generic choice of J. The work of Wendl in [21] implies that they are regular for all J.

Lemma 3.28 *For any tame almost complex structure J on*

$$(S^2\times S^2\setminus(\mathbb{L}\cup L_{1,1}),\pi_1^*\omega+\pi_2^*\omega)$$

that is adapted to both ψ and $\psi_{1,1}$, the top level curves of the buildings $\boldsymbol{F}$ and $\boldsymbol{G}$ are all regular.

Proof By [21, Theorem 1], any embedded J–holomorphic curve u mapping to $S^2\times S^2\setminus(\mathbb{L}\cup L_{1,1})$ is regular if its Fredholm index is greater than or equal to the number of its asymptotic ends. In our setting, we have

$$\operatorname{index}(u) = s-2+2c_1(u^*T(S^2\times S^2),u^*TL),$$

where s is the number of ends and c_1 is the relative first Chern class. It suffices to show that for each top level curve u of the buildings $\boldsymbol{F}$ and $\boldsymbol{G}$, we have $c_1(u^*T(S^2\times S^2),u^*TL)\geq 1$.

Since $\boldsymbol{F}$ and $\boldsymbol{G}$ both have Type 3, it follows from Lemma 3.23 that each top level curve u is either a J–holomorphic plane or cylinder. If u is a plane, then $2c_1(u^*T(S^2\times S^2),u^*TL)$ is just the Maslov class. By monotonicity, this is equal to 2 since, by Lemma 3.23, our top level curves all have area 1.

If u is a J–holomorphic cylinder, we can then produce a disk v from it by compactifying the ends of u and smoothly gluing a disk w to one of them. If the disk w has area A then, by monotonicity, it has Maslov index $2A$. By additivity of the area and the Chern class, v has area $A+1$ and Maslov index $2c_1(u^*T(S^2\times S^2),u^*TL)+2A$. Since the area is $A+1$, by monotonicity v must have Maslov index $2(A+1)$. This implies that $2c_1(u^*T(S^2\times S^2),u^*TL)=2$, as required. □

For $\boldsymbol{v}=(a,b)$ with $\|\boldsymbol{v}\|$ sufficiently small we now define the deformed building $\boldsymbol{F}(\boldsymbol{v})$ as follows. The top level curves of $\boldsymbol{F}(\boldsymbol{v})$ are obtained by applying Lemma 3.27 to those top level curves of $\boldsymbol{F}$ with ends on $\mathbb{L}$, and leaving the others unchanged. That is, the top level curves of $\boldsymbol{F}(\boldsymbol{v})$ are

$$\{u_{\mathbb{L}}(\boldsymbol{v}),\underline{u}(\boldsymbol{v}),u_{L_{1,1}},u_1(\boldsymbol{v}),\dots,u_{d-1}(\boldsymbol{v}),\mathfrak{u}_1,\dots,\mathfrak{u}_d\}.$$

The middle and bottom level curves of $\boldsymbol{F}(\boldsymbol{v})$ are the same as those of $\boldsymbol{F}$ except they are now considered to map to copies of $\mathbb{R}\times S^*\mathbb{T}^2$ and $T^*\mathbb{T}^2$ that correspond to $\mathbb{L}(\boldsymbol{v})$ rather than $\mathbb{L}$.

Note that $\boldsymbol{F}(\boldsymbol{v})$ still has a continuous compactification $\bar{\boldsymbol{F}}(\boldsymbol{v})\colon S^2\to S^2\times S^2$, which can be deformed arbitrarily close to $\mathbb{L}(\boldsymbol{v})$ to obtain a smooth sphere $F=F(\boldsymbol{v})\colon S^2\to S^2\times S^2$ which is $J_{\boldsymbol{v}}$–holomorphic away from a small neighborhood of $\mathbb{L}(\boldsymbol{v})$.

Lemma 3.29 *Set $\boldsymbol{v}=(a,b)$ and $\boldsymbol{V}=\{(0,0),\boldsymbol{v}\}$ and suppose that $\|\boldsymbol{v}\|$ is small enough for $\boldsymbol{F}(\boldsymbol{v})$ to exist. If a and b are both nonzero and $|a|$ is sufficiently small with respect to $|b|$, then each top level curve of $\boldsymbol{F}(\boldsymbol{v})$ is $J_{\boldsymbol{V}}$–holomorphic for some $J_{\boldsymbol{V}}$ in $J_{\mathcal{U}(\mathbb{L})}$.*

Proof By Lemma 3.26, for any adapted almost complex structure, the leaves of the corresponding foliation intersect $\mathcal{U}(\mathbb{L})$ in the annuli $\{P_1=\delta, Q_1=\theta, |P_2|<\epsilon\}$. Hence if $b\neq 0$ and $a=0$, the preimages of the regions A_0 and B for $\mathbb{L}(\boldsymbol{v})$ intersect $\mathcal{U}(\mathbb{L})$ in the subsets $\{P_1<0\}$ and $\{P_1>0\}$, respectively (since they consist of the leaves which are not broken along $\mathbb{L}(\boldsymbol{v})$). It follows that the closures of the essential curves of $\boldsymbol{F}(\boldsymbol{v})$ are disjoint from $\mathbb{L}$: the curves themselves are disjoint since they project to the regions A_0, B or A_∞, and they are compactified by circles in $\mathbb{L}(0,b)$ or $L_{1,1}$.

By continuity, these essential curves remain disjoint from $\mathbb{L}$ also for sufficiently small a when we deform using Lemma 3.27. Therefore, for all $|a|$ sufficiently small, the essential curves of $\boldsymbol{F}(\boldsymbol{v})$ are $J_{\boldsymbol{V}}$–holomorphic for any $J_{\boldsymbol{V}}$ which only differs from $J_{\boldsymbol{v}}$ in a small enough neighborhood of $\mathbb{L}$. Meanwhile, any top level curves of $\boldsymbol{F}(\boldsymbol{v})$ that cover broken leaves intersect $\mathcal{U}(\mathbb{L})$ in annuli lying in $\{P_1=a\}$, and these annuli are holomorphic for any adapted almost complex structure. □

Arguing in a similar fashion we can assert that the top level curves of $\boldsymbol{F}(\boldsymbol{v})$ are $J_{\boldsymbol{V}}$–holomorphic for more general collections $\boldsymbol{V}$. For example, we have the following statement.

Lemma 3.30 *Set $\boldsymbol{v}_1=(a_1,b_1)$, $\boldsymbol{v}_2=(a_2,b_2)$ and $\boldsymbol{V}=\{(0,0),\boldsymbol{v}_1,\boldsymbol{v}_2\}$, and suppose that $\|\boldsymbol{v}_1\|$ is small enough for $\boldsymbol{F}(\boldsymbol{v}_1)$ to exist. If a_1 and b_1 are both nonzero, $|a_1|$ is sufficiently small with respect to $|b_1|$, and $\|\boldsymbol{v}_2\|$ is sufficiently small with respect to $|a_1|$, then each top level curve of $\boldsymbol{F}(\boldsymbol{v}_1)$ is $J_{\boldsymbol{V}}$–holomorphic for some $J_{\boldsymbol{V}}$ in $J_{\mathcal{U}(\mathbb{L})}$.*

The deformed building $\boldsymbol{G}(\boldsymbol{v})$ is defined analogously, and satisfies the analogues of Lemmas 3.29 and 3.30.

Second deformation process Consider $\boldsymbol{V}=\{(0,0),(a_1,b_1),(a_2,b_2)\}$ with b_1 and b_2 nonzero. For suitable choices of a_i and b_i, our second deformation process deforms the essential J–holomorphic curve $u_{\mathbb{L}}$ of $\boldsymbol{F}$ into a curve, $u_{\mathbb{L}}^{\boldsymbol{V}}$, which has the same asymptotics as $u_{\mathbb{L}}$ but is $J_{\boldsymbol{V}}$–holomorphic for some $J_{\boldsymbol{V}}$ that is adapted to $\mathbb{L}(\boldsymbol{V})$ with respect to ψ.

The primary deformation result is as follows.

Lemma 3.31 *Set $\boldsymbol{v} = (0, b)$, $\boldsymbol{V} = \{(0,0), \boldsymbol{v}\}$ and suppose that $0 < |b| < \epsilon$. For $s \in [0,1]$, let J_s be a smooth family of almost complex structures in $\mathcal{J}_{\mathcal{U}(\mathbb{L})}$ such that*

- *$J_0 = J$,*
- *J_s is adapted to $\mathbb{L}$, with respect to ψ, for all $s \in [0,1)$, and*
- *J_1 is adapted to $\mathbb{L}(\boldsymbol{V})$ with respect to ψ.*

Then the essential curve $u_{\mathbb{L}}$ of $\boldsymbol{F}$ belongs to a smooth family of J_s–holomorphic planes $u_{\mathbb{L}}(s)$ for $s \in [0,1]$. Moreover, the $J_{\boldsymbol{V}}$–holomorphic plane

$$u_{\mathbb{L}}(1) \colon \mathbb{C} \to S^2 \times S^2 \setminus (\mathbb{L}(\boldsymbol{V}) \cup L_{1,1})$$

is disjoint from the region $\{P_1 > 0\}$ and is essential with respect to $\mathcal{F}$, and the closure of the image of $p \circ u_{\mathbb{L}}(1)$ is A_0.

Proof By Lemma 3.23, the initial curve $u_{\mathbb{L}}$ has area equal to 1. Since $\mathbb{L}$ is monotone, no degenerations are possible until $s = 1$. In other words, the family of deformed curves $u_{\mathbb{L}}(s)$ exists for all $s \in [0,1)$ and it suffices to show that it extends to $s = 1$. To prove the first assertion of Lemma 3.31 we argue by contradiction, and assume that there is a sequence $s_j \to 1$ such that the curves $u_{\mathbb{L}}(s_j)$ converge to a nontrivial $J_{\boldsymbol{V}}$–holomorphic building $\boldsymbol{H}$ which includes curves with punctures asymptotic to $\mathbb{L}(\boldsymbol{v})$. We will show that this implies that, unlike $u_{\mathbb{L}}$, none of the curves of $\boldsymbol{H}$ intersect T_0, a contradiction.

Claim 1 *Let v be a $J_{\boldsymbol{V}}$–holomorphic curve of $\boldsymbol{H}$. Any puncture of v asymptotic to $\mathbb{L}(\boldsymbol{v})$ must cover a closed geodesic in a class $(k, l) \in H_1(\mathbb{L}(\boldsymbol{v}); \mathbb{Z})$ with $k \leq 0$.*

Since the closure of $p \circ u_{\mathbb{L}}$ is A_0, by our choice of coordinates in Section 3.5, $u_{\mathbb{L}}$ is disjoint from the leaves of $\mathcal{F}$ which intersect $\mathcal{U}(\mathbb{L})$ in the region $\{P_1 > 0\}$. The same is true of the curves $u_{\mathbb{L}}(s)$ for all $s < 1$. Hence, v must also be disjoint from these leaves. The curve v can be extended smoothly to the oriented blow-up of the relevant puncture, such that the resulting map $\bar{v}$ acts on the corresponding boundary circle as

$$\theta \mapsto (0, b, Q_1 + k\theta, Q_2 + l\theta)$$

for some $Q_1, Q_2 \in S^1$. The tangent space to the image of $\bar{v}$ at a boundary point on the circle is spanned by $\{k\,\partial/\partial Q_1 + l\,\partial/\partial Q_2, k\,\partial/\partial P_1 + l\,\partial/\partial P_2\}$. If k were positive, this would contradict the fact that v is disjoint from the leaves through $\{P_1 > 0\}$ since $\boldsymbol{v} = (0, b)$. This proves Claim 1.

Claim 2 *Let v be a $J_{\boldsymbol{V}}$–holomorphic curve with a puncture that is asymptotic to $\mathbb{L}(\boldsymbol{v})$ along a geodesic in a class which is a multiple of the foliation class, ie of the form $(0, l) \in H_1^{\psi(\boldsymbol{v})}(\mathbb{L}(\boldsymbol{v}); \mathbb{Z})$. Then v must cover a plane or cylinder of a twice broken leaf of the foliation $\mathcal{F}(\boldsymbol{V})$.*

This follows, as in [8, Lemma 6.2], from the asymptotic properties of holomorphic curves and the fact that v lies in $\{P_1 \leq 0\}$. Let w be a broken plane asymptotic (modulo taking to covers) to the same Reeb orbit as an end of v. Then if v does not cover w it must intersect all nearby leaves of the foliation, including those which lie in the region $\{P_1 > 0\}$. This gives a contradiction as in Claim 1, proving Claim 2.

We can now complete the proof of the first assertion of Lemma 3.31. Let $\boldsymbol{H}_{\text{top}}$ denote the collection of top level curves of $\boldsymbol{H}$, let $\boldsymbol{H}_{\mathbb{L}}$ be the subbuilding consisting of the middle and bottom level curves of $\boldsymbol{H}$ that map to the copies of $\mathbb{R} \times S^*\mathbb{T}^2$ and $T^*\mathbb{T}^2$ corresponding to $\mathbb{L}$, and let $\boldsymbol{H}_{\boldsymbol{v}}$ be the subbuilding consisting of the middle and bottom level curves of $\boldsymbol{H}$ that map to the copies of $\mathbb{R} \times S^*\mathbb{T}^2$ and $T^*\mathbb{T}^2$ corresponding to $\mathbb{L}(\boldsymbol{v})$.

Consider the classes $(k_1, l_1), \dots, (k_m, l_m) \in H_1(\mathbb{L}(\boldsymbol{v}); \mathbb{Z})$ of the geodesics determined by all of the punctures of top level curves of $\boldsymbol{H}$ that are asymptotic to $\mathbb{L}(\boldsymbol{v})$. These constitute the boundary of the cycle in $\mathbb{L}(\boldsymbol{v})$ that is obtained by gluing together the compactifications of the curves of $\boldsymbol{H}_{\boldsymbol{v}}$. Hence, the sum of the classes $(k_1, l_1), \dots, (k_m, l_m)$ must be $(0, 0)$ and, by Claim 1, each k_i must be zero. It then follows from Claim 2 that any curve of $\boldsymbol{H}$ with an end on $\mathbb{L}(\boldsymbol{v})$ must cover a plane or cylinder of a broken leaf of $\mathcal{F}(\boldsymbol{v})$.

Partition the curves of $\boldsymbol{H}_{\text{top}} \cup \boldsymbol{H}_{\boldsymbol{v}} = \boldsymbol{H} \setminus \boldsymbol{H}_{\mathbb{L}}$ into connected components based on the matching of their ends in the copies of $\mathbb{R} \times S^*\mathbb{T}^2$ and $T^*\mathbb{T}^2$ corresponding to $\mathbb{L}(\boldsymbol{v})$. Denote these components by $\boldsymbol{H}_1, \dots, \boldsymbol{H}_k$. The compactification of each $\boldsymbol{H}_j$ is a cycle representing a class in $\pi_2(S^2 \times S^2, \mathbb{L})$. By monotonicity, the symplectic area of this cycle is a positive integer. Since the area of $u_{\mathbb{L}}$ is one, we must have $k = 1$ and the area of the cycle determined by $\boldsymbol{H}_1$ must be one. Assuming the limit is a building including curves asymptotic to $\mathbb{L}(\boldsymbol{v})$, by definition all curves of $\boldsymbol{H}_1$ have ends on $\mathbb{L}(\boldsymbol{v})$. By the discussion above, this implies that all the curves of $\boldsymbol{H}_1$ must cover a plane or cylinder of a broken leaf of $\mathcal{F}(\boldsymbol{V})$ through $\mathbb{L}(\boldsymbol{v})$. None of these leaves intersect T_0, and neither do the curves of $\boldsymbol{H}_{\mathbb{L}}$. Hence, no curve of $\boldsymbol{H} = \boldsymbol{H}_1 \cup \boldsymbol{H}_{\mathbb{L}}$ intersects T_0, which is the desired contradiction.

The remaining assertions of Lemma 3.31 follow easily from positivity of intersection. To see that $u_{\mathbb{L}}(1)$ is disjoint from the region $\{P_1 > 0\}$ note that the initial curve $u_{\mathbb{L}}$ is disjoint from the leaves of the foliation that intersect this region since its image under p is A_0 and our choice of coordinates has $\{P_1 > 0\}$ projecting to B. Positivity of intersection implies that no new intersections of the $u_{\mathbb{L}}(s)$ with these fibers can appear during the deformation.

Finally, since $u_{\mathbb{L}}(1)$ does not cover a leaf of the foliation, it also follows from positivity of intersection that $u_{\mathbb{L}}(1)$ is disjoint from the hypersurface $\{P_1 = 0\}$. In particular, any intersection with leaves in $\{P_1 = 0\}$ would imply intersections with the region $\{P_1 > 0\}$. Hence the closure of $p \circ u_{\mathbb{L}}(1)$ is equal to A_0. □

Translating the Lagrangian tori of Lemma 3.31 slightly in the P_1–direction, we get the following generalization.

Corollary 3.32 *Let $u_{\mathbb{L}}$ be the essential curve of $\boldsymbol{F}$ which is mapped by p onto A_0. Choose nonzero constants b_1 and b_2 in $(-\epsilon,\epsilon)$. If $\delta>0$ is sufficiently small, then for any a_1 and a_2 in $(-\delta,\delta)$ and*

$$\boldsymbol{V}=\{(0,0),(a_1,b_1),(a_2,b_2)\},$$

there is a $J_{\boldsymbol{V}}$–holomorphic curve

$$u_{\mathbb{L}}^{\boldsymbol{V}}\colon \mathbb{C}\to S^2\times S^2\setminus(\mathbb{L}(\boldsymbol{V})\cup L_{1,1})$$

in the class of $u_{\mathbb{L}}$ such that $u_{\mathbb{L}}^{\boldsymbol{V}}$ is disjoint from the region $\{P_1>0\}$ and is essential with respect to $\mathcal{F}$, and the closure of the image of $p\circ u_{\mathbb{L}}^{\boldsymbol{V}}$ is A_0.

Proof This has been established in the case when $a_1=a_2=0$. But then if the a_i are sufficiently small we can appeal to Lemma 3.27 to see that still there are no degenerations. □

Intersections near $\mathbb{L}$ Consider translation data

$$\boldsymbol{V}=\{\mathbf{0},\boldsymbol{v}_1,\boldsymbol{v}_2\}=\{(0,0),(a_1,b_1),(a_2,b_2)\}.$$

In what follows we will always assume that $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ are distinct and the a_i and b_i are as small as necessary but not zero. If $\|\boldsymbol{v}_1\|$ is sufficiently small then, as described in Lemma 3.30, the deformed building $\boldsymbol{F}(\boldsymbol{v}_1)$ is well defined and its top level curves

$$\{u_{\mathbb{L}}(\boldsymbol{v}_1),\underline{u}(\boldsymbol{v}_1),u_{L_{1,1}},u_1(\boldsymbol{v}_1),\dots,u_{d-1}(\boldsymbol{v}_1),\mathfrak{u}_1,\dots,\mathfrak{u}_d\}$$

are all $J_{\boldsymbol{V}}$–holomorphic for some $J_{\boldsymbol{V}}$ in $J_{\mathcal{U}(\mathbb{L})}$.

We also assume that Corollary 3.32 holds for $\boldsymbol{V}$. This yields a $J_{\boldsymbol{V}}$–holomorphic curve $u_{\mathbb{L}}^{\boldsymbol{V}}$ which is disjoint from the region $\{P_1>0\}$ and intersects the leaves of $\mathcal{F}(\boldsymbol{V})$ that pass through the planes $\{P_1=c<0,\,Q_1=\theta\}$ exactly once.

The intersection number between each top level curve of $\boldsymbol{F}(\boldsymbol{v}_1)$ and the curve $u_{\mathbb{L}}^{\boldsymbol{V}}$ is well defined since the curves of $\boldsymbol{F}(\boldsymbol{v}_1)$ are disjoint from $\mathbb{L}$, as established in Lemma 3.30; see Figure 4. We denote the total of these intersection numbers by $\boldsymbol{F}(\boldsymbol{v}_1)\bullet u_{\mathbb{L}}^{\boldsymbol{V}}$.

Similarly, the intersection number of each top level curve of $\boldsymbol{F}(\boldsymbol{v}_1)$ with any of the planes in either $\mathfrak{r}_0$ or $\mathfrak{r}_\infty$ is well defined and all such intersections are positive. Since this number is the same for any plane in the family, we denote these numbers by $\boldsymbol{F}(\boldsymbol{v}_1)\bullet\mathfrak{r}_0$ and $\boldsymbol{F}(\boldsymbol{v}_1)\bullet\mathfrak{r}_\infty$, respectively.

Let $\bar{\boldsymbol{F}}(\boldsymbol{v}_1)\colon S^2\to S^2\times S^2$ be the compactification of $\boldsymbol{F}(\boldsymbol{v}_1)$, let $\mathbb{E}\colon (D^2,S^1)\to(S^2\times S^2,\mathbb{L})$ be the compactification of the curve $u_{\mathbb{L}}^{\boldsymbol{V}}$, and let $\bar{\mathfrak{r}}_0$ and $\bar{\mathfrak{r}}_\infty$ be the solid tori obtained by compactifying the planes of $\mathfrak{r}_0$ and $\mathfrak{r}_\infty$. Deforming $\bar{\boldsymbol{F}}(\boldsymbol{v}_1)$ in a neighborhood of $\mathbb{L}(\boldsymbol{v}_1)$, we obtain a smooth map $F=F(\boldsymbol{v}_1)\colon S^2\to S^2\times S^2$ such that

$$F\bullet\mathbb{E}=\bar{\boldsymbol{F}}(\boldsymbol{v}_1)\bullet\mathbb{E}=\boldsymbol{F}(\boldsymbol{v}_1)\bullet u_{\mathbb{L}}^{\boldsymbol{V}}, \tag{3-2}$$

$$F\bullet\bar{\mathfrak{r}}_*=\bar{\boldsymbol{F}}(\boldsymbol{v}_1)\bullet\bar{\mathfrak{r}}_*=\boldsymbol{F}(\boldsymbol{v}_1)\bullet\mathfrak{r}_*\quad\text{for } *=0,\infty, \tag{3-3}$$

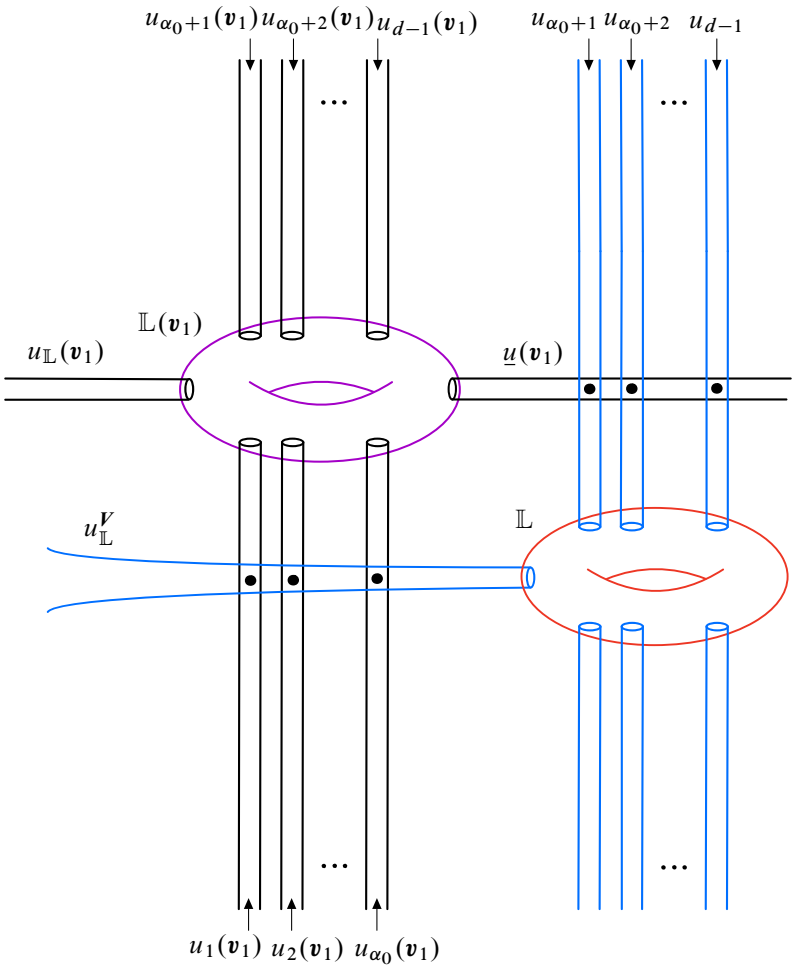


Figure 4: The intersection pattern of Lemma 3.33 for the case $b_1 > 0$. The large dots in the figure represent the isolated intersections points, in $\mathcal{U}(\mathbb{L})$, of the relevant pairs of curves.

where $\bar{\boldsymbol{F}}(\boldsymbol{v}_1) \bullet \bar{\mathfrak{r}}_*$ denotes the intersection number with any disk in the family. Moreover, the intersection points that determine the equal intersection numbers in (3-2) and (3-3) are identical.

Recall that α_0 is the number of top level curves of $\boldsymbol{F}$ lying in $\mathfrak{r}_0$. Hence, by Proposition 3.20, there are $d-1-\alpha_0$ top level curves lying in $\mathfrak{r}_\infty$.

Lemma 3.33 *Consider $\boldsymbol{V} = \{\mathbf{0}, \boldsymbol{v}_1, \boldsymbol{v}_2\} = \{(0,0), (a_1, b_1), (a_2, b_2)\}$ such that $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ are distinct, a_1 is negative, and b_1 and b_2 are nonzero. Suppose that $|a_1|$ is sufficiently small with respect to $|b_1|$.*

If $b_1 > 0$, then $F \bullet \bar{\mathfrak{r}}_0 = 0$, $F \bullet \bar{\mathfrak{r}}_\infty = 1$ and $F \bullet \mathbb{E} = \alpha_0$.

If $b_1 < 0$, then $F \bullet \bar{\mathfrak{r}}_0 = 1$, $F \bullet \bar{\mathfrak{r}}_\infty = 0$ and $F \bullet \mathbb{E} = d-1-\alpha_0$.

Proof Here we give the proof of the case when b_1 is positive. The proof for $b_1 < 0$ is identical and is left to the reader.

The situation for $b_1 > 0$ is illustrated in Figure 4, where the black curves can be deformed near $\mathbb{L}(\boldsymbol{v}_1)$ to form our sphere F. As the figure suggests, the contribution to $F \bullet \bar{\mathfrak{r}}_0$ from intersections in $\mathcal{U}(\mathbb{L})$ is zero, the contribution to $F \bullet \bar{\mathfrak{r}}_\infty$ from intersections in $\mathcal{U}(\mathbb{L})$ is one, and the contribution to $F \bullet \mathbb{E}$ from intersections in $\mathcal{U}(\mathbb{L})$ is α_0. These assertions are proven below along with the fact that there are no other contributions to these numbers.

The map F represents the class $(1, d)$. For each disk in $\bar{\mathfrak{r}}_0$ there is a companion disc in $\bar{\mathfrak{r}}_\infty$ such that the pair can be glued together, along $\mathbb{L}$, to form a sphere in the class $(0, 1)$. Hence,

$$F \bullet \bar{\mathfrak{r}}_0 + F \bullet \bar{\mathfrak{r}}_\infty = 1.$$

Since all intersections are positive, in order to prove that $F \bullet \bar{\mathfrak{r}}_0 = 0$, and $F \bullet \bar{\mathfrak{r}}_\infty = 1$, it suffices to prove that $F \bullet \bar{\mathfrak{r}}_\infty \geq 1$. In particular, it suffices to show that for the curve $\underline{u}(\boldsymbol{v}_1)$ of $\boldsymbol{F}(\boldsymbol{v}_1)$, we have $\underline{u}(\boldsymbol{v}_1) \bullet \mathfrak{r}_\infty \geq 1$.

The curve $\underline{u}(\boldsymbol{v}_1)$ is essential and projects under p to the region in S_∞ bounded by $p(\mathbb{L}(\boldsymbol{v}_1))$ and $p(L_{1,1})$. Thus it intersects $\mathcal{U}(\mathbb{L})$ in the region $\{P_1 > a_1\}$ and intersects all leaves of the foliation which meet $\mathcal{U}(\mathbb{L})$ in this set. Also, if $\mathcal{U}(\mathbb{L})$ is sufficiently small, it intersects $\mathcal{U}(\mathbb{L})$ inside $\{P_2 > 0\}$. This is true when $a_1 = 0$ because $b_1 > 0$, and remains true for small a_1 by continuity. As $\mathfrak{r}_\infty$ intersects $\mathcal{U}(\mathbb{L})$ in the region $\{P_1 = 0,\ P_2 > 0\}$ and $\mathfrak{r}_0$ in the region $\{P_1 = 0,\ P_2 < 0\}$, we see that $\underline{u}(\boldsymbol{v}_1)$ intersects the planes in $\mathfrak{r}_\infty$ rather than those in $\mathfrak{r}_0$, as required.

It remains to prove that $F \bullet \mathbb{E} = \alpha_0$ when $|a_1|$ is sufficiently small with respect to $|b_1|$. By (3-2), and the fact that the top level curves of $\boldsymbol{F}(\boldsymbol{v}_1)$ are

$$\{u_{\mathbb{L}}(\boldsymbol{v}_1), \underline{u}(\boldsymbol{v}_1), u_{L_{1,1}}, u_1(\boldsymbol{v}_1), \dots, u_{d-1}(\boldsymbol{v}_1), \mathfrak{u}_1, \dots, \mathfrak{u}_d\},$$

it suffices to prove that for $|a_1|$ sufficiently small with respect to $|b_1|$, we have

$$u_i(\boldsymbol{v}_1) \bullet u_{\mathbb{L}}^{\boldsymbol{V}} = 1 \quad \text{for } 1 \leq i \leq \alpha_0, \tag{3-4}$$

and $u_{\mathbb{L}}^{\boldsymbol{V}}$ is disjoint from all the other top level curves of $\boldsymbol{F}(\boldsymbol{v}_1)$.

By Corollary 3.32, the curve $u_{\mathbb{L}}^{\boldsymbol{V}}$ is essential for $\mathcal{F}$, and the closure of the image of $p \circ u_{\mathbb{L}}^{\boldsymbol{V}}$ is A_0. So if w is another curve in $S^2 \times S^2$ and $p \circ w$ is disjoint from A_0, then $u_{\mathbb{L}}^{\boldsymbol{V}}$ is disjoint from w. This observation implies that $u_{\mathbb{L}}^{\boldsymbol{V}}$ is disjoint from $u_{L_{1,1}}$ and the $\mathfrak{u}_j$ for $j = 1, \dots, d$, since these curves all project into A_∞.

Another consequence of $u_{\mathbb{L}}^{\boldsymbol{V}}$ being essential with respect to $\mathcal{F}$ is that it intersects any fiber of $\mathcal{F}$ either once or not at all. The curve $u_{\mathbb{L}}^{\boldsymbol{V}}$ intersects $\mathcal{U}(\mathbb{L})$ in the region $\{P_1 < 0\}$ and has an end asymptotic to a circle in $\mathbb{L} = \{P_1 = P_2 = 0\}$. Since $b_1 > 0$, this implies that for all $a_1 < 0$ such that $|a_1|$ is sufficiently small with respect to b_1, $u_{\mathbb{L}}^{\boldsymbol{V}}$ must intersect the annuli of the form $\{P_1 = a_1, Q_1 = \theta, P_2 < b_1\}$ exactly once. Now the planes $u_i(\boldsymbol{v}_1)$ all belong to broken fibers of $\mathcal{F}$ that intersect $\mathcal{U}(\mathbb{L})$. For $1 \leq i \leq \alpha_0$, the curves $u_i(\boldsymbol{v}_1)$ intersect $\mathcal{U}(\mathbb{L})$ in annuli of the form $\{P_1 = a_1, Q_1 = \theta, P_2 < b_1\}$. For $i > \alpha_0$, the $u_i(\boldsymbol{v}_1)$ intersect $\mathcal{U}(\mathbb{L})$ in annuli of the form $\{P_1 = a_1, Q_1 = \theta, P_2 > b_1\}$. Hence, for $1 \leq i \leq \alpha_0$, $u_{\mathbb{L}}^{\boldsymbol{V}}$ intersects the fiber of $\mathcal{F}$ containing $u_i(\boldsymbol{v}_1)$ at a point on $u_i(\boldsymbol{v}_1)$. This yields equation (3-4). On the other hand,

for $i > \alpha_0$, $u_{\mathbb{L}}^{V}$ intersects the fiber of $\mathcal{F}$ containing $u_i(\boldsymbol{v}_1)$ at a point in the complement of $u_i(\boldsymbol{v}_1)$. Hence, $u_{\mathbb{L}}^{V}$ is disjoint from these curves.

Next we show that, when $|a_1|$ is sufficiently small with respect to $|b_1|$, $u_{\mathbb{L}}^{V}$ is disjoint from $\underline{u}(\boldsymbol{v}_1)$. Considering projections, it is clear that the part of $\underline{u}(\boldsymbol{v}_1)$ in the complement of $\mathcal{U}(\mathbb{L})$ is disjoint from $u_{\mathbb{L}}^{V}$ since its projection is contained in the interior of $B \cup A_\infty$.

Suppose that $a_1 = 0$. Then $\underline{u}((0, b_1)) \cap \mathcal{U}(\mathbb{L})$ is contained in $\{P_1 > 0\}$ and is asymptotic to $\mathbb{L}(0, b_1)$. This is disjoint from $u_{\mathbb{L}}^{V} \cap \mathcal{U}(\mathbb{L})$, which is contained in $\{P_1 < 0\}$ and is asymptotic to $\mathbb{L}$. By continuity, $\underline{u}((a_1, b_1)) \cap \mathcal{U}(\mathbb{L})$ is then disjoint from $u_{\mathbb{L}}^{V} \cap \mathcal{U}(\mathbb{L})$ for all $a_1 < 0$ with $|a_1|$ sufficiently small with respect to $|b_1|$.

Lastly, we must prove that

$$u_{\mathbb{L}}(\boldsymbol{v}_1) \bullet u_{\mathbb{L}}^{V} = 0$$

when $|a_1|$ is sufficiently small with respect to $|b_1|$. Following Lemma 3.31 the compactifications of $u_{\mathbb{L}}^{V}$ and $u_{\mathbb{L}}$ are homotopic in the space of smooth maps $(D^2, S^1) \to (p^{-1}(A_0), \mathbb{L})$, so for a_1 sufficiently small it suffices to show that

$$u_{\mathbb{L}}(\boldsymbol{v}_1) \bullet u_{\mathbb{L}} = 0.$$

Let $\bar{u}_{\mathbb{L}}(\boldsymbol{v}_1)$ and $\bar{u}_{\mathbb{L}}$ be compactifications of $u_{\mathbb{L}}(\boldsymbol{v}_1)$ and $u_{\mathbb{L}}$. We claim that $u_{\mathbb{L}}(\boldsymbol{v}_1) \bullet u_{\mathbb{L}} = 0$ is equivalent to the fact that the Maslov index of $\bar{u}_{\mathbb{L}}$ is equal to 2. To see this we recall that

$$\mu(\bar{u}_{\mathbb{L}}) = 2c_1(\bar{u}_{\mathbb{L}}), \tag{3-5}$$

where $c_1(\bar{u}_{\mathbb{L}})$ is the relative Chern number of $\bar{u}_{\mathbb{L}}$, which is equal to the number of zeros of a generic section ξ of $\bar{u}_{\mathbb{L}}^*(\Lambda^2(T(S^2 \times S^2)))$ such that $\xi|_{S^1}$ is nonvanishing and is tangent to $\Lambda^2(T\mathbb{L})$.

Let $\nu(\bar{u}_{\mathbb{L}})$ be the normal bundle to the embedding $\bar{u}_{\mathbb{L}}$ and fix an identification of $\bar{u}_{\mathbb{L}}^*(T(S^2 \times S^2))$ with the Whitney sum $\nu(\bar{u}_{\mathbb{L}}) \oplus T(D^2)$. For polar coordinates (r, θ) on D^2 consider the section $r\,\partial/\partial\theta$ of $\bar{u}_{\mathbb{L}}^*(T(S^2 \times S^2))$. The restriction $r\,\partial/\partial\theta|_{S^1}$ is nonvanishing and tangent to $T\mathbb{L}$.

Replacing $\boldsymbol{v}_1$ by $t\boldsymbol{v}_1$ for some small $t > 0$, if necessary, we may assume that $\bar{u}_{\mathbb{L}}(\boldsymbol{v}_1)$ is close enough $\bar{u}_{\mathbb{L}}$, in the C^1–topology, to be identified with a section, $\sigma_{\mathbb{L}}(\boldsymbol{v}_1)$, of $\nu(\bar{u}_{\mathbb{L}}) \subset \bar{u}_{\mathbb{L}}^*(T(S^2 \times S^2))$. The restriction $\sigma_{\mathbb{L}}(\boldsymbol{v}_1)|_{S^1}$ is roughly parallel to the vector field $\partial/\partial P_2$. By rotating in the normal bundle this section is homotopic through nonvanishing sections of the normal bundle to a section of $T\mathbb{L}$ along ∂D^2 which is orthogonal to $\partial/\partial\theta$.

Set $\xi = r\,\partial/\partial\theta \wedge \sigma_{\mathbb{L}}(\boldsymbol{v}_1)$. It follows from the discussion above that $\xi|_{S^1}$ is nonvanishing and is tangent to $\Lambda^2(T\mathbb{L})$. Moreover, the zeroes of ξ correspond to the union of the zeros of $r\,\partial/\partial\theta$ and $\sigma_{\mathbb{L}}(\boldsymbol{v}_1)$. Since $\bar{u}_{\mathbb{L}}$ is embedded, the zeros of $\sigma_{\mathbb{L}}(\boldsymbol{v}_1)$ exactly correspond to the intersections $u_{\mathbb{L}}(\boldsymbol{v}_1) \bullet u_{\mathbb{L}}$. By (3-5), we have

$$\mu(\bar{u}_{\mathbb{L}}) = 2(1 + u_{\mathbb{L}}(\boldsymbol{v}_1) \bullet u_{\mathbb{L}}).$$

As $\mu(\bar{u}_{\mathbb{L}}) = 2$ (as it has area 1 by Lemma 3.23, and $\mathbb{L}$ is monotone) we have $u_{\mathbb{L}}(\boldsymbol{v}_1) \bullet u_{\mathbb{L}} = 0$. □

Assuming that $\boldsymbol{v}_2 = (a_2, b_2)$ is sufficiently small, we can deform the building $\boldsymbol{G}$ to obtain a new building $\boldsymbol{G}(\boldsymbol{v}_2)$ with top level curves

$$\{v_{\mathbb{L}}(\boldsymbol{v}_2), \underline{v}(\boldsymbol{v}_2), u_{L_{1,1}}, v_1(\boldsymbol{v}_2), \dots, v_d(\boldsymbol{v}_2), \mathfrak{v}_1, \dots, \mathfrak{v}_{d-1}\}.$$

Let $\bar{\boldsymbol{G}}(\boldsymbol{v}_2)\colon S^2 \to S^2 \times S^2$ be the compactification of $\boldsymbol{G}(\boldsymbol{v}_2)$. Again we can deform $\bar{\boldsymbol{G}}(\boldsymbol{v}_2)$, arbitrarily close to $\mathbb{L}(\boldsymbol{v}_2)$, to get a smooth map $G = G(\boldsymbol{v}_2)\colon S^2 \to S^2 \times S^2$ such that

$$\begin{aligned} G \bullet \mathbb{E} &= \bar{\boldsymbol{G}}(\boldsymbol{v}_2) \bullet \mathbb{E} = \boldsymbol{G}(\boldsymbol{v}_2) \bullet u_{\mathbb{L}}^{\boldsymbol{V}}, \\ G \bullet \bar{\mathfrak{r}}_* &= \bar{\boldsymbol{G}}(\boldsymbol{v}_2) \bullet \bar{\mathfrak{r}}_* = \boldsymbol{G}(\boldsymbol{v}_2) \bullet \mathfrak{r}_* \quad \text{for } * = 0, \infty. \end{aligned}$$

Arguing as in the proof of Lemma 3.33 we get the following.

Lemma 3.34 *Consider $\boldsymbol{V} = \{\mathbf{0}, \boldsymbol{v}_1, \boldsymbol{v}_2\} = \{(0,0), (a_1, b_1), (a_2, b_2)\}$ such that a_2 is negative, and b_1 and b_2 are nonzero. Suppose that $|a_2|$ is sufficiently small with respect to $|b_2|$.*

If $b_2 > 0$, then

$$G \bullet \mathbb{E} = \gamma_0 + v_{\mathbb{L}}(\boldsymbol{v}_2) \bullet u_{\mathbb{L}}^{\boldsymbol{V}}, \quad G \bullet \bar{\mathfrak{r}}_0 = 0 \quad \text{and} \quad G \bullet \bar{\mathfrak{r}}_\infty = 1.$$

If $b_2 < 0$, then

$$G \bullet \mathbb{E} = d - \gamma_0 + v_{\mathbb{L}}(\boldsymbol{v}_2) \bullet u_{\mathbb{L}}^{\boldsymbol{V}}, \quad G \bullet \bar{\mathfrak{r}}_0 = 1 \quad \text{and} \quad G \bullet \bar{\mathfrak{r}}_\infty = 0.$$

The term $v_{\mathbb{L}}(\boldsymbol{v}_2) \bullet u_{\mathbb{L}}^{\boldsymbol{V}}$ is not necessarily equal to zero. Instead we have the following identity.

Lemma 3.35 *For $\boldsymbol{V} = \{\mathbf{0}, \boldsymbol{v}_1, \boldsymbol{v}_2\} = \{(0,0), (a_1, b_1), (a_2, b_2)\}$, where b_1 and b_2 have opposite signs, and a_1 and a_2 are sufficiently small relative to b_1 and b_2, we have*

$$v_{\mathbb{L}}(\boldsymbol{v}_2) \bullet u_{\mathbb{L}}^{\boldsymbol{V}} = v_{\mathbb{L}}(\boldsymbol{v}_2) \bullet u_{\mathbb{L}}(\boldsymbol{v}_1).$$

Proof First we consider the case when $a_1 = a_2 = 0$. The image of the map $v_{\mathbb{L}}(\boldsymbol{v}_2)$ projects to A_0 and its boundary lies in $\mathbb{L}(\boldsymbol{v}_2)$. Hence, using our assumption on sign, the family of Lagrangians $\mathbb{L}(t\boldsymbol{v}_1)$ for $0 \le t \le 1$ are disjoint from the compactification of $v_{\mathbb{L}}(\boldsymbol{v}_2)$. It then follows from the proof of Lemma 3.27 that the compactification of $u_{\mathbb{L}}$ is connected to that of $u_{\mathbb{L}}(\boldsymbol{v}_1)$ by a path of smooth maps $u_t\colon (D^2, S^1) \to (S^2 \times S^2, \mathbb{L}(t\boldsymbol{v}_1))$. Therefore we have, as required,

$$v_{\mathbb{L}}(\boldsymbol{v}_2) \bullet u_{\mathbb{L}} = v_{\mathbb{L}}(\boldsymbol{v}_2) \bullet u_{\mathbb{L}}(\boldsymbol{v}_1).$$

For the general case we use the fact that the maps vary continuously with the parameters and so the intersection numbers remain unchanged for a_1 and a_2 sufficiently small. □

When $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ are distinct, with $b_1 \neq b_2$ and $a_1 \neq a_2$ sufficiently small, the intersection numbers of the top level curves of $\boldsymbol{F}(\boldsymbol{v}_1)$ and $\boldsymbol{G}(\boldsymbol{v}_2)$ are also well defined. The following results concerning these intersections will be useful.

Lemma 3.36 *For $\boldsymbol{v}_1 = (a_1, b_1)$ and $\boldsymbol{v}_2 = (a_2, b_2)$, suppose that $a_1 < a_2 < 0$, with a_1 and a_2 sufficiently small.*

If $b_1 > b_2$, then

$$u_i(\boldsymbol{v}_1) \bullet v_{\mathbb{L}}(\boldsymbol{v}_2) = 1 \quad \text{for } i = 1, \dots, \alpha_0,$$
$$v_i(\boldsymbol{v}_2) \bullet \underline{u}(\boldsymbol{v}_1) = 1 \quad \text{for } i = \gamma_0 + 1, \dots, d.$$

If $b_1 < b_2$, then

$$u_i(\boldsymbol{v}_1) \bullet v_{\mathbb{L}}(\boldsymbol{v}_2) = 1 \quad \text{for } i = \alpha_0 + 1, \dots, d-1,$$
$$v_i(\boldsymbol{v}_2) \bullet \underline{u}(\boldsymbol{v}_1) = 1 \quad \text{for } i = 1, \dots, \gamma_0.$$

Moreover, all the intersection points here project to A_0.

Proof Since the curves $\underline{u}(\boldsymbol{v}_1)$ and $v_{\mathbb{L}}(\boldsymbol{v}_2)$ are essential with respect to $\mathcal{F}$, they intersect a leaf of the foliation either once or not at all. Hence it suffices to detect a single intersection of the relevant pairs of curves listed. We detect an intersection for the first type of pair above and leave the other cases to the reader. For $1 \le i \le \alpha_0$ the planes $u_i(\boldsymbol{v}_1)$ intersect $\mathcal{U}(\mathbb{L})$ in annuli $\{P_1 = a_1, Q_1 = \theta, P_2 < b_1\}$. As $v_{\mathbb{L}}((0, b_2))$ is asymptotic to $\mathbb{L}((0, b_2)) = \{P_1 = 0, P_2 = b_2\}$ it intersects $u_i(\boldsymbol{v}_1)$ provided a_1 is sufficiently small (since the boundary of $v_{\mathbb{L}}((0, b_2))$ intersects all annuli $\{P_1 = 0, Q_1 = \theta, P_2 < b_1\}$). For a_2 sufficiently small, the plane $v_{\mathbb{L}}(\boldsymbol{v}_2)$ is a deformation of $v_{\mathbb{L}}((0, b_2))$ and so the intersection persists. As $v_{\mathbb{L}}(\boldsymbol{v}_2)$ intersects fibers at most once, the intersection number is equal to 1. Since $a_1 < 0$, the intersection point projects to A_0. □

Corollary 3.37 *For $\boldsymbol{v}_1 = (a_1, b_1)$ and $\boldsymbol{v}_2 = (a_2, b_2)$, suppose that $a_1 < a_2 < 0$, with a_1 and a_2 sufficiently small.*

If $b_1 > b_2$, then $F \cap G$ contains at least $\alpha_0 + d - \gamma_0$ points in $\mathcal{U}(\mathbb{L})$ that project to A_0.

If $b_1 < b_2$, then $F \cap G$ contains at least $d - 1 - \alpha_0 + \gamma_0$ points in $\mathcal{U}(\mathbb{L})$ that project to A_0.

Remark 3.38 It follows from Lemma 3.35 that any *excess* intersection points between F and G in $p^{-1}(A_0)$, that is, more than described by Corollary 3.37, correspond to intersection points between G and $\mathbb{E}$, at least if the b_i have opposite sign and the a_i are sufficiently small.

Adding deformations near $L_{1,1}$ To completely resolve the intersections of F and G we must also apply deformations in the Weinstein neighborhood

$$\mathcal{U}(L_{1,1}) = \{|p_1| < \epsilon, |p_2| < \epsilon\}.$$

Here we consider nearby Lagrangian tori of the form

$$L_{1,1}(\boldsymbol{w}) := \{p_1 = c, p_2 = d\} \quad \text{for } \boldsymbol{w} = (c, d) \in (-\epsilon, \epsilon) \times (-\epsilon, \epsilon).$$

The space of almost complex structures that are adapted to collections of these translated Lagrangian tori near $L_{1,1}$, with respect to $\psi_{1,1}$, is defined analogously to $\mathcal{J}_{\mathcal{U}(\mathbb{L})}$ and is denoted by $\mathcal{J}_{\mathcal{U}(L_{1,1})}$.

Given nontrivial collections

$$\boldsymbol{V} = \{\boldsymbol{v}_1, \dots, \boldsymbol{v}_k\} = \{(a_1, b_1), \dots, (a_k, b_k)\} \quad \text{and} \quad \boldsymbol{W} = \{\boldsymbol{w}_1, \dots, \boldsymbol{w}_l\} = \{(c_1, d_1), \dots, (c_l, d_l)\},$$

set $\boldsymbol{X} = \{\boldsymbol{V}, \boldsymbol{W}\}$. Let $J_{\boldsymbol{X}}$ denote the corresponding (doubly) adapted almost complex structures in $\mathcal{J}_{\mathcal{U}(\mathbb{L})} \cap \mathcal{J}_{\mathcal{U}(L_{1,1})}$.

Lemma 3.27 generalizes to this setting as follows.

Lemma 3.39 *Let u be a regular J–holomorphic curve with $k \geq 0$ ends on $\mathbb{L}$ and $l \geq 0$ ends on $L_{1,1}$. For all $\boldsymbol{x} = \{\boldsymbol{v}, \boldsymbol{w}\} = \{(a, b), (c, d)\}$ with $\|\boldsymbol{x}\|$ sufficiently small, there is a $J_{\boldsymbol{x}}$–holomorphic curve $u(\boldsymbol{x})$ that represents the class in $\pi_2(S^2 \times S^2, \mathbb{L}(\boldsymbol{v}) \cup L_{1,1}(\boldsymbol{w}))$ and which corresponds to the class $[u]$ in $\pi_2(S^2 \times S^2, \mathbb{L} \cup L_{1,1})$ under the obvious identification. The curve $u(\boldsymbol{x})$ has k ends on $\mathbb{L}(\boldsymbol{v})$ and these represent the identical classes in $H_1^{\psi(\boldsymbol{v})}(\mathbb{L}; \mathbb{Z})$ as do those of u in $H_1^{\psi}(\mathbb{L}; \mathbb{Z})$. The curve also has l ends on $L_{1,1}(\boldsymbol{w})$ which represent the identical classes in $H_1^{\psi_{1,1}(\boldsymbol{w})}(L_{1,1}; \mathbb{Z})$ as do those of u in $H_1^{\psi_{1,1}}(L_{1,1}; \mathbb{Z})$.*

Corollary 3.32 generalizes as follows.

Lemma 3.40 *Let $u_{\mathbb{L}}$ and $u_{L_{1,1}}$ be the essential curves of a building $\boldsymbol{F}$ as in Proposition 3.20. Let $\boldsymbol{X} = \{\boldsymbol{V}, \boldsymbol{W}\}$, where*

$$\boldsymbol{V} = \{(0, 0), (a_1, b_1), (a_2, b_2)\} \quad \text{and} \quad \boldsymbol{W} = \{(0, 0), (c_1, d_1), (c_2, d_2)\}.$$

If b_1, b_2, d_1 and d_2 are in $(-\epsilon, \epsilon)$ and a_1, a_2, c_1 and c_2 are in $(-\delta, \delta)$, then for all sufficiently small δ there is a $J_{\boldsymbol{X}}$–holomorphic curve

$$u_{\mathbb{L}}^{\boldsymbol{X}} : \mathbb{C} \to S^2 \times S^2 \setminus (\mathbb{L}(\boldsymbol{V}) \cup L_{1,1}(\boldsymbol{W}))$$

in the class of $u_{\mathbb{L}}$ such that $u_{\mathbb{L}}^{\boldsymbol{X}}$ is disjoint from the region $\{P_1 > 0\}$, the closure of the image of $p \circ u_{\mathbb{L}}^{\boldsymbol{X}}$ is A_0, and $u_{\mathbb{L}}^{\boldsymbol{X}}$ intersects, exactly once, the leaves of $\mathcal{F}(\boldsymbol{X})$ that pass through the planes $\{P_1 = c < 0,\, Q_1 = \theta\}$.

There is also a $J_{\boldsymbol{X}}$–holomorphic curve

$$u_{L_{1,1}}^{\boldsymbol{X}} : \mathbb{C} \to S^2 \times S^2 \setminus (L(\boldsymbol{V}) \cup L_{1,1}(\boldsymbol{W}))$$

in the class of $u_{L_{1,1}}$ such that $u_{L_{1,1}}^{\boldsymbol{X}}$ is disjoint from the region $\{p_1 < 0\}$, the closure of the image of $p \circ u_{L_{1,1}}^{\boldsymbol{X}}$ is A_∞, and $u_{L_{1,1}}^{\boldsymbol{X}}$ intersects, exactly once, the leaves of $\mathcal{F}(\boldsymbol{X})$ that pass through the planes $\{p_1 = c > 0, q_1 = \theta\}$.

Completion of the proof of Proposition 3.24 Let $\boldsymbol{F}$ be a building as in Proposition 3.20 and let $\boldsymbol{G}$ be a building as in Proposition 3.22. Set

$$\boldsymbol{x}_1 = \{\boldsymbol{v}_1, \boldsymbol{w}_1\} = \{(a_1, b_1), (c_1, d_1)\}, \qquad \boldsymbol{x}_2 = \{\boldsymbol{v}_2, \boldsymbol{w}_2\} = \{(a_2, b_2), (c_2, d_2)\},$$
$$\boldsymbol{V} = \{\boldsymbol{0}, \boldsymbol{v}_1, \boldsymbol{v}_2\} = \{(0, 0), (a_1, b_1), (a_2, b_2)\}, \quad \boldsymbol{W} = \{\boldsymbol{0}, \boldsymbol{w}_1, \boldsymbol{w}_2\} = \{(0, 0), (c_1, d_1), (c_2, d_2)\},$$

and set

$$\boldsymbol{X} = \{\boldsymbol{V}, \boldsymbol{W}\}.$$

We assume that $\|\boldsymbol{x}_1\|$ and $\|\boldsymbol{x}_2\|$ are small enough for Lemma 3.39 to yield the deformed buildings $\boldsymbol{F}(\boldsymbol{x}_1)$ and $\boldsymbol{G}(\boldsymbol{x}_2)$. We also assume that $|a_1|^2+|a_2|^2+|c_1|^2+|c_2|^2$ is small enough with respect to $|b_1|^2+|b_2|^2+|d_1|^2+|d_2|^2$ for Lemma 3.40 to yield the deformations $u_{\mathbb{L}}^{\boldsymbol{X}}$ and $u_{L_{1,1}}^{\boldsymbol{X}}$.

Let $\mathbb{E}\colon (D^2,S^1)\to(S^2\times S^2,\mathbb{L})$ be the compactification of $u_{\mathbb{L}}^{\boldsymbol{X}}$, and $E_{1,1}\colon (D^2,S^1)\to(S^2\times S^2,L_{1,1})$ be the compactification of $u_{L_{1,1}}^{\boldsymbol{X}}$. Since the homology classes represented by the ends of $u_{\mathbb{L}}^{\boldsymbol{X}}$ and $u_{L_{1,1}}^{\boldsymbol{X}}$ are identical to those of the essential curves $u_{\mathbb{L}}$ and $u_{L_{1,1}}$, the maps $\mathbb{E}$ and $E_{1,1}$ satisfy conditions (2) and (3) of Proposition 3.24.

Consider compactifications $\bar{\boldsymbol{F}}(\boldsymbol{x}_1)\colon S^2\to S^2\times S^2$ of $\boldsymbol{F}(\boldsymbol{x}_1)$, and $\bar{\boldsymbol{G}}(\boldsymbol{x}_2)\colon S^2\to S^2\times S^2$ of $\boldsymbol{G}(\boldsymbol{x}_2)$. Arguing as before, we can perturb these maps, arbitrarily close to the Lagrangians $\mathbb{L}(\boldsymbol{v}_1)$, $L_{1,1}(\boldsymbol{w}_1)$, $\mathbb{L}(\boldsymbol{v}_2)$ and $L_{1,1}(\boldsymbol{w}_2)$, to obtain smooth spheres F and G such that condition (1) of Proposition 3.24 holds.

It remains to verify the conditions (4) through (9) of Proposition 3.24, which involve intersections.

In the current setting, Lemma 3.33 holds as stated and the proof is unchanged.

Lemma 3.41 *Suppose a_1 is negative, and b_1 and b_2 are nonzero. Suppose that $|a_1|$ is sufficiently small with respect to $|b_1|$.*

If $b_1>0$, then $F\bullet\bar{\mathfrak{r}}_0=0$, $F\bullet\bar{\mathfrak{r}}_\infty=1$ and $F\bullet\mathbb{E}=\alpha_0$.

If $b_1<0$, then $F\bullet\bar{\mathfrak{r}}_0=1$, $F\bullet\bar{\mathfrak{r}}_\infty=0$ and $F\bullet\mathbb{E}=d-1-\alpha_0$.

Lemmas 3.34 and 3.35 and Corollary 3.37 change only in notation, and yield the following.

Lemma 3.42 *Suppose that a_2 is negative, b_1 and b_2 are nonzero, and $|a_2|$ is sufficiently small with respect to $|b_2|$.*

If $b_2>0$, then $G\bullet\bar{\mathfrak{r}}_0=0$, $G\bullet\bar{\mathfrak{r}}_\infty=1$ and

$$G\bullet\mathbb{E}=\gamma_0+v_{\mathbb{L}}(\boldsymbol{x}_2)\bullet u_{\mathbb{L}}^{\boldsymbol{X}}.$$

If $b_2<0$, then $G\bullet\bar{\mathfrak{r}}_0=1$, $G\bullet\bar{\mathfrak{r}}_\infty=0$ and

$$G\bullet\mathbb{E}=d-\gamma_0+v_{\mathbb{L}}(\boldsymbol{x}_2)\bullet u_{\mathbb{L}}^{\boldsymbol{X}}.$$

Lemma 3.43 *If b_1 and b_2 have opposite sign, and a_1 and a_2 are sufficiently small, then*

$$v_{\mathbb{L}}(\boldsymbol{x}_2)\bullet u_{\mathbb{L}}^{\boldsymbol{X}}=v_{\mathbb{L}}(\boldsymbol{x}_2)\bullet u_{\mathbb{L}}(\boldsymbol{x}_1).$$

Lemma 3.44 *Suppose that $a_1<a_2<0$, and a_1 and a_2 are sufficiently small.*

If $b_1>b_2$, then $F\cap G$ contains at least $\alpha_0+d-\gamma_0$ points in $\mathcal{U}(\mathbb{L})$ that project to A_0.

If $b_1<b_2$, then $F\cap G\cap\mathcal{U}(\mathbb{L})$ contains at least $d-1-\alpha_0+\gamma_0$ points in $\mathcal{U}(\mathbb{L})$ that project to A_0.

The following analogous results follow from similar arguments.

Lemma 3.45 *Suppose c_1 is positive, d_1 and d_2 are nonzero, and $|c_1|$ is sufficiently small with respect to $|d_1|$.*

If $d_1 > 0$, then $F \bullet \bar{\mathfrak{s}}_0 = 0$, $F \bullet \bar{\mathfrak{s}}_\infty = 1$ and $F \bullet E_{1,1} = \beta_0$.

If $d_1 < 0$, then $F \bullet \bar{\mathfrak{s}}_0 = 1$, $F \bullet \bar{\mathfrak{s}}_\infty = 0$ and $F \bullet E_{1,1} = d - \beta_0$.

Lemma 3.46 *Suppose c_2 is positive, d_1 and d_2 are nonzero, and $|c_2|$ is sufficiently small with respect to $|d_2|$.*

If $d_2 > 0$, then $G \bullet \bar{\mathfrak{s}}_0 = 0$, $G \bullet \bar{\mathfrak{s}}_\infty = 1$ and

$$G \bullet E_{1,1} = \delta_0 + v_{L_{1,1}}(\boldsymbol{x}_2) \bullet u^{\boldsymbol{X}}_{L_{1,1}}.$$

If $d_2 < 0$, then $G \bullet \bar{\mathfrak{s}}_0 = 1$, $G \bullet \bar{\mathfrak{s}}_\infty = 0$ and

$$G \bullet E_{1,1} = d - 1 - \delta_0 + v_{L_{1,1}}(\boldsymbol{x}_2) \bullet u^{\boldsymbol{X}}_{L_{1,1}}.$$

Lemma 3.47 *If d_1 and d_2 have opposite sign, and c_1 and c_2 are sufficiently small, then*

$$v_{L_{1,1}}(\boldsymbol{x}_2) \bullet u^{\boldsymbol{X}}_{L_{1,1}} = v_{L_{1,1}}(\boldsymbol{x}_2) \bullet u_{L_{1,1}}(\boldsymbol{x}_1).$$

Lemma 3.48 *Suppose that $c_1 > c_2 > 0$, and c_1 and c_2 are sufficiently small.*

If $d_1 > d_2$, then $F \cap G$ contains at least $\beta_0 + d - 1 - \delta_0$ points in $\mathcal{U}(L_{1,1})$ that project to A_∞.

If $d_1 < d_2$, then $F \cap G$ contains at least $d - \beta_0 + \delta_0$ points in $\mathcal{U}(L_{1,1})$ that project to A_∞.

With F and G fixed as above, the remaining analysis can be organized using the following two alternatives:

- **Alternative 1** Either $\alpha_0 \geq \gamma_0$ or $\gamma_0 \geq \alpha_0 + 1$.
- **Alternative 2** Either $\beta_0 \geq \delta_0 + 1$ or $\delta_0 \geq \beta_0$.

Case 1 ($\alpha_0 \geq \gamma_0$ and $\beta_0 \geq \delta_0 + 1$) In this case, we choose our translations so that

$$a_1 < a_2 < 0, \quad b_2 < 0 < b_1, \quad 0 < c_2 < c_1, \quad d_2 < 0 < d_1.$$

For these conditions on b_1 and b_2, Lemmas 3.41 and 3.42 yield $F \bullet \mathfrak{r}_0 = 0$, $F \bullet \mathfrak{r}_\infty = 1$, $G \bullet \mathfrak{r}_0 = 1$ and $G \bullet \mathfrak{r}_\infty = 0$. This implies condition (4) of Proposition 3.24.

Similarly, for these conditions on d_1 and d_2, Lemmas 3.45 and 3.46 imply that $F \bullet \mathfrak{s}_0 = 0$, $F \bullet \mathfrak{s}_\infty = 1$, $G \bullet \mathfrak{s}_0 = 1$ and $G \bullet \mathfrak{s}_\infty = 0$. This gives condition (5) of Proposition 3.24.

The maps F and G both represent the class $(1, d)$ in $H_2(S^2 \times S^2; \mathbb{Z})$, so $F \bullet G = (1, d) \bullet (1, d) = 2d$. On the other hand, for the choices above, Lemmas 3.44 and 3.48 imply that

$$F \bullet G \geq (\alpha_0 + d - \gamma_0) + (\beta_0 + d - 1 - \delta_0).$$

In the current case, with $\alpha_0 \geq \gamma_0$ and $\beta_0 \geq \delta_0 + 1$, these two summands are each at least d, and so we must have

$$\alpha_0 = \gamma_0, \tag{3-6}$$

$$\beta_0 = 1 + \delta_0. \tag{3-7}$$

It follows that $F \cap G$ consists of exactly $2d$ points, d of which are contained in $\mathcal{U}(\mathbb{L})$ and project to A_0, and d of which are contained in $\mathcal{U}(L_{1,1})$ and project to A_∞. This yields conditions (8) and (9) of Proposition 3.24.

Since $F \bullet G = \boldsymbol{F}(\boldsymbol{x}_1) \bullet \boldsymbol{G}(\boldsymbol{x}_2)$, it follows from the equalities above that there can be no intersections between the essential curves of $\boldsymbol{F}(\boldsymbol{x}_1)$ and those of $\boldsymbol{G}(\boldsymbol{x}_2)$. In particular, we must have

$$v_{\mathbb{L}}(\boldsymbol{x}_2) \bullet u_{\mathbb{L}}(\boldsymbol{x}_1) = 0, \tag{3-8}$$

$$v_{L_{1,1}}(\boldsymbol{x}_2) \bullet u_{L_{1,1}}(\boldsymbol{x}_1) = 0. \tag{3-9}$$

Equation (3-8) and Lemma 3.43 imply that

$$v_{\mathbb{L}}(\boldsymbol{x}_2) \bullet u_{\mathbb{L}}^{\boldsymbol{X}} = 0.$$

By Lemmas 3.41 and 3.42 and equation (3-6), we then have

$$F \bullet \mathbb{E} + G \bullet \mathbb{E} = \alpha_0 + d - \gamma_0 = d,$$

which yields condition (6) of Proposition 3.24.

Similarly, Lemmas 3.45, 3.46 and 3.47, together with equations (3-7) and (3-9), imply that

$$F \bullet E_{1,1} + G \bullet E_{1,1} = d$$

and hence condition (7) of Proposition 3.24. This completes the proof of Proposition 3.24 in the present case.

Other cases The proofs in the other cases follow along identical lines. For the sake of completeness we list the inequalities for the components of the translations that lead to the desired intersection patterns of Proposition 3.24, in the remaining scenarios. For the case $\alpha_0 \geq \gamma_0$ and $\delta_0 \geq \beta_0$, we choose

$$a_1 < a_2 < 0, \quad b_2 < 0 < b_1, \quad 0 < c_2 < c_1, \quad d_1 < 0 < d_2.$$

For $\gamma_0 \geq \alpha_0 + 1$ and $\beta_0 \geq \delta_0 + 1$, we choose

$$a_1 < a_2 < 0, \quad b_1 < 0 < b_2, \quad 0 < c_2 < c_1, \quad d_2 < 0 < d_1.$$

Finally, for the case $\gamma_0 \geq \alpha_0 + 1$ and $\delta_0 \geq \beta_0$, we choose

$$a_1 < a_2 < 0, \quad b_1 < 0 < b_2, \quad 0 < c_2 < c_1, \quad d_1 < 0 < d_2.$$

To complete the proof of Proposition 3.24, we remark that the smoothings F and G can be replaced by smooth symplectic spheres without changing the various intersection numbers. To do this, it is enough

to replace F and G by symplectic spheres which coincide with F and G away from neighborhoods of $\mathbb{L}(\boldsymbol{v}_1)$ and $L_{1,1}(\boldsymbol{w}_1)$, respectively $\mathbb{L}(\boldsymbol{v}_2)$ and $L_{1,1}(\boldsymbol{w}_2)$; that is, the new spheres differ only away from all intersection points.

Now, we know that the asymptotic ends of the top level curves of $\boldsymbol{F}(\boldsymbol{x}_1)$ and $\boldsymbol{G}(\boldsymbol{x}_2)$ are simply covered, either because the curves are essential, or for covers of leaves by applying Lemma 3.23. Generically the asymptotic limits are distinct. Then, for small perturbations, we may assume that the top level curves restricted to a neighborhood of the Lagrangians are symplectically isotopic to the corresponding top level curves of our original buildings $\boldsymbol{F}$ and $\boldsymbol{G}$. (In the case of $\boldsymbol{F}(\boldsymbol{x}_1)$ the isotopy maps $\mathbb{L}(\boldsymbol{v}_1)$ and $L_{1,1}(\boldsymbol{w}_1)$ to $\mathbb{L}$ and $L_{1,1}$, respectively.) Finally, recall that the buildings $\boldsymbol{F}$ and $\boldsymbol{G}$ are limits of sequences of smooth embedded holomorphic spheres as our almost complex structures are stretched along the Lagrangians. Therefore, after a small perturbation, we may assume the top level curves of these buildings restricted to a compact subset of the complement of $\mathbb{L} \cup L_{1,1}$ extend to smooth symplectic spheres in $S^2 \times S^2$. Combining the isotopies and these extensions gives our symplectic spheres as required.

3.9 Scene change

Consider $(S^2 \times S^2, \pi_1^*\omega + \pi_2^*\omega)$ with our disjoint Lagrangian tori $\mathbb{L}$ and $L_{1,1}$ and the various symplectic spheres and disks constructed in Proposition 3.24: F, G, $\mathbb{E}$ and $E_{1,1}$.

To prepare for the proof of our main theorem, we specify our choice of almost complex structure. Near their various intersection points, the listed spheres and disks are already complex for a suitable almost complex structure. We can correct this almost complex structure, without perturbing F or G, to an almost complex structure J which is compatible with our symplectic form at the intersection points (not just tame) and extends to make our symplectic spheres and planes (the interiors of the disks) complex. Also we may assume J remains adapted to the parametrizations ψ and $\psi_{1,1}$ of $\mathbb{L}$ and $L_{1,1}$, respectively. As the spheres and planes from Proposition 3.24 were already holomorphic near the axes T_0, T_∞, S_0 and S_∞, and also near the broken planes in $\mathfrak{t}_0$, $\mathfrak{t}_\infty$, $\mathfrak{s}_0$ and $\mathfrak{s}_\infty$, we may assume that these curves all remain complex. In other words the only correction from the J used in Proposition 3.24 occurs near the intersection points to ensure compatibility, and near the regions where the F and G are symplectic but not complex.

Given this choice of J we have an associated foliation $\mathcal{F}$ and projection $p\colon S^2 \times S^2 \to S_\infty$. As the broken curves are the same as in Proposition 3.24, the subsets A_0, B and A_∞ of S_∞ are the same as in the proposition, and in particular property (9) continues to hold.

Let H be a sphere of $\mathcal{F}$ which is disjoint from $F \cap G$, and H_i for $1 \le i \le 2d$ the spheres of $\mathcal{F}$ intersecting the $2d$ points $\{p_1, \dots, p_{2d}\}$ of $F \cap G$. We note that these H_i are distinct since F and G both represent a homology class $(1, d)$ and so intersect fibers of $\mathcal{F}$ each in a single point.

One other small perturbation is required. We may choose Darboux charts about each p_i mapping an open set B_i to the round open ball about the 0 of capacity ϵ in $\mathbb{R}^4 = \mathbb{C}^2$, such that J is pushed forward to

	F	G	$\mathbb{E}$	$E_{1,1}$	H
F	$2d$				
G	$2d$	$2d$			
$\mathbb{E}$	k	$d-k$	$*$		
$E_{1,1}$	l	$d-l$	0	$*$	
H	1	1	$*$	$*$	0

	$\pi_1^*\omega+\pi_2^*\omega$–area
F	$2+2d$
G	$2+2d$
$\mathbb{E}$	1
$E_{1,1}$	1
H	2

Table 1: Initial intersection numbers, left, and initial symplectic areas, right.

the standard complex structure at 0 (this requires compatibility of J). We may assume these charts are disjoint from H. In these charts, F, G and H_i intersect the origin and are tangent to distinct complex planes. Making ϵ smaller if necessary, we are able to perturb our symplectic spheres so that they actually coincide with their tangent plane in the open chart. Finally we adjust J so that it is pushed forward to the standard structure on the whole ball, while F, G and H_i remain complex.

Given this, we proceed with the main proof. We start with the intersection pattern and area profile in Table 1.

For pairs of distinct curves the intersection numbers here just denote a signed count of intersection points with multiplicity. The self intersection number of closed spheres are defined as usual. The asterisks denote undefined quantities. The numbers come from the fact that F and G represent the class $(1,d)$, and from the properties listed in Proposition 3.24. The integers $0\le k\le d$ and $0\le l\le d$ are undetermined.

We now alter $(S^2\times S^2,\pi_1^*\omega+\pi_2^*\omega)$, away from $\mathbb{L}$ and $L_{1,1}$, to obtain a new ambient symplectic manifold in which the disjointness of these Lagrangians is a contradiction.

Step 1 Blow up the balls B_i of capacity ϵ around each of the $2d$ points p_i in $F\cap G$.

Denote the new manifold by (W,Ω_1). It follows from the analysis of the blow-up procedure from [15], see also Proposition 9.3.3 of [16], that (W,Ω_1) contains $2d$ exceptional divisors $\mathscr{E}_i$ each of area ϵ. Since the H_i intersect the balls B_i in J–holomorphic planes, (W,Ω_1) also contains the proper transforms of the H_i. These are denoted here by $\widehat{H}_i$ and are symplectic spheres of area $2-\epsilon$. By property (9) of Proposition 3.24, d of the $\widehat{H}_i$ intersect $\mathbb{E}$ once, and the other d of the $\widehat{H}_i$ intersect $E_{1,1}$ once.

Again, since they intersect the B_i in planes, the proper transforms of F and G, denoted by $\widehat{F}$ and $\widehat{G}$, are also well defined. These are spheres of area $2d+2-2d\epsilon$ which are now disjoint. The sphere H of $\overline{\mathscr{F}}$ is disjoint from the balls, but we denote its image in the blow up by $\widehat{H}$, which remains of area 2. After this, the relevant intersection numbers and areas are as in Table 2.

Step 2 Inflate both $\widehat{F}$ and $\widehat{G}$ by adding a tubular neighborhood of capacity d.

Here we recall that since $\widehat{F}$ and $\widehat{G}$ are symplectic spheres of self intersection 0 they have tubular neighborhoods which can be identified symplectically with $S^2\times D^2(\delta)$, where S^2 is a sphere of area

	$\widehat{F}$	$\widehat{G}$	$\mathbb{E}$	$E_{1,1}$	$\widehat{H}$	$\mathscr{E}_i$	$\widehat{H}_i$
$\widehat{F}$	0						
$\widehat{G}$	0	0					
$\mathbb{E}$	k	$d-k$	$*$				
$E_{1,1}$	l	$d-l$	0	$*$			
$\widehat{H}$	1	1	$*$	$*$	0		
$\{\mathscr{E}_i\}$	$2d$	$2d$	0	0	0	-1	
$\{\widehat{H}_i\}$	0	0	d	d	0	1	-1

	Ω_1–area
$\widehat{F}$	$2+2d-2d\epsilon$
$\widehat{G}$	$2+2d-2d\epsilon$
$\mathbb{E}$	1
$E_{1,1}$	1
$\widehat{H}$	2
$\mathscr{E}_i$	ϵ
$\widehat{H}_i$	$2-\epsilon$

Table 2: Intersection numbers after Step 1, left, and areas after Step 1, right.

$2+2d-2d\epsilon$ and $D^2(\delta)$ a disk of area δ. In this case inflation means replacing the symplectic form Ω_1 on this neighborhood by another one, Ω_2, such that $\Omega_2-\Omega_1$ is a compactly supported area form of total area d on the disk factor, $D^2(\delta)$.

Applying the inflation result from [12], we may assume, by Lemma 3.1 in [14], that J is also tame with respect to Ω_2. This means that all of our J–holomorphic curves which intersect $\widehat{F}$ and $\widehat{G}$, namely $\mathbb{E}$, $E_{1,1}$, $\widehat{H}$ and $\mathscr{E}_i$, remain J–holomorphic and, in particular, symplectic.

The inflation procedure does not change the intersection pattern, and the Ω_2–area of curves increases, from the previous step, by d times the sum of the intersection numbers with $\widehat{F}$ and $\widehat{G}$ leaving us with the area profile in Table 3, left.

Step 3 Apply the negative inflation procedure from [2], of size ϵ, to each $\mathscr{E}_i$.

This negative inflation procedure yields a new symplectic form, Ω_3, such that the Ω_3–area of each $\mathscr{E}_i$ is less than its Ω_2–area by ϵ. That is, the Ω_3–area of each $\mathscr{E}_i$ is $2d$. One way to visualize this is to blow-down the $\mathscr{E}_i$ giving balls of capacity $\epsilon+2d$ and then blow-up slightly smaller balls of capacity $2d$.

Negative inflation by ϵ also increases the area of homology classes by ϵ times the sum of the intersection numbers with the $\mathscr{E}_i$. The Ω_3 area profile is given in Table 3, right.

	Ω_2–area
$\widehat{F}$	$2+2d-2d\epsilon$
$\widehat{G}$	$2+2d-2d\epsilon$
$\mathbb{E}$	$1+d^2$
$E_{1,1}$	$1+d^2$
$\widehat{H}$	$2+2d$
$\mathscr{E}_i$	$\epsilon+2d$
$\widehat{H}_i$	$2-\epsilon$

	Ω_3–area
$\widehat{F}$	$2+2d$
$\widehat{G}$	$2+2d$
$\mathbb{E}$	$1+d^2$
$E_{1,1}$	$1+d^2$
$\widehat{H}$	$2+2d$
$\mathscr{E}_i$	$2d$
$\widehat{H}_i$	2

Table 3

	$\widehat{F}$	$\widehat{G}$	$\mathbb{E}^X$	$E^X_{1,1}$	$\widehat{H}$	$\mathscr{H}_i$
$\widehat{F}$	0					
$\widehat{G}$	0	0				
$\mathbb{E}^X$	k	$d-k$	$*$			
$E^X_{1,1}$	l	$d-l$	0	$*$		
$\widehat{H}$	1	1	$*$	$*$	0	
$\{\mathscr{H}_i\}$	$2d$	$2d$	d	d	0	0

	Ω–area
$\widehat{F}$	$2+2d$
$\widehat{G}$	$2+2d$
$\mathbb{E}^X$	$1+d^2+2d$
$E^X_{1,1}$	$1+d^2+2d$
$\widehat{H}$	$2+2d$
$\mathscr{H}_i$	$2+2d$

Table 4: Intersection numbers after Step 4, left, and areas after Step 4, right.

Step 4 Blow down each $\widehat{H}_i$.

We denote the symplectic manifold resulting from this final step by (X,Ω). Each of the exceptional divisors $\mathscr{E}_i$ in (W,Ω_3) is transformed, by Step 4, into a sphere $\mathscr{H}_i$ in X which has Ω–area equal to $2d+2$ and now lies in the same class as $\widehat{H}$. The disks $\mathbb{E}$ and $E_{1,1}$ each intersect d of the $\widehat{H}_i$ and so are transformed by Step 4 into disks $\mathbb{E}^X$ and $E^X_{1,1}$, whose symplectic areas have each been increased by $2d$. See Table 4.

Lemma 3.49 (X,Ω) *is symplectomorphic to*

$$(S^2\times S^2,(d+1)\omega\oplus(d+1)\omega).$$

Proof The presence of the embedded symplectic spheres $\widehat{F}$ and $\widehat{H}$, with the same Ω–area and satisfying

$$\widehat{F}\bullet\widehat{F}=\widehat{H}\bullet\widehat{H}=0 \quad\text{and}\quad \widehat{F}\bullet\widehat{H}=1,$$

implies that either (X,ω) is symplectomorphic to

$$(S^2\times S^2,(d+1)\omega\oplus(d+1)\omega),$$

or there are finitely many symplectically embedded spheres with self-intersection number -1 in the complement of $\widehat{F}$ and $\widehat{H}$ in X, and X can be blown down to a copy of $S^2\times S^2$. This follows from the proof of Theorem 9.4.7 of [16]. As a consequence, if $H_2(X;\mathbb{Z})$ has rank 2 then X is symplectomorphic to $S^2\times S^2$.

A simple analysis of the construction of (X,Ω) from $(S^2\times S^2,\pi_1^*\omega+\pi_2^*\omega)$ allows us to compute this rank. The $2d$ blow ups in Step 1 imply that the rank of $H_2(W;\mathbb{Z})$ is $2+2d$. The subsequent $2d$ blow down operations in Step 4 imply that the rank of $H_2(X;\mathbb{Z})$ is 2, as required. □

Henceforth, we may identify (X,Ω) with $(S^2\times S^2,(d+1)\omega\oplus(d+1)\omega)$. The Lagrangian tori $\mathbb{L}$ and $L_{1,1}$ are untouched, as submanifolds, by the four steps above. They remain Lagrangian and disjoint in (X,Ω). Note that $L_{1,1}$ is not equal to the Clifford torus in (X,Ω) with respect to the identification above. In what follows we denote the Clifford torus in (X,Ω) by L_X.

The manifold (X,Ω) also inherits an almost complex structure, denoted here by $\widehat{J}$, which equals J away from the collection $\{\widehat{\mathcal{H}}_i\}$. In particular, $\widehat{J}$ is adapted to the original parametrizations ψ and $\psi_{1,1}$ of L and $L_{1,1}$. As in Section 3.5, $\widehat{J}$ determines a straightened foliation $\widehat{\mathcal{F}} = \mathcal{F}(\mathbb{L}, L_{1,1}, \psi, \psi_{1,1}, \widehat{J})$ of $X \setminus (\mathbb{L} \cup L_{1,1})$. The original collections of planes $\mathfrak{s}_0$, $\mathfrak{s}_\infty$, $\mathfrak{r}_0$ and $\mathfrak{r}_\infty$ still comprise the broken leaves of this new foliation. The symplectic spheres $\widehat{F}$ and $\widehat{G}$ now represent the class $(1,0) \in H_2(X;\mathbb{Z}) = H_2(S^2\times S^2;\mathbb{Z})$. As in Proposition 3.24, it is still true that exactly one of $\widehat{F}$ and $\widehat{G}$ intersects the planes of $\mathfrak{s}_0$ and the other intersects the planes of $\mathfrak{s}_\infty$, and exactly one of $\widehat{F}$ and $\widehat{G}$ intersects the planes of $\mathfrak{r}_0$ and the other intersects the planes of $\mathfrak{r}_\infty$.

Lemma 3.50 *The Lagrangian tori $\mathbb{L}$ and $L_{1,1}$ are both monotone in (X,Ω).*

Proof Let $D_\infty : (D^2, S^1) \to (S^2\times S^2, \mathbb{L})$ be a compactification of one of the planes of $\mathfrak{r}_\infty$. The disk D_∞ has Maslov index equal to 2 and symplectic area equal to 1 with respect to $\pi_1^*\omega + \pi_2^*\omega$. The map $D_\infty|_{S^1}$ represents the foliation class $\beta_{\mathbb{L}}$. The image of the map D_∞ is unaffected by the four steps defining the passage from $(S^2\times S^2, \pi_1^*\omega+\pi_2^*\omega)$ to (X,Ω). Viewed as a map from (D^2,S^1) to $(X,\mathbb{L})$, D_∞ still has Maslov index 2, and $D_\infty|_{S^1}$ still represents $\beta_{\mathbb{L}}$. The Ω–area of D_∞, as a map into (X,Ω), is $d+1$. This follows from the fact that exactly one of F and G intersect D_∞ and so the inflations in Step 2 increase the symplectic area by d.

By assertion (4) of Proposition 3.24, the boundary $\mathbb{E}|_{S^1}$ represents a class which, together with $\beta_{\mathbb{L}}$, forms an integral basis of $H_1(\mathbb{L};\mathbb{Z})$. The same holds for $\mathbb{E}^X|_{S^1}$. To prove that $\mathbb{L}$ is a monotone Lagrangian torus in (X,Ω) it then suffices, by Lemma 3.1, to prove that the Maslov index of $\mathbb{E}^X : (D^2,S^1) \to (X,\mathbb{L})$ is equal to

$$\frac{2}{d+1}(1+d^2+2d) = 2d+2,$$

where $1+d^2+2d$ is the area of $\mathbb{E}^X$. This follows from the fact that, in (W,Ω_3), $\mathbb{E}$ has Maslov index 2, intersects exactly d of the $\widehat{H}_i$, and each of the corresponding intersection numbers is 1. In blowing down the $\widehat{H}_i$, and passing from $\mathbb{E}$ to $\mathbb{E}^X$, each of these intersection points yields an increase of 2 in the Maslov index.

The proof that $L_{1,1}$ is monotone in (X,Ω) is identical. □

Lemma 3.51 *The Lagrangians $\mathbb{L}$ and $L_{1,1}$ are both Hamiltonian isotopic to the Clifford torus L_X in (X,Ω).*

Proof This follows from the main result of Cieliebak and Schwingenheuer in [4]. In the language of that paper, the compactification of the straightened foliation $\widehat{\mathcal{F}} = \mathcal{F}(\mathbb{L}, L_{1,1}, \psi, \psi_{1,1}, \widehat{J})$ yields a fibering of $\mathbb{L}$ and a fibering of $L_{1,1}$. For the fibering of $\mathbb{L}$, the spheres $\widehat{F}$ and $\widehat{G}$ are disjoint sections in the class $(1,0)$ and exactly one of them intersects the (compactification of the) planes of $\mathfrak{r}_0$ and the other intersects the those of $\mathfrak{r}_\infty$. The main theorem of [4] then implies that L is Hamiltonian isotopic to the Clifford torus L_X in (X,Ω). An identical argument holds for $L_{1,1}$. □

With this, the contradiction to Assumption 2 becomes apparent. Since the Lagrangian Floer homology of L_X is nontrivial by [17], any Lagrangian tori Hamiltonian isotopic to L_X must intersect nontrivially. Hence, $\mathbb{L}$ and $L_{1,1}$ cannot be disjoint in (X, Ω).

Remark 3.52 The assumption that $\mathbb{L}$ and $L_{1,1}$ are disjoint is used twice in the proof of Theorem 1.1: at the very end, and in the proof of Refinement 3 in Section 3.3.

Remark 3.53 The fact that $L_{1,1}$ is the Clifford torus (and not just another monotone Lagrangian torus) is crucial (only) in the proof of the existence results in Propositions 3.20 and 3.22.

Remark 3.54 There is an alternative to the argument used at the end of the proof of Theorem 1.1 that avoids appealing to Lagrangian Floer homology. Instead, one can use the fact that the symplectomorphism in Lemma 3.49 can be chosen to map $\widehat{F}$, $\widehat{G}$ and the transforms $\widehat{T}_0$ and $\widehat{T}_\infty$ to the axes $S^2\times\{0\}$, $S^2\times\{\infty\}$, $\{0\}\times S^2$ and $\{\infty\}\times S^2$, respectively. The complement of these axes in $S^2\times S^2$ can be identified with a domain in T^*T^2 in which the Clifford torus is identified with the zero section. We can check that $\mathbb{L}$ and $L_{1,1}$ are homologically nontrivial in this copy of T^*T^2 and so, by Theorem 3.9, are Hamiltonian isotopic to constant sections. The monotonicity condition then implies the constant section must be the zero section. Finally Gromov's intersection theorem for exact Lagrangians in cotangent bundles, from Section 2.3.B_4'' of [6], implies that they must intersect.

4 Proof of Theorem 1.2

It suffices to prove the following.

Theorem 4.1 *For any $\epsilon > 0$, there is a $\delta > 0$ and a symplectic embedding of the polydisk $P(1+\delta, 1+\delta)$ into $P(2+\epsilon, 2+\epsilon)$ whose image is disjoint from the product Lagrangians $L_{k,l}$ for $k, l \in \{1, 2\}$.*

The desired additional integral Lagrangian torus L^+ is the one on (the image of) the boundary of $P(1,1) \subset P(1+\delta, 1+\delta)$.

4.1 Proof of Theorem 4.1

We will use rescaled polar coordinates θ_i, R_i on $\mathbb{R}^4 = \mathbb{C}^2$, where $R_i = \pi|z_i|^2$ and $\theta_i \in \mathbb{R}/\mathbb{Z}$. In these coordinates the standard symplectic form is

$$\omega = \sum_{i=1}^{2} dR_i \wedge d\theta_i,$$

and $L_{k,l} = \{(\theta_1, k, \theta_2, l)\}$.

4.1.1 A polydisk For $\epsilon > 0$ fixed, choose positive numbers ℓ, w such that

$$2 < \ell < 2+\epsilon, \quad w < 2, \quad \frac{1}{\ell} + \frac{1}{w} < 1.$$

Then choose positive constants σ and δ such that

$$\ell+\sigma<2+\epsilon, \quad w+\sigma<2, \quad \frac{1+\delta}{\ell}+\frac{1+\delta}{w}<1.$$

Set

$$S=\{\sigma<R_1<\ell+\sigma, \sigma<R_2<w+\sigma\} \quad \text{and} \quad T=\left\{0<\theta_1<\frac{1+\delta}{\ell}, 0<\theta_2<\frac{1+\delta}{w}\right\}.$$

Note that $S\times T$ is a subset of $P(2+\epsilon,2+\epsilon)$ and is symplectomorphic to $P(1+\delta,1+\delta)$. Both $L_{1,1}$ and $L_{2,1}$ intersect $S\times T$, while $L_{1,2}$ and $L_{2,2}$ do not.

4.1.2 The plan To prove Theorem 4.1 it suffices to find a Hamiltonian diffeomorphism of $P(2+\epsilon,2+\epsilon)$ that displaces $S\times T$ from the $L_{k,l}$. Equivalently, we construct a Hamiltonian diffeomorphism Ψ of $P(2+\epsilon,2+\epsilon)$ such that each of the images $\Psi(L_{k,l})$ is disjoint from $S\times T$.

To construct Ψ we use Hamiltonian functions which are of the form $F(\theta_1,\theta_2)$. The Hamiltonian flow ϕ_F^t of such a function preserves θ_1 and θ_2 and generates a Hamiltonian vector field parallel to the R_1R_2–plane. In particular, the only points of $\phi_F^t(L_{k,l})$ which could possibly intersect $S\times T$ are those whose (θ_1,θ_2) coordinates lie in T.

Since we only need to control the images of the $L_{k,l}$, we can cut off autonomous functions like F in (moving) neighborhoods of $\phi_F^t(L_{k,l})$ for specific values of k and l. After this cutting off, the new Hamiltonian will depend on all variables and be time dependent. In general, for a closed subset V, we denote the function obtained by cutting of F along $\phi_F^t(V)$ by $F_{[V]}$. Note that

$$\phi_{F_{[V]}}^t(v)=\phi_F^t(v) \quad \text{for all } v\in V \text{ and } t\in[0,1].$$

Also, each map $\phi_{F_{[V]}}^t$ is equal to the identity away from an arbitrarily small neighborhood of

$$\bigcup_{t\in[0,1]}\phi_F^t(V).$$

4.1.3 A diagonal move Let $g:\mathbb{R}/\mathbb{Z}\to\mathbb{R}$ be a smooth function such that for some positive real number $c(g)>0$ we have

$$g'(s)=c(g) \quad \text{for } s\in\left[0,\frac{1+\delta}{\ell}+\frac{1+\delta}{w}\right],$$

$\max(g')=c(g)$, and $\min(g')$ is less than and arbitrarily close to

$$-c(g)\left(\frac{\frac{1+\delta}{\ell}+\frac{1+\delta}{w}}{1-\frac{1+\delta}{\ell}-\frac{1+\delta}{w}}\right).$$

Letting $G(\theta_1,\theta_2)=g(\theta_1+\theta_2)$, we have

$$\phi_G^t(\theta_1,R_1,\theta_2,R_2)=(\theta_1,R_1+tg'(\theta_1+\theta_2),\theta_2,R_2+tg'(\theta_1+\theta_2)). \tag{4-1}$$

The image $\phi_G^1(L_{1,2})$ is well defined as long as

$$c(g) < \frac{1 - \frac{1+\delta}{\ell} - \frac{1+\delta}{w}}{\frac{1+\delta}{\ell} + \frac{1+\delta}{w}}, \tag{4-2}$$

and is contained in $P(2+\epsilon, 2+\epsilon)$ as long as $c(g) < \epsilon$. Henceforth, we will assume that ℓ, w and δ have been chosen such that the first constraint on $c(g)$ implies the second.

It follows from (4-1) and (4-2) that $\phi_G^t(L_{1,2})$ is contained in

$$\{R_1 \leq 1 + c(g)\} \cap \{R_2 > 1\}$$

for all $t \in [0,1]$. Hence, each image $\phi_G^t(L_{1,2})$ is disjoint from the other $L_{k,l}$. Since $g' = c(g) > 0$ on T, each $\phi_G^t(L_{1,2})$ is also disjoint from $S \times T$.

4.1.4 A vertical move Let $h\colon \mathbb{R}/\mathbb{Z} \to \mathbb{R}$ be a smooth function such that for some positive real number $0 < c(h) < \sigma$ we have

$$h'(s) = -(1 - c(h)) \quad \text{for } s \in \left[0, \frac{1+\delta}{w}\right],$$

$\min(h') = -(1 - c(h))$ and $\max(h')$ is greater than and arbitrarily close to

$$\frac{(1 - c(h))\frac{1+\delta}{w}}{1 - \frac{1+\delta}{w}} = \frac{1 - c(h)}{\frac{w}{1+\delta} - 1},$$

which is greater than one since $w + \sigma < 2$ and $c(h) < \sigma$.

Letting $H(\theta_1, \theta_2) = h(\theta_2)$, we have

$$\phi_H^t(\theta_1, R_1, \theta_2, R_2) = (\theta_1, R_1, \theta_2, R_2 + t h'(\theta_2)). \tag{4-3}$$

Clearly, $L_{2,1}$ and $L_{2,2}$ are disjoint from $\phi_H^t(L_{1,1})$ for all $t \in [0,1]$. Moreover, for θ_2 in $[0, (1+\delta)/w]$ we have

$$\phi_H^1(\theta_1, 1, \theta_2, 1) = (\theta_1, 1, \theta_2, c(h)).$$

So $\phi_H^1(L_{1,1})$ is disjoint from $T \times S$ by our choice of $c(h)$.

Some points of $L_{1,1}$, with values of θ_2 in $((1+\delta)/w, 1)$, are mapped by ϕ_H^1 to points having R_2 coordinate greater than and arbitrarily close to

$$1 + \frac{1 - c(h)}{\frac{w}{1+\delta} - 1} > 2.$$

Choosing w sufficiently close to 2 and δ sufficiently small ensures that $\phi_H^1(L_{1,1})$ lies in $P(2+\epsilon, 2+\epsilon)$.

4.1.5 A time delay The Hamiltonian diffeomorphism $\phi^1_{H_{[L_{1,1}]}}$ cannot be used to move $L_{1,1}$ off of $S \times T$ while leaving $L_{1,2}$ undisturbed. For, as described in the discussion above, $\phi^1_{H_{[L_{1,1}]}}(L_{1,2})$ will intersect $S \times T$.

The Hamiltonian diffeomorphism

$$\phi^1_{H_{[L_{1,1}]}} \circ \phi^1_{G_{[L_{1,2}]}}$$

has the same problem. By (4-1) and (4-3), the image of $(\theta_1, 1, \theta_2, 1) \in L_{1,1}$ under ϕ^t_H belongs to $\phi^1_G(L_{1,2})$ if and only if $g'(\theta_1+\theta_2) = 0$ and $th'(\theta_2) = 1$. Since $\max(h') > 1$, these intersections occur and so the map above will again push $L_{1,2}$ into $S \times T$.

We can fix this by adding a time delay. The first intersection between $\phi^t_H(L_{1,1})$ and $\phi^1_G(L_{1,2})$ occurs at $t = (\max(h'))^{-1}$. Let τ be less than and arbitrarily close to $(\max(h'))^{-1}$. Hence, τ is also less than and arbitrarily close to

$$\frac{\frac{w}{1+\delta} - 1}{1 - c(h)}.$$

Consider the Hamiltonian diffeomorphism

$$\widetilde{\Psi} = \phi^{1-\tau}_{H_{[\phi^\tau_H(L_{1,1}) \cup \phi^1_G(L_{1,2})]}} \circ \phi^\tau_{H_{[L_{1,1}]}} \circ \phi^1_{G_{[L_{1,2}]}}.$$

It follows from the analysis above that the map $\widetilde{\Psi}$ is compactly supported in $P(2+\epsilon, 2+\epsilon)$. In fact, it is supported in an arbitrarily small neighborhood of the subset $\{R_1 \le 1 + c(g)\}$. Hence, $\widetilde{\Psi}(L_{2,1}) = L_{2,1}$ and $\widetilde{\Psi}(L_{2,2}) = L_{2,2}$. By the definitions of τ and the cut-off operation, we have $\widetilde{\Psi}(L_{1,1}) = \phi^1_H(L_{1,1})$ and thus $\widetilde{\Psi}(L_{1,1})$ is disjoint from $S \times T$. In addition, we now have the following.

Lemma 4.2 *The image $\widetilde{\Psi}(L_{1,2})$ is disjoint from $S \times T$ when $c(h)$ is sufficiently close to σ and δ is sufficiently small.*

Proof By construction, for $(\theta_1, \theta_2) \in T$ we have

$$\begin{aligned}\widetilde{\Psi}(\theta_1, 1, \theta_2, 2) &= (\theta_1, 1 + g'(\theta_1+\theta_2), \theta_2, 2 + g'(\theta_1+\theta_2) + (1-\tau)h'(\theta_2)) \\ &= \big(\theta_1, 1 + c(g), \theta_2, 2 + c(g) - (1-\tau)(1-c(h))\big).\end{aligned}$$

It suffices to show that we can choose $c(g)$ and $c(h)$ so that

$$2 + c(g) - (1-\tau)(1-c(h)) > w + \sigma. \tag{4-4}$$

Since τ is less than and arbitrarily close to

$$\frac{\frac{w}{1+\delta} - 1}{1 - c(h)},$$

is also suffices to show that we can choose $c(g)$ and $c(h)$ so that

$$c(g) > w\Big(1 - \frac{1}{1+\delta}\Big) + (\sigma - c(h)).$$

The right-hand side can be made arbitrarily small by taking $c(h)$ to be close to σ and δ to be small. Since the choice of $c(g)$ is independent of the choice of $c(h)$ and the constraint (4-2) on $c(g)$ relaxes as δ goes to zero, we are done. □

Henceforth, we will assume that the conditions of Lemma 4.2 hold.

4.1.6 A final (horizontal) adjustment The images $\widetilde{\Psi}(L_{1,1})$, $\widetilde{\Psi}(L_{1,2})$ and $\widetilde{\Psi}(L_{2,2})$ are disjoint from $S\times T$ but $\widetilde{\Psi}$ still fixes $L_{1,2}$, which intersects $S\times T$. Since $L_{1,2}$ is close to the boundary of $S\times T$, we can make a simple adjustment to obtain the desired map, Ψ, which moves $L_{1,2}$ off of $S\times T$ as well.

Let $f:\mathbb{R}/\mathbb{Z}\to\mathbb{R}$ be a smooth function such that for some positive real number $c(f)$ greater than and arbitrarily close to $\ell+\sigma-2$ we have

$$f'(s)=c(f)\quad \text{for } s\in\Big[0,\frac{1+\delta}{\ell}\Big],$$

$\max(f')=c(f)$ and $\min(f')$ is less than and arbitrarily close to

$$-\frac{c(f)}{\dfrac{\ell}{1+\delta}-1}.$$

Setting $F(\theta_1,\theta_2)=f(\theta_1)$, we have

$$\phi_F^t(\theta_1,R_1,\theta_2,R_2)=(\theta_1,R_1+tf'(\theta_1),\theta_2,R_2).$$

Our lower bound for $c(f)$ implies that $\phi_F^1(L_{2,1})$ is disjoint from $S\times T$. Looking at the R_2–component, it is clear that $\phi_F^1(L_{2,1})$ is disjoint from $L_{2,2}=\widetilde{\Psi}(L_{2,2})$. To prove that $\phi_F^1(L_{2,1})$ is also disjoint from $\widetilde{\Psi}(L_{1,1})$ and $\widetilde{\Psi}(L_{1,2})$, it suffices to prove the following.

Lemma 4.3 *The sets $\{R_1\le 1+c(g)\}$ and $\phi_F^1(L_{2,1})$ are disjoint.*

Proof It suffices to prove that

$$2-\frac{c(f)}{\dfrac{\ell}{1+\delta}-1}>1+c(g)$$

or, even more, that

$$1>c(g)+\frac{\ell+\sigma-2}{\dfrac{\ell}{1+\delta}-1}.$$

The latter inequality clearly holds for all sufficiently small values of $c(g)$ and $\ell+\sigma-2$. □

The Hamiltonian diffeomorphism

$$\Psi=\phi^1_{F_{[L_{2,1}]}}\circ\phi^{1-\tau}_{H_{[\phi^\tau_H(L_{1,1})\cup\phi^1_G(L_{1,2})]}}\circ\phi^\tau_{H_{[L_{1,1}]}}\circ\phi^1_{G_{[L_{1,2}]}}$$

now has all the desired properties. With its construction, the proof of Theorem 4.1 is complete.

Question 4.4 *Can Ψ, or any other Hamiltonian diffeomorphism which displaces the $L_{k,l}$ from $S\times T$, be generated by an autonomous Hamiltonian*?

References

[1] **F Bourgeois**, **Y Eliashberg**, **H Hofer**, **K Wysocki**, **E Zehnder**, *Compactness results in symplectic field theory*, Geom. Topol. 7 (2003) 799–888 MR Zbl

[2] **O Buşe**, *Negative inflation and stability in symplectomorphism groups of ruled surfaces*, J. Symplectic Geom. 9 (2011) 147–160 MR Zbl

[3] **K Cieliebak**, **K Mohnke**, *Punctured holomorphic curves and Lagrangian embeddings*, Invent. Math. 212 (2018) 213–295 MR Zbl

[4] **K Cieliebak**, **M Schwingenheuer**, *Hamiltonian unknottedness of certain monotone Lagrangian tori in* $S^2 \times S^2$, Pacific J. Math. 299 (2019) 427–468 MR Zbl

[5] **G Dimitroglou Rizell**, **E Goodman**, **A Ivrii**, *Lagrangian isotopy of tori in* $S^2 \times S^2$ *and* $\mathbb{C}P^2$, Geom. Funct. Anal. 26 (2016) 1297–1358 MR Zbl

[6] **M Gromov**, *Pseudo holomorphic curves in symplectic manifolds*, Invent. Math. 82 (1985) 307–347 MR Zbl

[7] **J Hicks**, **C Y Mak**, *Some cute applications of Lagrangian cobordisms towards examples in quantitative symplectic geometry*, preprint (2022) arXiv 2208.14498

[8] **R Hind**, **S Lisi**, *Symplectic embeddings of polydisks*, Selecta Math. 21 (2015) 1099–1120 MR Zbl Correction in 23 (2017) 813–815

[9] **R Hind**, **E Opshtein**, *Squeezing Lagrangian tori in dimension* 4, Comment. Math. Helv. 95 (2020) 535–567 MR Zbl

[10] **H Hofer**, **K Wysocki**, **E Zehnder**, *Properties of pseudoholomorphic curves in symplectisations, I: Asymptotics*, Ann. Inst. H. Poincaré C Anal. Non Linéaire 13 (1996) 337–379 MR Zbl

[11] **A Ivrii**, *Lagrangian unknottedness of tori in certain symplectic* 4*–manifolds*, PhD thesis, Stanford University (2003) Available at `https://www.proquest.com/docview/305300870`

[12] **F Lalonde**, **D McDuff**, *J–curves and the classification of rational and ruled symplectic* 4*–manifolds*, from "Contact and symplectic geometry" (C B Thomas, editor), Publ. Newton Inst. 8, Cambridge Univ. Press (1996) 3–42 MR Zbl

[13] **C Y Mak**, **I Smith**, *Non-displaceable Lagrangian links in four-manifolds*, Geom. Funct. Anal. 31 (2021) 438–481 MR Zbl

[14] **D McDuff**, *Symplectomorphism groups and almost complex structures*, from "Essays on geometry and related topics, II" (É Ghys, P de la Harpe, V F R Jones, V Sergiescu, T Tsuboi, editors), Monogr. Enseign. Math. 38, Enseign. Math., Geneva (2001) 527–556 MR Zbl

[15] **D McDuff**, **L Polterovich**, *Symplectic packings and algebraic geometry*, Invent. Math. 115 (1994) 405–434 MR Zbl

[16] **D McDuff**, **D Salamon**, *J–holomorphic curves and symplectic topology*, Amer. Math. Soc. Colloq. Publ. 52, Amer. Math. Soc., Providence, RI (2004) MR Zbl

[17] **Y-G Oh**, *Addendum to: "Floer cohomology of Lagrangian intersections and pseudo-holomorphic disks, I"*, Comm. Pure Appl. Math. 48 (1995) 1299–1302 MR Zbl

[18] **L Polterovich**, **E Shelukhin**, *Lagrangian configurations and Hamiltonian maps*, Compos. Math. 159 (2023) 2483–2520 MR Zbl

[19] **A F Ritter**, **I Smith**, *The monotone wrapped Fukaya category and the open-closed string map*, Selecta Math. 23 (2017) 533–642 MR Zbl

[20] **R Vianna**, *Infinitely many monotone Lagrangian tori in del Pezzo surfaces*, Selecta Math. 23 (2017) 1955–1996 MR Zbl

[21] **C Wendl**, *Automatic transversality and orbifolds of punctured holomorphic curves in dimension four*, Comment. Math. Helv. 85 (2010) 347–407 MR Zbl

Department of Mathematics, University of Notre Dame
Notre Dame, IN, United States

Department of Mathematics, University of Illinois at Urbana-Champaign
Urbana, IL, United States

hind.1@nd.edu, ekerman@illinois.edu

Proposed: Leonid Polterovich
Seconded: Dmitri Burago, Yakov Eliashberg

Received: 15 February 2022
Revised: 27 October 2022

msp

Geometry & Topology **28**:5 (2024) 2259–2311
DOI: 10.2140/gt.2024.28.2259
Published: 24 August 2024

The parabolic Verlinde formula: iterated residues and wall-crossings

ANDRÁS SZENES
OLGA TRAPEZNIKOVA

We give a new proof for the parabolic Verlinde formula in all ranks based on a comparison of wall-crossings in geometric invariant theory and certain iterated residue functionals. On the way, we develop a tautological variant of Hecke correspondences, calculate the Hilbert polynomials of the moduli spaces, and present a new, transparent, local approach to the ρ–shift problem of the theory.

14D20, 14H60

1 Introduction

1.1 The Verlinde formula

The Verlinde formula [31] is a strikingly beautiful statement in enumerative geometry motivated by quantum physics. Our focus in this paper will be the more difficult, parabolic variant, which we briefly describe below.

Let C be a smooth, complex projective curve of genus $g \geq 1$, and fix an auxiliary point $p \in C$. We will call a vector $c = (c_1 > c_2 > \cdots > c_r) \in \mathbb{R}^r$ satisfying $\sum c_i = 0$ and $c_1 - c_r < 1$ *regular* if no nontrivial subset of its coordinates sums to an integer. For such a $c \in \mathbb{R}^r$, there exists a smooth projective moduli space $P_0(c)$ (see Seshadri [21], Mehta and Seshadri [15] and Bhosle [4]), whose points are in one-to-one correspondence with the equivalence classes of pairs (W, F_*), where $W \to C$ is a vector bundle of

rank r on C with trivial determinant, F_* is a full flag of the fiber W_p, and the pair satisfies a certain parabolic stability condition depending on c; see Section 2.1. This condition roughly states that for a proper subbundle $W' \subset W$, the degree $\deg(W')$ is strictly smaller than the sum of a subset of the coordinates of c depending on the position of W'_p with respect to F_*.

There is a natural way to associate to a positive integer k and an integer vector $\lambda \in \mathbb{Z}^r$ satisfying $\lambda_1 + \cdots + \lambda_r = 0$ a line bundle $\mathscr{L}(k;\lambda)$ on $P_0(c)$, in such a way that if $c = \lambda/k$, then $\mathscr{L}(k;\lambda)$ is ample. The *parabolic Verlinde formula* is the following expression for the Euler characteristic of the ample line bundle $\mathscr{L}(k;\lambda)$: assume $c = \lambda/k$ is regular; then

$$\chi(P_0(c), \mathscr{L}(k;\lambda)) = N_{r,k} \cdot \sum \frac{(-i)^{\binom{r}{2}} \exp(2\pi i \widehat{\lambda} \cdot x)}{\prod_{i<j} (2 \sin \pi (x_i - x_j))^{2g-1}}, \tag{1}$$

where $N_{r,k} = r(r(k+r)^{r-1})^{g-1}$, $\widehat{\lambda} = \lambda + \frac{1}{2}(r-1, r-3, \ldots, 1-r)$, and the sum is taken over the finite set of those points in the interior of the parallelopiped

$$\{x = (x_1, x_2, \ldots, x_r = 0) \mid 0 < x_i - x_{i+1} < 1 \text{ for } i = 1, \ldots, r-1\},$$

which satisfy the conditions

- $(k+r)x \in \mathbb{Z}^r$, and
- $x_i - x_j \notin \mathbb{Z}$ for $1 \le i < j < r$.

Remark 1.1 This finite set is a set of lattice points in the interior of $(r-1)!$ identical simplices. (These are the orange-colored points in the rhombus on Figure 1.) By symmetrizing with respect to the group of permutations of the r coordinates, one obtains the same function on each of these simplices. Using the Weyl character formula, this allows one to rewrite (1) in a more familiar form as

$$\chi(P_0(c), \mathscr{L}(k;\lambda)) = (r(k+r)^{r-1})^{g-1} \cdot \sum \frac{\chi_\lambda(x)}{\prod_{i<j} (2 \sin \pi (x_i - x_j))^{2g-2}},$$

where χ_λ is the character of the irreducible $\mathrm{SU}(r)$–representation of highest weight λ, and the sum is now taken over the lattice points of the form $(k+r)x \in \mathbb{Z}^r$ in the interior of a single simplex $\{x = (1 > x_1 > x_2 > \cdots > x_{r-1} > x_r = 0)\}$.

Remark 1.2 Equality (1) remains valid in greater generality, for certain cases when λ/k is nonregular. This slightly more technical statement will be given in Theorems 4.7 and 4.8.

Notation We denote the discrete sum in (1) depending on k, λ, r and g by $\mathrm{Ver}(k,\lambda)$. In what follows, the shift $\frac{1}{2}(r-1, r-3, \ldots, 1-r)$ will be denoted by ρ, and thus we have $\widehat{\lambda} = \lambda + \rho$.

Equality (1), the *parabolic Verlinde formula*, has attracted a lot of attention over the years, and there are a number of different proofs. There is a generalization of this formula associated to a simply connected compact Lie group, and the form presented here corresponds to the case of the group $\mathrm{SU}(r)$.

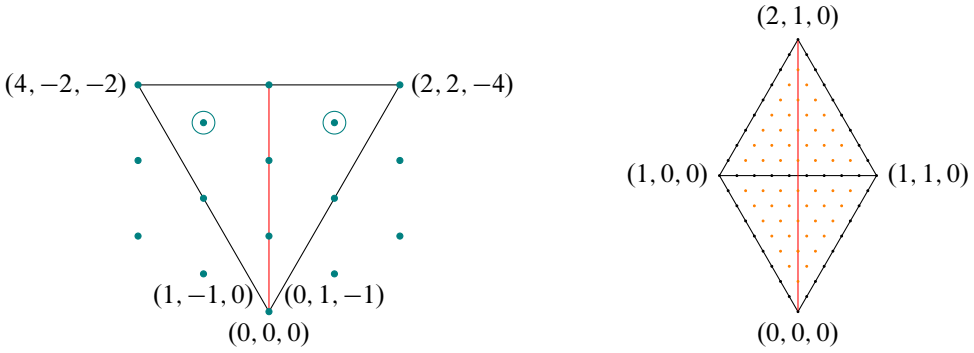


Figure 1: The set of λ vectors, left, and the finite set from (1), right, for $k = 6$, $r = 3$.

In this article, we give a novel proof of this result, which stands out with its technical simplicity. We believe the methods and ideas described in the paper will have other applications in geometric invariant theory and the study of moduli spaces.

Below, we give a quick sketch of the strategy of the proof, treating the example of the case of rank 3 in Sections 1.2–1.4. Next, in Section 1.5, we give a short guide to the contents of the paper.

Our work has a close relationship with several earlier approaches, and we describe these links in Section 1.6

Acknowledgements The authors gratefully acknowledge the help and insights of Michael Thaddeus at several stages of this work, as well as the advice and encouragement of Tamás Hausel and Michèle Vergne. We had useful discussions with Frances Kirwan, Eckhard Meinrenken, Gábor Tardos and Chris Woodward. This research was supported by SNF grant 175799, and the NCCR SwissMAP.

1.2 The residue formula

The proof is based on 3 ideas. We will follow the arguments below for the case $r = 3$. We thus fix an integer $k > 1$ and an integer vector $\lambda = (\lambda_1 > \lambda_2 > \lambda_3)$ such that $\lambda_1 + \lambda_2 + \lambda_3 = 0$ and $\lambda_1 - \lambda_r < k$.

We start with the study of the right-hand side of (1), which, for $r = 3$, may be written in the somewhat simplified form

$$\mathrm{Ver}(k,\lambda) = N_{3,k} \sum_{0<n_2<n_1<k+3} \frac{-2\sin 2\pi \dfrac{(\lambda_1+1)n_1+\lambda_2 n_2}{k+3}}{\left(8\sin\pi\dfrac{n_1-n_2}{k+3}\sin\pi\dfrac{n_2}{k+3}\sin\pi\dfrac{n_1}{k+3}\right)^{2g-1}},$$

where n_1 and n_2 are integers.

Remark 1.3 This is essentially the form of the formula given in Remark 1.1. Since we are considering representations with trivial central character, in this case, χ_λ is 3 identical sine terms, and this factor of 3 is hidden in $N_{3,k}$.

Using Theorem 4.7 and Remark 4.6 one can show that

$$\operatorname{Ver}(k,\lambda) = \begin{cases} p_+(k;\lambda) & \text{if } \lambda_2 > 0, \\ p_-(k;\lambda) & \text{if } \lambda_2 < 0, \end{cases}$$

where p_+ and p_- are two polynomials, given by the right-hand sides of the expressions of Example 4.11 on page 2278. We note two properties of $p_\pm(k,\lambda)$:

(A) The wall-crossing difference $p_- - p_+$ has a relatively simple form (cf Example 4.15 with $\lambda_1+\lambda_3$ replaced by $-\lambda_2$):

$$p_-(k;\lambda) - p_+(k;\lambda) = \operatorname*{Res}_{y=0}\operatorname*{Res}_{x=0} \frac{(-3(k+3)^2)^g \cdot e^{(\lambda_1+1)x-\lambda_2 y}}{(1-e^{x(k+3)})w_\Phi(x,y)^{2g-1}}\,dx\,dy,$$

where $w_\Phi(x,y) = 2\sinh\big(\frac{1}{2}x\big)\cdot 2\sinh\big(\frac{1}{2}y\big)\cdot 2\sinh\big(\frac{1}{2}(x+y)\big)$.

(B) An easy calculation via substitution shows that for any permutation on three elements $\sigma \in \Sigma_3$, our polynomials have the symmetries

$$p_+(k;\sigma\cdot\lambda + \theta_1[k]) = (-1)^\sigma p_+(k;\lambda+\theta_1[k]), \tag{2}$$

$$p_-(k;\sigma\cdot\lambda + \theta_{-1}[k]) = (-1)^\sigma p_-(k;\lambda+\theta_{-1}[k]), \tag{3}$$

where

$$\theta_1[k] = \tfrac{1}{3}k(1,1,-2) + (0,1,-1) \quad \text{and} \quad \theta_{-1}[k] = \tfrac{1}{3}k(2,-1,-1) + (1,-1,0).$$

1.3 Wall-crossings in moduli spaces

Now consider the left-hand side of (1). It is easy to check that the set of isomorphism classes of parabolic bundles in $P_0(c)$ remains unchanged as long as c_2 does not change sign. Hence, effectively, we have two moduli spaces $P_0(>)$ and $P_0(<)$, corresponding to the two chambers separated by the red ($c_2 = 0$) line in Figure 1. Introduce the notation

$$q_+(k;\lambda) = \chi(P_0(>), \mathscr{L}(k;\lambda)) \quad \text{and} \quad q_-(k;\lambda) = \chi(P_0(<), \mathscr{L}(k;\lambda))$$

for the generalized Hilbert polynomials of these two spaces.

In Section 5, we derive a simple formula (31) for the wall-crossing difference in geometric invariant theory. The formula has the form of a residue of an equivariant integral, taken with respect to the equivariant parameter. In our case, the space on which we integrate is the space of rank-3 parabolic bundles which split into a direct sum of a rank-2 and a rank-1 bundle. This equivariant integral may be evaluated using induction on the rank (cf the detailed calculation in Example 6.15 on page 2295), and the result is

$$q_-(k;\lambda) - q_+(k;\lambda) = \operatorname*{Res}_{u=0}\operatorname*{Res}_{z=0} \frac{(-3(k+3)^2)^g \cdot e^{\lambda_1 z+\lambda_2 u+z}}{-w_\Phi(z,-u)^{2g-1}(1-e^{(k+3)z})}\,dz\,du, \tag{4}$$

where u plays the role of the equivariant parameter, the generator of $H^*_{\mathbb{C}^*}(\mathrm{pt})$. This iterated residue coincides with the expression above after changing (z,u) to $(x,-y)$, and thus we have

$$p_+ - p_- = q_+ - q_-. \tag{5}$$

1.4 Hecke correspondences, Serre duality and the symmetry argument

Hecke correspondences between moduli spaces of bundles of different degrees were introduced by Narasimhan and Ramanan in [19]. In Section 7 of our paper, we describe a "tautological" variant of this construction, which identifies the same space with several moduli spaces of parabolic bundles with different degrees and weights. Using this construction we can fiber our two moduli spaces, $P_0(>)$ and $P_0(<)$ over the moduli spaces of stable bundles (without parabolic structure) of degrees 1 and -1:

$$\mathrm{Flag}_3 \to P_0(<) \to N_{-1} \quad \text{and} \quad N_1 \leftarrow P_0(>) \leftarrow \mathrm{Flag}_3,$$

where the fibers are full flags of 3–dimensional vector spaces. Serre duality applied to a Flag_3–bundle implies a Σ_3–antisymmetry of the Euler characteristics of line bundles on this space, and after careful identification of these bundles, we derive the same symmetry properties for the functions $q_\pm$ as we did for the polynomials $p_\pm$: $q_+(k;\lambda)$ satisfies (2), while $q_-(k;\lambda)$ satisfies (3).

The final argument is elegant: we can rearrange equation (5) describing the equality of wall-crossings as

$$p_+ - q_+ = p_- - q_-,$$

and we introduce the notation $\Theta(k;\lambda)$ for this polynomial. Then Θ satisfies both (2) and (3), and thus it is anti-invariant with respect to an affine Weyl group action in the plane for each fixed k. This implies that $\Theta(k;\lambda)$ vanishes and this completes the proof.

1.5 Contents of the paper

There are a number of complications which arise when $r > 3$. We will highlight these in this section, and also give a brief guide to the contents of the paper.

We start with a quick introduction into the theory of parabolic bundles in Section 2. Here we describe the line bundles we are considering, as well as the chamber structure of the space of parabolic weights induced by the stability condition. The combinatorics of the iterated residue formulas mentioned in Section 1.2 above is considerably more complicated in the higher rank case, and is best treated using the notion of *diagonal bases* of hyperplane arrangements introduced in Szenes [23]; we review this construction in the special case of the A_r root arrangement in Section 3.

Using this notion, in Section 4, we present a residue formula for the Verlinde sums on the right-hand side of (1) obtained in Szenes [24] (Theorems 4.4 and 4.7). It turns out that because of a standard ρ–shift type effect in the theory, this residue formula does not have a manifestly polynomial form on our chambers, and thus, we formulate our main result, Theorem 4.8 in two parts: in part (I) we state the equality of the Euler characteristics of line bundles with a modified residue formula, which is manifestly polynomial on our chambers, and in part (II) we state the equality of the modified formula with the original residue formula from [24]. Part (II) is proved in Section 10, while the proof of part (I) takes up the rest of the paper.

At the end of Section 4, we present our wall-crossing formula for Verlinde sums in Proposition 4.18, which uses in an essential manner the yoga of diagonal bases; cf property (A) above for the case of $r = 3$.

The geometric part of our work starts in Section 5, where we derive a simple general result, formula (31), for wall-crossings in GIT. We apply this result to parabolic moduli spaces in Section 6 and, using induction on the rank, obtain Theorem 6.13, the higher-rank version of formula (4) above.

It is downhill from here: in Section 7 we describe the tautological Hecke correspondences we need in several places in the paper, and in Section 8 we derive the Weyl-symmetries of the polynomials $q_{\pm}$, and finish the proof along the lines sketched above.

We are essentially done, but we hit a snag when checking the beginning of our induction on the rank: our argument does not work for $r = 2$. Roughly, the reason for this is that we need our simplex of parabolic weights to have at least two regular vertices, and for $r = 2$, we have only one. The way out is to consider the moduli space with two punctures and then all the pieces fall in place. This argument is carried out in Section 9.

1.6 Historical remarks

There is a long list of proofs of the Verlinde formula, and we cannot do justice to all the approaches in this short introduction. We will thus focus on the historical lineage of our paper, and the works that are closest in spirit to what we do; see Sorger [22] for a more comprehensive overview.

The proofs of the Verlinde formula fall in two categories: proofs of the fusion rules and proofs that find some interpretation of the "Fourier-transformed" discrete sum on the right-hand side of (1); our work belongs to this second group. Another line of division concerns the model which one uses for the moduli spaces: via the Narasimhan–Seshadri correspondence, the moduli spaces of vector bundles may equally be presented as symplectic manifolds of certain types of flat connections on punctured Riemann surfaces, and this opens the way of using the methods of symplectic geometry. While these symplectic approaches lead to results equivalent to the ones coming out of the algebrogeometric setup, the fields of applications of the two approaches seem to be very different.

The idea of proving the Verlinde formula via wall-crossings appeared in the seminal paper of Michael Thaddeus [27]. He used a geometric approach and managed to prove the Verlinde formula in rank 2 by crossing walls in the moduli of stable pairs. The *master space construction*, which plays a central role in our paper, also first appeared in his work [28]. In a sense, our paper may be thought of as the completion of his program.

A paper closely related to our work is that of Jeffrey and Kirwan [13], which approaches the problem from a symplectic/cohomological point of view (see also Jeffrey and Kirwan [12]), and has a somewhat different angle form ours. This paper also uses the residue calculus introduced in [23; 24], but not quite as consistently as our work, and the parabolic case was not resolved from this point of view; see Jeffrey [11]. The geometric model used in [13] to represent the moduli spaces as quotients is rather complicated.

In a comprehensive paper covering the case of all compact groups, Bismut and Labourie [5] used a differential-geometric approach to find the generating function for the parabolic Verlinde formula. This work was the motivation for the residue formula in [24], which is also used in the present paper.

In a remarkable series of papers Alekseev, Meinrenken and Woodward [1], again approaching the subject from the symplectic point of view, gave a direct proof of (1), using reduction in infinite dimensions. A general approach related to twisted K-theory was introduced by Meinrenken in [16]. We should also mention recent work by Loizides and Meinrenken in [14], which employs the residue techniques of [24].

Finally, we drew motivation from the paper of Teleman and Woodward [26], where the Verlinde formula is put in the framework of localization in K-theory of stacks. This very impressive work is probably accessible to a small number of experts only. In the present article we demonstrate, in particular, that the sophisticated tools employed in [26], at least in this instance, may be replaced by a simple combinatorial device.

In summary, the virtues of this article are:

- A proof of the parabolic Verlinde formula which needs as background only the basics of GIT.
- The discrete sum, and the generating function giving the coefficients of the Hilbert polynomial are treated at the same time, and the ρ–shift is dealt with explicitly.
- A few technical innovations, such as an efficient wall-crossing formula in GIT (Theorem 5.7) and the tautological Hecke correspondences, keep the arguments simple, and the technical difficulties related to infinite-dimensional quotients or singularities, in our approach, are absorbed by a combinatorial device: the theory of iterated residues.

2 Parabolic bundles

2.1 Definitions

Let C be a smooth complex projective curve of genus $g \geq 2$, and fix a point $p \in C$.

- A *parabolic bundle* on C is a vector bundle W of rank r with a full flag F_* in the fiber over p,

$$W_p = F_r \supsetneq \cdots \supsetneq F_1 \supsetneq F_0 = 0,$$

and *parabolic weights* $c = (c_1, \ldots, c_r)$ assigned to $F_r, F_{r-1}, \ldots, F_1$, satisfying the conditions

$$c_1 > c_2 > \cdots > c_r \quad \text{and} \quad c_1 - c_r < 1.$$

- The *parabolic degree*[1] and the *parabolic slope* of W are defined as

$$\operatorname{par\,deg} W = \deg W - \sum_{i=1}^{r} c_i \quad \text{and} \quad \operatorname{par\,slope} W = \frac{\operatorname{par\,deg} W}{\operatorname{rank} W}.$$

[1]For technical reasons, we have chosen a sign convention opposite to that in the majority of treatments in the literature.

- A *morphism* $f : W \to W'$ of parabolic bundles is a morphism of vector bundles satisfying $f_p(F_i) \subset F'_{j-1}$ if $c_{r-i+1} < c'_{r-j+1}$. In particular, an *endomorphism* of a parabolic bundle W is a vector bundle endomorphism preserving the flag F_*.
- Denote by $\mathrm{ParHom}(W, W')$ the sheaf of parabolic morphisms from W to W'. Then there is a short exact sequence of sheaves

$$0 \to \mathrm{ParHom}(W, W') \to \mathrm{Hom}(W, W') \to T_p \to 0, \tag{6}$$

where T_p is a torsion sheaf supported at p. The rank of T_p is the number of pairs (i, j), such that $c_i < c'_j$; cf Boden and Hu [6].

If $W' \subset W$ is a subbundle of W, then both W' and the quotient W/W' inherit a parabolic structure from W in a natural way; cf Mehta and Seshadri [15, Definition 1.7].

- A parabolic bundle W is *stable of weight c* if any proper subbundle $W' \subset W$ satisfies $\mathrm{par\,slope}(W') < \mathrm{par\,slope}(W)$; and W is *semistable of weight c*, if the inequality is not strict.

Remark 2.1 The parabolic stability condition depends on the parabolic weights only up to adding the same constant to all weights c_i.

2.2 Construction of the moduli spaces

We start with a quick review of the construction of Mehta and Seshadri [15] of the moduli space of stable parabolic bundles. It follows from Remark 2.1 that, without loss of generality, we can assume that the parabolic weights of a rank-r degree-d bundle belong to the simplex

$$\Delta_d = \left\{ (c_1, c_2, \dots, c_r) \;\middle|\; c_1 > c_2 > \cdots > c_r,\ c_1 - c_r < 1,\ \sum_i c_i = d \right\}.$$

Definition 2.2 We will call a vector $c = (c_1, \dots, c_r) \in \mathbb{R}^r$ such that $\sum_i c_i \in \mathbb{Z}$ *regular* if for any nontrivial subset $\Psi \subset \{1, 2, \dots, r\}$, we have $\sum_{i \in \Psi} c_i \notin \mathbb{Z}$.

Now choose an integer $d \gg 0$ such that $H^1(W) = 0$ and W is generated by global sections for any rank-r degree-d semistable parabolic bundle W of parabolic degree 0. Put $N = r(1-g) + d$.

- Consider the Grothendieck quot scheme $\mathrm{Quot}(N, r)$ [9] parametrizing quotients $\mathbb{C}^N \twoheadrightarrow W$, where W is a coherent sheaf of degree d and rank r.
- This space is endowed with a universal bundle UQ, and a generically free action of the group $G = \mathrm{PSL}(N)$ which does not, however, lift to UQ.
- Let $\mathrm{LFQuot} \subset \mathrm{Quot}(N, r)$ be the open subscheme consisting of locally free quotients W such that the induced map $H^0(\mathbb{C}^N) \to H^0(W)$ is an isomorphism.

- Denote by XQ the total space of the flag bundle $\mathrm{Flag}(UQ_p)$ on $\mathrm{LFQuot}\times p$. This space is endowed with the flag of vector bundles $\mathrm{Fl}_1 \subset \cdots \subset \mathrm{Fl}_{r-1} \subset \mathrm{Fl}_r = UQ_p$.
- Let $k \in \mathbb{Z}$ and $(\lambda_1, \ldots, \lambda_r) \in \mathbb{Z}^r$ be such that $\sum_{i=1}^r \lambda_i = kd$, and consider the line bundle
$$L(k;\lambda) = \det(UQ_p)^{k(1-g)} \otimes \det(\pi_* UQ)^{-k} \otimes (\mathrm{Fl}_r/\mathrm{Fl}_{r-1})^{\lambda_1} \otimes \cdots \otimes (\mathrm{Fl}_1)^{\lambda_r}$$
on XQ, which does carry a G–linearization (lift of the G–action from XQ).
- Finally, assume $c \in \Delta_d$ is regular (cf Definition 2.2 above) and define $\widetilde{P}_d(c)$, the moduli space of stable parabolic weight-c vector bundles on C, as the GIT quotient $XQ/\!/^c G$ of XQ with respect to any linearization $L(k;\lambda)$ such that $\lambda/k = c$.

Theorem 2.3 [21] *Assume that $c \in \Delta_d$ is a regular weight vector. Then the moduli space $\widetilde{P}_d(c)$ is a smooth projective variety of dimension $r^2(g-1)+\binom{r}{2}+1$, whose points are in one-to-one correspondence with the set of isomorphism classes of stable parabolic bundles of weight c* (cf Section 2.1).

Remark 2.4 Via the determinant map, the moduli space $\widetilde{P}_d(c)$ fibers over the Jacobian of degree-d line bundles with isomorphic fibers, and in this paper, we will focus on the moduli space
$$P_d(c) = \{W \in \widetilde{P}_d(c) \mid \det W \simeq \mathcal{O}(dp)\},$$
which is smooth, projective and has dimension $(r^2-1)(g-1)+\binom{r}{2}$.

Remark 2.5 Tensoring with the line bundle $\mathcal{O}(mp)$ induces an isomorphism
$$\otimes\mathcal{O}(mp)\colon P_d(c) \to P_{d+rm}(c),$$
so the moduli spaces $P_d(c)$, essentially, depend only on d modulo r.

2.3 The Picard group of $P_d(c)$

For a regular $c \in \Delta_d$, there exist universal bundles U over $P_d(c)\times C$ endowed with a flag $\mathcal{F}_1 \subset \cdots \subset \mathcal{F}_{r-1} \subset \mathcal{F}_r = U_p$, and satisfying the obvious tautological properties. In general, such universal bundles U, and hence the flag line bundles $\mathcal{F}_{i+1}/\mathcal{F}_i$, are unique only up to tensoring by the pullback of a line bundle from $P_d(c)$. Nevertheless, we have the following statement.

Lemma 2.6 *For $k \in \mathbb{Z}$ and $\lambda = (\lambda_1, \ldots, \lambda_r) \in \mathbb{Z}^r$ such that $\sum_{i=1}^r \lambda_i = kd$, the line bundle*
$$\mathcal{L}_d(k;\lambda) = \det(U_p)^{k(1-g)} \otimes \det(\pi_* U)^{-k} \otimes (\mathcal{F}_r/\mathcal{F}_{r-1})^{\lambda_1} \otimes \cdots \otimes (\mathcal{F}_1)^{\lambda_r} \tag{7}$$
on $P_d(c)$ is independent of the choice of the universal bundle U.

Proof Note that tensoring U with a pullback $\pi^*\mathcal{L}$ of a line bundle $\mathcal{L}$ on $P_0(c)$ changes $\det(U_p)$ by $\mathcal{L}^r$ and $\det(\pi_* U)$ by $\mathcal{L}^{d-r(g-1)}$. □

Remark 2.7 The line bundle $L(k;\lambda)$ defined in Section 2.2 descends to the line bundle $\mathcal{L}_d(k;\lambda)$ on the GIT quotient $P_d(c)$.

Notation We will say that U is *normalized* if the line subbundle $\mathcal{F}_1 \subset U_p$ is trivial. The parameter k is often called the *level*.

Let $\omega \in H^2(C)$ be the fundamental class of our curve C, and $e_1, \dots, e_{2g}$ a basis of $H^1(C)$ such that $e_i e_{i+g} = \omega$ for $1 \le i \le g$, and all other intersection numbers $e_i e_j$ equal 0. For a class $\delta \in H^*(P \times C)$ of a product, we introduce notation for its Künneth components (cf [32]):

$$\delta = \delta_{(0)} \otimes 1 + \sum_i \delta_{(e_i)} \otimes e_i + \delta_{(2)} \otimes \omega \in \bigoplus_{i=0}^{2} H^{*-i}(P) \otimes H^i(C). \tag{8}$$

We will need the following formula.

Lemma 2.8 *The equality* $2\,c_1(\mathcal{L}_d(r; d, \dots, d)) = c_2(\mathrm{End}_0(U_d))_{(2)}$ *holds, where* End_0 *stands for traceless endomorphisms.*

Proof Taking the first Chern class on both sides of (7), we obtain

$$c_1(\mathcal{L}_d(r; d, \dots, d)) = r(1-g)c_1(U_d)_{(0)} - r c_1(\pi_*(U_d)) + d c_1(U_d)_{(0)},$$

where we evaluate the middle term using the Grothendieck–Riemann–Roch theorem, and $c_1(U_d)_{(2)} = d$:

$$\begin{aligned} c_1(\pi_*(U_d)) &= \mathrm{ch}_1(\pi_!(U_d)) = \pi_* \,\mathrm{ch}_2(U_d) - (g-1)c_1(U_d)_{(0)} \\ &= c_1(U_d)_{(0)} d - c_2(U_d)_{(2)} - (g-1)c_1(U_d)_{(0)}. \end{aligned}$$

This leads to the formula

$$c_1(\mathcal{L}_d(r; d, \dots, d)) = -d(r-1)c_1(U_d)_{(0)} + r c_2(U_d)_{(2)},$$

which is easily seen to equal $\frac{1}{2} c_2(\mathrm{End}_0(U_0))_{(2)}$. □

2.4 Walls and chambers

The central question we address in this paper is how the moduli space of stable parabolic bundles depends on the choice of parabolic weights. Let W be a vector bundle of degree d with a fixed full flag F_* of the fiber W_p, and let us try to determine the structure of the set of parabolic weights $c \in \Delta_d$ for which W is stable. Clearly, for this we need to study the set of parabolic weights $c = (c_1, c_2, \dots c_r)$ for which one can find a proper subbundle $W' \subset W$ such that

$$\mathrm{par\,slope}(W') = \mathrm{par\,slope}(W) = 0. \tag{9}$$

A subbundle $W' \subset W$ determines a short exact sequence of parabolic bundles

$$0 \to W' \to W \to W'' \to 0,$$

and the position of W'_p with respect to F_* gives rise to a nontrivial partition of the set $\{1, 2, \dots, r\}$ into two sets, Π' and Π'', cf [15, Definition 1.7]; the parabolic weights of W' and W'' are then $c' = (c_i)_{i \in \Pi'}$ and $c'' = (c_i)_{i \in \Pi''}$, correspondingly. The slope condition (9) translates into a pair of equivalent equalities

$$d' = \sum_{i \in \Pi''} c_i \quad \text{and} \quad d'' = \sum_{i \in \Pi''} c_i, \tag{10}$$

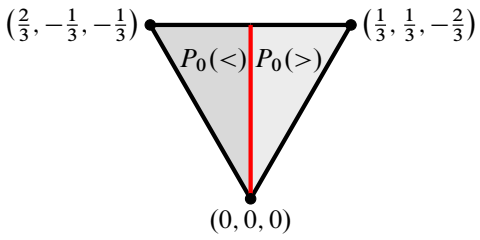


Figure 2: The space of admissible parabolic weights for rank $r = 3$.

where d' and $d'' = d - d'$ are the degrees of W' and W'', respectively. This means that the critical values of $c \in \Delta_d$ for which (9) is possible lie on the union of affine hyperplanes (or *walls*) defined by the equations

$$\sum_{i \in \Pi'} c_i = l, \quad \text{where } l \in \mathbb{Z}, \text{ and } \Pi' \subset \{1, 2, \dots, r\} \text{ is nontrivial.}$$

As only finitely many of these walls intersect the simplex Δ_d, their complement is a finite union of open polyhedral *chambers*. It is easy to verify that as we vary c inside one of these chambers, the stability condition, and thus the moduli space $P_d(c)$, does not change.

Example 2.9 Consider the case of rank-3 degree-0 stable parabolic bundles with parabolic weights $c = (c_1, c_2, c_3) \in \Delta_0$. The set Δ_0 is an open triangle with vertices $(0, 0, 0)$, $\left(\frac{2}{3}, -\frac{1}{3}, -\frac{1}{3}\right)$ and $\left(\frac{1}{3}, \frac{1}{3}, -\frac{2}{3}\right)$ (see Figure 2), and there exist only two essentially different stability conditions. The wall separating the two regimes is given by the condition $c_2 = 0$. We write $P_0(>)$ for the moduli space $P_0(c_1, c_2, c_3)$ with $c_2 > 0$, and $P_0(<)$ for $P_0(c_1, c_2, c_3)$ with $c_2 < 0$.

3 Wall-crossing in the Verlinde formula

A key component of our approach is the notion of *diagonal basis* and the associated generalized Bernoulli polynomials introduced for general hyperplane arrangements in [23]. Using this formalism, we will be able to formulate our main result, Theorem 4.8.

3.1 Notation

We begin by setting up some extra notation for the space of parabolic weights introduced in Section 2.1.

- Let $V = \mathbb{R}^r/\mathbb{R}(1, 1, \dots, 1)$ be the $(r-1)$–dimensional vector space, obtained as the quotient of $\mathbb{R}^r$. The dual space V^* is then naturally represented as

$$V^* = \{a = (a_1, \dots, a_r) \in \mathbb{R}^r \mid a_1 + \cdots + a_r = 0\}.$$

  Let $x_1, x_2, \dots, x_r$ be the coordinates on $\mathbb{R}^r$; given $a \in V^*$, we will write $\langle a, x \rangle$ for the linear function $\sum_i a_i x_i$ on V. We will sometimes identify this linear function with the vector a itself.

- The vector space V^* is endowed with a lattice Λ of full rank:
$$\Lambda = \{\lambda = (\lambda_1, \dots, \lambda_r) \in \mathbb{Z}^r \mid \lambda_1 + \cdots + \lambda_r = 0\}.$$
In particular, for $1 \le i \ne j \le r$, we can define the element $\alpha^{ij} = x_i - x_j$ in Λ.
- Our arrangement is the set of hyperplanes $\{x_i = x_j\} \subset V$, with $1 \le i < j \le r$. It will be convenient for us to think about this set as the set of roots of the A_{r-1} root system with the opposite roots identified:
$$\Phi = \{\ \pm\alpha^{ij} \mid 1 \le i < j \le r\}.$$
Note that V^* carries a natural action of the permutation group Σ_r, permuting the coordinates x_j for $j = 1, \dots, r$, and this action restricts to an action on Φ as well.
- The basic object of the theory is an *ordered* linear basis $\boldsymbol{B}$ of V^* consisting of the elements of Φ. Let us denote the set of these objects by $\mathcal{B}$:
$$\mathcal{B} = \{\boldsymbol{B} = (\beta^{[1]}, \dots, \beta^{[r-1]}) \in \Phi^{r-1} \mid \boldsymbol{B} \text{ a basis of } V^*\}.$$
- For $\boldsymbol{B} \in \mathcal{B}$, we will write $\mathrm{Fl}(\boldsymbol{B})$ for the full flag
$$[V^* = \langle\beta^{[1]}, \beta^{[2]}, \dots, \beta^{[r-1]}\rangle_{\mathrm{lin}}, \dots, \langle\beta^{[r-1]}, \beta^{[r-2]}\rangle_{\mathrm{lin}}, \langle\beta^{[r-1]}\rangle_{\mathrm{lin}}],$$
where $\langle\,\cdot\,\rangle_{\mathrm{lin}}$ stands for linear span.

3.2 Diagonal bases

Definition 3.1
- For $\tau \in \Sigma_{r-1}$ and $\boldsymbol{B} \in \mathcal{B}$, we will write $\boldsymbol{B} \circlearrowleft \tau$ for the permuted sequence
$$(\beta^{[\tau(1)]}, \beta^{[\tau(2)]}, \dots, \beta^{[\tau(r-1)]}).$$
- For two elements $\boldsymbol{B}, \boldsymbol{C} \in \mathcal{B}$ we will write $\boldsymbol{B} \dashv \boldsymbol{C}$ if for any $\tau \in \Sigma_{r-1}$, we have $\mathrm{Fl}(\boldsymbol{B} \circlearrowleft \tau) \ne \mathrm{Fl}(\boldsymbol{C})$.
- A subset $\mathcal{D} \subset \mathcal{B}$ of $(r-1)!$ elements is called a *diagonal basis* if for any two different elements $\boldsymbol{B}, \boldsymbol{C} \in \mathcal{D}$, we have $\boldsymbol{B} \dashv \boldsymbol{C}$.

Remark 3.2 This definition is motivated by a construction in [23], which associates to each diagonal basis $\mathcal{D}$ a pair of dual bases of the middle homology and the cohomology of the complexified hyperplane arrangement on $V \otimes_{\mathbb{R}} \mathbb{C}$ defined by Φ. The dimension of these (co)homology spaces is $(r-1)!$.

3.3 Combinatorial interpretation

This notion has the following purely combinatorial form.

- We can think of Φ as the edges of the complete graph on r vertices.
- Then the set $\mathcal{B}$ may be thought of as the set of spanning trees of this graph with edges enumerated from 1 to $r-1$. We will introduce the notation
$$\boldsymbol{B} \mapsto \mathrm{Tree}(\boldsymbol{B})$$
for this ordered tree.

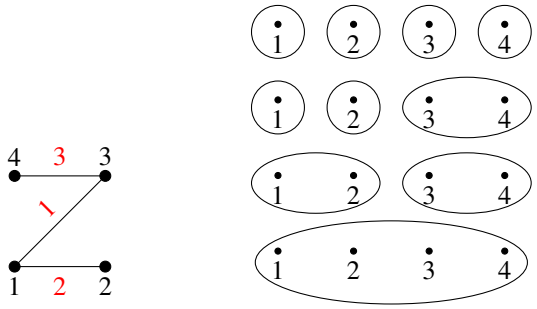


Figure 3: $\boldsymbol{B} = (\alpha^{1,3}, \alpha^{1,2}, \alpha^{3,4})$.

- In this language, the flag $\mathrm{Fl}(\boldsymbol{B})$ corresponds to a sequence of r nested partitions of the vertices (starting with the total partition into one-element sets and ending with the trivial partition) associated to $\mathrm{Tree}(\boldsymbol{B})$, the j^{th} partition being the one induced by the first $j-1$ edges. For example, the ordered tree $[(2,4)(1,3),(1,2)]$ induces the same sequence of partitions as $[(1,4),(2,3),(1,2)]$; see Figure 3.
- A diagonal basis $\mathscr{D}$ is then a set of $(r-1)!$ ordered trees such that the $(r-1)!$ partition sequences obtained by reordering the edges of any one of the ordered trees are different from $(r-1)!-1$ sequences of partitions obtained from the remaining elements of $\mathscr{D}$.

3.4 Examples

There are essentially two known constructions of diagonal bases [23].

I. **The Hamiltonian basis** For each permutation $\sigma \in \Sigma_r$, we can define

$$\sigma(\boldsymbol{B}) = (\alpha^{\sigma(r-1),\sigma(r)}, \alpha^{\sigma(r-2),\sigma(r-1)}, \dots, \alpha^{\sigma(1),\sigma(2)}) \in \mathscr{B}. \tag{11}$$

The set $\mathscr{H}_m = \{\sigma(\boldsymbol{B}) \mid \sigma \in \Sigma_r, \sigma(1) = m\}$ is then a diagonal basis. In the combinatorial description, this diagonal basis corresponds to the set of Hamiltonian paths starting at vertex m, and endowed with the reversed natural ordering of edges.

Example 3.3 Here are some examples of Hamiltonian bases:

- For $r = 3$: $\mathscr{H}_1 = \{(\alpha^{2,3}, \alpha^{1,2}), (\alpha^{3,2}, \alpha^{1,3})\}$.
- For $r = 4$: $\mathscr{H}_1 = \big\{(\alpha^{3,4}, \alpha^{2,3}, \alpha^{1,2}), (\alpha^{2,4}, \alpha^{3,2}, \alpha^{1,3}), (\alpha^{4,3}, \alpha^{2,4}, \alpha^{1,2}), (\alpha^{3,2}, \alpha^{4,3}, \alpha^{1,4}), (\alpha^{4,2}, \alpha^{3,4}, \alpha^{1,3}), (\alpha^{2,3}, \alpha^{4,2}, \alpha^{1,4})\big\}$.

II. **The no-broken-circuit (nbc) bases** Let $\upsilon\colon \{1, \dots, \frac{1}{2}r(r-1)\} \to \Phi$ be a total ordering, which we will represent as an order relation $\overset{\upsilon}{<}$ on Φ. To this ordering, one can associate the so-called *noncommutative no-broken-circuit diagonal basis* [23]

$$\mathscr{D}[\upsilon] = \big\{(\beta^{[1]}, \dots, \beta^{[r-1]}) \in \mathscr{B} \mid \beta^{[1]} \overset{\upsilon}{<} \dots \overset{\upsilon}{<} \beta^{[r-1]}, \text{ and } \alpha^{ij} \overset{\upsilon}{<} \beta^{[m]} \implies (\alpha^{ij}, \beta^{[m]}, \dots, \beta^{[r-1]}) \text{ linearly independent}\big\}.$$

Example 3.4 Let $\alpha^{1,3} \overset{\upsilon}{<} \alpha^{1,4} \overset{\upsilon}{<} \alpha^{2,3} \overset{\upsilon}{<} \alpha^{2,4} \overset{\upsilon}{<} \alpha^{1,2} \overset{\upsilon}{<} \alpha^{3,4}$ be the ordering of the positive roots for rank $r = 4$. Then

$$\mathcal{D}[\upsilon] = \big\{(\alpha^{1,3}, \alpha^{1,2}, \alpha^{3,4}), (\alpha^{1,3}, \alpha^{1,4}, \alpha^{2,3}), (\alpha^{1,3}, \alpha^{1,4}, \alpha^{2,4}), (\alpha^{1,3}, \alpha^{1,4}, \alpha^{1,2}), \\ (\alpha^{1,3}, \alpha^{2,3}, \alpha^{3,4}), (\alpha^{1,3}, \alpha^{2,3}, \alpha^{2,4})\big\}$$

is the corresponding no-broken-circuit diagonal basis.

Remark 3.5 The hyperplane arrangement induced by Φ is invariant under the natural action of Σ_r on the vector space V. It follows easily from the definition that if $\mathcal{D}$ is a diagonal basis and $\sigma \in \Sigma_r$ is a permutation, then $\sigma(\mathcal{D})$ is also a diagonal basis.

4 The residue formula and the main result

In this section, we recall the residue formula from [23] for $\mathrm{Ver}(k, \lambda)$, the discrete Verlinde sum on the right-hand side of (1). The key feature of this formula is that it exposes the piecewise polynomial nature of $\mathrm{Ver}(k, \lambda)$, which is key for our wall-crossing analysis. While the objects are relatively simple, the formalism is heavy with notation, so we begin by describing the one-dimensional case.

4.1 The residue formula in dimension 1

The story begins with the Fourier series

$$\frac{1}{(2\pi i)^m} \sum_{n \in \mathbb{Z} \setminus 0} \frac{\exp(2\pi i a n)}{n^m} \tag{12}$$

for $m \geq 2$, which is a periodic, piecewise polynomial function given by the formula

$$\operatorname*{Res}_{x=0} \frac{\exp(\{a\}x)}{1 - \exp(x)} \frac{dx}{x^m},$$

where $\{a\}$ is the fractional part of the real number a. The polynomial functions thus obtained on the interval $[0, 1]$ are called *Bernoulli polynomials*. The polynomial on the interval containing the real number $c \in \mathbb{R} \setminus \mathbb{Z}$ is given by

$$\operatorname*{Res}_{x=0} \frac{\exp((a - [c])x)}{1 - \exp(x)} \frac{dx}{x^m},$$

where $[c]$ is the integer part of c.

Now we pass to a trigonometric version of this formula, calculating finite sums of values of rational trigonometric functions over rational points with denominators equal to an integer k.

We thus replace the rational function x^{-m} by the (hyperbolic) trigonometric function

$$f(x) = \big(2 \sinh\big(\tfrac{1}{2}x\big)\big)^{-2m},$$

and introduce an integer parameter λ related to a via $ka = \lambda$. We consider the sum of values of the function f over a finite set of rational points in analogy with (12),

$$\sum_{n=1}^{k-1} \frac{\exp(2\pi i\,\lambda n/k)}{(2\sin(\pi n/k))^{2m}},$$

where $\lambda, k \in \mathbb{Z}$. This sum is again periodic in $\lambda \bmod k$, and for $m \geq 2$ we can evaluate it via the residue theorem as

$$(-1)^m \operatorname*{Res}_{z=1} \frac{z^{k\{\lambda/k\}}}{(z^{1/2}-z^{-1/2})^{2m}} \cdot \frac{k\,dz}{z(1-z^k)} \overset{z=\exp(x/k)}{=} (-1)^m \operatorname*{Res}_{x=0} \frac{\exp(\{\lambda/k\}\cdot x)}{1-\exp(x)} \cdot f\left(\frac{x}{k}\right) dx.$$

Again, this is a piecewise polynomial function in the pair (k, λ), which is polynomial in the cones bounded by the lines $\lambda = qk$ for $q \in \mathbb{Z}$.

Note that in these calculations, a key role is played by the Bernoulli operator

$$f \mapsto \mathrm{Ber}[f](a) = \frac{f(x)\exp(ax)\,dx}{1-\exp(x)}, \tag{13}$$

which transforms meromorphic functions in the variable x into polynomials in a, and plays the role of a generalized Fourier operator.

4.2 The multidimensional case

We return to the setup of Section 3 with the vector space V endowed with the hyperplane arrangement Φ. We introduce the notation $\mathcal{F}_\Phi$ *for the space of meromorphic functions defined in a neighborhood of* 0 *in* $V \otimes_{\mathbb{R}} \mathbb{C}$ *with poles on the union of hyperplanes*

$$\bigcup_{1\leq i<j\leq r} \{x \mid \langle \alpha^{ij}, x\rangle = 0\}.$$

In particular, the function

$$w_\Phi = \prod_{i<j} \big(2\sinh(\pi(x_i - x_j))\big)$$

is an element of $\mathcal{F}_\Phi$.

To write down our residue formula, we need a multidimensional generalization of the notions of integer and fractional parts. Given a basis $\boldsymbol{B} = (\beta^{[1]}, \dots, \beta^{[r-1]}) \in \mathcal{B}$ of V^*, and an element $a \in V^*$, we define $[a]_{\boldsymbol{B}}$ and $\{a\}_{\boldsymbol{B}}$ to be the unique elements of V^* satisfying

$$[a]_{\boldsymbol{B}} = a - \{a\}_{\boldsymbol{B}} \in \Lambda \quad \text{and} \quad \{a\}_{\boldsymbol{B}} \in \sum_{j=1}^{r-1} [0,1)\beta^{[j]}.$$

This notion naturally induces a chamber structure on V^*: we will call $a \in V^*$ *regular* if a is a point of continuity for the functions $a \mapsto [a]_{\boldsymbol{B}}, \{a\}_{\boldsymbol{B}}$ for all $\boldsymbol{B} \in \mathcal{B}$, ie when $\{a\}_{\boldsymbol{B}} \in \sum_{j=1}^{r-1}(0,1)\beta^{[j]}$. Now, for regular a and b we define the equivalence relation

$$a \sim b \quad \text{when } [a]_{\boldsymbol{B}} = [b]_{\boldsymbol{B}} \text{ for all } \boldsymbol{B} \in \mathcal{B}. \tag{14}$$

The equivalence classes for this relation form a Λ–periodic system of *chambers* in V^*.

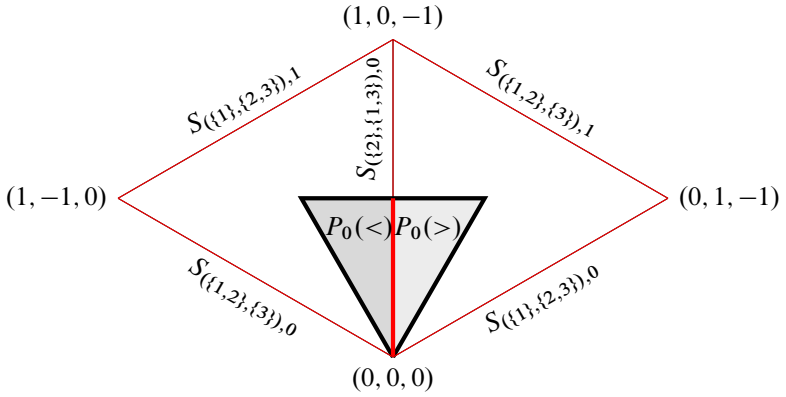


Figure 4: Chambers for rank $r = 3$.

Convention We will think of a partition Π of $\{1, 2, \dots, r\}$ into two nonempty sets as an ordered partition $\Pi = (\Pi', \Pi'')$ such that $r \in \Pi''$, and we will call these objects *nontrivial partitions* for short.

The following statement is straightforward.

Lemma 4.1 *The equivalence classes of the relation $\sim$ are precisely the chambers in V^* created by the walls parametrized by a nontrivial partition $\Pi = (\Pi', \Pi'')$ of the first r positive integers, and an integer l:*

$$S_{\Pi,l} = \Big\{ c \in V^* \;\Big|\; \sum_{j \in \Pi'} c_j = l \Big\}. \tag{15}$$

Remark 4.2 The walls given in (10) are precisely the same as the ones given in (10) for the case $d = 0$, where they play the role of walls separating the chambers of parabolic weights c in which the parabolic moduli spaces $P_0(c)$ are naturally the same. This "coincidence" is precisely what we need for our comparative wall-crossing strategy. There is a small terminological issue here: the "chambers" in Section 2.4 are the intersections of the equivalence classes of $\sim$ defined above with the open simplex $\Delta_0 \overset{\text{def}}{=} \Delta$, where the parabolic weights live; see Figures 2 and 4. We will use the term "chamber" in both cases if this causes no confusion.

Each element $\boldsymbol{B} = (\beta^{[1]}, \dots, \beta^{[r-1]}) \in \mathcal{B}$ defines an *iterated* version of the Bernoulli operator (13) on the space of functions $\mathcal{F}_\Phi$: interpreting the elements $a, \beta^{[j]} \in V^*$ as linear functions on V, we define

$$\operatorname*{iBer}_{\boldsymbol{B}}[f(x)](a) = \frac{1}{(2\pi i)^{r-1}} \int_{Z(\boldsymbol{B})} \frac{f(x) \exp\langle a, x\rangle \, d\langle \beta^{[1]}, x\rangle \wedge \cdots \wedge d\langle \beta^{[r-1]}, x\rangle}{(1 - \exp(\langle \beta^{[1]}, x\rangle)) \cdots (1 - \exp(\langle \beta^{[r-1]}, x\rangle))}, \tag{16}$$

where the naturally oriented cycle $Z_{\boldsymbol{B}}$ is defined by

$$Z_{\boldsymbol{B}} = \{ v \in V \otimes_{\mathbb{R}} \mathbb{C} \mid |\langle \beta^{[j]}, x\rangle| = \varepsilon_j \text{ for } j = 1, \dots, r-1 \} \subset V \otimes_{\mathbb{R}} \mathbb{C} \setminus \{ w_\Phi(x) = 0 \},$$

with real constants ε_j satisfying $0 \le \varepsilon_{r-1} \ll \cdots \ll \varepsilon_1$. Thus again, $\operatorname{iBer}_{\boldsymbol{B}}$ is a linear operator which associates to a function in $\mathcal{F}_\Phi$ a polynomial on V^*.

Remark 4.3 Let us make a small remark about the computational aspects of the operator $\mathrm{iBer}_{\mathbf{B}}$. Denoting the coordinate $\langle \beta^{[j]}, x\rangle$ by y_j for $j = 1, \dots, r-1$, and writing f and a in these coordinates: $f(x) = \widehat{f}(y)$, $\langle a, x\rangle = \langle \widehat{a}, y\rangle$, we can rewrite (16) as

$$\underset{\boldsymbol{B}}{\mathrm{iBer}}[f(x)](a) = \operatorname*{Res}_{y_1=0} \cdots \operatorname*{Res}_{y_{r-1}=0} \frac{\widehat{f}(y)\exp\langle \widehat{a}, y\rangle\, dy_1 \wedge \cdots \wedge dy_{r-1}}{(1-\exp(y_1))\cdots(1-\exp(y_{r-1}))},$$

where *iterating* the residues here means that we keep the variables with lower indices as unknown constants, and then use geometric series expansions of the type

$$\frac{1}{1-\exp(y_1-y_2)} = \frac{y_1-y_2}{1-\exp(y_1-y_2)} \cdot \frac{1}{y_1-y_2} = \frac{y_1-y_2}{1-\exp(y_1-y_2)} \cdot \sum_{n=0}^{\infty} \frac{y_2^n}{y_1^{n+1}}.$$

4.3 Invariance of diagonal bases and the main results

Diagonal bases have the following key invariance property.

Theorem 4.4 [23] *Let $f \in \mathcal{F}_\Phi$, and $c \in V^*$ be regular; let $\mathcal{D}$ be a diagonal basis of Φ. Then the functional*

$$f \mapsto \sum_{\boldsymbol{B}\in\mathcal{D}} \underset{\boldsymbol{B}}{\mathrm{iBer}}[f(x)](a - [c]_{\boldsymbol{B}})$$

(see (16) above) *transforming a meromorphic function $f \in \mathcal{F}_\Phi$ into a polynomial in the variable $a \in V^*$ is independent of the choice of the diagonal basis $\mathcal{D}$. In particular, for regular $a \in V^*$, the functional*

$$f \mapsto \sum_{\boldsymbol{B}\in\mathcal{D}} \underset{\boldsymbol{B}}{\mathrm{iBer}}[f(x)](\{a\}_{\boldsymbol{B}}) \tag{17}$$

transforms f into a well-defined piecewise polynomial function on V^*, which is polynomial in each chamber.

As this functional is invariantly defined, it is not surprising that it is equivariant with respect to the symmetries of our hyperplane arrangement. For $\sigma \in \Sigma_r$, we define, as usual

$$\sigma \cdot f(x) = f(\sigma^{-1}x). \tag{18}$$

This convention is consistent with (11).

Lemma 4.5 *Let $f \in \mathcal{F}_\Phi$ and $\sigma \in \Sigma_r$, and pick any diagonal basis $\mathcal{D}$. Then*

$$\sum_{\boldsymbol{B}\in\mathcal{D}} \underset{\boldsymbol{B}}{\mathrm{iBer}}[f(x)](\sigma\cdot a - [\sigma\cdot c]_{\boldsymbol{B}}) = \sum_{\boldsymbol{B}\in\mathcal{D}} \underset{\boldsymbol{B}}{\mathrm{iBer}}[\sigma^{-1}\cdot f(x)](a - [c]_{\boldsymbol{B}}).$$

Proof Indeed, recall that $\sigma \in \Sigma$ takes a diagonal basis to another diagonal basis (see Remark 3.5), so

$$\sum_{\boldsymbol{B}\in\mathcal{D}} \underset{\boldsymbol{B}}{\mathrm{iBer}}[f(x)](\sigma\cdot a - [\sigma\cdot c]_{\boldsymbol{B}}) = \sum_{\boldsymbol{B}\in\mathcal{D}} \underset{\sigma\boldsymbol{B}}{\mathrm{iBer}}[f(x)](\sigma\cdot a - [\sigma\cdot c]_{\sigma\boldsymbol{B}}).$$

Now we perform the linear substitution $x=\sigma(y)$, and obtain

$$\sum_{\boldsymbol{B}\in\mathcal{D}} \operatorname*{iBer}_{\sigma\boldsymbol{B}}[f(x)](\sigma\cdot a-[\sigma\cdot c]_{\sigma\boldsymbol{B}})=\sum_{\boldsymbol{B}\in\mathcal{D}} \operatorname*{iBer}_{\boldsymbol{B}}[\sigma^{-1}\cdot f(y)](a-[c]_{\boldsymbol{B}}). \qquad \square$$

Remark 4.6 By picking the Hamiltonian diagonal basis $\mathcal{H}_1=\{\sigma\cdot\boldsymbol{B}_0 \mid \sigma\in \mathrm{Stab}(1,\Sigma_r)\}$, we can turn the argument in the proof above around, and obtain the formula

$$\begin{aligned}\sum_{\boldsymbol{B}\in\mathcal{H}_1} \operatorname*{iBer}_{\boldsymbol{B}}[f(x)](a-[c]_{\boldsymbol{B}}) &= \sum_{\sigma\in\mathrm{Stab}(1,\Sigma_r)} \operatorname*{iBer}_{\boldsymbol{B}_0}[\sigma\cdot f(x)](\sigma\cdot a-[\sigma\cdot c]_{\boldsymbol{B}})\\ &= \operatorname*{Res}_{y_1=0}\cdots\operatorname*{Res}_{y_{r-1}=0}\sum_{\sigma\in\mathrm{Stab}(1,\Sigma_r)} \frac{\sigma\cdot f(y)\exp\langle\sigma\cdot a-[\sigma\cdot c]_{\boldsymbol{B}},y\rangle\, dy_1\wedge\cdots\wedge dy_{r-1}}{(1-\exp(y_1))\cdots(1-\exp(y_{r-1}))},\end{aligned}$$

where

$$\boldsymbol{B}_0=(y_1=x_{r-1}-x_r,\dots,y_{r-2}=x_2-x_3,\,y_{r-1}=x_1-x_2)\in\mathcal{B}.$$

Now we are ready to write down the residue formula for the Verlinde sums proved in [24, Theorem 4.2]. Recall that we denoted by $\mathrm{Ver}(k,\lambda)$ the finite sum on the right-hand side of (1).

Theorem 4.7 *Let $g\geq 1$, $k\in\mathbb{Z}^{>0}$, $\lambda\in\Lambda$, and let $\mathcal{D}$ be any diagonal basis of Φ. Introducing the notation $\widehat{k}=k+r$ and $\widehat{\lambda}=\lambda+\rho$, we have*

$$\mathrm{Ver}(k,\lambda)=\widetilde{N}_{r,k}\sum_{\boldsymbol{B}\in\mathcal{D}} \operatorname*{iBer}_{\boldsymbol{B}}[w_\Phi^{1-2g}(x/\widehat{k})](\widehat{\lambda}/\widehat{k}-[\widehat{\zeta}]_{\boldsymbol{B}}), \tag{19}$$

where $\widetilde{N}_{r,k}=(-1)^{\binom{r}{2}(g-1)}N_{r,k}$ (see (1)) and $\widehat{\zeta}\in V^*$ is a regular point in a chamber that contains $\widehat{\lambda}/\widehat{k}$ in its closure.

Now, if we look at our main goal (1): proving the equality

$$\mathrm{Ver}(k,\lambda)=\chi(P_0(\lambda/k),\mathscr{L}_0(k;\lambda)), \tag{20}$$

we discover a rather embarrassing mismatch. Both sides are piecewise polynomial functions; however,

- according to the HRR theorem, $\chi(P(\lambda/k),\mathscr{L}_0(k;\lambda))$ is polynomial on the cones over the equivalence classes (see (14)) of λ/k, while
- according to (19), $\mathrm{Ver}(k,\lambda)$ is polynomial on the cones over the equivalence classes of $\widehat{\lambda}/\widehat{k}$,

and these conic partitions of $\{(k,\lambda)\mid\lambda/k\in\Delta\}$ could clearly be different; see Figure 5 for a sketch of this problem.

Thus for (1) to be true, some miracle needs to occur, and these miracles are well-known in the area of "quantization commutes with reduction" [17; 29; 30; 25]. We will return to this problem in Section 10, but for now, we will be satisfied to use (19) to write down a (conjectural for the moment) formula for $\chi(P_0(\lambda/k),\mathscr{L}_0(k;\lambda))$, which is manifestly polynomial on the cones where λ/k is in a fixed equivalence class.

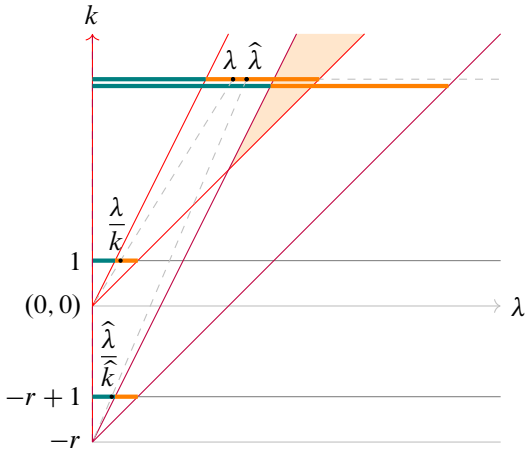


Figure 5: The vector λ/k is in the orange chamber, while $\widehat{\lambda}/\widehat{k}$ is in the green chamber.

Let us fix a regular $c \in \Delta$ marking a particular chamber in Δ. The two cones $\{(k;\lambda) \mid \lambda/k \sim c\}$ and $\{(k;\lambda) \mid \widehat{\lambda}/\widehat{k} \sim c\}$ intersect along an open cone (this cone is shaded in orange in Figure 5), and on this intersection, the expression

$$\sum_{\boldsymbol{B}\in\mathcal{D}} \mathrm{iBer}_{\boldsymbol{B}}[w_\Phi^{1-2g}(x/\widehat{k})](\widehat{\lambda}/\widehat{k} - [\lambda/k]_{\boldsymbol{B}}) \tag{21}$$

coincides with the right-hand side of (19). As (21) is manifestly polynomial on each cone where λ/k is in a particular chamber in Δ, this expression will be then our main candidate for $\chi(P_0(\lambda/k), \mathscr{L}_0(k;\lambda))$.

Our plan is thus to split the proof of (20) into three parts: the first is equality (19), and the other two are given in our main theorem below. We formulated all our statements in a manner that allows us to treat the cases when λ/k or $\widehat{\lambda}/\widehat{k}$ are on a boundary separating two of our chambers in Δ.

Theorem 4.8 *Let $\lambda \in \Lambda$ and $k \in \mathbb{Z}^{>0}$ be such that $\lambda/k \in \Delta$. Let ς and $\widehat{\varsigma} \in \Delta$ be regular elements, specifying two chambers in Δ, which contain λ/k and $\widehat{\lambda}/\widehat{k}$ in their closures, correspondingly. Then for any diagonal basis $\mathcal{D}$, the following two equalities hold*:

(I) $$\chi(P_0(\varsigma), \mathscr{L}(k;\lambda)) = \widetilde{N}_{r,k} \sum_{\boldsymbol{B}\in\mathcal{D}} \mathrm{iBer}_{\boldsymbol{B}}[w_\Phi^{1-2g}(x/\widehat{k})](\widehat{\lambda}/\widehat{k} - [\varsigma]_{\boldsymbol{B}}),$$

(II) $$\sum_{\boldsymbol{B}\in\mathcal{D}} \mathrm{iBer}_{\boldsymbol{B}}[w_\Phi^{1-2g}(x/\widehat{k})]4\widehat{\lambda}/\widehat{k} - [\varsigma]_{\boldsymbol{B}}) = \sum_{\boldsymbol{B}\in\mathcal{D}} \mathrm{iBer}_{\boldsymbol{B}}[w_\Phi^{1-2g}(x/\widehat{k})](\widehat{\lambda}/\widehat{k} - [\widehat{\varsigma}]_{\boldsymbol{B}}).$$

Remark 4.9 Part (I) of the theorem implies that if $\lambda/k \in \Delta$ is not regular, then

$$\chi(P_0(c^+), \mathscr{L}(k;\lambda)) = \chi(P_0(c^-), \mathscr{L}(k;\lambda))$$

for regular $c^\pm \in \Delta$ in two neighboring chambers that contain λ/k in their closure; see Proposition 10.1 and Remark 10.4.

Before we proceed, we formulate a mild generalization of part (I) of our theorem. As observed above, if we fix a generic $c \in \Delta$, and vary (λ, k) in such a way that $\lambda/k \sim c$, then both sides of the equality (I) are manifestly polynomial, and thus we can extend the validity of this equality as follows.

Corollary 4.10 *Let $c \in \Delta$ be a regular element, which thus specifies a chamber in Δ and a parabolic moduli space $P_0(c)$ as well. Then for a diagonal basis $\mathcal{D}$, an arbitrary weight $\lambda \in \Lambda$, and a positive integer k, we have*

$$\chi(P_0(c), \mathcal{L}(k;\lambda)) = \widetilde{N}_{r,k} \sum_{\boldsymbol{B} \in \mathcal{D}} \operatorname{iBer}_{\boldsymbol{B}}[w_\Phi^{1-2g}(x/\widehat{k})](\widehat{\lambda}/\widehat{k} - [c]_{\boldsymbol{B}}). \tag{22}$$

Example 4.11 Let us write down these formulas in case of $r = 3$ explicitly. Let $\mathcal{D}$ be the diagonal basis from Example 3.3; then using Remark 4.6, we obtain

$$\chi(P_0(<), \mathcal{L}(k;\lambda)) = (-1)^{g-1}(3(k+3)^2)^g \operatorname*{Res}_{y=0}\operatorname*{Res}_{x=0} \frac{e^{\lambda_1 x+(\lambda_1+\lambda_2)y+x+y} - e^{\lambda_1 x+(\lambda_1+\lambda_3)y+x}}{(1-e^{x(k+3)})(1-e^{y(k+3)})w_\Phi(x,y)^{2g-1}}\,dx\,dy,$$

$$\chi(P_0(>), \mathcal{L}(k;\lambda)) = (-1)^{g-1}(3(k+3)^2)^g \operatorname*{Res}_{y=0}\operatorname*{Res}_{x=0} \frac{e^{\lambda_1 x+(\lambda_1+\lambda_2)y+x+y} - e^{\lambda_1 x+(\lambda_1+\lambda_3)y+x+y(k+3)}}{(1-e^{x(k+3)})(1-e^{y(k+3)})w_\Phi(x,y)^{2g-1}}\,dx\,dy,$$

where $w_\Phi(x,y) = 2\sinh\left(\frac{1}{2}x\right)2\sinh\left(\frac{1}{2}y\right)2\sinh\left(\frac{1}{2}(x+y)\right)$.

4.4 The walls

Our first step is to identify the wall-crossing terms of the residue formula (22), which originate in the discontinuities of the function $c \mapsto \{c\}_{\boldsymbol{B}}$. These discontinuities occur on "walls": the affine hyperplanes (15). The following is straightforward:

Lemma 4.12 *Let $S_{\Pi,l}$ be the wall defined by* (15)*, and $\boldsymbol{B} = (\beta^{[1]}, \dots, \beta^{[r-1]}) \in \mathcal{B}$ an ordered basis of V^*. Then, as a function of c, the fractional part function $\{c\}_{\boldsymbol{B}}$ has a discontinuity at a generic point of the wall $S_{\Pi,l}$ exactly when* Tree($\boldsymbol{B}$) (see page 2270) *is a union of a tree on Π', a tree on Π'' (the enumeration of the edges is irrelevant here) and a single edge (which we will call the* **link**) *connecting Π' and Π''.*

Notation We will denote the element of $\boldsymbol{B}$ corresponding to this edge by β_{link}; this vector thus depends on $\boldsymbol{B}$ and the partition Π.

Proof We fix $\boldsymbol{B}$, and note that for our purposes, $c \in S_{\Pi,l}$ will be considered *generic* if it belongs to only this one wall in Δ; this is equivalent to the condition that the only nontrivial subsets of coordinates of c which sum up to an integer are Π' and Π''.

Note that an element

$$c = \sum_{j=1}^{r-1} b_j \beta^{[j]} \in \Delta$$

is a point of discontinuity of the fractional part function $\{\cdot\}_{\boldsymbol{B}}$ if and only if $b_j \in \mathbb{Z}$ for some $1 \le j \le r-1$.

Next, we express the coefficient b_j via the coordinates of c: we show that for all $1 \le j \le r-1$ we have

$$b_j = \sum_{i \in \Psi_j} c_i \quad \text{for some subset } \Psi_j \subset \{1, \dots, r\}. \tag{23}$$

Now we orient the edges of Tree($\boldsymbol{B}$) in such way that they are all directed "away" from the root vertex r, and, without loss of generality (recall that $\sum c_i = 0$), we can assume that this orientation agrees with the signs of the elements $\beta^{[j]} \in \boldsymbol{B}$. It is easy to verify then that the subset

$$\Psi_j = \big\{k \in \{1, \dots, r\} \mid \text{the unique directed path in Tree}(\boldsymbol{B}) \text{ from } r \text{ to } k \text{ contains the edge corresponding to } \beta^{[j]}\big\}$$

satisfies (23).

Hence we can conclude that if $c \in S_{\Pi,l}$ is generic and the coefficient b_j is an integer, then necessarily $\Psi_j = \Pi'$, and thus $\Pi'' = \{1, \dots, r\} \setminus \Psi_j$, and cutting the edge corresponding to $\beta^{[j]}$ from Tree($\boldsymbol{B}$) results in two disjoint trees, on Π' and on Π'', respectively. □

Now choose two regular elements $c^+, c^- \in V^*$ in two neighboring chambers separated by the wall $S_{\Pi,l}$, in such a way that

$$[c^+_{\Pi'}] = l \quad \text{and} \quad [c^-_{\Pi'}] = l - 1, \tag{24}$$

where

$$c_{\Pi'} \stackrel{\text{def}}{=} \sum_{i \in \Pi'} c_i,$$

and, as usual, $[q]$ stands for the integer part of the real number q. Now introduce the notation

$$p_\pm(k;\lambda) = \widetilde{N}_{r,k} \sum_{\boldsymbol{B} \in \mathcal{D}} \operatorname*{iBer}_{\boldsymbol{B}}[w_\Phi^{1-2g}(x/\widehat{k})](\widehat{\lambda}/\widehat{k} - [c^\pm]_{\boldsymbol{B}})$$

for the two polynomial functions in (k, λ) corresponding to c^+ and c^-, respectively. We define the *wall-crossing term* in our residue formula (22) as the difference between these two polynomials:

$$p_+(k;\lambda) - p_-(k;\lambda).$$

Using Lemma 4.1 and (24), we obtain the following simple residue formula for this difference.

Lemma 4.13 *Let (Π, l), c^+ and c^- be as above, and let us fix a diagonal basis $\mathcal{D} \subset \mathcal{B}$. Denote by $\mathcal{D}|\Pi$ the subset of those elements of $\mathcal{D}$, which satisfy the condition described in Lemma 4.12. Then*

$$p_+(k,\lambda) - p_-(k,\lambda) = \widetilde{N}_{r,k} \sum_{\boldsymbol{B} \in \mathcal{D}|\Pi} \operatorname*{iBer}_{\boldsymbol{B}}\big[\big(1 - \exp(\beta_{\text{link}}(x))\big) w_\Phi^{1-2g}(x/\widehat{k})\big](\lambda/\widehat{k} - [c^+]_{\boldsymbol{B}}), \tag{25}$$

where β_{link} is the "link" element of $\boldsymbol{B}$ (depending on Π and $\boldsymbol{B}$) defined after Lemma 4.12.

Remark 4.14 The multiplication by $1 - \exp(\beta_{\text{link}}(x))$ in (25) has the effect of canceling one of the factors in the denominator in the definition (16) of the operation iBer.

Example 4.15 Calculating the difference of two polynomials from Example 4.11, we obtain the wall-crossing term for the rank-3 case:

$$p_-(k;\lambda) - p_+(k;\lambda) = (-3(k+3)^2)^g \operatorname*{Res}_{y=0} \operatorname*{Res}_{x=0} \frac{e^{\lambda_1 x + (\lambda_1+\lambda_3)y + x}}{(1-e^{x(k+3)})w_\Phi(x,y)^{2g-1}}\, dx\, dy.$$

4.5 Wall-crossing and diagonal bases

Now we pass to the study of the combinatorial object $\mathcal{D}|\Pi$ defined in Lemma 4.13. One thing we will discover is that even though each diagonal basis consists of $(r-1)!$ elements and the right-hand side of (25) does not depend on the choice of $\mathcal{D}$, the number of elements in $\mathcal{D}|\Pi$ might vary with $\mathcal{D}$.

First we look at the case of the Hamiltonian basis $\mathcal{H}_1$. Form now on, we will use the notation $|\Pi'| = r'$ and $|\Pi''| = r''$ for a nontrivial partition $\Pi = (\Pi', \Pi'')$ — recall the convention $r \in \Pi''$. The following statement is easy to verify.

Lemma 4.16 *Let $\Pi = (\Pi', \Pi'')$ be a nontrivial partition such that $1 \in \Pi'$ (the other case is analogous). Then*

$$\mathcal{H}_1|\Pi = \{\sigma(\boldsymbol{B}) \mid \sigma(1) = 1 \text{ and } \sigma(\Pi') \in \Pi'\}.$$

In particular, $\big|\mathcal{H}_1|\Pi\big| = (r'-1)!\cdot r''!$.

It turns out that for our geometric applications, instead of $\mathcal{H}_1$, we will need to choose a particular nbc basis, where the ordering is chosen to be consistent with Π.

To simplify our terminology, we will use the language of graphs and edges introduced in Section 3.3, and we will think of $\alpha^{ij} \in \Phi$ as an edge in the complete graph on r vertices. To define the ordering υ, we need to choose an edge between Π', Π''; the choice is immaterial, but for simplicity we settle for $m \stackrel{\text{def}}{=} \max\{i \in \Pi'\}$ and $r \in \Pi''$, and set $\beta_{\text{link}} = \alpha^{m,r}$ to be the smallest element according to υ.

The υ–ordered list of edges thus starts with β_{link}, and then continues with the remaining $r'\cdot r'' - 1$ edges connecting Π' and Π''. Next we list the $\frac{1}{2}r'(r'-1)$ edges connecting vertices in Π' in any order, and finally, we list the remaining edges, those connecting vertices in Π''.

Notation We introduce the natural notation Φ' and Φ'' for the $A_{r'}$ and $A_{r''}$ root systems corresponding to Π' and Π'', and we denote by $\mathcal{D}[\upsilon]$, $\mathcal{D}'[\upsilon]$ and $\mathcal{D}''[\upsilon]$ the diagonal nbc bases induced by the ordering υ on Φ, Φ' and Φ'', respectively.

The following is easy to verify.

Lemma 4.17 *Given elements $\boldsymbol{B}' \in \mathcal{D}'[\upsilon]$ and $\boldsymbol{B}'' \in \mathcal{D}''[\upsilon]$, we can define an element of $\mathcal{D}[\upsilon]$ as follows: we start with β_{link}, then append $\boldsymbol{B}'$, and then continue with $\boldsymbol{B}''$. This construction creates a one-to-one correspondence*

$$\mathcal{D}'[\upsilon] \times \mathcal{D}''[\upsilon] \to \mathcal{D}[\upsilon]|\Pi. \tag{26}$$

In particular, $\big|\mathcal{D}[\upsilon]|\Pi\big| = (r'-1)!\cdot(r''-1)!$.

Finally, putting Lemmas 4.13 and 4.17 together, we arrive at the following elegant statement.

Proposition 4.18 *Let (Π, l), c^+ and c^- be as in Lemma 4.13, and let $\mathscr{D}'$ and $\mathscr{D}''$ be diagonal bases of Φ' and Φ'', correspondingly. Then*

$$p_+(k;\lambda)-p_-(k;\lambda)=(k+r)\widetilde{N}_{r,k}\sum_{\boldsymbol{B}'\in\mathscr{D}'}\sum_{\boldsymbol{B}''\in\mathscr{D}''}\operatorname*{Res}_{\beta_{\mathrm{link}}=0}\operatorname*{iBer}_{\boldsymbol{B}'}\operatorname*{iBer}_{\boldsymbol{B}''}[w_\Phi^{1-2g}(x/\widehat{k})](\widehat{\lambda}/\widehat{k}-[c^+]_{\boldsymbol{B}})\,d\beta_{\mathrm{link}}, \tag{27}$$

where $\operatorname{Res}_{\beta_{\mathrm{link}}=0}\operatorname{iBer}_{\boldsymbol{B}'}\operatorname{iBer}_{\boldsymbol{B}''}\,d\beta_{\mathrm{link}}$ is simply $\operatorname{iBer}_{\boldsymbol{B}}$ (see (16)) with $\boldsymbol{B}$ obtained by appending $\boldsymbol{B}'$, and then $\boldsymbol{B}''$ to β_{link}, and with the factor $(1-\exp\langle\beta_{\mathrm{link}},x\rangle)$ removed from the denominator.

Remark 4.19 The expression

$$\operatorname*{Res}_{\beta_{\mathrm{link}}=0}\operatorname*{iBer}_{\boldsymbol{B}'}\operatorname*{iBer}_{\boldsymbol{B}''}[w_\Phi^{1-2g}(x/\widehat{k})](\widehat{\lambda}/\widehat{k}-[c^+]_{\boldsymbol{B}})\,d\beta_{\mathrm{link}}$$

may equally be interpreted as follows. We write

$$\Lambda\ni\widehat{\lambda}/\widehat{k}-[c^+]_{\boldsymbol{B}}=m_{\mathrm{link}}\beta_{\mathrm{link}}+n'+n''$$

according to the splitting of $\boldsymbol{B}$, think of $w(x/\widehat{k})$ as a function in $\mathscr{F}_{\Phi''}$ with some fixed values of the parameters from $\boldsymbol{B}'$ and β_{link}, and then calculate

$$\operatorname*{iBer}_{\boldsymbol{B}''}[w_\Phi^{1-2g}(x/\widehat{k})](n'').$$

The result will be a rational function Q in the variables from $\boldsymbol{B}'$ and β_{link}, and we proceed to calculate $\operatorname{iBer}_{\boldsymbol{B}'}[Q](n')$ to obtain a function F in the variable β_{link}, and finally the answer is

$$\operatorname*{Res}_{\beta_{\mathrm{link}}=0}\exp(\beta_{\mathrm{link}})F(\beta_{\mathrm{link}})\,d\beta_{\mathrm{link}}.$$

We observe that since the trees $\operatorname{Tree}(\boldsymbol{B}')$ and $\operatorname{Tree}(\boldsymbol{B}'')$ are disjoint, the order of the application of the operations $\operatorname{iBer}_{\boldsymbol{B}'}$ and $\operatorname{iBer}_{\boldsymbol{B}''}$ is immaterial.

5 Wall-crossing in master space

Master spaces were introduced by Thaddeus in [28] in order to understand the dependence of GIT quotients on their linearizations. Following his footsteps, in this section we describe a simple but very effective method to control the changes in the Euler characteristics of line bundles when crossing a wall in the space of linearizations. (Similar results appeared in [8].)

5.1 Wall-crossing and holomorphic Euler characteristics

We begin by recalling the basic notions of geometric invariant theory.

Let X be a smooth projective variety over $\mathbb{C}$, and G a reductive group acting on X. A *linearization* of this action is a line bundle L on X with a lifting of the G–action to a linear action on L. An ample linearization is G–effective if L^n has a nonzero G–invariant section for some $n>0$; the space of such linearizations $\operatorname{Cone}_G(X)$ is called the *G–effective ample cone*.

For $L \in \mathrm{Cone}_G(X)$, we define the invariant-theoretic quotient $M_L = X /\!\!/^L G$ as the Proj of the graded ring of invariant sections of the powers of L:

$$M_L = \mathrm{Proj} \bigoplus_n H^0(X, L^n)^G.$$

According to Mumford's geometric invariant theory [18], there is a partition of X (depending on L)

$$X = X^{\mathrm{s}}[L] \cup X^{\mathrm{sss}}[L] \cup X^{\mathrm{us}}[L] \tag{28}$$

into the set of stable, strictly semistable, and unstable points such that there is a surjective map

$$(X^{\mathrm{s}}[L] \cup X^{\mathrm{sss}}[L])/G \to M_L.$$

When $X^{\mathrm{sss}}[L]$ is empty, this map is a bijection, and the quotient $M_L = X^{\mathrm{s}}[L]/G$ is a smooth orbifold.

In [7], Dolgachev and Hu studied the dependence of the GIT quotient $M_L = X /\!\!/^L G$ on L. They showed that $\mathrm{Cone}_G(X)$ is divided by hyperplanes, called walls, into finitely many convex chambers such that when L varies within a chamber, the partition (28) and thus the GIT quotient M_L remains unchanged. Moreover, an ample effective linearization lies on a wall precisely when it possesses a strictly semistable point.

Now let us consider two neighboring chambers, with smooth GIT quotients M_+ and M_-. We pick an arbitrary linearization $\mathscr{L}$ of the G–action on X, which descends to M_+ and M_-. This last condition means that if $S \subset G$ is the stabilizer of a generic point in X, then S acts trivially on the fibers of $\mathscr{L}$. We will call such linearizations *descending*.

Thus, given such a descending linearization $\mathscr{L}$ of the G–action on X, we obtain two line bundles: one on M_+ and one on M_-, which, by abuse of notation, we will denote by the same letter $\mathscr{L}$. Via taking Chern classes, this construction creates a correspondence between classes in $H^2(M_+, \mathbb{Z})$ and $H^2(M_-, \mathbb{Z})$, which we will assume to be an isomorphism of free $\mathbb{Z}$–modules. We will thus identify these lattices, and introduce the notation Γ for them:

$$\Gamma = H^2(M_+, \mathbb{Z}) \simeq H^2(M_-, \mathbb{Z}).$$

The walls mentioned above can be thought of as hyperplanes in $\Gamma_{\mathbb{R}} = \Gamma \otimes_{\mathbb{Z}} \mathbb{R}$.

Our goal in this section is to compare the holomorphic Euler characteristics $\chi(M_+, \mathscr{L})$ and $\chi(M_-, \mathscr{L})$, which are given by the Hirzebruch–Riemann–Roch theorem:

$$\chi(M_\pm, \mathscr{L}) = \int_{M_\pm} \exp(c_1(\mathscr{L}))\, \mathrm{Todd}(M_\pm).$$

As this expression is manifestly polynomial in $c_1(\mathscr{L})$, we obtain thus two polynomials on Γ, and our goal is to calculate their difference, the *wall-crossing term*

$$\chi(M_+, \mathscr{L}) - \chi(M_-, \mathscr{L}). \tag{29}$$

5.2 The master space construction

To simplify our setup, we will make some additional assumptions.

Assumptions 5.1 (a) The generic stabilizer of X is trivial.

(b) Let L_+ and L_- be two ample linearizations of the G–action on X from the adjacent chambers corresponding to the quotients M_+ and M_-. Without loss of generality, we can assume that the linearization $L_0 = L_+ \otimes L_-$ lies on the single wall separating the two chambers, and that the interval connecting $c_1(L_+)$ and $c_1(L_-)$ in $\Gamma_{\mathbb{R}} = \Gamma \otimes_{\mathbb{Z}} \mathbb{R}$ does not intersect any other walls.

(c) Let X^0 be the set of those semistable points $x \in X^{\mathrm{ss}}[L_0]$ which are not stable for $L_\pm$:

$$X^0 := X^{\mathrm{ss}}[L_0] \setminus (X^{\mathrm{s}}[L_+] \cup X^{\mathrm{s}}[L_-])$$

We assume that X^0 is smooth, and that for $x \in X^0$ the stabilizer subgroup $G_x \subset G$ is isomorphic to $\mathbb{C}^*$.

(d) Assume that there is a linearization $\vec{L}$ of the G–action on X such that $L_+ = L_- \otimes \vec{L}^n$ for some positive integer n, and such that for each $x \in X^0$, the stabilizer subgroup G_x acts freely on $\vec{L}_x \setminus 0$.

Now we introduce the *master space* construction of Thaddeus [28]. Consider the variety $Y = \mathbb{P}(\mathbb{O} \oplus \vec{L})$, which is a $\mathbb{P}^1$–bundle over X endowed with the additional $\mathbb{C}^*$–action $(1, t^{-1})$. As Y is a projectivization of a vector bundle on X, it comes equipped with $\mathbb{O}(1)$, which is the standard $G \times \mathbb{C}^*$–equivariant line bundle. To simplify our notation, we will denote the same way the linearizations of the G–action on X and their pullbacks (with tautological G–action) to Y.

The *master space* Z then is the GIT quotient of Y with respect to the linearization $L_-(n) = L_- \otimes \mathbb{O}(n)$:

$$Z = Y /\!\!/^{L_-(n)} G,$$

which inherits a $\mathbb{C}^*$–action from Y. Some additional notation:

- We denote this copy of $\mathbb{C}^*$ by T.
- We denote the projection $Y \to X$ by π, and the quotient map $Y^{\mathrm{s}} \to Z$ by ψ.
- Introduce the notation $Y(0:\cdot)$ and $Y(\cdot:0)$ for the two copies of X in Y, corresponding to the two poles of the projective line; then Y is partitioned into three sets

$$Y = Y(0:\cdot) \sqcup Y(\cdot:0) \sqcup \vec{L}^\circ,$$

where $\vec{L}^\circ$ is the line bundle $\vec{L}$ with the zero-section removed. We will write $\pi_\circ$ for the restriction of π to $\vec{L}^\circ$. We collect our maps in the following diagram:

(30)
$$\begin{array}{ccccc} \vec{L}^\circ & \hookrightarrow & Y = \mathbb{P}(\mathbb{O} \oplus \vec{L}) \supset Y^{\mathrm{s}} & \xrightarrow{\psi} & Z \\ & \searrow^{\pi_\circ} & \downarrow \pi & & \\ & & X & & \end{array}$$

Proposition 5.2 (i) *There are embeddings*

$$\iota_-\colon M_- \to Z \quad \text{and} \quad \iota_+\colon M_+ \to Z$$

obtained as the quotients $Y^{\mathrm{s}} \cap Y(\cdot\colon 0)/G$ and $Y^{\mathrm{s}} \cap Y(0\colon\cdot)/G$, correspondingly.

(ii) *The strictly semistable locus of Y with respect to the linearization $L_-(n)$ is empty, and the GIT quotient $Z = Y^{\mathrm{s}}/G$ is smooth.*

(iii) *There is an embedding $\iota_0\colon X^0/G \to Z$, obtained via $\psi(\pi_\circ{}^{-1}(X^0))$. We denote the image of ι_0 by Z^0.*

(iv) *The fixed point locus Z^T is the disjoint union of $\iota_+(M_+)$, $\iota_-(M_-)$ and Z^0.*

Proof Statements (i)–(iii) follow from [28, 4.2, 4.3]. To prove (iv), first note that $Y(\cdot\colon 0)$ and $Y(0\colon\cdot)$ are fixed by T, so we immediately obtain that $M_\pm \subset Z$ are fixed components. Also the G–action on Y commutes with the T–action, so a point $\psi(y) \in \psi(\pi_\circ^{-1}(X))$ is fixed by T if and only if the T–orbit $T\cdot y \subset \pi_\circ^{-1}(X)$ is contained in the G–orbit $G\cdot y \subset \pi_\circ^{-1}(X)$. Since $T\cdot y \subset \pi_\circ^{-1}(x)$ for some $x \in X$, we need $y \in \pi_\circ^{-1}(X^0)$. Moreover, for any $y \in \pi_\circ^{-1}(x) \subset \pi_\circ^{-1}(X^0)$, $T\cdot y = \pi_\circ^{-1}(x) = G_x \cdot y$, so a point $\psi(y) \in \psi(\pi_\circ^{-1}(X))$ is fixed by T if and only if $\psi(y) \in \psi(\pi_\circ^{-1}(X^0)) = Z^0$. □

Construction Given a G–equivariant vector bundle E on X, we can construct a T–equivariant vector bundle $\zeta(E) \to Z$ on Z by first pulling E back from X to Y, and endowing the resulting bundle π^*E with the trivial action of T, and the action of G pulled back from X. We then obtain $\zeta(E) \to Z$ by descending π^*E to Z.

Before we formulate our wall-crossing formula, we need one more ingredient: the identification of the normal bundles of the fixed-point components of Z.

Lemma 5.3 (i) *The normal bundle on the component M_+ of Z^T is $\zeta(\vec{L}^{-1})|_{M_+}$, and the normal bundle of M_- is $\zeta(\vec{L})|_{M_-}$.*

(ii) *Let $x \in X^0$, denote by G_x the stabilizer of x in the group G, and consider the point $\iota_0(x) \in Z^0$* (see Proposition 5.2(iii)). *Then the normal vector space of $Z^0 \subset Z$ at the point $\iota_0(x)$ is canonically T–equivariantly isomorphic to the T–vector space $\vec{L}_x^\circ \times_{G_x} N_x X^0$, where $N_x X^0$ is the vector space normal to $X^0 \subset X$ at x, and the $T \simeq \mathbb{C}^*$–action is induced by left multiplication by t^{-1} on $\vec{L}_x$.*

Proof Part (i) immediately follows from the formula for the tangent space of the projective line: $T\mathbb{P}(V) \simeq \mathrm{Hom}(S, Q)$, where $S \to V \to Q$ is the tautological sequence on $\mathbb{P}(V)$, the projectivization of the vector space V.

For part (ii), consider diagram (30); our goal is to identify the descent to Z_0 of the normal bundle $N_{\pi_\circ^{-1}X^0}$ to $\pi_\circ^{-1}X^0$ in $\vec{L}^\circ$. We only need to observe that this bundle may be identified with the pullback $\pi_\circ^* N X^0$ of the normal bundle to X^0 in X, endowed with the natural G–action and a T–action, which is trivial on the fibers. □

Remark 5.4 Restricting the operator ζ to X^0, we can construct a T–equivariant vector bundle on Z^0 from a G–equivariant vector bundle on X^0. Then the normal bundle N_{Z^0} of $Z^0 \subset Z$ may be also described as $\zeta|_{X^0}(NX^0)$. The T–weights of the action may be computed by fixing $x \in X^0$, identifying the stabilizer subgroup $G_x \subset G$ with T via its action on the fiber $\vec{L}_x$, and then considering the action of G_x on $N_x X^0$.

Lemma 5.5 *The restriction of the line bundle $\zeta(\vec{L})$ to Z^0 is trivial with T–weight* 1.

Proof Note that $\pi_\circ^* \vec{L}$ admits a G–equivariant tautological nonvanishing section. For calculating the weight, we observe that while T acts on $\vec{L}_x$ with weight -1, the T–weight of $\vec{L}_x^\circ \times_{G_x} \vec{L}$ is $+1$. □

Definition 5.6 Given a T–vector bundle V on a manifold on which T acts trivially, the T–equivariant *K-theoretical Euler class* of V^*, which we denote by $E_t(V)$, may be described as follows: let $x_1, \dots, x_n$ be the Chern roots of V, and $l_1, \dots l_n \in \mathbb{Z}$ be the corresponding T–weights. Then

$$E_t(V) = \prod_{j=1}^{n} (1 - t^{-l_j} \exp(-x_j)).$$

Now we are ready to write down our wall-crossing formula for (29). A key role will be played by the following notion: given a rational differential one-form on the Riemann sphere, let us denote taking the sum of residues at 0 and at infinity by $\mu \mapsto \operatorname{Res}_{t=0,\infty} \mu$:

$$\operatorname*{Res}_{t=0,\infty} \stackrel{\text{def}}{=} \operatorname*{Res}_{t=0} + \operatorname*{Res}_{t=\infty}.$$

Theorem 5.7 *Let $\mathscr{L}$ be a linearization of the G–action on X and denote by $\zeta(\mathscr{L})$, as above, the T–equivariant line bundle on Z obtained by pullback to Y and descent to Z. If Assumptions 5.1 hold, then*

$$\chi(M_+, \mathscr{L}) - \chi(M_-, \mathscr{L}) = \operatorname*{Res}_{t=0,\infty} \int_{Z^0} \frac{\operatorname{ch}_t(\zeta(\mathscr{L})|_{Z^0})}{E_t(N_{Z^0})} \operatorname{Todd}(Z^0) \frac{dt}{t}, \tag{31}$$

where N_{Z^0} is the T–equivariant bundle on Z^0 described in Lemma 5.3, ch_t is the T–equivariant Chern character, and $E_t(N_{Z^0})$ is the K-theoretical Euler class of $N_{Z^0}^*$.

Proof The Atiyah–Bott fixed-point formula [2] applied to the line bundle $\mathscr{L}$ on our master space Z yields

$$\chi_t(Z, \mathscr{L}) = \sum_{F \subset Z^T} \int_F \frac{\operatorname{ch}_t(\zeta(\mathscr{L})|_F)}{E_t(N_F)} \operatorname{Todd}(F), \tag{32}$$

where the sum is taken over the connected components of the fixed-point locus Z^T.

In Proposition 5.2, we identified these components as M_+, M_- and Z^0. Lemma 5.3 identifies the equivariant normal bundles of M_+ and M_-, and thus the corresponding contributions are

$$\int_{M_+} \frac{\mathrm{ch}(\mathscr{L})\,\mathrm{Todd}(M_+)}{1-t^{-1}\exp(c_1(\vec{L}))} \quad \text{and} \quad \int_{M_-} \frac{\mathrm{ch}(\mathscr{L})\,\mathrm{Todd}(M_-)}{1-t\exp(-c_1(\vec{L}))}.$$

We observe that $\chi_t(Z,\mathscr{L})$ is a Laurent polynomial in t since it is the alternating sum of T–characters of finite-dimensional vector spaces. Thus, as a function of t, $\chi_t(Z,\mathscr{L})$ has poles only at $t=0,\infty$, and by the residue theorem, we have

$$\operatorname*{Res}_{t=0,\infty} \chi_t(Z,\mathscr{L})\frac{dt}{t} = 0.$$

On the other hand, since

$$\operatorname*{Res}_{t=0,\infty} \frac{A}{1-t^{-1}B}\frac{dt}{t} = -A \quad \text{and} \quad \operatorname*{Res}_{t=0,\infty} \frac{A}{1-tB}\frac{dt}{t} = A,$$

we have

$$\operatorname*{Res}_{t=0,\infty} \int_{M_+} \frac{\mathrm{ch}(\mathscr{L})\,\mathrm{Todd}(M_+)}{1-t^{-1}\exp(c_1(\vec{L}))}\frac{dt}{t} = -\chi(M_+,L),$$
$$\operatorname*{Res}_{t=0,\infty} \int_{M_-} \frac{\mathrm{ch}(\mathscr{L})\,\mathrm{Todd}(M_-)}{1-t\exp(-c_1(\vec{L}))}\frac{dt}{t} = \chi(M_-,L).$$

Now, applying the functional $\mathrm{Res}_{t=0,\infty}$ to the two sides of (32) multiplied by dt/t, we obtain the desired result (31). □

6 Wall-crossings in parabolic moduli spaces

In this section, we apply Theorem 5.7 to wall-crossings in the moduli space of parabolic bundles.

From now on, we assume that $d=0$, and we write Δ for the corresponding set of admissible parabolic weights Δ_0. Recall from Section 2.2 that for regular $c\in\Delta$, the moduli space of stable parabolic bundles $P_0(c)$ is the GIT quotient $XQ/\!/^c\mathrm{PSL}(\chi)$, where XQ is a subspace of the total space of a flag bundle over the Quot scheme. Let us fix a partition $\Pi=(\Pi',\Pi'')$ and an integer l, and introduce the notation Δ'_l and Δ''_{-l} for the simplices of parabolic weights of Π' and Π''. Let $\phi\in\Sigma_r$ be the unique permutation which sends $\{1,\dots,r'\}$ to Π' preserving the order of the first r' and the last r'' elements. We choose $c^0=(c^0_1,\dots,c^0_r)\in S_{\pi,l}$ and two regular elements $c^+,c^-\in\Delta$ in two neighboring chambers separated by the wall $S_{\Pi,l}$, such that

$$c^\pm = c^0 \pm \epsilon(\dots,0,1,0,\dots,0,-1)$$

for some positive $\epsilon\in\mathbb{Q}$, where 1 and -1 are on the $\phi(r')^{\text{th}}$ and r^{th} places, respectively. Let

$$c' = \sum_{i\in\Pi'} c^0_i x_i \in \Delta'_l \quad \text{and} \quad c'' = \sum_{i\in\Pi''} c^0_i x_i \in \Delta''_{-l}.$$

For $(k,\lambda)\in\mathbb{Z}\times\Lambda$, consider the polynomials

$$q_\pm(k,\lambda) = \chi(P_0(c^\pm),\mathscr{L}_0(k;\lambda)).$$

Our goal is to calculate the difference of these two polynomials.

Notation To simplify our notation, from now on, we omit the index t from the symbols for equivariant characteristic classes.

6.1 The master space construction

We construct the master space Z from Section 5.2 using the following data:

- A smooth variety $X = XQ$; see Section 2.2.
- Linearizations $L^{\pm} = L(k; \lambda^{\pm})$ of the G–action on X (see Section 2.2) such that $\lambda^{\pm}/k = c^{\pm}$.
- The linearization $\vec{L} = L(0; x_{\phi(r')} - x_r)$ of the G–action on X.

The following statement is easy to verify.

Lemma 6.1 [6, Section 3.2] *The subset $X^0 \subset X$ is the set of points representing vector bundles W on C such that W splits as a direct sum $W' \oplus W''$, where W' and W'' are, respectively, c'– and c''–stable parabolic bundles. Therefore, we have the following description of the locus Z^0:*

$$Z^0 = \{W = W' \oplus W'' \mid W' \in \widetilde{P}_l(c'),\ W'' \in \widetilde{P}_{-l}(c''),\ \det(W) \simeq \mathbb{O}\}.$$

Remark 6.2 The locus Z^0 is fibered over Jac^l with fiber $P_l(c') \times P_{-l}(c'')$ by the determinant map $\widetilde{P}_l(c') \to \mathrm{Jac}^l$, and

$$H^*(Z^0, \mathbb{Q}) \simeq H^*(P_l(c') \times P_{-l}(c''), \mathbb{Q}) \otimes H^*(\mathrm{Jac}^l, \mathbb{Q}). \tag{33}$$

Remark 6.3 If the rank of the vector bundle $W \in \widetilde{P}_l(c)$ is 1, then $c = l$ and $\widetilde{P}_l(l)$ is isomorphic to Jac^l, while $P_l(l)$ is a point.

Now we need to verify the hypotheses of Theorem 5.7. Note that in our present construction X is not projective; however, it contains all semisimple points of the flag bundle over the open subscheme of the Quot scheme parametrizing locally free quotients (see Section 2.2) for all possible polarizations, and hence the missing points of the Quot scheme have no effect on any of our constructions (a similar argument appeared in [28]).

Assumptions 5.1(a)–(b) are trivially satisfied, so we study the action of the stabilizer $G_x \subset \mathrm{PSL}(N)$ of point $x \in X$ on the fiber $\vec{L}_x \setminus 0$.

- For a general point $x \in X$ the stabilizer of x is the center $\mathbb{Z}_N \subset \mathrm{SL}(N)$, which acts trivially on the fiber $\vec{L}_x \setminus 0$.

- For $x \in X^0$, any element of the stabilizer of x induces an automorphism of the corresponding vector bundle $W = W' \oplus W''$, so the stabilizer of x in $\mathrm{GL}(N)$ is isomorphic to $\mathbb{C}^* \times \mathbb{C}^* \subset \mathrm{GL}(N)$. An element $(t_1, t_2) \in \mathbb{C}^* \times \mathbb{C}^*$ is in $\mathrm{SL}(N)$ if and only if $t_1^{N'} t_2^{N''} = 1$, where $N' = \chi(W')$ and $N'' = \chi(W'')$. Note that (t_1, t_2) acts on $\vec{L}_x$ as $t_1 t_2^{-1}$, and we need $t_1 = t_2$ (hence $t_1^N = 1$) for this action to be trivial, so the stabilizer of any point in $\vec{L}_x \setminus 0$ is the center $\mathbb{Z}_N \subset \mathrm{SL}(N)$.

Then the action of $G = \mathrm{PSL}(N)$ is free on $Y \setminus (Y(0:\cdot) \cup Y(\cdot:0))$, and the action of $G_x \subset \mathrm{PSL}(N)$ on $\vec{L}_x \setminus 0$ induces an isomorphism $G_x \simeq \mathbb{C}^* \simeq T$.

Now by Theorem 5.7, the wall-crossing polynomial $q_-(k;\lambda) - q_+(k;\lambda)$ is equal to

$$\operatorname*{Res}_{t=0,\infty} \int_{Z^0} \frac{\mathrm{ch}(\mathscr{L}_0(k;\lambda)|_{Z^0})}{E(N_{Z^0})} \,\mathrm{Todd}(Z^0)\, \frac{dt}{t}. \tag{34}$$

Note that in our case, the T–action on Z is free outside the fixed locus Z^T, so as a function in $t \in T$, the integral in (34) may have poles only at $t = 0, 1, \infty$. Then, using the residue theorem and substituting $t = e^u$, we conclude that (34) equals

$$-\operatorname*{Res}_{u=0} \int_{Z^0} \frac{\mathrm{ch}(\mathscr{L}_0(k;\lambda)|_{Z^0})}{E(N_{Z^0})} \,\mathrm{Todd}(Z^0)\, du, \tag{35}$$

and thus our goal is to calculate this integral.

Our first step is to identify the characteristic classes under the integral sign; see Proposition 6.11 for the result.

We start with the study of the restriction of the line bundle $\mathscr{L}_0(k;\lambda)$ to the fixed locus $Z^0 \subset Z$. First, we describe a parametrization of the factor $H^*(\mathrm{Jac}^l, \mathbb{Q})$ in (33). Let $\mathscr{P}$ be the Poincaré bundle over $\mathrm{Jac} \times C$ such that $c_1(\mathscr{P})_{(0)} = 0$; define $\eta \in H^2(\mathrm{Jac})$ by $(\sum_i c_1(\mathscr{P})_{(e_i)} \otimes e_i)^2 = -2\eta \otimes \omega$ (see Section 2.3), then (see [33]) for any $m \in \mathbb{Z}$,

$$\int_{\mathrm{Jac}} e^{\eta m} = m^g. \tag{36}$$

As Z^0 is a connected component of the fixed locus of the T–action on Z, its equivariant cohomology factors: $H_T^*(Z^0) \simeq H^*(Z^0) \otimes \mathbb{C}[u]$. In particular, there are canonical embeddings $H^*(Z^0) \hookrightarrow H_T^*(Z^0)$ and $\mathbb{C}[u] \hookrightarrow H_T^*(Z^0)$.

Remark 6.4 It follows from Lemma 5.5 that $c_1(\zeta(\vec{L})|_{Z^0}) = u$.

Recall that for a parabolic weight $c = (c_1, \dots, c_r) \in \Delta$, we have set $c_{\Pi'} = \sum_{i \in \Pi'} c_i$.

Lemma 6.5 *Let $\lambda = (\lambda_1, \dots, \lambda_r) \in \Lambda$, $k \in \mathbb{Z}^{>0}$ and let $\Pi = (\Pi', \Pi'')$ be a nontrivial partition with $r \in \Pi''$. Let*

$$\lambda' = \sum_{i \in \Pi'} \lambda_i x_i \quad \text{and} \quad \lambda'' = \sum_{i \in \Pi''} \lambda_i x_i,$$

and define δ by $(\lambda/k)_{\Pi'} = l + \delta$. Then

$$\mathrm{ch}(\mathscr{L}_0(k;\lambda)|_{Z^0}) = e^{k\delta u}\exp\left(\frac{\eta k}{r'} + \frac{\eta k}{r''}\right)\mathrm{ch}(\mathscr{L}_l(k;\lambda'_1,\dots,\lambda'_{r'} - k\delta)\boxtimes\mathscr{L}_{-l}(k;\lambda''_1,\dots,\lambda''_{r''} + k\delta)),$$

where $\boxtimes$ denotes the external tensor product of line bundles on $P_l(c')\times P_{-l}(c'')$.

Proof First, note that

$$\mathrm{ch}(\mathscr{L}_0(0;\lambda)|_{Z^0}) = e^{k(l+\delta)u}\,\mathrm{ch}(\mathscr{L}_l(0;\lambda'_1,\dots,\lambda'_{r'} - kl - k\delta)\boxtimes\mathscr{L}_{-l}(0;\lambda''_1,\dots,\lambda''_{r''} + kl + k\delta)),$$

and thus it will be sufficient to identify the restriction of $\mathscr{L}(k;0)$. It follows from Lemma 2.8 that

$$c_1(\mathscr{L}_0(k;0)) = \frac{k}{2r}c_2(\mathrm{End}_0(U))_{(2)}.$$

Note that

$$c_2(\mathrm{End}_0(U))_{(2)} = -k\,\mathrm{ch}_2(U)_{(2)} + c_1^2(U)_{(2)} = -k\,\mathrm{ch}_2(U)_{(2)},$$

and thus

$$c_1(\mathscr{L}_0(k;0)) = -k\,\mathrm{ch}_2(U)_{(2)}.$$

Denote by $\widetilde{U}'$ and $\widetilde{U}''$ the normalized (see Section 2.3) universal bundles over $\widetilde{P}_l(c')\times C$ and $\widetilde{P}_{-l}(c'')\times C$, respectively. Since

$$\mathrm{ch}_2(U|_{Z^0})_{(2)} = \mathrm{ch}_2(\widetilde{U}'\otimes\zeta(\vec{L})|_{Z^0})_{(2)} + \mathrm{ch}_2(\widetilde{U}''|_{Z^0})_{(2)},$$

we have (see Remark 6.4)

$$\begin{aligned}(37)\qquad c_1(\mathscr{L}_0(k;0)|_{Z^0}) &= -k\,\mathrm{ch}_2(\widetilde{U}')_{(2)} - k\,u\,c_1(\widetilde{U}')_{(2)} - k\,\mathrm{ch}_2(\widetilde{U}'')_{(2)}\\ &= \frac{k}{2r'}c_2(\widetilde{U}')_{(2)} - \frac{k}{2r'}c_1^2(\widetilde{U}')_{(2)} + \frac{k}{2r''}c_2(\widetilde{U}'')_{(2)} - \frac{k}{2r''}c_1^2(\widetilde{U}'')_{(2)} - klu.\end{aligned}$$

Now, since

$$c_1^2(\widetilde{U}')_{(2)} = 2l\,c_1(U') - 2\eta \quad\text{and}\quad c_1^2(\widetilde{U}'')_{(2)} = -2l\,c_1(U'') - 2\eta,$$

by Lemma 2.8 we have

$$\begin{aligned}c_1(\mathscr{L}_0(k;0)|_{Z^0}) &= \frac{k}{r'}c_1(\mathscr{L}_l(r';l,\dots,l)) - \frac{kl}{r'}c_1(U') + \eta\frac{k}{r'} + \frac{k}{r''}c_1(\mathscr{L}_{-l}(r';-l,\dots,-l))\\ &\qquad + \frac{kl}{r''}c_1(U'') + \eta\frac{k}{r''} - klu\\ &= c_1\big(\mathscr{L}_l(k;(0,\dots,0,kl))\big) + c_1\big(\mathscr{L}_{-l}(k;(0,\dots,0,-kl))\big) + \eta\left(\frac{k}{r'} + \frac{k}{r''}\right) - klu,\end{aligned}$$

and this completes the proof. □

Lemma 6.6 *Denote by $\widetilde{U}'$ and $\widetilde{U}''$ the normalized* (see Section 2.3) *universal bundles over $\widetilde{P}_l(c')\times C$ and $\widetilde{P}_{-l}(c'')\times C$, and denote by π the projections along C. Then the T–equivariant normal bundle to the fixed locus $Z^0\subset Z$ is*

$$(38)\qquad N_{Z^0} = R^1_T\pi_*(\mathrm{ParHom}(\widetilde{U}',\widetilde{U}'')) \oplus R^1_T\pi_*(\mathrm{ParHom}(\widetilde{U}'',\widetilde{U}')),$$

where the $T\simeq\mathbb{C}^*$–action has weights -1 and $+1$ on the two summands, respectively.

Remark 6.7 As we are working with fixed determinant moduli spaces, the pushforwards in (38) are to be taken along the curve C in the part of $\widetilde{P}_l(c')\times\widetilde{P}_{-l}(c'')\times C$, where $\det(W')\det(W'')\simeq\mathbb{C}$; see Lemma 6.1.

Proof According to Lemma 5.3, for any point $x\in X^0$, the normal bundle N_{Z^0} at the point $\iota_0(x)\in Z^0$ may be identified with the T–vector space $\vec{L}_x^\circ\times_{G_x}N_xX^0$, where N_xX^0 is the normal bundle to $X^0\subset X$ at x, with the T–action induced by left multiplication by t^{-1} on $\vec{L}_x$.

Denote by UQ the universal bundle over X, which descends to the normalized universal bundles on $P_0(c^\pm)$. Recall that any point $x\in X^0$ represents a vector bundle which splits as a direct sum of two subbundles, hence we have $UQ_x=U_x^+\oplus U_x^-$, and

$$N_xX^0=H^1(C,\mathrm{ParHom}(U_x^+,U_x^-))\oplus H^1(C,\mathrm{ParHom}(U_x^-,U_x^+)).$$

(See [20, Proposition 1.13] for the description of the deformation space of parabolic bundles.) A simple calculation (see Remark 5.4 and Lemma 5.5) shows we have a T–module isomorphism

$$\vec{L}_x^\circ\times_{G_x}H^1(C,\mathrm{ParHom}(U_x^+,U_x^-))\simeq\vec{L}_x\otimes H^1(C,\mathrm{ParHom}(U_x^+,U_x^-))$$

with T–weight -1 induced by multiplication on $\vec{L}_x$ and trivial action on U_x^+ and U_x^-; applying the projection formula we obtain that

$$\vec{L}_x\otimes H^1(C,\mathrm{ParHom}(U_x^+,U_x^-))\simeq H^1_T(C,\mathrm{ParHom}(U_x^+\otimes\vec{L}_x^{-1},U_x^-)).$$

Similarly, we have

$$\begin{aligned}\vec{L}_x^\circ\times_{G_x}H^1(C,\mathrm{ParHom}(U_x^-,U_x^+))&\simeq\vec{L}_x^{-1}\otimes H^1(C,\mathrm{ParHom}(U_x^-,U_x^+))\\&\simeq H^1_T(C,\mathrm{ParHom}(U_x^-,U_x^+\otimes\vec{L}_x^{-1}))\end{aligned}$$

with T–action of weight 1.

Finally, we observe that according to our normalizations, the bundles $U^+\otimes\vec{L}^{-1}$ and U^- descend to the normalized universal bundles $\widetilde{U}'$ and $\widetilde{U}''$ over $\widetilde{P}_l(c')\times C$ and $\widetilde{P}_{-l}(c'')\times C$, respectively, and this completes the proof. □

6.2 Calculation of the characteristic classes of N_{Z^0}

Before we calculate the equivariant K-theoretical Euler class of the conormal bundle $N^*_{Z^0}$, we need to introduce some notation. Recall that for $1\le i,j\le r$, the differences $x_i-x_j\in V^*$ are linear functions on V, and the function x_i-x_j corresponds to the linearization $L_0(0;x_i-x_j)$ on X, which descends to the line bundle $\mathscr{L}_0(0;x_i-x_j)$ on the moduli space $P_0(c)$; see Section 2.2. As in Section 5.2, we denote by $\zeta(L_0(0;x_i-x_j))$ the line bundle on Z obtained by the pullback and then descent. This way, we obtain a correspondence between the linear functions x_i-x_j and the T–equivariant line bundles on Z.

Recall the definition of the permutation $\phi \in \Sigma_r$ given at the beginning of this chapter: ϕ takes the first r' numbers to Π', preserving the order of the first r' and the last r'' elements. We introduce the symbols

$$\begin{aligned} z'_i - z'_j &= c_1\big(\zeta(L_0(0; x_{\phi(i)} - x_{\phi(j)}))|_{Z^0}\big) && \text{for } 1 \le i, j \le r', \\ z''_i - z''_j &= c_1\big(\zeta(L_0(0; x_{\phi(r'+i)} - x_{\phi(r'+j)}))|_{Z^0}\big) && \text{for } 1 \le i, j \le r'', \\ u = (z'_{r'} - z''_r) &= c_1\big(\zeta(L_0(0; x_{\phi(r')} - x_r))|_{Z^0}\big) \end{aligned} \tag{39}$$

for the equivariant cohomology classes in $H^2_T(Z^0)$. The last equalities are consistent with Lemma 5.5.

Remark 6.8 Letting $\mathscr{F}'_i$ and $\mathscr{F}''_i$ denote the flag bundles (see Section 2.3) on $P_0(c')$ and $P_0(c'')$, correspondingly, we have (see Remark 6.2)

$$\begin{aligned} z'_i - z'_j &= c_1(\mathscr{F}'_{r-i+1}/\mathscr{F}'_{r-i} \otimes (\mathscr{F}'_{r-j+1}/\mathscr{F}'_{r-j})^*) \in H^2(P_l(c')), \\ z''_i - z''_j &= c_1(\mathscr{F}''_{r-i+1}/\mathscr{F}''_{r-i} \otimes (\mathscr{F}''_{r-j+1}/\mathscr{F}''_{r-j})^*) \in H^2(P_{-l}(c'')). \end{aligned}$$

Taking into account these identifications, functions on V give rise to equivariant cohomology classes on Z^0. To make the splitting $H^*_T(Z^0) \simeq H^*(Z^0) \otimes \mathbb{C}[u]$, explicit, however, we will write these classes in the form $f_u(z', z'')$, thinking of them as functions of the differences of the z'_is and the differences of the z''_i, depending on the parameter u. With this convention, we introduce the notation

$$\begin{aligned} w^\times_u(z', z'') &= \prod_{\substack{i,j \\ \phi(i)<\phi(r'+j)}} 2\,\sinh(z'_i - z''_j) \prod_{\substack{i,j \\ \phi(r'+j)<\phi(i)}} 2\,\sinh(z''_j - z'_i), \\ \rho^\times_u(z', z'') &= \frac{1}{2} \sum_{\substack{i,j \\ \phi(i)<\phi(r'+j)}} (z'_i - z''_j) + \frac{1}{2} \sum_{\substack{i,j \\ \phi(r'+j)<\phi(i)}} (z''_j - z'_i), \end{aligned}$$

where, according to (39),

$$z'_i - z''_j = (z'_i - z'_{r'}) + u - (z''_j - z''_r) = c_1\big(\zeta(\mathscr{L}_0(0; x_{\phi(i)} - x_{\phi(r'+j)}))|_{Z^0}\big) \in H^2_T(Z^0).$$

Proposition 6.9 *The K-theoretical Euler class $E(N_{Z^0})$* (see Definition 5.6 with $t = e^u$) *is given by the formula*

$$\begin{aligned} E(N_{Z^0})^{-1} = (-1)^{lr+r'r''(g-1)} e^{-rlu} \exp\left(\frac{\eta r}{r'} + \frac{\eta r}{r''}\right) w^\times_u(z', z'')^{1-2g} \exp(\rho^\times_u(z', z'')) \\ \cdot \operatorname{ch}(\mathscr{L}_l(r''; -l, \dots, -l, -l + rl) \boxtimes \mathscr{L}_{-l}(r'; l, \dots, l, l - rl)). \end{aligned}$$

Proof It follows from the short exact sequence (6) for parabolic morphisms that

$$\begin{aligned} \operatorname{ch}\big(-\pi_!(\operatorname{ParHom}(\widetilde{U}'', \widetilde{U}'))\big) &= -\operatorname{ch}\big(\pi_!(\operatorname{Hom}(\widetilde{U}'', \widetilde{U}'))\big) + \sum_{\substack{i,j \\ \phi(i)<\phi(r'+j)}} e^{z'_i - z''_j}, \\ \operatorname{ch}\big(-\pi_!(\operatorname{ParHom}(\widetilde{U}', \widetilde{U}''))\big) &= -\operatorname{ch}\big(\pi_!(\operatorname{Hom}(\widetilde{U}', \widetilde{U}''))\big) + \sum_{\substack{i,j \\ \phi(r'+j)<\phi(i)}} e^{z''_j - z'_i}, \end{aligned}$$

so by Lemma 6.6,

$$\mathrm{ch}(N_{Z^0}) = \mathrm{ch}\big(-\pi_!(\mathrm{Hom}(\widetilde{U}'',\widetilde{U}')) \oplus -\pi_!(\mathrm{Hom}(\widetilde{U}'',\widetilde{U}')^*)\big) + \sum_{\substack{i,j \\ \phi(i)<\phi(r'+j)}} e^{z_i'-z_j''} + \sum_{\substack{i,j \\ \phi(r'+j)<\phi(i)}} e^{z_j''-z_i'}. \tag{40}$$

Let $f(x)$ be a power series in one variable, and W a vector bundle of rank r with (equivariant) Chern roots $y_1,\dots,y_r$. Then we denote by $[f(x)]^W$ the multiplicative (equivariant) characteristic class of W given by the function $f(x)$ in Chern roots of W:

$$[f(x)]^W = \prod_{j=1}^{r} f(y_j).$$

Lemma 6.10 *Let P be a smooth variety, and let S be a T–vector bundle on $P\times C$ with T–weight 1; pick a point $p\in C$ and denote by $\pi\colon P\times C\to P$ the projection along the curve. Then*

$$E(-\pi_!S\oplus -\pi_!S^*)^{-1} = (-1)^{\mathrm{rk}(-\pi_!S)}\frac{\exp(-\mathrm{ch}_2(S)_{(2)})}{\big[\big(2\sinh\big(\frac{1}{2}x\big)\big)^{2g-2}\big]^{S_p}}.$$

Proof Note that

$$E(-\pi_!S)^{-1} = \left[\frac{1}{1-t^{-1}e^{-x}}\right]^{-\pi_!S} = \left[\frac{-te^x}{1-te^x}\right]^{-\pi_!S},$$
$$E(-\pi_!S^*)^{-1} = \left[\frac{1}{1-te^{-x}}\right]^{-\pi_!S^*} = \left[\frac{1}{1-te^x}\right]^{(-\pi_!S^*)^*}.$$

Applying Serre duality and the Grothendieck–Riemann–Roch theorem we obtain

$$\begin{aligned}\mathrm{ch}(-\pi_!S)+\mathrm{ch}((-\pi_!S^*)^*) &= \mathrm{ch}(-\pi_!S)+\mathrm{ch}(\pi_!(S\otimes K_C)) = \mathrm{ch}(-\pi_!S)+\pi_*(\mathrm{ch}(S\otimes K_C)\,\mathrm{Todd}(C))\\ &= \mathrm{ch}(-\pi_!S)+\mathrm{ch}(\pi_!S)+(2g-2)\,\mathrm{ch}(S_p) = (2g-2)\,\mathrm{ch}(S_p),\end{aligned}$$

where K_C is the canonical sheaf on the curve C, hence

$$\left[\frac{1}{1-te^x}\right]^{-\pi_!S\oplus(-\pi_!S^*)^*} = \left[\frac{1}{(1-te^x)^{2g-2}}\right]^{S_p} = \frac{\exp(-c_1(S_p)(g-1))}{\big[\big(2\sinh\big(\frac{1}{2}x\big)\big)^{2g-2}\big]^{S_p}}.$$

Since

$$[-te^x]^{-\pi_!S} = (-1)^{\mathrm{rk}(-\pi_!S)}\exp(c_1(-\pi_!S)),$$

and, by the Grothendieck–Riemann–Roch theorem,

$$\mathrm{ch}_1(-\pi_!S) = \mathrm{ch}_1(S_p)(g-1) - \mathrm{ch}_2(S)_{(2)},$$

we conclude that

$$[-te^x]^{-\pi_!S} = (-1)^{\mathrm{rk}(-\pi_!S)}\exp(c_1(S_p)(g-1))\exp(-\mathrm{ch}_2(S)_{(2)}),$$

which finishes the proof of Lemma 6.10. □

Note that the last two terms in (40) are the sums of Chern characters of line bundles, so they contribute the multiplicative factor

$$\frac{\exp(\rho_u^\times(z',z''))}{w_u^\times(z',z'')}$$

to the equivariant class $E(N_{Z^0})^{-1}$; and using Lemma 6.10 with $S=\operatorname{Hom}(\widetilde{U}'',\widetilde{U}')$, we obtain that the inverse of the K-theoretical Euler class of the first term in (40) is

$$(-1)^{lr+r'r''(g-1)}w_u^\times(z',z'')^{2-2g}\exp\bigl(-\operatorname{ch}_2(\operatorname{Hom}(\widetilde{U}'',\widetilde{U}'))_{(2)}\bigr).$$

Note that

$$\begin{aligned}-\operatorname{ch}_2(\operatorname{Hom}(\widetilde{U}'',\widetilde{U}'))_{(2)} &= \tfrac{1}{2}c_2(\operatorname{End}_0(\widetilde{U}'\oplus\widetilde{U}''))_{(2)}-\tfrac{1}{2}c_2(\operatorname{End}_0(\widetilde{U}'))_{(2)}-\tfrac{1}{2}c_2(\operatorname{End}_0(\widetilde{U}''))_{(2)}\\ &= c_1(\mathscr{L}(r;0)|_{Z^0}\otimes\mathscr{L}_l(-r';-l,\dots,-l)\boxtimes\mathscr{L}_{-l}(-r'';l,\dots,l)).\end{aligned}$$

The latter equality follows from Lemma 2.8. Finally, using Lemma 6.5 to calculate the Chern character of $\mathscr{L}(r;0)|_{Z^0}$, we obtain the formula for the class $E(N_{Z^0})^{-1}$, and the proof of the lemma is complete. □

6.3 The wall-crossing formula

Putting Lemma 6.5 and Proposition 6.9 together, we obtain the following.

Proposition 6.11 *The wall-crossing term* (35) *is equal to*

$$\begin{aligned}K\operatorname*{Res}_{u=0}e^{(k\delta-rl)u}\int_{P_l(c')\times P_{-l}(c'')}\Bigl[&(w_u^\times(z',z''))^{1-2g}\exp(\rho_u^\times(z',z''))\\ &\cdot\operatorname{ch}\bigl(\mathscr{L}_l(k+r'';\lambda_1'-l,\dots,\lambda_{r'-1}'-l,\lambda_{r'}'-l-k\delta+rl)\\ &\quad\boxtimes\mathscr{L}_{-l}(k+r';\lambda_1''+l,\dots,\lambda_{r''-1}''+l,\lambda_{r''}''+l+k\delta-rl)\bigr)\\ &\quad\operatorname{Todd}(P_l(c')\times P_{-l}(c''))\Bigr]\,du,\end{aligned}$$

where δ is a parameter depending on λ and the wall $S_{\Pi,l}$ (see Lemma 6.5) *and K is the constant*

$$(-1)^{lr+r'r''(g-1)}\frac{(r(k+r))^g}{(r'r'')^g}.$$

Now all that is left to do is to perform the integral, using an induction on the rank based on Corollary 4.10. We will begin with the case $l=0$, as it is simpler. For $l=0$, the integral from Proposition 6.11 has the form

$$\begin{aligned}(41)\quad\int_{P_0(c')\times P_0(c'')}\bigl[&w_u^\times(z',z'')^{1-2g}e^{\rho_u^\times(z',z'')}\operatorname{Todd}(P_0(c'))\operatorname{Todd}(P_0(c''))\\ &\cdot\operatorname{ch}(\mathscr{L}_0(k+r'';\lambda_1',\dots,\lambda_{r'-1}',\lambda_{r'}'-k\delta)\boxtimes\mathscr{L}_0(k+r';\lambda_1'',\dots,\lambda_{r''-1}'',\lambda_{r''}''+k\delta))\bigr].\end{aligned}$$

The inductive hypothesis (22) may be cast in the form

$$(42)\quad\int_{P_0(c)}\operatorname{ch}(\mathscr{L}_0(k;\lambda))\operatorname{Todd}(P_0(c))=\widetilde{N}_{r,k}\sum_{\boldsymbol{B}\in\mathscr{D}}\operatorname{iBer}_{\boldsymbol{B}}[\exp\langle\lambda,x/\widehat{k}\rangle\cdot w_\Phi(x/\widehat{k})^{1-2g}](\rho/\widehat{k}-[c]_{\boldsymbol{B}}).$$

Now let us fix k, and allow λ to vary. We can extend this equality by linearity to arbitrary linear combinations of Chern characters of line bundles of the form

$$\sum_i \mathrm{ch}(\mathscr{L}_0(k;\lambda^i)) = \mathrm{ch}(\mathscr{L}_0(k;0)) \cdot \sum_i \mathrm{ch}(\mathscr{L}_0(0;\lambda^i)).$$

Since any polynomial on V up to a fixed degree may be represented as a linear combination of exponential functions of the form $\exp\langle\lambda, x/\widehat{k}\rangle$, formula (42) may be generalized in the following way.

Lemma 6.12 *Let $G(x)$ be a formal power series on V, and denote by $G(z)$ the characteristic class in $H^*(P_0(c))$ obtained by the identification of functions on V and cohomology classes of $P_0(c)$, described before the equation* (39)*. Then we have*

$$\int_{P_0(c)} \mathrm{ch}(\mathscr{L}_0(k;0))G(z)\,\mathrm{Todd}(P_0(c)) = \widetilde{N}_{r,k} \sum_{\boldsymbol{B}\in\mathbb{D}} \operatorname*{iBer}_{\boldsymbol{B}}[G(x/\widehat{k})\cdot w_\Phi^{1-2g}(x/\widehat{k})](\rho/\widehat{k} - [c]_{\boldsymbol{B}}). \tag{43}$$

Finally, let $\mathbb{D}'$ and $\mathbb{D}''$ be Hamiltonian bases (see Section 4.5). Since

$$w_{\Phi'}(x/\widehat{k})w_{\Phi''}(x/\widehat{k})w_u^\times(x/\widehat{k}) = w_\Phi(x/\widehat{k}) \quad \text{and} \quad \rho'(x/\widehat{k})\rho''(x/\widehat{k})\rho_u^\times(x/\widehat{k}) = \rho(x/\widehat{k}),$$

where $w_{\Phi'}$, $w_{\Phi''}$ and ρ', ρ'' are naturally defined for the root systems Φ' and Φ'' (see Section 4.5), the integral (41) is equal to

$$\begin{aligned}&\widetilde{N}_{r',k+r''}\widetilde{N}_{r'',k+r'}\\ &\cdot \sum_{\boldsymbol{B}'\in\mathbb{D}'}\sum_{\boldsymbol{B}''\in\mathbb{D}''} \operatorname*{iBer}_{\boldsymbol{B}'}\operatorname*{iBer}_{\boldsymbol{B}''}[w_\Phi(x/\widehat{k})^{1-2g}e^{\rho(x/\widehat{k})}]\\ &\qquad\cdot((\lambda'_1,\dots,\lambda'_{r'-1},\lambda'_{r'}-k\delta)/\widehat{k} - [c']_{\boldsymbol{B}'} + (\lambda''_1,\dots,\lambda''_{r''-1},\lambda''_{r''}+k\delta)/\widehat{k} - [c'']_{\boldsymbol{B}''}).\end{aligned}$$

Identifying u (see (39)) with the "link" element of the diagonal basis $\mathbb{D} = (\alpha^{\phi(r'),r}\,\mathbb{D}'\,\mathbb{D}'')$ (see Section 4.5), and moving the factor $e^{k\delta u}$ from Proposition 6.11 inside the argument of iBer, we obtain the proof of the following theorem for $l = 0$.

Theorem 6.13 *Let $c^\pm \in \Delta$ be in the neighboring chambers; then the wall-crossing term*

$$\chi(P_0(c^+),\mathscr{L}_0(k;\lambda)) - \chi(P_0(c^-),\mathscr{L}_0(k;\lambda))$$

is equal to

$$(k+r)\widetilde{N}_{r,k}\sum_{\boldsymbol{B}'\in\mathbb{D}'}\sum_{\boldsymbol{B}''\in\mathbb{D}''}\operatorname*{Res}_{\alpha^{\phi(r'),r}=0}\operatorname*{iBer}_{\boldsymbol{B}'}\operatorname*{iBer}_{\boldsymbol{B}''}[w_\Phi(x/\widehat{k})^{1-2g}](\widehat{\lambda}/\widehat{k} - [c^+]_{\boldsymbol{B}})\,d\alpha^{\phi(r'),r},$$

where $\mathbb{D}'$ and $\mathbb{D}''$ are the diagonal bases of Φ' and Φ'' (see Section 4.5) *correspondingly.*

Remark 6.14 This wall-crossing term coincides with the one from Proposition 4.18.

Example 6.15 It follows from Example 2.9 that in case of rank 3, the permutation $\phi \in \Sigma_3$ sends $(1,2,3)$ to $(1,3,2)$. Then $u = c_1(\mathscr{F}'_1 \otimes \mathscr{F}''^*_1)$ and let $z = z''_1 - z''_2 = c_1(\mathscr{F}''_2/\mathscr{F}''_1 \otimes \mathscr{F}''^*_1)$. Then the inverse of the K-theoretical Euler class of the conormal bundle is (see Proposition 6.9)

$$\operatorname{ch}(\mathscr{L}) e^{9\eta/2} e^{z/2} \big(2\sinh\big(\tfrac{1}{2}u\big)\, 2\sinh\big(\tfrac{1}{2}(z-u)\big)\big)^{1-2g},$$

where $\mathscr{L} = \mathscr{L}_0(2;0,0)$ is a line bundle on the moduli space P_0 of rank-2 degree-0 stable parabolic bundles. The Chern character of the restriction of the line bundle $\mathscr{L}_0(k;\lambda_1,\lambda_2,\lambda_3)$ to Σ is

$$e^{3k\eta/2} \operatorname{ch}(\mathscr{L}_0^k) e^{\lambda_1 z + \lambda_2 u}.$$

Hence the wall-crossing term

$$\chi(P_0(<), \mathscr{L}_0(k,\lambda)) - \chi(P_0(>), \mathscr{L}_0(k,\lambda))$$

is equal to

$$-\big(\tfrac{3}{2}(k+3)\big)^g \operatorname*{Res}_{u=0} \frac{e^{\lambda_2 u}}{\big(2\sinh\big(\frac{1}{2}u\big)\big)^{2g-1}} \cdot \int_{P_0} \frac{\operatorname{ch}\big(\mathscr{L}_0\big(k+1;\lambda_1+\frac{1}{2},-\lambda_1-\frac{1}{2}\big)\big)}{\big(2\sinh\big(\frac{1}{2}(z-u)\big)\big)^{2g-1}} \operatorname{Todd}(P_0)\, du.$$

The integral is the Euler characteristics of a line bundle on a moduli space of degree-0 rank-2 stable parabolic bundles, so we can calculate it using the induction by rank. It is equal to

$$(-1)^{g-1}(2(k+3))^g \operatorname*{Res}_{z=0} \frac{e^{(\lambda_1+1)z}}{\big(2\sinh\big(\frac{1}{2}(z-u)\big)\, 2\sinh\big(\frac{1}{2}z\big)\big)^{2g-1}(1-e^{(k+3)z})}\, dz,$$

so the wall-crossing term is

$$(-3(k+3)^2)^g \operatorname*{Res}_{u=0} \operatorname*{Res}_{z=0} \frac{e^{\lambda_1 z + \lambda_2 u + z}}{\widetilde{w}_\phi(z,u)^{2g-1}(1-e^{(k+3)z})}\, dz\, du,$$

where $\widetilde{w}_\phi(z,u) = 2\sinh\big(\frac{1}{2}(z-u)\big)\, 2\sinh\big(\frac{1}{2}u\big)\, 2\sinh\big(\frac{1}{2}z\big)$. Note that this is exactly the same polynomial as in Example 4.15, after changing (z,u) to $(x,-y)$.

7 Tautological Hecke correspondences

If $l \neq 0$, then we need one more step in our proof, which uses the Hecke correspondence to calculate the wall-crossing term (35).

7.1 The Hecke correspondence

Given a rank-r degree-d vector bundle W with a full flag $0 \subsetneq F_1 \subsetneq \cdots \subsetneq F_r = W_p$ at p, one can obtain a rank-r degree-$(d-1)$ vector bundle W' with a full flag $0 \subsetneq G_1 \subsetneq \cdots \subsetneq G_r = W'_p$ using the tautological Hecke correspondence construction as follows.

The evaluation map $W \to W_p$ induces the short exact sequence of the associated sheaves of sections

$$0 \to \mathcal{W}' \xrightarrow{\widetilde{\alpha}} \mathcal{W} \to W_p/F_{r-1} \to 0 \tag{44}$$

on curve C. Since $\mathcal{W}'$ is a kernel of $\widetilde{\alpha}$, it is a locally free sheaf, thus gives a rank-r vector bundle W' over C with $\det(W') \simeq \det(W) \otimes \mathcal{O}(-p)$. The image of the associated morphism of vector bundles α at the point p is $F_{r-1} \subset W_p$, so $\alpha_p \colon W'_p \to W_p$ has a one-dimensional kernel $G_1 \subset W'_p$. Moreover, compositions of α_p with the quotient morphisms $F_{r-1} \to F_{r-1}/F_i$ induce a full flag of the corresponding kernels $G_1 \subsetneq \cdots \subsetneq G_{r-1} \subsetneq G_r = W'_p$ in W'_p.

Denote this operator between the sets of isomorphism classes of degree-d and $d-1$ vector bundles with a flag at p by

$$\mathcal{H} \colon (W, F_*) \mapsto (W', G_*).$$

Similarly, for any $m \geq 0$, one can define the operator $\mathcal{H}^m$ between the sets of isomorphism classes of degree-d and $d-m$ vector bundles with a flag at the point p by iterating the above construction m times. Clearly, these maps are independent of the parabolic weights.

Proposition 7.1 *Let $c \in \Delta$ be a regular* (see page 2266) *point. Then the operator $\mathcal{H}$ induces an isomorphism between the moduli spaces $P_d(c_1, \dots, c_r)$ and $P_{d-1}(c_2, \dots, c_r, c_1 - 1)$.*

Proof First, we need to show that if $W \in P_d(c_1, \dots, c_r)$ is a parabolic stable bundle with parabolic weights $(c_1, \dots, c_r)$, then W', its image under the Hecke operator $\mathcal{H}$, is parabolic stable with respect to parabolic weights $(c_2, \dots, c_r, c_1 - 1)$. For this, consider the subbundle $V' \subset W'$ and let $\alpha(V') = V \subset W$ (see (44)) be its image. Since W is parabolic stable,

$$\operatorname{par\,slope}(V) < \operatorname{par\,slope}(W) = \operatorname{par\,slope}(W').$$

We need to prove that $\operatorname{par\,slope}(V') < \operatorname{par\,slope}(W')$. There are two possible cases:

- If α maps V' to V isomorphically, then $\deg(V') = \deg(V)$ and $V_p \subset F_{r-1}$, hence $\operatorname{par\,slope}(V') = \operatorname{par\,slope}(V) < \operatorname{par\,slope}(W')$.
- Otherwise, $\deg(V') = \deg(V) - 1$, and V_p is not contained in F_{r-1}, so one of the parabolic weights of V' is $c_1 - 1$. Then, as in the previous case, $\operatorname{par\,slope}(V') = \operatorname{par\,slope}(V)$, and the result follows.

To show that the map $\mathcal{H}$ is an isomorphism, note that $\mathcal{H}^r$ maps

$$P_d(c_1, c_2, \dots, c_r) \to P_{d-r}(c_1 - 1, c_2 - 1, , \dots, c_r - 1). \tag{45}$$

It is easy to check that given $\mathcal{W}$ and iterating the associated morphism of locally free sheaves of sections (44) r times, we obtain a subsheaf $\mathcal{W}' \subset \mathcal{W}$ of sections of W which vanishes at the point p. So the map (45) is just tensoring by $\mathcal{O}(-p)$, and hence it is an isomorphism. □

Now we can define an operator $\mathscr{H}^m$ for any $m \in \mathbb{Z}$, taking the inverse map if necessary. We will need the following statement, which follows from Proposition 7.1 and the construction of $\mathscr{H}^m$.

Corollary 7.2 *Let $m \geq 0$. Then under the isomorphism $\mathscr{H}^m$ the line bundle $\mathscr{L}_d(k; \lambda_1, \dots, \lambda_r)$ corresponds to the line bundle $\mathscr{L}_{d-m}(k; \lambda_{r-m+1}, \dots, \lambda_r, \lambda_1 - k, \dots, \lambda_{r-m} - k)$.*

7.2 The effect of the Hecke correspondence on the integral

Recall that our goal is to calculate the wall-crossing term from Proposition 6.11. For simplicity, we assume that l is positive (the other case is analogous). We apply the Hecke operators $\mathscr{H}^l$ and $\mathscr{H}^{-l}$ to the moduli spaces $P_l(c')$ and $P_{-l}(c'')$ to obtain

$$\begin{aligned} P_0' &= P_0(c'_{l+1}, \dots, c'_{r'}, c'_1 - 1, \dots, c'_l - 1) \simeq P_l(c'), \\ P_0'' &= P_0(c''_{r''-l+1} + 1, \dots, c''_{r''} + 1, c''_1, \dots, c''_{r''-l}) \simeq P_{-l}(c''). \end{aligned}$$

Recall (see page 2270) that there is a natural action of the group Σ_r on V^*, and hence (see page 2291) on $H^2(P_l(c') \times P_{-l}(c''))$. Let $\tau' \in \Sigma_{r'}$ and $\tau'' \in \Sigma_{r''}$ be the cyclic permutations defined by

$$\begin{aligned} \tau' \cdot (c'_1 - 1, \dots, c'_l - 1, c'_{l+1}, \dots, c'_{r'}) &= (c'_{l+1}, \dots, c'_{r'}, c'_1 - 1, \dots, c'_l - 1), \\ \tau'' \cdot (c''_1, \dots, c''_{r''-l}, c''_{r''-l+1} + 1, \dots, c''_r + 1) &= (c''_{r''-l+1} + 1, \dots, c''_{r''} + 1, c''_1, \dots, c''_{r''-l}). \end{aligned}$$

Now set $\tau = (\tau', \tau'') \in \Sigma_{r'} \times \Sigma_{r''} \subset \Sigma_r$. Note that

$$\tau' \cdot (-l + r', \dots, -l + r', -l, \dots, -l) = \tau' \cdot \rho' - \rho' \quad \text{and} \quad \tau'' \cdot (l, \dots, l, l - r'', \dots, l - r'') = \tau'' \cdot \rho'' - \rho'',$$

so applying the Hecke operator $\mathscr{H}^l \times \mathscr{H}^{-l}$ to the wall-crossing term from Proposition 6.11 and using Corollary 7.2, we obtain that the wall-crossing term (35) is equal to

$$\begin{aligned} (46)\ K \operatorname*{Res}_{u=0} e^{(k\delta - rl)u} \int_{P_0' \times P_0''} \Big(& \tau \cdot w_u^{\times}(z', z'')^{1-2g} e^{\tau \cdot \rho_u^{\times}(z', z'')} \\ & \cdot \operatorname{ch}\big(\mathscr{L}_0(k + r''; \tau' \cdot (\lambda'_1 - \widehat{k}, \dots, \lambda'_l - \widehat{k}, \lambda'_{l+1}, \dots, \lambda'_{r'-1}, \lambda'_{r'} - k\delta + rl))\big) \\ & \cdot \operatorname{ch}\big(\mathscr{L}_0(k + r'; \tau'' \cdot (\lambda''_1, \dots, \lambda''_{r'-l}, \lambda''_{r''-l+1} + \widehat{k}, \dots, \lambda''_{r''} + \widehat{k} + k\delta - rl))\big) \\ & e^{\tau'' \cdot \rho''(z', z'') - \rho''(z', z'')} e^{\tau' \cdot \rho'(z', z'') - \rho'(z', z'')} \operatorname{Todd}(P_0') \operatorname{Todd}(P_0'') \Big)\, du. \end{aligned}$$

As in Section 6.3, according to Lemma 6.12, we can calculate this integral using the induction on rank. Let $\mathscr{D}'$ and $\mathscr{D}''$ be two Hamiltonian diagonal bases. Then $\tau'(\mathscr{D}')$ and $\tau''(\mathscr{D}'')$ are also Hamiltonian diagonal bases (see Remark 3.5) and the integral in (46) is equal to

$$\begin{aligned} (47)\quad & (-1)^{lr} \widetilde{N}_{r', k+r''} \widetilde{N}_{r'', k+r'} \\ & \sum_{\substack{\boldsymbol{B}' \in \tau'(\mathscr{D}') \\ \boldsymbol{B}'' \in \tau''(\mathscr{D}'')}} \operatorname{iBer}_{\boldsymbol{B}'} \operatorname{iBer}_{\boldsymbol{B}''} [\tau \cdot w_u^{\times}(x/\widehat{k})^{1-2g} (w_{\Phi'}(x/\widehat{k}) w_{\Phi''}(x/\widehat{k}))^{1-2g} e^{\tau \cdot \rho(x/\widehat{k})}] \\ & \qquad \big(\tau' \cdot (\lambda'_1 - \widehat{k}, \dots, \lambda'_l - \widehat{k}, \lambda'_{l+1}, \dots, \lambda'_{r'-1}, \lambda'_{r'} - k\delta + rl)/\widehat{k} \\ & \qquad\quad - [\tau' \cdot (c'_1 - 1, \dots, c'_l - 1, c'_{l+1}, \dots, c'_{r'})]_{\boldsymbol{B}'} \\ & \qquad\quad + \tau'' \cdot (\lambda''_1, \dots, \lambda''_{r'-l}, \lambda''_{r''-l+1} + \widehat{k}, \dots, \lambda''_{r''-1} + \widehat{k}, \lambda''_{r''} + \widehat{k} + k\delta - rl)/\widehat{k} \\ & \qquad\quad - [\tau'' \cdot (c''_1, \dots, c''_{r'-l}, c''_{r''-l+1} + 1, \dots, c''_{r''} + 1)]_{\boldsymbol{B}''} \big). \end{aligned}$$

To arrive at Theorem 6.13, we need to make additional transformations of formula (47): first, we shift λ' and λ'', and then we apply Lemma 4.5 to eliminate the cyclic permutation τ.

Note that given an ordered basis $\boldsymbol{B} \in \mathcal{B}$ and an element $v \in V^*$ such that $\{v\}_{\boldsymbol{B}} = 0$, for any weight $\lambda \in \Lambda$ and positive integer k we have

$$(\lambda + \widehat{k}v)/\widehat{k} - [c+v]_{\boldsymbol{B}} = \lambda/\widehat{k} - [c]_{\boldsymbol{B}}. \tag{48}$$

In particular, to perform the shift of λ' in (47), we use the following equality for any $\boldsymbol{B}' \in \mathcal{D}'$:

$$\begin{aligned}(\lambda'_1 - \widehat{k}, \dots, \lambda'_l - \widehat{k}, \lambda'_{l+1}, \dots, \lambda'_{r'-1}, \lambda'_{r'} - k\delta + rl)/\widehat{k} \\ - [(c'_1, \dots, c'_{r'-1}, c'_{r'} - l) - (1, \dots, 1, 0, \dots 0, -l)]_{\boldsymbol{B}'} \\ = (\lambda'_1, \dots, \lambda'_{r'-1}, \lambda'_{r'} - k\delta + rl - l\widehat{k})/\widehat{k} - [(c'_1, \dots, c'_{r'-1}, c'_{r'} - l)]_{\boldsymbol{B}'},\end{aligned} \tag{49}$$

which clearly remains true after changing $\mathcal{D}'$ to $\tau'(\mathcal{D}')$ and applying τ' to both sides of the equation. Similarly, shifting the last terms of (47) by $\tau''(0, \dots, 0, -1, \cdots -1, -1+l)$, we can rewrite (47) as

$$\begin{aligned}&(-1)^{lr} \widetilde{N}_{r', k+r''} \widetilde{N}_{r'', k+r'} \\ &\sum_{\substack{\boldsymbol{B}' \in \tau'(\mathcal{D}') \\ \boldsymbol{B}'' \in \tau''(\mathcal{D}'')}} \operatorname{iBer}_{\boldsymbol{B}'} \operatorname{iBer}_{\boldsymbol{B}''} [\tau \cdot w_u^{\times}(x/\widehat{k})^{1-2g} (w_{\Phi'}(x/\widehat{k}) w_{\Phi''}(x/\widehat{k}))^{1-2g} e^{\tau \cdot \rho(x/\widehat{k})}] \\ &\cdot \big(\tau' \cdot (\lambda'_1, \dots, \lambda'_{r'-1}, \lambda'_{r'} - k\delta + rl - l\widehat{k})/\widehat{k} - [\tau' \cdot (c'_1, \dots, c'_{r'-1}, c'_{r'} - l)]_{\boldsymbol{B}'} \\ &\quad + \tau'' \cdot (\lambda''_1, \dots, \lambda''_{r''-1}, \lambda''_{r''} + k\delta - rl + l\widehat{k})/\widehat{k} - [\tau'' \cdot (c''_1, \dots, c''_{r''-1}, c''_{r''} + l)]_{\boldsymbol{B}''}\big).\end{aligned} \tag{50}$$

Finally, identifying u (see (39)) with the "link" element of the diagonal basis

$$\tau(\mathcal{D}) = (\alpha^{\tau\phi(r'), \tau(r)} \tau'(\mathcal{D}') \tau''(\mathcal{D}''))$$

(see Section 4.5) and

- moving the factor $e^{(k\delta - rl)u}$ from (46) inside the argument of $\operatorname{iBer}_{\boldsymbol{B}}$, where
$$\boldsymbol{B} = (\alpha^{\tau\phi(r'), \tau(r)} \boldsymbol{B}' \boldsymbol{B}''),$$
- applying (48) with $\boldsymbol{B} = (\alpha^{\tau\phi(r'), \tau(r)}\, \boldsymbol{B}'\, \boldsymbol{B}'')$ and $v = l\alpha^{\tau\phi(r'), \tau(r)}$,
- applying Lemma 4.5, and
- using the fact that
$$\tau^{-1} \cdot (w_{\Phi'}(x/\widehat{k}) w_{\Phi''}(x/\widehat{k})) = (-1)^{lr} w_{\Phi'}(x/\widehat{k}) w_{\Phi''}(x/\widehat{k}),$$

we obtain the formula of Theorem 6.13 for arbitrary $l \in \mathbb{Z}$.

8 Affine Weyl symmetry and the proof of Theorem 4.8(I)

In this section, we prove certain symmetry properties of our Hilbert polynomials on the left-hand side of equation (1), and we finish the proof of Theorem 4.8(I). We start with the basic instance of symmetry of Hilbert polynomials: relative Serre duality.

8.1 Serre duality

Proposition 8.1 *Let $\mathscr{E} \to X$ be a rank-2 vector bundle over a smooth variety X, let $\pi : Y = \mathbb{P}(\mathscr{E}) \to X$ be its projectivization and $\omega_{X/Y}$ the relative cotangent line bundle. Then*

$$\chi(Y, \pi^*\mathscr{L} \otimes \omega_{X/Y}^m) = -\chi(Y, \pi^*\mathscr{L} \otimes \omega_{X/Y}^{-m+1})$$

for any line bundle $\mathscr{L} \in \mathrm{Pic}(X)$.

Proof By Serre duality for families of curves [10, Chapter III, Sections 7–8], for any integer n,

$$\chi(Y, \pi^*\mathscr{L} \otimes \mathbb{O}(n)) = -\chi(Y, \pi^*(\mathscr{L} \otimes (\wedge^2\mathscr{E})^{n+1}) \otimes \mathbb{O}(-n-2)). \tag{51}$$

Denote by $\Omega_{X/Y}$ the sheaf of relative differentials on Y; the short exact sequence

$$0 \to \Omega_{X/Y} \otimes \mathbb{O}_X \to \pi^*\mathscr{E}(-1) \to \mathbb{O}_X \to 0$$

implies that

$$\omega_{X/Y} = \wedge^2(\pi^*\mathscr{E}(-1)) = \pi^*(\wedge^2\mathscr{E}) \otimes \mathbb{O}(-2).$$

Then the statement follows from (51) by substituting $n = -2m$. □

Now we can generalize this statement to the case of flag bundles.

Proposition 8.2 *Let $\pi : Y = \mathrm{Flag}(\mathscr{E}) \to X$ be a rank-r flag bundle over X. Let $\mathscr{L}$ be a line bundle on X, and $\mathscr{F}_1, \mathscr{F}_2/\mathscr{F}_1, \dots, \mathscr{F}_r/\mathscr{F}_{r-1}$ the standard flag line bundles on Y. For $k \in \mathbb{Z}$ and $\lambda = (\lambda_1, \dots, \lambda_r) \in \Lambda$, write*

$$\mathscr{L}(k;\lambda) = (\pi^*\mathscr{L})^k \otimes (\mathscr{F}_r/\mathscr{F}_{r-1})^{\lambda_1} \otimes (\mathscr{F}_{r-1}/\mathscr{F}_{r-2})^{\lambda_2} \otimes \cdots \otimes \mathscr{F}_1^{\lambda_r}.$$

Consider the polynomial

$$q(k;\lambda_1,\lambda_2,\dots,\lambda_r) = \chi(Y, \mathscr{L}(k;\lambda_1,\lambda_2,\dots,\lambda_r))$$

in $(k,\lambda) \in \mathbb{Z} \times \Lambda$, and extend this definition to $\mathbb{R} \times V^*$. Then $q(k;\lambda-\rho)$ is anti-invariant under the permutations of $\lambda_1, \lambda_2, \dots, \lambda_r$.

Proof For $1 \le k < r$, let $\mathrm{Flag}_{\widehat{k}}(\mathscr{E}) \to X$ be the flag bundle over X obtained from Y by forgetting the k–dimensional subspace. Then $Y \simeq \mathbb{P}(\mathscr{F}_{k+1}/\mathscr{F}_{k-1}) \to \mathrm{Flag}_{\widehat{k}}(\mathscr{E})$ is a $\mathbb{P}^1$–bundle over $\mathrm{Flag}_{\widehat{k}}(\mathscr{E})$, and thus applying Proposition 8.1 we obtain

$$\chi(Y, \mathscr{L}(k;\lambda_1,\dots,\lambda_{r-k},\lambda_{r-k+1},\dots,\lambda_r)) = -\chi(Y, \mathscr{L}(k;\lambda_1,\dots,\lambda_{r-k+1}-1,\lambda_{r-k}+1,\dots,\lambda_r)),$$

and the result follows. □

8.2 The Weyl antisymmetry of the functions q_1 and q_{-1}

Armed with this statement, we are ready to take on the symmetries of the Hilbert polynomial of our parabolic moduli spaces. We note that the two sets $\Delta_{\pm 1}$ of weights for degree-± 1 stable parabolic bundles are simplices with one of their vertices at $(1/r, \dots, 1/r)$ and $(-1/r, \dots, -1/r)$, correspondingly; see Section 2.2.

Denote by $N_{\pm 1}$ the moduli spaces of rank-r degree-± 1 stable vector bundles and by UN any universal bundle over $N_{\pm 1} \times C$; see eg [3].

Lemma 8.3 *Let $c = (c_1, \dots, c_r)$ be a parabolic weight from the chamber in Δ_1, which has as one of its vertices the (regular) point $(1/r, \dots, 1/r)$. Then the moduli space $P_1(c)$ of rank-r degree-1 stable parabolic bundles is isomorphic to the flag bundle $\mathrm{Flag}(UN_p)$ over N_1. An analogous statement holds in the case of degree -1 and the point $(-1/r, \dots, -1/r) \in \Delta_{-1}$.*

Proof A simple calculation shows that a point $(c_1, \dots, c_r) \in \Delta_1$ such that all $c_i > 0$ lies inside the chamber in Δ_1 with the vertex $(1/r, \dots, 1/r)$. Hence it is enough to prove the first statement for the moduli space $P_1(c_1, \dots, c_r)$ with positive parabolic weights.

Moreover, it is sufficient to show that if (W, F_*) is a parabolic stable vector bundle which represents a point in $P_1(c_1, \dots, c_r)$, then W is stable as an ordinary bundle. Assume that W admits a proper subbundle W' with $\mathit{slope}(W') \geq \mathit{slope}(W) = 1/r$, then $\deg(W') \geq 1$. Since all parabolic weights of W are positive, this implies that $\operatorname{par\,slope}(W') > 0 = \operatorname{par\,slope}(W)$, and therefore W is parabolic unstable. The proof for degree-(-1) bundles is analogous. □

Denote the moduli spaces described above by $P_1(>)$ and $P_{-1}(<)$, correspondingly, and their images under the Hecke isomorphisms $\mathscr{H}$ and $\mathscr{H}^{-1}$ by $P_0(>)$ and $P_0(<)$.

The following statement is straightforward; see Lemma 2.8.

Lemma 8.4 *The line bundles $\mathscr{L}_1(r; 1, \dots, 1)$ and $\mathscr{L}_{-1}(r; -1, \dots, -1)$ on $P_1(>)$ and $P_{-1}(<)$ defined in Lemma 2.6 may be obtained as pullbacks of the ample generators of the Picard groups $\mathrm{Pic}(N_{\pm 1})$.*

Example 8.5 In the case of rank-3 parabolic bundles the moduli space $P_1(c_1, c_2, c_3)$ with $2c_3 > c_1 + c_2 - 1$ is a flag bundle over N_1 and it is isomorphic to the moduli space $P_0(>)$ from Example 2.9, while the moduli space $P_{-1}(c_1, c_2, c_3)$ with $2c_1 < c_2 + c_3 + 1$ is a flag bundle over N_{-1} and it is isomorphic to $P_0(<)$.

Now we establish the Weyl antisymmetry of the polynomials

$$q_{-1}(k; \lambda_1, \dots, \lambda_r) = \chi(P_0(<), \mathscr{L}_0(k; \lambda_1, \dots, \lambda_r)),$$
$$q_1(k; \lambda_1, \dots, \lambda_r) = \chi(P_0(>), \mathscr{L}_0(k; \lambda_1, \dots, \lambda_r)),$$

defined on $\mathbb{R}\times\Lambda$, as in Proposition 8.2. Let $\tau\in\Sigma_r$ be the cyclic permutation such that $\tau\cdot(c_1,\dots,c_r)=(c_2,\dots,c_r,c_1)$, and consider two points in V^*:

$$\begin{aligned}\theta_1[k]&=\frac{k+r}{r}\cdot(1,1,\dots,1)-(k+r)x_r-\rho=\tau\cdot\Big(\frac{k}{r}-k,\frac{k}{r},\dots,\frac{k}{r}\Big)-\tau\cdot(\rho)\\&=\Big(\frac{k}{r}-\tfrac{1}{2}(r-1)+1,\frac{k}{r}-\tfrac{1}{2}(r-1)+2,\dots,\frac{k}{r}-\tfrac{1}{2}(r-1)+r-1,-k+\frac{k}{r}-\tfrac{1}{2}(r-1)\Big),\\\theta_{-1}[k]&=-\frac{k+r}{r}\cdot(1,1,\dots,1)+(k+r)x_1-\rho=\tau^{-1}\cdot\Big(-\frac{k}{r},\dots,-\frac{k}{r},-\frac{k}{r}+k\Big)-\tau^{-1}\cdot(\rho)\\&=\Big(k-\frac{k}{r}+\tfrac{1}{2}(r-1),-\frac{k}{r}-\tfrac{1}{2}(r-1),-\frac{k}{r}-\tfrac{1}{2}(r-1)+1,\dots,-\frac{k}{r}-\tfrac{1}{2}(r-1)+r-2\Big).\end{aligned}$$

Proposition 8.6 *The polynomials $q_1(k;\lambda+\theta_1[k])$ and $q_{-1}(k;\lambda+\theta_{-1}[k])$ are anti-invariant under the action of the Weyl group by permutations of $\lambda_1,\dots,\lambda_r$.*

Proof Recall that the moduli space $P_0(>)$ is isomorphic to the flag bundle $P_1(>)$ over N_1 under the Hecke isomorphism $\mathcal{H}^{-1}$. Then using Corollary 7.2, Proposition 8.2 and Lemma 8.4, for any permutation $\sigma\in\Sigma_r$ we obtain

$$\begin{aligned}q_1(k;\sigma\cdot\lambda+\theta_1[k])&\stackrel{\text{def}}{=}\chi\big(P_0(>),\mathcal{L}_0(k;\sigma\cdot\lambda+\theta_1[k])\big)\\&=\chi\Big(P_1(>),\mathcal{L}_1\Big(k;\tau^{-1}\cdot\sigma\cdot\lambda+\Big(\frac{k}{r},\dots,\frac{k}{r}\Big)-\rho\Big)\Big)\\&=(-1)^\sigma\chi\Big(P_1(>),\mathcal{L}_1\Big(k;\tau^{-1}\cdot\lambda+\Big(\frac{k}{r},\dots,\frac{k}{r}\Big)-\rho\Big)\Big)\\&=(-1)^\sigma\chi\big(P_0(>),\mathcal{L}_0(k;\lambda+\theta_1[k])\big)\\&\stackrel{\text{def}}{=}(-1)^\sigma q_1(k;\lambda+\theta_1[k]),\end{aligned}$$

where the second and fourth equalities hold by Corollary 7.2, and the third equality follows from Proposition 8.2 and Lemma 8.4. The proof for q_{-1} is similar. □

The two group actions in Proposition 8.6 may be combined in the following manner. For $k\geq 0$, we define an action of the *affine Weyl group* $\Sigma\rtimes\Lambda$ on $\Lambda\times\mathbb{Z}$, which acts trivially on the second factor, the level, and the action at level k is given by setting

$$\sigma.\lambda=\sigma\cdot(\lambda+\rho)-\rho\quad\text{and}\quad\gamma.\lambda=\lambda+(k+r)\gamma\quad\text{for }\sigma\in\Sigma,\ \gamma\in\Lambda.$$

We denote the resulting group of affine-linear transformations of V^* by $\widetilde{\Sigma}[k]$, and note that the action is defined in such a way that

$$\sigma.\lambda+\rho=\sigma\cdot(\lambda+\rho)\quad\text{and}\quad(\gamma.\lambda+\rho)/\widehat{k}=\gamma+(\lambda+\rho)/\widehat{k}. \tag{52}$$

It is easy to verify that the stabilizer subgroup

$$\Sigma_r^+\stackrel{\text{def}}{=}\operatorname{Stab}(\theta_1[k],\widetilde{\Sigma}[k])\subset\widetilde{\Sigma}[k]$$

is generated by the transpositions $s_{i,i+1}$ for $1 \le i \le r-2$, and the reflection $\alpha^{r-1,r} \circ s_{r-1,r}$; similarly,

$$\Sigma_r^- \stackrel{\text{def}}{=} \mathrm{Stab}(\theta_{-1}[k], \widetilde{\Sigma}[k]) \subset \widetilde{\Sigma}[k]$$

is generated by $s_{i,i+1}$ for $2 \le i \le r-1$, and the reflection $\alpha^{1,2} \circ s_{1,2}$.

Then Proposition 8.6 maybe recast in the following form: the polynomial $q_1(k;\lambda)$ is anti-invariant with respect to the copy Σ_r^+ of the symmetric group Σ_r, while $q_{-1}(k;\lambda)$ is anti-invariant with respect to the copy Σ_r^- of the symmetric group Σ_r.

The following statement is straightforward:

Lemma 8.7 *Both subgroups $\Sigma_r^\pm$ are isomorphic to Σ_r, and for $r > 2$, the two subgroups generate the affine Weyl group $\widetilde{\Sigma}[k]$.*

8.3 The Weyl antisymmetry of the polynomials p_1 and p_{-1}

Following (22), we define the two polynomials

$$p_{\pm 1}(k;\lambda) = \sum_{\boldsymbol{B} \in \mathbb{D}} \mathrm{iBer}_{\boldsymbol{B}}[w_\Phi^{1-2g}(x/\widehat{k})](\widehat{\lambda}/\widehat{k} - [\theta_{\pm 1}]_{\boldsymbol{B}}),$$

where

$$\theta_1 = \frac{1}{r} \cdot (1,1,\dots,1) - x_r \quad \text{and} \quad \theta_{-1} = -\frac{1}{r} \cdot (1,1,\dots,1) + x_1.$$

Proposition 8.8 *The polynomial $p_1(k;\lambda)$ is anti-invariant with respect to Σ_r^+, and $p_{-1}(k;\lambda)$ is anti-invariant with respect to Σ_r^-.*

Proof The points $\theta_{\pm 1}[k]$ are the fixed points of the actions of $\Sigma^\pm$, and clearly $\lim_{k\to\infty} \theta_{\pm 1}[k]/k = \theta_{\pm 1}$. This means that we can fix a small open ball $D \subset V^*$ centered at θ_1 such that

$$\lambda/k \in D \implies (\sigma.\lambda + \rho)/\widehat{k} \sim \theta_1 \text{ for all } \sigma \in \Sigma^+. \tag{53}$$

Then for $\lambda/k \in D$ we have

$$p_1(k;\lambda) = \sum_{\boldsymbol{B} \in \mathbb{D}} \mathrm{iBer}_{\boldsymbol{B}}[w_\Phi^{1-2g}(x/\widehat{k})](\{\widehat{\lambda}/\widehat{k}\}_{\boldsymbol{B}}).$$

Now let us consider a generator of Σ^+ of the type $\sigma = s_{i,i+1}$ for some $1 \le i \le r-2$. Using (52) and Lemma 4.5, and the fact that $\sigma \cdot w_\Phi = -w_\Phi$, we obtain

$$p_1(k;\sigma.\lambda) = \sum_{\boldsymbol{B} \in \mathbb{D}} \mathrm{iBer}_{\boldsymbol{B}}[w_\Phi^{1-2g}(x/\widehat{k})](\sigma \cdot \{\widehat{\lambda}/\widehat{k}\}_{\boldsymbol{B}}) = \sum_{\boldsymbol{B} \in \mathbb{D}} \mathrm{iBer}_{\boldsymbol{B}}[(-w_\phi)^{1-2g}(x/\widehat{k})](\{\widehat{\lambda}/\widehat{k}\}_{\boldsymbol{B}}) = -p_1(k;\lambda).$$

The case of the last generator $\alpha^{r-1,r} \circ s_{r-1,r}$ is similar, but after the substitution we need to use the equality $\{\alpha^{r-1,r} + \widehat{\lambda}/\widehat{k}\}_{\boldsymbol{B}} = \{\widehat{\lambda}/\widehat{k}\}_{\boldsymbol{B}}$ to obtain $p_1(k;\widehat{k}\alpha^{r-1,r} + s_{r-1,r}.\lambda) = -p_1(k;\lambda)$. □

8.4 Proof of Theorem 4.8(I)

Recall that in Lemma 4.1 we introduced a chamber structure on $\Delta \subset V^*$ created by the walls $S_{\Pi,l}$, where $\Pi = (\Pi', \Pi'')$ is a nontrivial partition, and $l \in \mathbb{Z}$. Before we proceed, we introduce some extra notation. Denote by

$$\check{\Delta} = \{(k; a) \mid a/k \in \Delta\} \subset \mathbb{R}^{>0} \times V^*$$

the cone over $\Delta \subset V^*$, and let

$$\check{\Delta}^{\mathrm{reg}} = \{(k; a) \mid a/k \in \Delta \text{ is regular}\} \subset \check{\Delta}$$

be the set of its regular points. Denote by $\check{S}_{\Pi,l} \subset \check{\Delta}$ the cone over the wall $S_{\Pi,l} \subset \Delta$; then $\check{\Delta}^{\mathrm{reg}}$ is the complement of the union of walls $\check{S}_{\Pi,l}$ in $\check{\Delta}$. Finally, denote by $\check{\Delta}^{\mathrm{reg}}_{\Lambda}$ the intersection of the lattice $\mathbb{Z}^{>0} \times \Lambda$ with $\check{\Delta}^{\mathrm{reg}}$.

By substituting $\varsigma = \lambda/k$, we can consider the left-hand side and the right-hand side of formula (I) of Theorem 4.8 as functions in $(k, \lambda) \in \check{\Delta}^{\mathrm{reg}}_{\Lambda}$. We denote by $q(k; \lambda)$ and $p(k; \lambda)$ the left-hand side and the right-hand side, correspondingly.

We showed that $q(k; \lambda)$ and $p(k; \lambda)$ are *polynomials* on the cone over each chamber in Δ; see Theorem 4.4, Section 2.4. We proved that the wall-crossing terms — ie the differences between polynomials on neighboring chambers — for $q(k; \lambda)$ (see Theorem 6.13) and for $p(k; \lambda)$ (see Proposition 4.18) coincide, hence there exists a polynomial $\Theta(k; \lambda)$ on $\mathbb{Z}^{>0} \times \Lambda$ such that the restriction of $\Theta(k; \lambda)$ to $\check{\Delta}^{\mathrm{reg}}_{\Lambda}$ is equal to the difference $p(k; \lambda) - q(k; \lambda)$.

Now for $r > 2$, we can conclude that

$$\Theta(k; \lambda) = p_1(k; \lambda) - q_1(k; \lambda) = p_{-1}(k; \lambda) - q_{-1}(k; \lambda),$$

where $p_{\pm 1}(k; \lambda)$ and $q_{\pm 1}(k; \lambda)$ are the restrictions of $p(k; \lambda)$ and $q(k; \lambda)$ to two specific chambers defined in Sections 8.3 and 8.2. Then, according to Propositions 8.6 and 8.8, the polynomial $\Theta(k; \lambda)$ is anti-invariant with respect to the action of the subgroups $\Sigma_r^{\pm}$, and hence by Lemma 8.7, it is anti-invariant under the action of the entire affine Weyl group $\widetilde{\Sigma}[k]$. It is easy to see that any such polynomial function has to vanish, and thus $p(k; \lambda) = q(k; \lambda)$, and this completes the proof of part (I) of Theorem 4.8 for the case when $\lambda/k \in \Delta$ is regular.

As in Corollary 4.10, we can extend $p(k; \lambda)$ from the interior of each chamber to its boundary by polynomiality. Clearly, to prove part (I) of Theorem 4.8 for the cases when λ/k is not regular, it is sufficient to show that these extensions from the chambers containing λ/k in their closure give the same value on $(k; \lambda)$. It follows from Remark 10.4 that this is the case, and this completes the proof of part (I) of Theorem 4.8; see Remark 4.9.

9 Rank 2, two points

Unfortunately, the argument above does not work for $r = 2$ because in this case, $\theta_1[k] = \theta_{-1}[k]$, the groups Σ_r^- and Σ_r^+ coincide, and thus they do not generate the entire affine Weyl group. The way out is to pass to the 2–punctured case.

9.1 Wall-crossing

We will thus fix two points $p, s \in C$, and study the moduli space of rank-2, stable parabolic bundles W with fixed determinant isomorphic to $\mathcal{O}(pd)$, with parabolic structure given by a line $F_1 \subset W_p$ with weight $(c, -c)$, and a line $G_1 \subset W_s$ with weight $(a, -a)$.

Now we need to repeat the analysis of our work so far in this somewhat simpler case; some details will thus be omitted.

Set $d = 0$; then the space of admissible weights (see Figure 6) is a square

$$\square = \{(c, a) \mid 1 > 2c > 0,\ 1 > 2a > 0\},$$

which has two adjacent chambers defined by the conditions

$$c > a \quad \text{and} \quad c < a.$$

Denote the corresponding moduli spaces by $P_0(c > a)$ and $P_0(c < a)$. Again, we have universal bundles over $P_0(c > a) \times C$ and $P_0(c < a) \times C$, which we will denote by the same symbol U; this bundle is endowed with two flags, $\mathcal{F}_1 \subset \mathcal{F}_2 = U_p$ and $\mathcal{G}_1 \subset \mathcal{G}_2 = U_s$. For $\mu, \lambda \in \mathbb{Z}$, we introduce the line bundle

$$\mathcal{L}(k; \lambda, \mu) = \det(U_p)^{k(1-g)} \otimes \det(\pi_*(U))^{-k} \otimes (\mathcal{F}_2/\mathcal{F}_1)^{\lambda} \otimes (\mathcal{F}_1)^{-\lambda} \otimes (\mathcal{G}_2/\mathcal{G}_1)^{\mu} \otimes (\mathcal{G}_1)^{-\mu}.$$

We repeat the construction of the master space from Section 5.2, choosing a point (c^0, c^0) on the wall and two points

$$(c, a)^{\pm} = (c^0, c^0) \pm \epsilon(1, 0) \in \square, \quad \text{where } \epsilon \in \mathbb{Q}_{>0},$$

from the adjacent chambers. We can identify the fixed-point set Z^0 as follows.

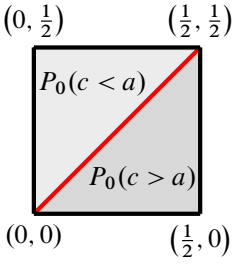


Figure 6: The space of admissible weights in the case of rank $r = 2$, two points.

Lemma 9.1 *The locus Z^0 defined in Proposition 5.2 is*

$$Z^0 \simeq \mathrm{Jac}^0 \simeq \{V = L \oplus L^{-1} \mid L_p = F_1,\, L_s^{-1} = G_1\}.$$

As in Section 6.1, denote by $\mathcal{J}$ the universal bundle over $\mathrm{Jac}^0 \times C$ normalized in such a way that $c_1(\mathcal{J})_{(0)} = 0$; see (8). Define

$$\eta \in H^2(\mathrm{Jac}) \quad \text{by} \quad \Big(\sum_i c_1(\mathcal{J})_{(e_i)} \otimes e_i\Big)^2 = -2\eta \otimes \omega.$$

We have then $\int_{\mathrm{Jac}} e^{\eta m} = m^g$ for $m \in \mathbb{Z}$.

Let $\pi : \mathrm{Jac}^0 \times C \to \mathrm{Jac}^0$ be the projection and N_{Z^0} the equivariant normal bundle to Z^0 in Z. Then, as in Lemma 6.6, Proposition 6.9 and Lemma 6.5, we obtain the identifications

- $N_{Z^0} = R^1_T\pi_*(\mathrm{ParHom}(\mathcal{J}, \mathcal{J}^{-1})) \oplus R^1_T\pi_*(\mathrm{ParHom}(\mathcal{J}^{-1}, \mathcal{J}))$, where $T \simeq \mathbb{C}^*$–action has weights $(-1, 1)$,
- $E(N_{Z^0})^{-1} = (-1)^g \big(2\sinh\big(\frac{1}{2}u\big)\big)^{-2g} \exp(4\eta)$,
- $\mathrm{ch}_T(\mathcal{L}(k; \lambda, \mu)|_{Z^0}) = \exp(2k\eta)\exp(u(\lambda - \mu))$.

Now we define the polynomials

$$h_>(k; \lambda, \mu) \stackrel{\text{def}}{=} \chi(P_0(c > a), \mathcal{L}(k; \lambda, \mu)) \quad \text{and} \quad h_<(k; \lambda, \mu) \stackrel{\text{def}}{=} \chi(P_0(c < a), \mathcal{L}(k; \lambda, \mu)).$$

Applying Theorem 5.7, we obtain the following expression for their difference.

Lemma 9.2 *The wall-crossing term equals*

$$h_>(k; \lambda, \mu) - h_<(k; \lambda, \mu) = \ (-1)^g (2k + 4)^g \operatorname*{Res}_{u=0} \frac{\exp(u(\lambda - \mu))}{\big(2\sinh\big(\frac{1}{2}u\big)\big)^{2g}}\, du.$$

9.2 Symmetry

Denote by $P_{-1}(c > a)$ the image of the moduli space $P_0(c > a)$ under the Hecke isomorphism $\mathcal{H}$ (see Section 7) at the point p, and by $P_{-1}(c < a)$ the image of the moduli space $P_0(c < a)$ under the Hecke isomorphism $\mathcal{H}$ at the point s.

We have the following analogue of Lemma 8.3.

Lemma 9.3 *Denote by N_{-1} the moduli space of rank-2 degree-(-1) stable bundles on C, and by UN any universal bundle over $N_{-1} \times C$. Then the moduli spaces $P_{-1}(c > a)$ and $P_{-1}(c < a)$ are isomorphic to the bundle $\mathbb{P}(UN_p) \times \mathbb{P}(UN_s)$ over N_{-1}.*

Denote by $\mathcal{T}[p]$ and $\mathcal{T}[s]$ the vertical tangent lines of $\mathbb{P}(UN_p)$ and $\mathbb{P}(UN_s)$, respectively, and by $\mathcal{L}_{-1}$ the pullback of the ample generator of the Picard group of N_{-1} to $\mathbb{P}(UN_p) \times \mathbb{P}(UN_s)$; see Lemma 8.4. Then a simple calculation shows the following.

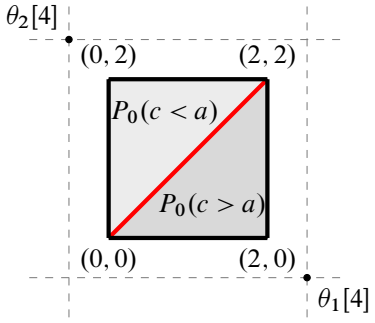


Figure 7: The space of admissible weights in the case of $k=4$, $r=2$, two points.

Lemma 9.4 *Under the Hecke isomorphism $\mathcal{H}$ at p, the line bundle $\mathcal{L}(2k;\lambda,\mu)$ on $P_0(c>a)$ corresponds to the line bundle $\mathcal{L}_{-1}^k\otimes\mathcal{T}[p]^{-\lambda+k}\otimes\mathcal{T}[s]^{\mu}$ on $P_{-1}(c>a)$.*

Under the Hecke isomorphism $\mathcal{H}$ at the point s, $\mathcal{L}(2k;\lambda,\mu)$ on $P_0(c<a)$ corresponds to the line bundle $\mathcal{L}_{-1}^k\otimes\mathcal{T}[p]^{\lambda}\otimes\mathcal{T}[s]^{-\mu+k}$ on $P_{-1}(c<a)$.

As in Section 8.2, applying Serre duality for families of curves (see Proposition 8.2) to the line bundles on the two $\mathbb{P}^1\times\mathbb{P}^1$ bundles over N_{-1}, we obtain that the polynomials $h_>(k;\lambda,\mu)$ and $h_<(k;\lambda,\mu)$ are anti-invariant under the action of the Weyl group $\Sigma_2\times\Sigma_2$ with the center at $\theta_1[k]=\left(\frac{1}{2}(k+1),-\frac{1}{2}\right)$ and $\theta_2[k]=\left(-\frac{1}{2},\frac{1}{2}(k+1)\right)$, correspondingly; see Figure 7. In other words, we obtain the following four identities.

Lemma 9.5
$$h_>(k;\lambda,\mu)=-h_>(k;\lambda,-\mu-1)=-h_>(k;-\lambda+k+1,\mu),$$
$$h_<(k;\lambda,\mu)=-h_<(k;-\lambda-1,\mu)=-h_<(k;\lambda,-\mu+k+1).$$

Now define the polynomials

$$\widetilde{h}_>(k;\lambda,\mu)=(-1)^{g-1}(2k+4)^g\operatorname*{Res}_{u=0}\frac{\exp(u(\lambda+\mu+1))-\exp(u(\lambda-\mu))}{\left(2\sinh\left(\frac{1}{2}u\right)\right)^{2g}(1-e^{u(k+2)})}\,du,$$

$$\widetilde{h}_<(k;\lambda,\mu)=(-1)^{g-1}(2k+4)^g\operatorname*{Res}_{u=0}\frac{\exp(u(\lambda+\mu+1))-\exp(u(\lambda-\mu+k+2))}{\left(2\sinh\left(\frac{1}{2}u\right)\right)^{2g}(1-e^{u(k+2)})}\,du,$$

and from here we can follow the logic of the proof of part (I) of Theorem 4.8.

Proposition 9.6 *The polynomials introduced above, in fact, coincide*:

$$h_>(k;\lambda,\mu)=\widetilde{h}_>(k;\lambda,\mu)\quad\text{and}\quad h_<(k;\lambda,\mu)=\widetilde{h}_<(k;\lambda,\mu).$$

Proof It is a simple exercise to show that $\widetilde{h}_>(k;\lambda,\mu)$ and $\widetilde{h}_<(k;\lambda,\mu)$ satisfy the identities appearing in Lemmas 9.2 and 9.5, and hence the polynomial

$$\Theta(k;\lambda,\mu)=h_>(k;\lambda,\mu)-\widetilde{h}_>(k;\lambda,\mu)=h_<(k;\lambda,\mu)-\widetilde{h}_<(k;\lambda,\mu)$$

satisfies all four Σ_2–symmetries listed in Lemma 9.5. These groups together generate a double action of the affine Weyl group $\widetilde{\Sigma}$ in λ and μ separately, and this implies the vanishing of Θ. □

As $P_0(c>a)$ is a $\mathbb{P}^1$–bundle over the moduli space of rank-2 degree-0 stable parabolic bundles $P_0(c,-c)$, substituting $\mu=0$ in $\widetilde{h}_>$, we obtain the Verlinde formula for rank 2.

Corollary 9.7 $\chi(P_0(c,-c),\mathscr{L}_0(k;\lambda)) = (-1)^{g-1}(2k+4)^g \operatorname{Res}_{u=0} \dfrac{\exp\big(u\big(\lambda+\frac{1}{2}\big)\big)}{\big(2\sinh\big(\frac{1}{2}u\big)\big)^{2g-1}(1-e^{u(k+2)})}\,du.$

10 The combinatorics of the $[Q,R]=0$

In this section, we give a proof of the second part of Theorem 4.8. Let $\lambda/k\in\Delta$, and fix a regular element $\varsigma\in\Delta$ in a chamber containing λ/k in its closure, and another regular element $\widehat{\varsigma}\in\Delta$ containing $\widehat{\lambda}/\widehat{k}$ in its closure. Our goal is to prove the equality $p_\varsigma(k;\lambda)=p_{\widehat{\varsigma}}(k;\lambda)$, where we define

$$p_c(k;\lambda)=\widetilde{N}_{r,k}\sum_{\boldsymbol{B}\in\mathscr{D}} \mathrm{iBer}_{\boldsymbol{B}}[w_\Phi^{1-2g}(x/\widehat{k})](\widehat{\lambda}/\widehat{k}-[c]_{\boldsymbol{B}}) \tag{54}$$

for a regular $c\in\Delta$ and diagonal basis $\mathscr{D}$. This is a subtle statement, which is a combinatorial-geometric projection of the idea of quantization commutes with reduction (or $[Q,R]=0$ for short; see [17; 25]).

If $\lambda/k\sim\widehat{\lambda}/\widehat{k}$, ie when λ/k and $\widehat{\lambda}/\widehat{k}$ are regular elements in the same chamber in Δ, then $p_\varsigma(k;\lambda)=p_{\widehat{\varsigma}}(k;\lambda)$ is a tautology. We assume thus that this is not the case, and denote by $\mathscr{S}(k,\lambda)$ the set of walls separating ς and $\widehat{\varsigma}$, or containing either λ/k or $\widehat{\lambda}/\widehat{k}$ or both. Equivalently, the wall $S_{\Pi,l}$ belongs to $\mathscr{S}(k,\lambda)$ if

$$(\lambda/k)_{\Pi'}\geq l\geq(\widehat{\lambda}/\widehat{k})_{\Pi'}\quad\text{or}\quad(\lambda/k)_{\Pi'}\leq l\leq(\widehat{\lambda}/\widehat{k})_{\Pi'},$$

where $c_{\Pi'}=\sum_{i\in\Pi'}c_i$ for an element $c=(c_1,\dots,c_r)\in V^*$. Clearly, there is a path in Δ connecting ς and $\widehat{\varsigma}$, which intersects only walls from $\mathscr{S}(k,\lambda)$ in a generic points. Then to prove the equality $p_\varsigma(k;\lambda)=p_{\widehat{\varsigma}}(k;\lambda)$, it is enough to show the following, at first sight somewhat surprising, fact.

Proposition 10.1 *Assume $g\geq 1$, $\lambda/k\in\Delta$, $S_{\Pi,l}\in\mathscr{S}(k,\lambda)$ and let $c^\pm\in\Delta$ be two points in two neighboring chambers separated by the wall $S_{\Pi,l}$. Then*

$$p_{c^+}(k;\lambda)=p_{c^-}(k;\lambda). \tag{55}$$

Proof The difference of the two sides of (55) is expressed as a residue in (34). The integral in (34) is a rational expression in the variable t, and our plan is to show by degree count in t and t^{-1} that its residues at zero and at ∞ vanish. We define the degree of the quotient of two polynomials $R=P/Q$ of the variable t as $\deg_t(R)=\deg_t(P)-\deg_t(Q)$, and we set $\deg_{t^{-1}}(R)=\deg_t(R(t^{-1}))$. Then, clearly,

$$\deg_t(R)<0\implies\operatorname*{Res}_{t=\infty}R\,\frac{dt}{t}=0\qquad\text{and}\qquad\deg_{t^{-1}}(R)<0\implies\operatorname*{Res}_{t=0}R\,\frac{dt}{t}=0.$$

A convenient expression for (34) will be (46), where we change variables via $t = e^u$. In what follows, we will always tacitly assume this substitution, and we will write, for example, $\deg_{t^{\pm 1}}(1/(e^u - e^{-u})) = -1$. We thus obtain a formula of the form $\mathrm{Res}_{t=0,\infty} f(t)\, dt/t$, and to show that this is zero, it is sufficient to show that $\deg_t(f) < 0$ and $\deg_{t^{-1}}(f) < 0$.

Now we observe that the variable u occurs only in the first line of (46), and thus, calculating the degrees in t and t^{-1} separately, we obtain the formula

$$\deg_{t^{\pm 1}}(f) = \pm(k\delta - rl) + (1-2g)\deg_{t^{\pm 1}}(\tau \cdot w_u^{\times}) + \deg_{t^{\pm 1}}(\exp(\tau \cdot \rho_u^{\times})). \tag{56}$$

Recall that here, δ represents the distance of λ/k from the wall $S_{\Pi,l}$, while $w_u^{\times}$ and $\rho_u^{\times}$, represent the parts of the Weyl denominator and the ρ–shift corresponding to roots connecting Π' and Π'', respectively.

We begin the study of this expression with some simple remarks. We recall that the permutation τ preserves the partition $\Pi = (\Pi', \Pi'')$, and thus we have

$$\deg_{t^{\pm 1}}(\tau \cdot w_u^{\times}) = \deg_{t^{\pm 1}}(w_u^{\times}) = \tfrac{1}{2} r' r''.$$

Using, in addition, that $\rho_u^{\times}$ is linear in u, we obtain

$$\deg_t(\exp(\tau \cdot \rho_u^{\times})) = -\deg_{t^{-1}}(\exp(\tau \cdot \rho_u^{\times})) = \deg_t(\exp(\rho_u^{\times})).$$

Combining these equalities, and assuming $g \geq 1$, we arrive at the following conclusion.

Lemma 10.2 *The inequality*

$$|(k\delta - rl) + \deg_t(\exp(\rho_u^{\times}))| < \tfrac{1}{2} r' r'' \tag{57}$$

implies the vanishing of the wall-crossing term: *equality* (55).

Before we proceed, we introduce some notation. Denote by

$$\mathrm{Inv}(\Pi) = \{(i,j) \mid \Pi' \ni i > j \in \Pi''\}$$

the set of "inverted" pairs of elements of the partition Π. The number of these pairs $|\mathrm{Inv}(\Pi)|$ coincides with the standard notion of length of the shuffle permutation $\phi \in \Sigma_r$ introduced in Section 6.

Each pair (i,j) which is not inverted contributes $+\frac{1}{2}u$ to $\rho_u^{\times}$, while each inverted pair contributes $-\frac{1}{2}u$, and thus we have

$$\deg_t(\exp(\rho_u^{\times})) = \tfrac{1}{2} r' r'' - |\mathrm{Inv}(\Pi)|. \tag{58}$$

Also, recall the notation $c_{\Pi'} = \sum_{i \in \Pi'} c_i$ for an element $c = (c_1, \dots, c_r) \in V^*$; in particular, we have $(\lambda/k)_{\Pi'} = l + \delta$ and

$$\rho_{\Pi'} = \sum_{i \in \Pi'} \tfrac{1}{2}(r+1) - i.$$

The following is a simple exercise, whose proof will be omitted:

$$\deg_t(\exp(\rho_u^{\times})) = \rho_{\Pi'}. \tag{59}$$

Now we come to a key point of our argument.

Lemma 10.3 *If the intersection of the wall $S_{\Pi,l}$ with Δ is nonempty, then*

$$-\tfrac{1}{2}r'r'' < lr - \rho_{\Pi'} < \tfrac{1}{2}r'r''. \tag{60}$$

Proof Pick a point $c = (c_1, \dots, c_r)$ in the intersection $S_{\Pi,l} \cap \Delta$, and recall that for any $1 \le i < j \le r$, we have $0 < c_i - c_j < 1$, and

$$\sum_{i\in\Pi'} c_i = -\sum_{i\in\Pi''} c_i = l.$$

Then

$$-|\mathrm{Inv}(\Pi)| < \sum_{(i,j)\in\mathrm{Inv}(\Pi)} (c_i - c_j) \le \sum_{i\in\Pi'}\sum_{j\in\Pi''} (c_i - c_j),$$

and, similarly,

$$\sum_{i\in\Pi'}\sum_{j\in\Pi''} (c_i - c_j) < r'r'' - |\mathrm{Inv}(\Pi)|.$$

Now, since

$$\sum_{i\in\Pi'}\sum_{j\in\Pi''} (c_i - c_j) = r''\sum_{i\in\Pi'} c_i - r'\sum_{j\in\Pi''} c_j = lr,$$

we can conclude

$$-|\mathrm{Inv}(\Pi)| < lr < r'r'' - |\mathrm{Inv}(\Pi)|.$$

In view of (58) and (59), these inequalities are equivalent to (60), and this completes the proof. □

Now we are ready to prove (55). The condition $S_{\Pi,l} \in \mathcal{S}(k,\lambda)$, ie that $S_{\Pi,l}$ separates λ/k and $\widehat{\lambda}/\widehat{k}$ or contains λ/k or $\widehat{\lambda}/\widehat{k}$, may occur in two ways.

- $(\lambda/k)_{\Pi'} \ge l \ge (\widehat{\lambda}/\widehat{k})_{\Pi'}$, which is equivalent to the two inequalities $\delta \ge 0$ and $lk + lr \ge \lambda_{\Pi'} + \rho_{\Pi'}$. After canceling lk and reordering the terms, we can rewrite these as

$$0 \ge k\delta - lr + \rho_{\Pi'} \ge \rho_{\Pi'} - lr. \tag{61}$$

  Using Lemma 10.3 then we can conclude that

$$0 \ge k\delta - lr + \rho_{\Pi'} > -\tfrac{1}{2}r'r'',$$

  which, in view of the equality (59), implies the necessary estimate (57).

- The second case is similar: $(\lambda/k)_{\Pi'} \le l \le (\widehat{\lambda}/\widehat{k})_{\Pi'}$ is equivalent to $\delta \le 0$ and $lk + lr \le \lambda_{\Pi'} + \rho_{\Pi'}$. This leads to

$$0 \le k\delta - lr + \rho_{\Pi'} \le \rho_{\Pi'} - lr, \tag{62}$$

  which, in turn, implies

$$0 \le k\delta - lr + \rho_{\Pi'} < \tfrac{1}{2}r'r'',$$

  and hence (57).

This completes the proof of Proposition 10.1: indeed, a simple calculation shows that if $\lambda/k \in \Delta$ then $\widehat{\lambda}/\widehat{k} \in \Delta$, so the conditions of Lemma 10.3 hold. We have just shown that this implies (57), and according to Lemma 10.2, we can conclude the vanishing of the wall-crossing term (55). □

Remark 10.4 If $\lambda/k \in \Delta$ is nonregular, then it belongs to some wall from the set $\mathcal{S}(k,\lambda)$. Hence Proposition 10.1 implies that the right-hand side of formula (I) of Theorem 4.8 is a well-defined function on the cone over Δ:

$$\{(k,\lambda) \in \mathbb{Z}^{>0} \times \Lambda \mid \lambda/k \in \Delta\}.$$

References

[1] **A Alekseev**, **E Meinrenken**, **C Woodward**, *The Verlinde formulas as fixed point formulas*, J. Symplectic Geom. 1 (2001) 1–46 MR Zbl

[2] **M F Atiyah**, **R Bott**, *A Lefschetz fixed point formula for elliptic complexes, II: Applications*, Ann. of Math. 88 (1968) 451–491 MR Zbl

[3] **M F Atiyah**, **R Bott**, *The Yang–Mills equations over Riemann surfaces*, Philos. Trans. Roy. Soc. A 308 (1983) 523–615 MR Zbl

[4] **U N Bhosle**, *Parabolic vector bundles on curves*, Ark. Mat. 27 (1989) 15–22 MR Zbl

[5] **J-M Bismut**, **F Labourie**, *Symplectic geometry and the Verlinde formulas*, from "Surveys in differential geometry: differential geometry inspired by string theory" (S-T Yau, editor), Surv. Differ. Geom. 5, International, Boston, MA (1999) 97–311 MR Zbl

[6] **H U Boden**, **Y Hu**, *Variations of moduli of parabolic bundles*, Math. Ann. 301 (1995) 539–559 MR Zbl

[7] **I V Dolgachev**, **Y Hu**, *Variation of geometric invariant theory quotients*, Inst. Hautes Études Sci. Publ. Math. 87 (1998) 5–56 MR Zbl

[8] **E González**, **C Woodward**, *Quantum Kirwan for quantum K–theory*, from "Facets of algebraic geometry, I" (P Aluffi, D Anderson, M Hering, M Mustaţă, S Payne, editors), Lond. Math. Soc. Lect. Note Ser. 472, Cambridge Univ. Press (2022) 265–332 MR Zbl

[9] **A Grothendieck**, *Techniques de construction et théorèmes d'existence en géométrie algébrique, IV: Les schémas de Hilbert*, from "Séminaire Bourbaki, 1960/1961", Benjamin, New York (1966) exposé 221 MR Reprinted in "Séminaire Bourbaki 6", Soc. Math. France, Paris (1995) 249–276

[10] **R Hartshorne**, *Algebraic geometry*, Graduate Texts in Math. 52, Springer (1977) MR Zbl

[11] **L C Jeffrey**, *The Verlinde formula for parabolic bundles*, J. Lond. Math. Soc. 63 (2001) 754–768 MR Zbl

[12] **L C Jeffrey**, **F C Kirwan**, *Localization for nonabelian group actions*, Topology 34 (1995) 291–327 MR Zbl

[13] **L C Jeffrey**, **F C Kirwan**, *Intersection theory on moduli spaces of holomorphic bundles of arbitrary rank on a Riemann surface*, Ann. of Math. 148 (1998) 109–196 MR Zbl

[14] **Y Loizides**, **E Meinrenken**, *The decomposition formula for Verlinde sums*, Ann. Inst. Fourier (Grenoble) 72 (2022) 1207–1248 MR Zbl

[15] **V B Mehta**, **C S Seshadri**, *Moduli of vector bundles on curves with parabolic structures*, Math. Ann. 248 (1980) 205–239 MR Zbl

[16] **E Meinrenken**, *Twisted K–homology and group-valued moment maps*, Int. Math. Res. Not. 2012 (2012) 4563–4618 MR Zbl

[17] **E Meinrenken**, **R Sjamaar**, *Singular reduction and quantization*, Topology 38 (1999) 699–762 MR Zbl

[18] **D Mumford**, **J Fogarty**, *Geometric invariant theory*, 2nd edition, Ergebnisse der Math. 34, Springer (1982) MR Zbl

[19] **M S Narasimhan**, **S Ramanan**, *Geometry of Hecke cycles, I*, from "C P Ramanujam: a tribute" (K G Ramanathan, editor), Tata Inst. Fundam. Res. Stud. Math. 8, Springer (1978) 291–345 MR Zbl

[20] **N Nitsure**, *Cohomology of the moduli of parabolic vector bundles*, Proc. Indian Acad. Sci. Math. Sci. 95 (1986) 61–77 MR Zbl

[21] **C S Seshadri**, *Moduli of vector bundles on curves with parabolic structures*, Bull. Amer. Math. Soc. 83 (1977) 124–126 MR Zbl

[22] **C Sorger**, *La formule de Verlinde*, from "Séminaire Bourbaki, 1994/1995", Astérisque 237, Soc. Math. France, Paris (1996) exposé 794, pages 87–114 MR Zbl

[23] **A Szenes**, *Iterated residues and multiple Bernoulli polynomials*, Int. Math. Res. Not. 1998 (1998) 937–956 MR Zbl

[24] **A Szenes**, *Residue theorem for rational trigonometric sums and Verlinde's formula*, Duke Math. J. 118 (2003) 189–227 MR Zbl

[25] **A Szenes**, **M Vergne**, $[Q, R] = 0$ *and Kostant partition functions*, Enseign. Math. 63 (2017) 471–516 MR Zbl

[26] **C Teleman**, **C T Woodward**, *The index formula for the moduli of G–bundles on a curve*, Ann. of Math. 170 (2009) 495–527 MR Zbl

[27] **M Thaddeus**, *Stable pairs, linear systems and the Verlinde formula*, Invent. Math. 117 (1994) 317–353 MR Zbl

[28] **M Thaddeus**, *Geometric invariant theory and flips*, J. Amer. Math. Soc. 9 (1996) 691–723 MR Zbl

[29] **M Vergne**, *Multiplicities formula for geometric quantization, I*, Duke Math. J. 82 (1996) 143–179 MR Zbl

[30] **M Vergne**, *Multiplicities formula for geometric quantization, II*, Duke Math. J. 82 (1996) 181–194 MR Zbl

[31] **E Verlinde**, *Fusion rules and modular transformations in 2D conformal field theory*, Nuclear Phys. B 300 (1988) 360–376 MR Zbl

[32] **E Witten**, *Two-dimensional gauge theories revisited*, J. Geom. Phys. 9 (1992) 303–368 MR Zbl

[33] **D Zagier**, *On the cohomology of moduli spaces of rank two vector bundles over curves*, from "The moduli space of curves" (R Dijkgraaf, C Faber, G van der Geer, editors), Progr. Math. 129, Birkhäuser, Boston, MA (1995) 533–563 MR Zbl

Section de mathématiques, Université de Genève
Geneva, Switzerland
Section de mathématiques, Université de Genève
Geneva, Switzerland
andras.szenes@unige.ch, olga.trapeznikova@unige.ch

Proposed: Frances Kirwan Received: 17 March 2022
Seconded: Jim Bryan, Mark Gross Revised: 31 December 2022

msp

Geometry & Topology **28**:5 (2024) 2313–2343
DOI: 10.2140/gt.2024.28.2313
Published: 24 August 2024

The signature and cusp geometry of hyperbolic knots

ALEX DAVIES
ANDRÁS JUHÁSZ
MARC LACKENBY
NENAD TOMAŠEV

We introduce a new real-valued invariant, called the natural slope of a hyperbolic knot in the 3–sphere, which is defined in terms of its cusp geometry. We show that twice the knot signature and the natural slope differ by at most a constant times the hyperbolic volume divided by the cube of the injectivity radius. This inequality was discovered using machine learning to detect relationships between various knot invariants. It has applications to Dehn surgery and to 4–ball genus. We also show a refined version of the inequality, where the upper bound is a linear function of the volume, and the slope is corrected by terms corresponding to short geodesics that link the knot an odd number of times.

57K10, 57K31, 57K32, 68T07

1 Introduction

In low-dimensional topology, there are two very different types of invariant: those derived from hyperbolic structures on 3–manifolds, and those with connections to 4–dimensional manifolds. Of the latter type, one of the most fundamental invariants is the signature of a knot. Our main goal in this paper is to establish a new and unexpected connection between these two fields. We will show that the cusp geometry of a hyperbolic knot in the 3–sphere encodes information about the signature of the knot.

One of the most important geometric features of a hyperbolic knot K is its maximal cusp. The boundary of this cusp is a Euclidean torus that forms the boundary of a regular neighbourhood of K. This torus is isometric to $\mathbb{C}/\Lambda$ for a lattice Λ in $\mathbb{C}$. The meridian and longitude of the knot give generators μ and λ for Λ. The parallelogram in $\mathbb{C}$ spanned by 0, μ, λ and $\mu+\lambda$ forms a fundamental domain for the action of Λ on $\mathbb{C}$. We introduce a new geometric quantity, called the *natural slope*, that measures how far this parallelogram is from being right-angled. It can be defined by the formula

$$\operatorname{slope}(K) = \operatorname{Re}\Big(\frac{\lambda}{\mu}\Big).$$

Alternatively, natural slope can be defined as follows. Pick a geodesic on the torus $\mathbb{C}/\Lambda$ that represents a meridian. Choose any point on such a geodesic and send off a geodesic orthogonally from this point. It runs along the knot and eventually it comes back to the initial meridian; see Figure 1. In doing so, it has

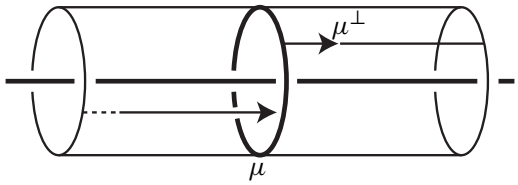


Figure 1: A geodesic running in the direction $\mu^\perp$ that is perpendicular to the meridian μ. By the time it returns to the meridian, it has travelled one longitude minus some multiple s of the meridian. This real number s is the natural slope of K.

gone along a longitude minus some number s of meridians. This number s is not necessarily an integer because the geodesic may return to a different point along the meridian from where it started. This real number s is the natural slope of K.

Quantities with a resemblance to the natural slope have been defined by [Benard et al. 2021; Degtyarev et al. 2022]. However, these other quantities do not seem to be directly related to natural slope, and none of these previous articles seems to provide a connection between hyperbolic geometry and signature.

Experimentally, starting from the plot in Figure 2, we have observed that the natural slope of K is very highly correlated with $2\sigma(K)$, where $\sigma(K)$ is the signature. See Figure 3 for plots of signature versus slope for knots up to 16 crossings in the Regina census [Burton et al. 1999–2021] and for random knots generated by SnapPy [Culler et al. 2021] having 10–80 crossings in their SnapPy-simplified forms. Our goal in this paper is to prove that such a surprising connection holds and to explore its consequences. Our first main result, which we prove in Section 4, establishes that slope(K) is approximately equal to $2\sigma(K)$, but with an additive error that can be bounded by geometric quantities.

Theorem 1.1 *There exists a constant c_1 such that, for any hyperbolic knot K,*

$$|2\sigma(K) - \text{slope}(K)| \leq c_1 \,\text{vol}(K)\,\text{inj}(K)^{-3}.$$

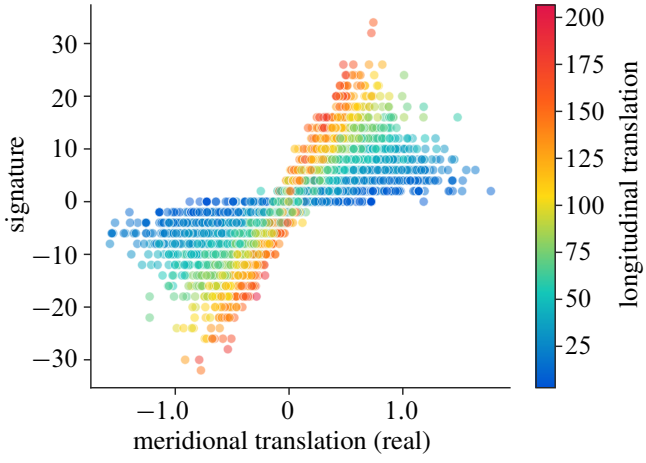


Figure 2: A plot of signature versus the real part of the meridional translation, Re(μ), coloured by longitudinal translation, for a dataset of knots randomly generated by SnapPy.

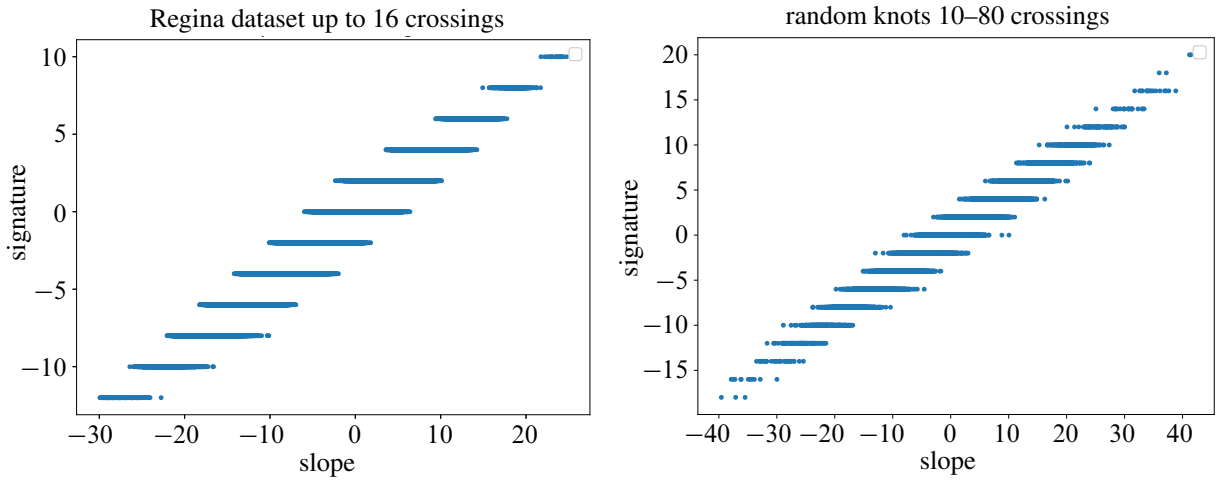


Figure 3: A plot of signature versus slope for knots up to 16 crossings in the Regina census (left) and for a dataset of knots randomly generated by SnapPy having 10–80 crossings in their SnapPy-simplified form (right).

Here $\mathrm{vol}(K)$ is the hyperbolic volume of the complement of K. Also, $\mathrm{inj}(K)$ is the *injectivity radius* of $S^3 \setminus K$, which we define to be

$$\mathrm{inj}(K) = \inf\{\mathrm{inj}_x(S^3 \setminus K) : x \in (S^3 \setminus K) \setminus N\},$$

where N is a maximal cusp and $\mathrm{inj}_x(S^3 \setminus K)$ denotes the injectivity radius of a point x in $S^3 \setminus K$. Note that, although $\mathrm{inj}(K)^{-3}$ appears in the inequality in Theorem 1.1, in practice $\mathrm{inj}(K)$ tends not to be particularly small. (See Figure 12, for example.) Experimental evidence, which we provide in Section 7, suggests that c_1 should be quite small: perhaps $c_1 = 0.3$ suffices. This is based on the largest value 0.234 of $|2\sigma(K) - \mathrm{slope}(K)|\,\mathrm{inj}(K)^3/\mathrm{vol}(K)$ that we managed to obtain by studying a class of knots that are closures of certain braids.

One might wonder whether there is a constant c_2 such that

$$|2\sigma(K) - \mathrm{slope}(K)| \le c_2\,\mathrm{vol}(K)$$

for every hyperbolic knot K. However, we show in Corollary 5.1 that there cannot exist such a constant. We achieve this by exhibiting a sequence of examples that are obtained by twisting three strands of a hyperbolic knot. Nevertheless, we can estimate $\sigma(K)$ in terms of geometric quantities, with an error that is at most a linear function of $\mathrm{vol}(K)$. The main term in this estimate is $\frac{1}{2}\,\mathrm{slope}(K)$, but there are also correction terms that are defined using the complex length of short geodesics. From the complex lengths, the following parameters are computed:

Definition 1.2 Let γ be a geodesic in a hyperbolic 3–manifold with complex length $\mathrm{cl}(\gamma)$. Here $\mathrm{cl}(\gamma)$ is chosen so that $\mathrm{Im}(\mathrm{cl}(\gamma)) \in (-\pi, \pi]$. The *twisting parameter* $\mathrm{tw}(\gamma) = (\mathrm{tw}_p(\gamma), \mathrm{tw}_q(\gamma))$ is the pair (p, q) of coprime integers satisfying the following:

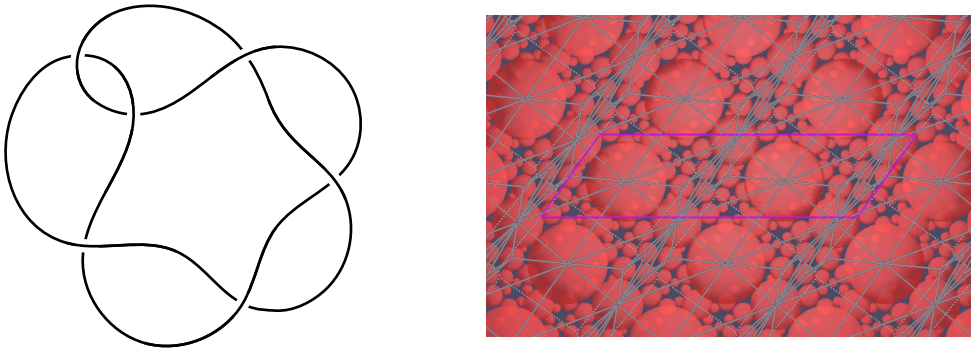


Figure 4: The stevedore knot 6_1 (left), which is a slice knot, and its cusp torus (right), as provided by SnapPy [Culler et al. 2021]. The longitude is 3.9279 and the meridian is $0.7237 + 1.0160i$. Its natural slope is 1.8267 and its signature is 0.

(1) p is even and q is odd and nonnegative.

(2) Subject to this condition, the quantity $|\mathrm{cl}(\gamma)p + 2\pi i q|$ is minimised.

(3) If there are several values of (p, q) for which this quantity is minimised, then choose the pair that is minimal with respect to lexicographical ordering.

Consider a hyperbolic knot K in S^3. For any $\varepsilon \in \mathbb{R}_+$ less than the Margulis constant ε_3, let $\mathrm{OddGeo}\left(\frac{1}{2}\varepsilon\right)$ denote the set of geodesics with length less than $\frac{1}{2}\varepsilon$ and having odd linking number with K. For $p, q \in \mathbb{Z}_+$, the signature correction term $\kappa(p, q)$ is given by Definition 4.2 and satisfies

$$\kappa(p, q) = -\sigma(T(p, q)) - \tfrac{1}{2}pq,$$

where $T(p, q)$ is the (p, q)–torus knot. Then we have the following refinement of Theorem 1.1, which we prove in Section 6, that does not depend on the injectivity radius:

Theorem 1.3 *Let ε_3 be the Margulis constant and let $\varepsilon \in (0, \varepsilon_3)$. Then there is a constant c_4 (depending on ε) such that, for any hyperbolic knot K, the quantities $\sigma(K)$ and*

$$\tfrac{1}{2}\,\mathrm{slope}(K) - \sum_{\gamma \in \mathrm{OddGeo}(\varepsilon/2)} \kappa(\mathrm{tw}_p(\gamma), \mathrm{tw}_q(\gamma))$$

differ by at most $c_4\,\mathrm{vol}(K)$.

Figures 4 and 5 illustrate the relationship between signature and slope in Theorem 1.1 for the knots 6_1 and $12a52$, respectively.

Theorem 1.1 has applications in low-dimensional topology. On the one hand, the signature of K controls the cusp shape, which in turn has consequences for the possible exceptional surgeries on K. On the other hand, the cusp shape controls the signature, which has consequence for the 4–ball genus of K. We now provide these applications.

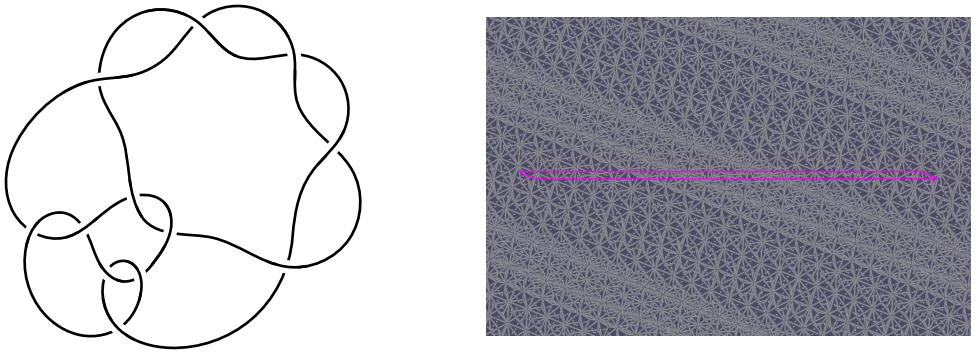


Figure 5: The knot 12a52 (left) and its cusp torus (right). The longitude is 27.7228 and the meridian is $-1.2838 + 0.5145i$. Its natural slope is -18.6064 and its signature is -8. Note how far the parallelogram is from being right-angled; this is the defining feature of having very positive or very negative slope.

1.1 An application to Dehn surgery

Cusp geometry is well known to control the exceptional surgeries on a knot K. Recall that a slope s on $\partial N(K)$ is said to be *exceptional* if the manifold $K(s)$ obtained by Dehn filling along s does not admit a hyperbolic structure.

The *length* of a slope $s = q/p \in \mathbb{Q}$, denoted by $\ell(s)$, is defined to be the length of any geodesic representative of $s = p\lambda + q\mu$ in the boundary of the maximal cusp. A theorem of Agol [2000] and Lackenby [2000] states that, if $\ell(s) > 6$, then s is not exceptional.

We relate slope length to natural slope, using the following simple geometric lemma, which we will prove in Section 2:

Lemma 1.4 *If K is a hyperbolic knot, then the length of the slope q/p satisfies*

$$\ell\Big(\frac{q}{p}\Big) \geq |p \operatorname{slope}(K) + q|.$$

Hence, if q/p is exceptional, then

$$\frac{q}{p} \in \Big[-\operatorname{slope}(K) - \frac{6}{p}, -\operatorname{slope}(K) + \frac{6}{p}\Big].$$

Given that $\operatorname{slope}(K)$ and $2\sigma(K)$ are highly correlated, one would expect that any exceptional slope q/p should lie within a short interval around $-2\sigma(K)$. It is also known that $|p| \leq 8$, by a theorem of Lackenby and Meyerhoff [2013]. Hence, we obtain a bounded set of slopes that contains all the exceptional ones, and that is defined in terms of the signature.

An interesting case is the $(-2, 3, 7)$–pretzel knot $12n242$. This has signature -8 and slope approximately -18.215. It has seven exceptional slopes: 16, 17, 18, $\frac{37}{2}$, 19 and 20. Observe that these slopes are

concentrated in a short interval $[16, 20]$ that contains both $-\text{slope}(K)$ and $-2\sigma(K)$. This close correlation between the exceptional slopes and $-2\sigma(K)$ seems to be a phenomenon that had not previously been observed. Specifically, we have the following consequence of our main theorem:

Corollary 1.5 *If K is a hyperbolic knot and q/p is a slope satisfying*

$$\left|\frac{q}{p} + 2\sigma(K)\right| > \frac{6}{|p|} + c_1 \operatorname{vol}(K) \operatorname{inj}(K)^{-3} \quad \text{or} \quad |p| > 8,$$

then the manifold $K(q/p)$ obtained by q/p Dehn surgery along K is hyperbolic.

Theorem 1.3 gives a similar bound on slopes resulting in hyperbolic surgeries that does not involve $\text{inj}(K)$.

1.2 An application to 4–ball genus

One of the most important 4–dimensional quantities associated to a knot K is its *4–ball genus* $g_4(K)$. This is defined to be the minimal possible genus of a smoothly embedded compact orientable surface in the 4–ball B^4 with boundary K. One can also define the *topological 4–ball genus* $g_4^{\text{top}}(K)$ by considering locally flat topologically embedded compact orientable surfaces with boundary K. The inequality $g_4(K) \geq g_4^{\text{top}}(K)$ is immediate.

The following result provides a lower bound on $g_4^{\text{top}}(K)$ in terms of purely hyperbolic data. This follows immediately from our main theorem together with the well-known inequality $g_4^{\text{top}}(K) \geq \frac{1}{2}|\sigma(K)|$.

Corollary 1.6 *The topological 4–ball genus $g_4^{\text{top}}(K)$ of a hyperbolic knot K satisfies*

$$g_4^{\text{top}}(K) \geq \tfrac{1}{4}|\text{slope}(K)| - \tfrac{1}{4}c_1 \operatorname{vol}(K) \operatorname{inj}(K)^{-3}.$$

This corollary seems to be the first time that information about the 4–ball genus has been obtained in terms of hyperbolic geometry. Again, Theorem 1.3 gives a similar lower bound on $g_4^{\text{top}}(K)$ that does not involve $\text{inj}(K)$.

1.3 Spanning surfaces

Theorem 1.1 is proved using a new construction of spanning surfaces with a specified slope. It is of independent interest.

Theorem 1.7 *There is a constant c_3 such that every hyperbolic knot K in S^3 has an unoriented spanning surface F satisfying*

$$|\chi(F)| \leq c_3 \operatorname{vol}(K) \operatorname{inj}(K)^{-3}.$$

Moreover, the boundary slope of this surface is $n/1$, where n is an even integer that is closest to $\text{slope}(K)$.

We prove this in Section 3. The *crosscap number* of a knot K is the minimum of $b_1(F)$ for F an unoriented spanning surface of K. When K is hyperbolic, the above theorem gives an upper bound on a version of the crosscap number where ∂F has slope $n/1$.

Theorem 1.1 is proved by combining this result with a theorem of Gordon and Litherland [1978], which asserts that one can compute the signature of a knot K using any spanning surface F for K; see Theorem 4.1.

Note that slope also gives a lower bound on the Seifert genus:

$$\frac{1}{4\pi}|\text{slope}(K)| + \tfrac{1}{2} \leq g(K);$$

see Proposition 2.5.

1.4 Highly twisted knots

In Section 5, we show the following result for highly twisted knots:

Theorem 1.8 *Let K be a knot in the 3–sphere and let $C_1, \dots, C_n$ be a collection of disjoint simple closed curves in the complement of K that bound disjoint discs. Suppose that $S^3 \setminus (K \cup C_1 \cup \cdots \cup C_n)$ is hyperbolic. Let $K(q_1, \dots, q_n)$ be the knot obtained from K by adding q_i full twists to the strings going through C_i for each $i \in \{1, \dots, n\}$. Let ℓ_i be the linking number between C_i and K, when they are both given some orientation. Suppose that $\ell_1, \dots, \ell_m$ are even and $\ell_{m+1}, \dots, \ell_n$ are odd. Then there is a constant k, depending on K and $C_1, \dots, C_n$, such that, provided each $|q_i|$ is sufficiently large,*

$$\left|\text{slope}(K(q_1, \dots, q_n)) + \sum_{i=1}^{n} \ell_i^2 q_i\right| \leq k, \quad \left|\sigma(K(q_1, \dots, q_n)) + \left(\frac{1}{2}\sum_{i=1}^{m} \ell_i^2 q_i + \frac{1}{2}\sum_{i=m+1}^{n} (\ell_i^2 - 1) q_i\right)\right| \leq k.$$

The slight difference between the behaviour of $\sigma(K(q_1, \dots, q_n))$ and that of $\frac{1}{2}\,\text{slope}(K(q_1, \dots, q_n))$ as the q_i tend to infinity enables us to construct families of knots that show the injectivity radius cannot be dropped from Theorem 1.1.

1.5 Methodology

One of the novel aspects of this work was the use of machine learning. We embarked with the aim of discovering new relationships between various 3–dimensional invariants. By using machine learning, we observed an unexpected nonlinear relationship between $\sigma(K)$ and $\text{Re}(\mu)$, the real part of the meridional translation μ. This led us to define the natural slope, which we observed to have a strong linear correlation with $\sigma(K)$. Theorems 1.1 and 1.3 are the results of our attempts to prove this correlation.

2 Hyperbolic knots and natural slope

A knot K is hyperbolic if its complement $S^3 \setminus K$ admits a complete finite-volume hyperbolic metric. By the Mostow rigidity theorem [1968], the hyperbolic structure is unique up to isometry; hence, every geometric invariant of the hyperbolic structure on $S^3 \setminus K$ is a topological invariant of the knot. For example, the volume $\text{vol}(K) := \text{vol}(S^3 \setminus K)$ and the injectivity radius $\text{inj}(K)$ defined in the introduction are such invariants.

For a pair of coprime integers (p,q), the *torus knot* $T(p,q)$ is one that can be drawn on the surface of the standard torus in the 3–sphere and winds p times in the longitude direction and q times along the meridian. Given a knot K in S^3 and a knot K' in the solid torus $S^1\times D^2$, one can form the satellite of K with pattern K' by mapping the solid torus in a neighbourhood of K, and considering the image of K'. By the work of Thurston [Morgan 1984], a knot is hyperbolic if and only if it is not a torus knot or a satellite knot. In particular, every hyperbolic knot is prime, ie not the connected sum of two nontrivial knots. In other words, one can build all knots from hyperbolic knots and torus knots using satellite operations.

Definition 2.1 For any hyperbolic knot K, the end of $S^3\setminus K$ has a neighbourhood called a *cusp*. The boundary ∂N of a maximal cusp neighbourhood $N\subset S^3\setminus K$ is a Euclidean torus. Identify ∂N with $\mathbb{C}/\Lambda$, where $\mathbb{C}$ is the complex plane and Λ is a lattice in $\mathbb{C}$. We arrange this identification so that the longitude lifts to a straight line in $\mathbb{C}$ starting at 0 and ending at some $\lambda\in\mathbb{R}_{>0}$. This is the knot's *longitudinal translation*. Given this normalisation, the meridian lifts to a straight line starting at 0 and ending at some complex number μ with $\mathrm{Im}(\mu)>0$. This is the *meridional translation* of K.

We remark that the real part of meridional translation $\mathrm{Re}(\mu)$ in the KnotInfo data set [Livingston and Moore 2021] for knots with at most 12 crossings is listed without signs. However, SnapPy [Culler et al. 2021] does compute the sign for hyperbolic knots.

Note that $|\mu|\leq 6$, where $|\mu|$ denotes the length of the meridian. Indeed, by work of Agol [2000] and Lackenby [2000], Dehn filling along a slope longer than 6 gives a hyperbolic 3–manifold, while Dehn filling along the meridian is S^3, which is not hyperbolic. Furthermore, any curve on the cusp torus ∂N has length at least 1. In particular, $|\mu|\geq 1$.

If S is an essential surface with connected boundary in a hyperbolic 3–manifold, then $\ell(\partial S)\leq -2\pi\chi(S)$; see Cooper and Lackenby [1998, Theorem 5.1] or Hass, Rubinstein and Wang [Hass et al. 1999, (6)]. If S is a Seifert surface for a knot K, then $\chi(S)=1-2g(S)$. Hence, if K is hyperbolic, then

$$|\lambda|\leq 4\pi g(K)-2\pi, \tag{2.2}$$

where $g(K)$ is the Seifert genus of K.

For the maximal cusp neighbourhood N, we have

$$\mathrm{vol}(\partial N)=2\,\mathrm{vol}(N)\leq 2\,\mathrm{vol}(K),$$

and $\mathrm{vol}(\partial N)\leq|\lambda||\mu|$. On the other hand, by a result of Lackenby and Purcell [2016], there is a constant C such that, for K alternating,

$$C\,\mathrm{vol}(K)\leq\mathrm{vol}(\partial N).$$

Based on experimental data, one might ask if this also holds for random knots.

Definition 2.3 The *natural slope* $\mathrm{slope}(K)$ of a hyperbolic knot K is defined as follows. Let $\mu^\perp$ be a unit vector at the origin of $\mathbb{C}$ orthogonal to μ. Then some multiple of $\mu^\perp$ is equal to $\lambda-s\mu$ for some $s\in\mathbb{R}$. Then $\mathrm{slope}(K):=s$.

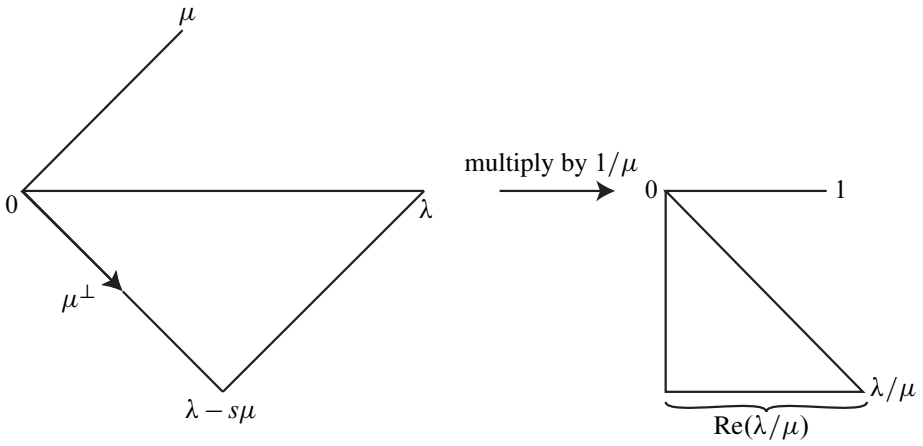


Figure 6: The calculation of natural slope.

Lemma 2.4
$$\operatorname{slope}(K) = \operatorname{Re}\Big(\frac{\lambda}{\mu}\Big) = \frac{\lambda \operatorname{Re}(\mu)}{|\mu|^2}.$$

Proof Figure 6 shows a lift of the cusp torus to the complex plane $\mathbb{C}$. The point $\lambda - s\mu$ is shown (which is a multiple of $\mu^\perp$). If we apply the transformation to $\mathbb{C}$ that is multiplication by $1/\mu$, then $\mu^\perp$ becomes purely imaginary. So $\lambda/\mu - s$ is purely imaginary. Hence, $s = \operatorname{Re}(\lambda/\mu)$. This is also equal to $\lambda \operatorname{Re}(\mu)/|\mu|^2$. □

We are now ready to prove Lemma 1.4 from the introduction:

Proof of Lemma 1.4 We have $\ell(q/p) = |p\lambda + q\mu|$. Since $\lambda \in \mathbb{R}$,
$$\ell\Big(\frac{q}{p}\Big)^2 = p^2\lambda^2 + 2pq\lambda \operatorname{Re}(\mu) + q^2|\mu|^2.$$
On the other hand, by Lemma 2.4, we have $\operatorname{slope}(K) = \lambda \operatorname{Re}(\mu)/|\mu|^2$. Hence,
$$|p \operatorname{slope}(K) + q|^2 = p^2\lambda^2 \frac{\operatorname{Re}(\mu)^2}{|\mu|^4} + 2pq\lambda \frac{\operatorname{Re}(\mu)}{|\mu|^2} + q^2 \le \ell\Big(\frac{q}{p}\Big)^2$$
since $|\mu| \ge 1$. □

Slope gives a lower bound on the Seifert genus:

Proposition 2.5 *If K is a hyperbolic knot in S^3, then*
$$\frac{1}{4\pi}|\operatorname{slope}(K)| + \tfrac{1}{2} \le g(K).$$

Proof By (2.2), we have $|\lambda| \le 4\pi g(K) - 2\pi$. Furthermore, $|\mu| \ge 1$. Together with Lemma 2.4, we obtain that
$$|\operatorname{slope}(K)| = |\lambda| \frac{|\operatorname{Re}(\mu)|}{|\mu|^2} \le \frac{|\lambda|}{|\mu|} \le 4\pi g(K) - 2\pi,$$
and the result follows. □

3 Proof of Theorem 1.7

The key to proving Theorem 1.7 is the construction of a nice triangulation of a hyperbolic knot complement:

Proposition 3.1 *There is a constant c_1 such that, for every hyperbolic knot K in S^3 with embedded cusp neighbourhood N, there is a triangulation $\mathcal{T}$ of $M := S^3 \setminus (K \cup \mathrm{int}(N))$ with the following properties*:

(1) *The number t of tetrahedra of $\mathcal{T}$ is at most $c_1 \,\mathrm{vol}(K)\,\mathrm{inj}(K)^{-3}$.*

(2) *If n is a closest even integer to* $\mathrm{slope}(K)$, *then $\nu := \lambda - n\mu$* (see Definition 2.3) *is a normal curve in ∂M that intersects each edge at most once.*

Proof We remark that the validity of the conclusion in the proposition does not depend on the choice of embedded cusp neighbourhood N. We will pick N as follows. Let $N_{\max}$ be the maximal cusp neighbourhood. Retract this to form the embedded cusp neighbourhood N, so each point of ∂N has distance 0.5 from $\partial N_{\max}$. Note that the Euclidean metric on ∂N is obtained from that of $\partial N_{\max}$ by scaling by the factor $e^{-0.5} = 1/\sqrt{e}$.

Let $\varepsilon := \frac{1}{2}\,\mathrm{inj}(K)$. We use a variation of Jørgensen's and Thurston's method [Thurston 1979, Section 5.11] to build the triangulation $\mathcal{T}$. (See also [Breslin 2009; Kobayashi and Rieck 2011].)

We pick a maximal collection of points in ∂M that are all at least $\frac{1}{8}\varepsilon$ from each other. We will extend this to a collection of points P in M without adding any new points in ∂M. Our aim is to ensure that the Voronoi diagram for P in M restricts to the Voronoi diagram for $P \cap \partial M$ in ∂M, where the latter is given its Euclidean metric. Recall that the Voronoi diagram [1908a; 1908b] corresponding to P is a cell structure of M where the interior of every 3–cell consists of the set of points in M that are closer to a specific point of P than any other point of P. Similarly, the Voronoi diagram for $P \cap \partial M$ is a cell structure of M where the interior of every 2–cell consists of the set of points in ∂M that are closer (in the Euclidean metric) to a specific point of $P \cap \partial M$ than any other point of $P \cap \partial M$.

The Voronoi diagram for M can be constructed as follows. The universal cover $\mathbb{H}^3 \to S^3 \setminus K$ restricts to the universal cover $\widetilde{M} \to M$. This set $\widetilde{M}$ is obtained from $\mathbb{H}^3$ by removing the interior of the inverse image of N. We may arrange that one component of this inverse image is a horoball $N_\infty = \{(x, y, z) : z \geq k\}$ in the upper half-space model for $\mathbb{H}^3$ for some $k > 0$. Let $\widetilde{P}$ denote the inverse image of P in $\widetilde{M}$. Each cell of the Voronoi diagram for M is the image of a cell for the Voronoi diagram for $\widetilde{P}$ in $\widetilde{M}$. Each 2–cell that does not lie in $\partial\widetilde{M}$ is equidistant from two points of $\widetilde{P}$. Hence, it is totally geodesic. Our aim is to ensure that each such 2–cell that intersects the horosphere ∂N_∞ is equidistant between two points of $\widetilde{P} \cap \partial N_\infty$. This will imply that the 2–cell intersects ∂N_∞ in a Euclidean geodesic arc. The union of these arcs forms the 1–skeleton of the Voronoi diagram for $\widetilde{P} \cap \partial N_\infty$ in ∂N_∞. Thus, we can deduce that the Voronoi diagram for P in M restricts to the Voronoi diagram for $P \cap \partial M$ in ∂M.

We now describe how the set P is chosen. We have already picked a maximal collection of points in ∂M that are all at least $\frac{1}{8}\varepsilon$ from each other. This set will be $P \cap \partial M$. We then add points to this set that lie in

the interior of M, but subject to the condition that each of these points in the interior of M has distance at least $\frac{1}{4}\varepsilon$ from the other points in the set. We stop when it is no longer possible to add any further points with this property. Let P be the resulting set of points.

By our choice of P, each point in ∂M has distance less than $\frac{1}{8}\varepsilon$ from some point of $P \cap \partial M$. It also has distance at least $\frac{1}{8}\varepsilon$ from each point of $P \cap \mathrm{int}(M)$. Thus, for each point of ∂M, each of its closest points in P also lies in ∂M.

Now consider a 2–cell of the Voronoi diagram for $\widetilde{M}$ that intersects ∂N_∞ but does not lie in ∂N_∞. This is equidistant between two points p_1 and p_2 of $\widetilde{P}$. The intersection between this 2–cell and ∂N_∞ is an arc. Let x be any point in the interior of this arc. Then x is equidistant between p_1 and p_2, and these are the closest two points of $\widetilde{P}$ to x. As argued above, any point of $\widetilde{P}$ that is closest to x must lie in $\partial\widetilde{M}$. We will show that, in fact, p_1 and p_2 lie in ∂N_∞. Suppose not. Then one of these points lies in $\partial\widetilde{M} \setminus \partial N_\infty$. The shortest arc from x to $\partial\widetilde{M} \setminus \partial N_\infty$ must run through the inverse image of $\partial N_{\max}$. One component of this inverse image is a horosphere about the point at infinity, with distance 0.5 from ∂N_∞. Hence, the length of this arc is at least 0.5. On the other hand, each point in $\partial\widetilde{M}$ has distance less than $\frac{1}{8}\varepsilon$ from some point of $P \cap \partial\widetilde{M}$. We will show below that $\frac{1}{8}\varepsilon < 0.12 < 0.5$, and hence this is a contradiction.

Thus, we have indeed guaranteed that the restriction to ∂M of the Voronoi diagram for P in M is the Voronoi diagram for $P \cap \partial M$ in ∂M, as claimed. We now subdivide each 2–cell of the Voronoi diagram for M into triangles without introducing any new vertices, and subdivide each 3–cell into tetrahedra by coning off from the point of P lying in it, obtaining the triangulation $\mathcal{T}$ of M. Since the restriction of the Voronoi diagram to ∂M agrees with that arising from its Euclidean metric, this implies that each triangle of $\mathcal{T}$ in ∂M is straight.

Since the open balls of radius $\frac{1}{16}\varepsilon$ about the points of P are pairwise disjoint,

$$|P|\,\mathrm{vol}\big(B\big(\tfrac{1}{16}\varepsilon\big)\big) \leq \mathrm{vol}(S^3 \setminus K),$$

where $B\big(\frac{1}{16}\varepsilon\big)$ is a ball in $\mathbb{H}^3$ of radius $\frac{1}{16}\varepsilon$.

We claim that the number t_p of tetrahedra of $\mathcal{T}$ incident to a point $p \in P$ is at most a universal constant k. Indeed, when p lies in the interior of M, t_p is exactly the number of triangles in the boundary 2–sphere S of the 3–cell of the Voronoi diagram containing p. When p lies in the boundary of M, t_p is the number of triangles in this sphere that are not incident to p. When a vertex of one of these triangles lies in the interior of M, it is equidistant from at least four points of P, one of which is p. When a vertex of the triangles lies on the boundary of M, it is equidistant from at least three points of P, one of which is p. So a vertex in S is specified by choosing two or three other points of P, each of which is at most $\frac{1}{2}\varepsilon$ from p. The ball $B\big(p, \frac{1}{2}\varepsilon\big)$ is embedded in $S^3 \setminus K$, since $\frac{1}{2}\varepsilon = \frac{1}{4}\,\mathrm{inj}(K)$, and hence lifts to a ball B in $\mathbb{H}^3$. The balls of radius $\frac{1}{16}\varepsilon$ about the inverse image of P in B are disjoint, and lie within $B\big(\frac{9}{16}\varepsilon\big)$. So the number

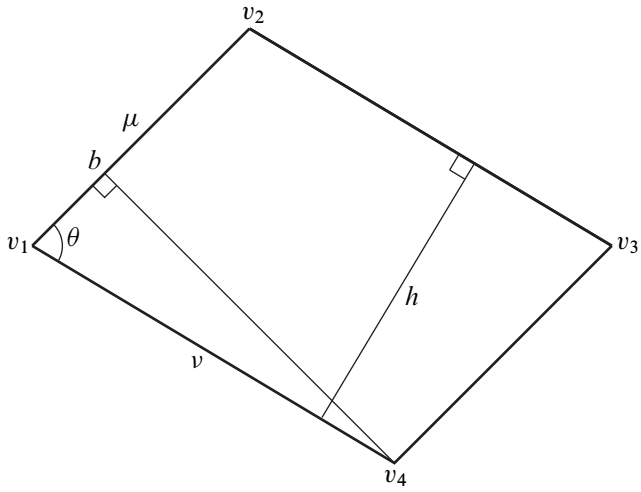


Figure 7: A fundamental domain D in $\partial N_{\max}$ with sides μ and ν.

of points of P at most $\frac{1}{2}\varepsilon$ from p is bounded above by

$$k_0 := \left\lfloor \frac{\operatorname{vol}\big(B\big(\frac{9}{16}\varepsilon\big)\big)}{\operatorname{vol}\big(B\big(\frac{1}{16}\varepsilon\big)\big)} \right\rfloor.$$

It follows that $t_p \leq k := \binom{k_0}{3}$. Then the total number of tetrahedra

$$t \leq k|P| \leq \frac{k \operatorname{vol}(K)}{\operatorname{vol}\big(B\big(\frac{1}{16}\varepsilon\big)\big)} \leq c_1 \operatorname{vol}(K) \operatorname{inj}(K)^{-3}$$

for a universal constant c_1.

We may pick the Euclidean geodesic representative for the slope ν so that it misses the vertices of $\mathcal{T}$. Hence, ν is a normal curve, because it is a Euclidean geodesic and each triangle of $\mathcal{T}$ in ∂M is straight. We now show ν does not intersect any triangle in ∂M more than once. Let D be a fundamental domain in $\partial N_{\max}$ with sides μ and ν. (See Figure 7.) We will show that the perpendicular distance h between the sides of D that are parallel to ν is at least 0.55. Hence, the perpendicular distance between sides of the corresponding fundamental domain in ∂N_∞ is at least $0.55/\sqrt{e} > 0.33$. On the other hand, we will show that the length of each edge of $\mathcal{T}$ in ∂M is at most 0.23. This will imply that, in the triangulation of ∂M, no triangle can run in D between these opposite sides, and hence that $\mathcal{T}$ satisfies property (2). This will complete the proof.

According to a theorem of Cao and Meyerhoff [2001], the area A of the boundary of the maximal cusp is at least 3.35. Let θ be the angle of two of the four corners of D satisfying $0 < \theta \leq \frac{1}{2}\pi$. Say that this angle is at the vertex v_1 of D, and label the remaining vertices v_2, v_3 and v_4 so that the line joining v_1 to v_2 has slope μ.

Let b be the perpendicular projection of v_4 onto the line joining v_1 and v_2. We claim that b lies between v_1 and v_2, or possibly equals one of these vertices. Place v_1 at the origin in the complex plane. Then

$v_2 = \pm\mu$ and $v_4 = \lambda - n\mu$. Now, by the definition of $s = \text{slope}(K)$, the perpendicular projection of $\lambda - s\mu$ onto the line through v_1 and v_2 is v_1. Hence, the perpendicular projection b of $\lambda - n\mu$ onto this line has distance $|n-s||\mu|$ from v_1. But n is a closest even integer to s, and so $|n-s| \leq 1$. Therefore, b lies between v_1 and v_2, or is equal to one of these points, as claimed.

Hence,

$$\tan\theta \geq \frac{A}{|\mu|^2}$$

and so

$$\sec^2\theta = 1 + \tan^2\theta \geq \frac{A^2 + |\mu|^4}{|\mu|^4}.$$

Therefore,

$$\sin^2\theta = 1 - \cos^2\theta \geq 1 - \frac{|\mu|^4}{A^2 + |\mu|^4} = \frac{A^2}{A^2 + |\mu|^4}.$$

So the distance h satisfies

$$h = |\mu|\sin\theta \geq \frac{|\mu|A}{\sqrt{A^2 + |\mu|^4}}.$$

The square of the reciprocal of this expression is

$$\frac{A^2 + |\mu|^4}{|\mu|^2 A^2} = \frac{1}{|\mu|^2} + \frac{|\mu|^2}{A^2}.$$

It is easy to check that this is a convex function of $|\mu|$ and hence its maximal value over the interval $1 \leq |\mu| \leq 6$ occurs when $|\mu| = 1$ or 6. It also is maximised by taking A as small as possible; in other words, $A = 3.35$. We deduce that h is at least

$$\frac{6 \cdot 3.35}{\sqrt{3.35^2 + 36^2}} \geq 0.55.$$

Hence, the perpendicular distance between sides of the corresponding fundamental domain in ∂N_∞ is at least $0.55/\sqrt{e} > 0.33$.

We now compare this to the maximal length of an edge of $\mathcal{T}$ in ∂M. Each triangle of $\mathcal{T}$ in ∂M lies within a disc centred at a point of $P \cap \partial M$ with radius at most $\frac{1}{8}\varepsilon$. Hence, each triangle has side length at most $\frac{1}{4}\varepsilon = \frac{1}{8}\,\text{inj}(K)$. Now the length L of the shortest slope s on $\partial N_{\max}$ is at most $|\mu| \leq 6$. This gives an upper bound on $\text{inj}(K)$, as follows. By applying an isometry to hyperbolic space, we may arrange that a component of the inverse image of $N_{\max}$ in upper half-space is $\{(x, y, z) : z \geq 1\}$. We may also arrange that a covering transformation corresponding to s is $(x, y, z) \mapsto (x + L, y, z)$. It therefore sends $(0, 0, 1)$ to $(L, 0, 1)$. The hyperbolic distance between these points is at most

$$2\ln\left(\tfrac{1}{2}(6 + \sqrt{40})\right) \leq 3.64.$$

Hence, $\text{inj}(K)$ is at most 1.82 and $\frac{1}{4}\varepsilon$ is at most 0.23. □

Proof of Theorem 1.7 Let the triangulation $\mathcal{T}$ and the curve ν be as in Propositions 3.1. Since $\nu = \lambda - n\mu$ for n even, $[\nu] = [\lambda] \in H_1(\partial M; \mathbb{Z}_2)$, so ν bounds an unoriented surface S in M. If we make S transverse to the 1–skeleton of $\mathcal{T}$, it defines a simplicial 1–cocycle $c \in C^1(M; \mathbb{Z}_2)$ via $c(e) = |S \cap e| \mod 2$ for each edge e of $\mathcal{T}$. If we connect the midpoints of the edges e of T such that $c(e) = 1$, we obtain a surface F that intersects each tetrahedron T of the triangulation $\mathcal{T}$ in at most one triangle or square. In particular, F is a normal surface. Furthermore, $\partial F = \nu$ as ν is a normal curve that intersects each triangle in ∂M at most once. Discard any closed components of F.

Let t be the number of tetrahedra of $\mathcal{T}$. Furthermore, write v, e and f for the number of vertices, edges and faces of F, respectively. By the above, $f \leq t$. Then $\chi(F) = v - e + f$ and, since F is not a disk, $|\chi(F)| = e - f - v$. Since every face of F is a triangle or a quadrilateral,

$$e \leq \tfrac{1}{2}(4f + e_\partial) \leq t + f + \tfrac{1}{2}e_\partial,$$

where e_∂ is the number of edges of F in ∂M. As $v \geq e_\partial$, we obtain that

$$|\chi(F)| \leq t \leq c_1 \operatorname{vol}(K) \operatorname{inj}(K)^{-3},$$

where the second inequality is property (1) of $\mathcal{T}$ in Proposition 3.1. □

4 The knot signature

Another fundamental knot invariant is the signature $\sigma(K)$. Given a Seifert surface S for K, ie a compact, oriented and connected surface with boundary K, one can define the Seifert form

$$Q_S \colon H_1(S) \times H_1(S) \to \mathbb{Z}$$

as follows: Given $a, b \in H_1(S)$, we write b^+ for the positive push-off of b into $S^3 \setminus S$. Then $Q_S(a, b) = \operatorname{lk}(a, b^+)$. If V is a matrix of Q_S, then $\sigma(K)$ is the signature of $V + V^T$. The signature is a 4–dimensional invariant, in the sense that it gives a lower bound on the topological 4–ball genus $g_4^{\text{top}}(K)$, which is the minimal genus of a compact, oriented, locally flat, connected surface bounded by K in the 4–ball B^4.

One can also compute the signature of a knot from unoriented surfaces using the work of Gordon and Litherland [1978]. Let F be an unoriented surface bounding a knot K in S^3. Let $\{b_1, \dots, b_n\}$ be a basis of $H_1(F)$, and let b_i' be the double push-off of b_i into $S^3 \setminus F$. Then the *Goeritz matrix* G_F is an $n \times n$ symmetric matrix with $(i, j)^{\text{th}}$ entry $\operatorname{lk}(b_i, b_j')$ for $i, j \in \{1, \dots, n\}$. Furthermore, the *normal Euler number* $e(F)$ of F is defined to be $-\operatorname{lk}(K, K')$, where K' is the framing of K given by F. Gordon and Litherland proved the following:

Theorem 4.1 *Let F be an unoriented surface bounding the knot K in S^3. Then*

$$\sigma(K) = \sigma(G_F) + \tfrac{1}{2}e(F),$$

where $\sigma(G_F)$ is the signature of the Goeritz matrix.

We are now ready to show how Theorem 1.1 follows from Theorem 1.7.

Proof of Theorem 1.1 Let F be the surface provided by Theorem 1.7, with boundary slope $\nu = \lambda - n\mu$, where n is a closest even integer to slope(K). Let G_F be the Goeritz matrix of F. Since

$$|\chi(F)| \leq c_1 \operatorname{vol}(K) \operatorname{inj}(K)^{-3},$$

we deduce that

$$b_1(F) \leq c_1 \operatorname{vol}(K) \operatorname{inj}(K)^{-3} + 1,$$

and so $|\sigma(G_F)| \leq c_1 \operatorname{vol}(K) \operatorname{inj}(K)^{-3} + 1$. Therefore,

$$\begin{aligned} |2\sigma(K) - \operatorname{slope}(K)| &\leq |2\sigma(K) - n| + 1 = |2\sigma(K) + \operatorname{lk}(K, \nu)| + 1 = |2\sigma(G_F)| + 1 \\ &\leq 2c_1 \operatorname{vol}(K) \operatorname{inj}(K)^{-3} + 3 \leq c_2 \operatorname{vol}(K) \operatorname{inj}(K)^{-3} \end{aligned}$$

for the absolute constant

$$c_2 := 2c_1 + \frac{3 \cdot 1.82^3}{2.0298} < 2c_1 + 8.92.$$

Indeed, for any hyperbolic knot K, we have $\operatorname{inj}(K) \leq 1.82$, as shown in the proof of Proposition 3.1, and $\operatorname{vol}(K) > 2.0298$, with the figure eight knot having the smallest volume, by [Cao and Meyerhoff 2001]. □

In the following definition, we introduce the signature correction $\kappa(p, q)$ for integers p and q, which is related to the signature of the (p, q)–torus knot. The correction terms in Theorem 1.3 are defined in terms of $\kappa(p, q)$.

Definition 4.2 For any pair of positive integers (p, q), we define the *signature correction* $\kappa(p, q)$ recursively as follows:

(1) If $p > 2q$ and q is odd, then $\kappa(p, q) = \kappa(p - 2q, q) - 1$.

(2) If $p > 2q$ and q is even, then $\kappa(p, q) = \kappa(p - 2q, q)$.

(3) If $p = 2q$, then $\kappa(p, q) = -1$.

(4) If $q \leq p < 2q$ and q is odd, then $\kappa(p, q) = -\kappa(q, 2q - p) - 1$.

(5) If $q \leq p < 2q$ and q is even, then $\kappa(p, q) = -\kappa(q, 2q - p) - 2$.

(6) If $p < q$, then $\kappa(p, q) = \kappa(q, p)$.

We extend κ to nonzero integers p and q by defining $\kappa(-p, q) = \kappa(p, -q) = -\kappa(p, q)$. When one of p or q is zero, $\kappa(p, q) = 0$.

It is reasonably clear that this gives a well-defined value of $\kappa(p, q)$. This is because it defines $\kappa(p, q)$ uniquely when $p = q$, and, when $p \neq q$, it defines $\kappa(p, q)$ in terms of some $\kappa(p', q')$ where either $q' < q$, or $q' = q$ and $p' < p$. However, the rationale for the definition comes from the following fact, due to Gordon, Litherland and Murasugi [Gordon et al. 1981]:

Theorem 4.3 *The signature of the (p,q)–torus link $T(p,q)$ satisfies*

$$\sigma(T(p,q)) = -\tfrac{1}{2}pq - \kappa(p,q).$$

The signature correction $\kappa(p,q)$ arises naturally as the signature of the Goeritz form of a surface bounding the (p,q)–torus knot, as follows:

Lemma 4.4 *Let V be the standard solid torus in S^3, and let $T(p,q)$ be the curve on ∂V that is the (p,q)–torus knot, where p is even and q is odd. Thus, p is the winding number of $T(p,q)$ in V. Then there is a compact unoriented surface F in V with boundary $T(p,q)$, and $\sigma(G_F) = -\kappa(p,q)$ for any such F.*

Proof Since p is even, $T(p,q)$ is trivial in $H_1(V;\mathbb{Z}_2)$. It therefore bounds an unoriented surface F in V. Applying Gordon and Litherland's signature formula (Theorem 4.1) to F, we deduce that

$$\sigma(T(p,q)) = \sigma(G_F) + \tfrac{1}{2}e(F).$$

The push-off K' of ∂F into F has linking number pq with ∂F. To see this, observe that K' is homologous in V to p times a core curve γ' of V. Similarly, ∂F is homologous in the solid torus $\mathrm{cl}(S^3 \setminus V)$ to q times its core curve γ, which is a meridian of γ'. Thus,

$$\mathrm{lk}(\partial F, K') = pq\,\mathrm{lk}(\gamma,\gamma') = pq;$$

hence, $e(F) = -pq$. So

$$\sigma(G_F) = \sigma(T(p,q)) + \tfrac{1}{2}pq = -\kappa(p,q),$$

where the final equality is Theorem 4.3. □

Lemma 4.5 *Let A be a nonsingular square matrix with real entries. Let A_+ be a nonsingular matrix obtained from A by adding a final row and final column. Then $\sigma(A_+)$ is either $\sigma(A)-1$ or $\sigma(A)+1$.*

Proof Let $\lambda_1 \le \cdots \le \lambda_n$ be the eigenvalues of A, and let $\lambda_1^+ \le \cdots \le \lambda_{n+1}^+$ be the eigenvalues of A_+. Cauchy's interlacing theorem states that

$$\lambda_1^+ \le \lambda_1 \le \lambda_2^+ \le \cdots \le \lambda_n \le \lambda_{n+1}^+.$$

Hence, the number of negative eigenvalues of A_+ is at least the number of negative eigenvalues of A and, similarly, the number of positive eigenvalues of A_+ is at least the number of positive eigenvalues of A. □

Lemma 4.6 *Let V be a solid torus embedded in S^3. Pick a slope λ on ∂V that has winding number 1 in V. Let K be the knot on ∂V that has slope $p\lambda + q\mu$, where μ is the meridian of V, and where p is even and q is odd. Then K bounds a compact unoriented surface F in V with the property that the Goeritz form G_F satisfies $|\sigma(G_F) + \kappa(p,q)| \le 2$.*

Proof Because p is even, K bounds a compact surface F in V. We may pick a basis $e_1,\dots,e_n$ for $H_1(F)$ so that $e_1,\dots,e_{n-1}$ have zero winding number around V. Let V' be an embedding of V in S^3 such that K is sent to $T(p,q)$. Let F' be the image of F.

We claim that the Goeritz forms G_F and $G_{F'}$ agree on the first $n-1$ rows and columns. To prove this, we view V' as the regular neighbourhood of a standard unknot embedded in the horizontal plane. Then, up to isotopy, V can be obtained from V' by applying Reidemeister moves and crossing changes to this unknot. None of these moves affects the first $n-1$ rows and columns of the Goeritz form, for the following reason. Any given entry of the Goeritz form is $\operatorname{lk}(b_i, b_j')$ for a suitable curve b_i in the surface and b_j' the double push-off of another curve in the surface. When the entry of the Goeritz form lies in the first $n-1$ rows and columns, these curves b_i and b_j' have zero winding number around the solid torus. Hence, geometrically, b_i winds an equal number of times around the solid torus in opposite directions, as does b_j'. So, when we perform a Reidemeister move or a crossing change to the solid torus and we compare the resulting projections of $b_i \cup b_j'$ to the horizontal plane, the sum of the signs of the crossings between b_i and b_j' remains unchanged. This sum is $2\operatorname{lk}(b_i, b_j')$. This proves the claim.

Hence, by Lemma 4.5, we have $|\sigma(G_F) - \sigma(G_{F'})| \leq 2$. But $\sigma(G_{F'}) = -\kappa(p,q)$ by Lemma 4.4. □

5 Highly twisted knots

The following is Theorem 1.8 from the introduction:

Theorem 1.8 *Let K be a knot in the 3–sphere and let $C_1, \dots, C_n$ be a collection of disjoint simple closed curves in the complement of K that bound disjoint discs. Suppose that $S^3 \setminus (K \cup C_1 \cup \cdots \cup C_n)$ is hyperbolic. Let $K(q_1, \dots, q_n)$ be the knot obtained from K by adding q_i full twists to the strings going through C_i for each $i \in \{1, \dots, n\}$. Let ℓ_i be the linking number between C_i and K, when they are both given some orientation. Suppose that $\ell_1, \dots, \ell_m$ are even and $\ell_{m+1}, \dots, \ell_n$ are odd. Then there is a constant k, depending on K and $C_1, \dots, C_n$, such that, provided each $|q_i|$ is sufficiently large,*

$$\left|\operatorname{slope}(K(q_1,\dots,q_n)) + \sum_{i=1}^{n} \ell_i^2 q_i\right| \leq k, \quad \left|\sigma(K(q_1,\dots,q_n)) + \left(\frac{1}{2}\sum_{i=1}^{m} \ell_i^2 q_i + \frac{1}{2}\sum_{i=m+1}^{n} (\ell_i^2 - 1) q_i\right)\right| \leq k.$$

One can use this to show that the factor $\operatorname{inj}(K)^{-3}$ cannot simply be dropped from Theorem 1.1 (see Conjecture 7.4 for what we expect for random knots):

Corollary 5.1 *There does **not** exist a constant c_2 such that*

$$|2\sigma(K) - \operatorname{slope}(K)| \leq c_2 \operatorname{vol}(K)$$

for every hyperbolic knot K.

Proof Pick $n = 1$ and $\ell_1 = 3$. Then $\operatorname{slope}(K(q_1)) \sim -9q_1$, whereas $2\sigma(K(q_1)) \sim -8q_1$. On the other hand, $\operatorname{vol}(K(q_1))$ is bounded. □

Proof of Theorem 1.8 The knot $K(q_1,\dots,q_n)$ is obtained by performing $-1/q_i$ surgery on C_i for each $i\in\{1,\dots,n\}$. Let L denote the link $K\cup C_1\cup\cdots\cup C_n$. By Thurston's hyperbolic Dehn surgery theorem, as all the $|q_i|$ tend to infinity, the hyperbolic structures on $S^3\setminus K(q_1,\dots,q_n)$ tend in the geometric topology to the hyperbolic structure on $S^3\setminus L$. In fact, more is true. Fix a horoball neighbourhood N of the cusps of $S^3\setminus L$ that is small enough that the cusp torus T surrounding K lies in the complement of N. Then, if all the $|q_i|$ are sufficiently large, the inclusion $(S^3\setminus L)\setminus N\to S^3\setminus K(q_1,\dots,q_m)$ is a bi-Lipschitz homeomorphism onto its image, with bi-Lipschitz constants that tend to 1 as all the $|q_i|$ tend to infinity. (See [Benedetti and Petronio 1992], for instance.)

Let $\lambda(K)$ be the longitude and $\mu(K)$ the meridian of K. These form a basis of the lattice $\Lambda(K)$, where the cusp torus of K in $S^3\setminus L$ is $\mathbb{C}/\Lambda(K)$. Let γ be the image of $\lambda(K)$ and μ the image of $\mu(K)$ on the cusp torus $\mathbb{C}/\Lambda(K(q_1,\dots,q_n))$ of $K(q_1,\dots,q_n)$. The curves γ and μ form a basis of the lattice $\Lambda(K(q_1,\dots,q_n))$. So we may assume that γ and μ are approximately constant complex numbers when the $|q_i|$ are large. However, we have *not* normalised the lattice so that γ is real. We know that there is some $N\in\mathbb{R}_+$ such that

$$N\mu^{\perp}=\gamma-s'\mu$$

for some $s'\in\mathbb{R}$. Here N, $\mu^{\perp}$, γ, s' and μ all depend on $q_1,\dots,q_n$. But N and s' tend to fixed real numbers as the $|q_i|$ go to infinity.

The key observation is that γ is *not* necessarily the longitude λ for $K(q_1,\dots,q_n)$. In fact, the linking number between γ and $K(q_1,\dots,q_n)$ is $\sum_i \ell_i^2 q_i$; see Figure 8. For suppose that the disc bounded by C_i

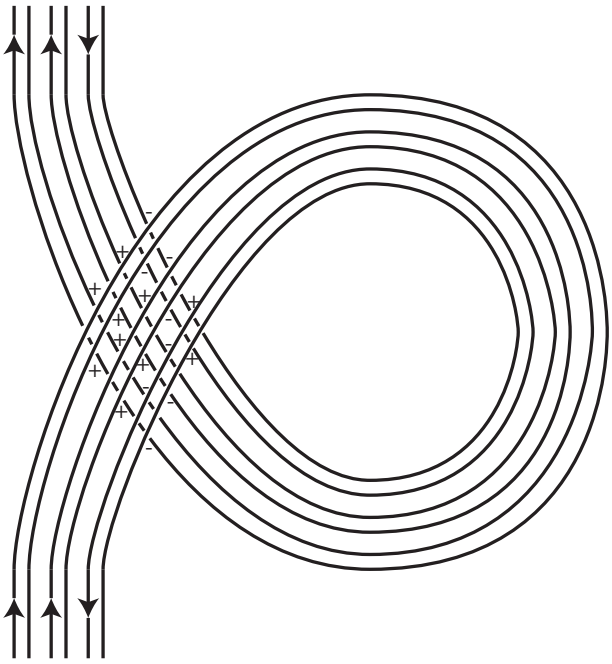


Figure 8: Each full twist about C_i changes the linking number between γ and $K(q_1,\dots,q_n)$ by ℓ_i^2.

intersects K in p_- points of negative sign and p_+ points of positive sign. So $\ell_i = p_+ - p_-$. Then, when we perform a full twist about C_i, we introduce $2(p_+ + p_-)^2$ new crossings between γ and $K(q_1, \dots, q_n)$. Of these, $2(p_+^2 + p_-^2)$ have positive sign and $4p_+p_-$ have negative sign. So the linking number between γ and $K(q_1, \dots, q_n)$ changes by

$$p_+^2 + p_-^2 - 2p_+p_- = \ell_i^2.$$

It follows that

$$\gamma = \lambda + \Big(\sum_{i=1}^{n} \ell_i^2 q_i\Big)\mu,$$

and hence

$$N\mu^\perp = \lambda - \Big(s' - \sum_{i=1}^{n} \ell_i q_i^2\Big)\mu.$$

We conclude that $\text{slope}(K(q_1, \dots, q_n)) = s' - \sum_{i=1}^n \ell_i q_i^2$. On the other hand, there is a constant k such that $|s'| \le k$ if $|q_1|, \dots, |q_n|$ are sufficiently large, which implies the first inequality of the theorem.

Recall that $\ell_{m+1}, \dots, \ell_n$ are odd. Suppose that $q_{m+1}, \dots, q_r$ are even and that $q_{r+1}, \dots, q_n$ are odd. Let $\mu_{m+1}, \dots, \mu_r$ be meridians for $C_{m+1}, \dots, C_r$, respectively. Let F be a spanning surface for

$$K \cup \mu_{m+1} \cup \cdots \cup \mu_r \cup C_{r+1} \cup \cdots \cup C_n.$$

Since this link has even linking number with each component of $C_1 \cup \cdots \cup C_r$, we may choose this spanning surface to be disjoint from these components. We can view this surface as properly embedded in the exterior of $K \cup C_1 \cup \cdots \cup C_n$. It is disjoint from $\partial N(C_1) \cup \cdots \cup \partial N(C_m)$. We have $F \cap \partial N(C_i) = \mu_i$ for $i \in \{m+1, \dots, r\}$. For $i \in \{r+1, \dots, n\}$, the curve $F \cap \partial N(C_i)$ has slope equal to a longitude plus an odd number of meridians. By choosing the surface appropriately, we can ensure that this odd number is 1.

Now perform surgery along $C_1, \dots, C_n$. The surface becomes a surface in the exterior of the new link. On $\partial N(C_i)$ for $i \in \{m+1, \dots, r\}$, it now has slope equal to a meridian plus q_i longitudes. On $\partial N(C_i)$ for $i \in \{r+1, \dots, n\}$, it is a meridian plus $q_i + 1$ longitudes. Since we are assuming that $|q_i|$ is sufficiently large, we can suppose that $q_i \neq 0, -1$ and hence that this slope is not meridional. Within each solid torus $N(C_{m+1}), \dots, N(C_n)$, we can now insert a surface, as shown in Figure 9. Let F' denote the resulting spanning surface of $K(q_1, \dots, q_n)$.

Also shown in Figure 9 is a collection of generators for $H_1(F' \cap N(C_i))$ for $i \ge m+1$. Note that $H_1(F' \cap N(C_i))$ for $i \ge m+1$ form direct summands of $H_1(F')$. So we can extend this set of generators to a basis of $H_1(F')$, by adding further elements of $H_1(F)$. The associated Goeritz form G_F is diagonal when restricted to the rows and columns corresponding to $H_1(F' \cap N(C_1 \cup \cdots \cup C_n))$. Each C_i gives rise to $\frac{1}{2}|q_i|$ diagonal entries when $m+1 \le i \le r$ and $\frac{1}{2}|q_i + 1|$ entries when $r+1 \le i \le n$. These entries are $+1$ when q_i is positive and -1 when q_i is negative. Hence, the signature of this matrix differs from

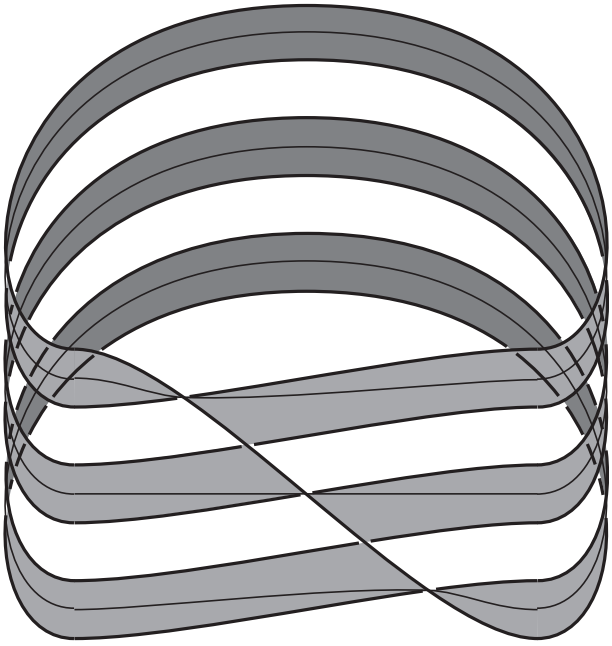


Figure 9: The part of the spanning surface in $N(C_i)$ for $i \geq m+1$. Here $q_i = 5$ or 6.

$\sum_{i=m+1}^{n} \frac{1}{2} q_i$ by at most $\frac{1}{2}(n-r)$. So, applying Lemma 4.5,

$$\left| \sigma(G_{F'}) - \sum_{i=m+1}^{n} \tfrac{1}{2} q_i \right|$$

is bounded.

Theorem 4.1, due to Gordon and Litherland, states that

$$\sigma(K(q_1, \dots, q_n)) = \sigma(G_{F'}) + \tfrac{1}{2} e(F').$$

Here

$$e(F') = -\mathrm{lk}(K(q_1, \dots, q_n), \partial F') = -\mathrm{lk}(K, \partial F') - \sum_{i=1}^{n} \ell_i^2 q_i.$$

The second inequality of the theorem follows immediately. □

6 Proof of Theorem 1.3

In this section, we prove Theorem 1.3 from the introduction:

Theorem 1.3 *Let ε_3 be the Margulis constant and let $\varepsilon \in (0, \varepsilon_3)$. Then there is a constant c_4 (depending on ε) such that, for any hyperbolic knot K, the quantities $\sigma(K)$ and*

$$\tfrac{1}{2} \operatorname{slope}(K) - \sum_{\gamma \in \mathrm{OddGeo}(\varepsilon/2)} \kappa(\mathrm{tw}_p(\gamma), \mathrm{tw}_q(\gamma))$$

differ by at most $c_4 \operatorname{vol}(K)$.

Note that, if we set $\varepsilon = \frac{1}{2}\varepsilon_3$, then c_4 becomes a universal constant. However, given the present uncertainty about the precise value of ε_3, we do not specify ε definitively.

Definition 6.1 Let γ be an embedded closed geodesic in the hyperbolic 3–manifold M, and let $N(\gamma)$ be a regular neighbourhood of γ consisting of points at most a certain distance r from γ. Let $\widetilde{\gamma}$ be a component of the inverse image of γ in $\mathbb{H}^3$, which we can take to be $\{(0,0,z) : z > 0\}$ in the upper half-space model. Let $N(\widetilde{\gamma})$ be the component of the inverse image of $N(\gamma)$ containing $\widetilde{\gamma}$. We let λ be the slope on $\partial N(\gamma)$ that has winding number one around $N(\gamma)$ and that lifts to a path in $N(\widetilde{\gamma})$ starting on the half-plane $\{(x,y,z) : y = 0,\ x \geq 0\}$ and with interior that is disjoint from the half-plane $\{(x,y,z) : y = 0,\ x \leq 0\}$. In the event that this path ends precisely on the half-plane $\{(x,y,z) : y = 0,\ x \leq 0\}$, λ is chosen so that it avoids $\{(x,y,z) : y \leq 0,\ x = 0\}$. Then λ is called the *canonical longitude* of γ. Note that it does not necessarily have zero linking number with γ.

There is the following alternative interpretation of the canonical longitude in terms of the complex length of γ. We give $T = \partial N(\gamma)$ its inherited Riemannian metric. This is homogeneous, since any two points of T differ by an isometry of T. The metric on T therefore has constant curvature, which must be zero by the Gauss–Bonnet theorem. It is therefore Euclidean. We can represent it as the quotient of the Euclidean plane $\mathbb{E}^2$ by a lattice $\mathcal{L}$. Each slope on T corresponds to a lattice point. We can assume that the lattice point corresponding to the meridian is a purely imaginary number μ. As the circumference of a radius r circle in the hyperbolic plane is $2\pi \sinh(r)$, we have

$$\mu = 2\pi \sinh(r) i,$$

where r is the radius of the tube around γ. Let ν be a geodesic in T that is perpendicular to a meridian and that starts and ends on the meridian (but not necessarily at the same point). Then

$$\ell(\nu) = \cosh(r)\,\mathrm{Re}(\mathrm{cl}(\gamma)),$$

where $\mathrm{cl}(\gamma)$ is the complex length of the geodesic γ and $\mathrm{Re}(\mathrm{cl}(\gamma)) = \ell(\gamma)$; see [Futer et al. 2019, equation (2.2)]. Then the canonical longitude of T is

$$\lambda = \cosh(r)\,\mathrm{Re}(\mathrm{cl}(\gamma)) + \sinh(r)\,\mathrm{Im}(\mathrm{cl}(\gamma))i.$$

The significance of the twisting parameter arises from the following lemma:

Lemma 6.2 *Let M be a hyperbolic 3–manifold and $\varepsilon \in (0, \varepsilon_3)$. Let γ be a geodesic in M with $\ell(\gamma) < \frac{1}{2}\varepsilon$. Let T be the toral boundary component of $M_{(0,3\varepsilon/4]}$ that encloses γ, let $\mu \subset T$ be a meridian of γ, and let λ be the canonical longitude. If $(p,q) = (\mathrm{tw}_p(\gamma), \mathrm{tw}_q(\gamma))$, then*

$$\ell(p\lambda + q\mu) \leq c_5\,\mathrm{Area}(T)$$

for some constant c_5 depending only on ε.

Proof By the Margulis lemma, the component V of $M_{(0,3\varepsilon/4]}$ containing T is a solid torus, with γ as a core curve. We claim that the tube radius r of V satisfies $r > \frac{1}{8}\varepsilon$. Indeed, note that γ has length $\ell(\gamma) < \frac{1}{2}\varepsilon$, whereas, at each point $y \in T$, the open ball $B\big(y, \frac{3}{8}\varepsilon\big)$ is embedded. If $r < \frac{3}{8}\varepsilon$ and $x \in \gamma$ satisfies $d(x, y) = r$, then $B\big(x, \frac{3}{8}\varepsilon - r\big) \subset B\big(y, \frac{3}{8}\varepsilon\big)$ is an embedded ball about x. So

$$\tfrac{3}{8}\varepsilon - r < \tfrac{1}{2}\ell(\gamma) < \tfrac{1}{4}\varepsilon,$$

and hence $r > \frac{1}{8}\varepsilon$, as claimed.

Suppose that γ_0 is a shortest geodesic on T, and let $L := \ell(\gamma_0)$. We claim that

$$L \in [k_0, k_0']$$

for constants $k_0, k_0' \in \mathbb{R}_+$ depending only on ε. Since $T \subset \partial M_{(0,3\varepsilon/4]}$, every point $p \in T$ has two lifts to $\mathbb{H}^3$ that are exactly $\frac{3}{4}\varepsilon$ apart, and no two lifts of p are less than $\frac{3}{4}\varepsilon$ apart. The meridian of T has length

$$\ell(\mu) = 2\pi \sinh(r) > 2\pi \sinh\big(\tfrac{1}{8}\varepsilon\big).$$

If s is a slope different from the meridian, then $[s] = m[\gamma] \in \pi_1(M)$ for $m \neq 0$. As $[\gamma]$ has infinite order in $\pi_1(M)$, the lift $\tilde{s}$ of s to $\mathbb{H}^3$ satisfies $\tilde{s}(0) \neq \tilde{s}(1)$. Then

$$\ell(s) \geq d_{\mathbb{H}^3}(\tilde{s}(0), \tilde{s}(1)) \geq \tfrac{3}{4}\varepsilon,$$

so we can set $k_0 := \min\big(\frac{3}{4}\varepsilon, 2\pi \sinh\big(\frac{1}{8}\varepsilon\big)\big)$.

We now give an upper bound on L. Let s be a slope on T whose lift $\tilde{s}$ to $\mathbb{H}^3$ satisfies $d_{\mathbb{H}^3}(\tilde{s}(0), \tilde{s}(1)) = \frac{3}{4}\varepsilon$. This is again possible since $T \subset \partial M_{(0,3\varepsilon/4]}$. If $r \leq 2\varepsilon$, then $L \leq |\mu| \leq 2\pi \sinh(2\varepsilon)$. Now suppose that $r > 2\varepsilon$. Let $N(\gamma) \subset V$ be a regular neighbourhood of γ of radius $r - \varepsilon$. Let $\widetilde{\beta}$ be a geodesic in $\mathbb{H}^3$ connecting $\tilde{s}(0)$ and $\tilde{s}(1)$, and let β be its projection to M. Then β is a geodesic homotopic to s of length $\frac{3}{4}\varepsilon$, which hence lies in $V \setminus N(\gamma)$. The nearest point projection $\varphi\colon V \setminus N(\gamma) \to T$ satisfies $\ell(\varphi(\beta)) \leq l_0 \ell(\beta) = l_0\big(\frac{3}{4}\varepsilon\big)$ for a constant l_0 depending only on ε. Hence,

$$L \leq k_0' := \max\big(2\pi \sinh(2\varepsilon), \tfrac{3}{4} l_0 \varepsilon\big),$$

as claimed.

A consequence of $L \geq k_0$ is that $\mathrm{Area}(T) \geq a_0$ for a constant a_0 depending only on ε. Indeed, a disc D of radius $\frac{1}{2}L$ on T about an arbitrary point of T is embedded, so

$$\mathrm{Area}(T) \geq \mathrm{Area}(D) = \big(\tfrac{1}{2}L\big)^2 \pi \geq \big(\tfrac{1}{2}k_0\big)^2 \pi =: a_0.$$

We claim the length of the shortest curve in any nontrivial class in $H_1(T; \mathbb{Z}_2)$ is at most $k_1 \,\mathrm{Area}(T)$ for a constant k_1 depending on ε. Indeed, let $\gamma_0^{\perp}\colon I \to T$ be a geodesic arc starting and ending on the shortest geodesic γ_0 and orthogonal to it. Then $\ell(\gamma_0^{\perp}) = \mathrm{Area}(T)/L$. The points $\gamma_0^{\perp}(0)$ and $\gamma_0^{\perp}(1)$ divide γ_0 into two arcs, one of which has length at most $\frac{1}{2}L$. Let γ_1 be a geodesic representative of the closed curve that runs along $\gamma_0^{\perp}$ and then along the shorter of the two arcs in γ_0. We obtain that

$$\ell(\gamma_1) \leq \tfrac{1}{2}L + \frac{\mathrm{Area}(T)}{L}.$$

The curves γ_0 and γ_1 give a basis for $H_1(T;\mathbb{Z}_2)$. Hence, the shortest representative of every nontrivial class in $H_1(T;\mathbb{Z}_2)$ is at most $L + \left(\frac{1}{2}L + \text{Area}(T)/L\right)$. As $L \in [k_0, k_0']$ and $\text{Area}(T) \geq a_0$, we have

$$L + \left(\tfrac{1}{2}L + \frac{\text{Area}(T)}{L}\right) \leq k_1 \,\text{Area}(T)$$

for $k_1 := \frac{3}{2}k_0'/a_0 + 1/k_0$. Indeed,

$$\tfrac{3}{2}L \leq \tfrac{3}{2}k_0' = \left(k_1 - \frac{1}{k_0}\right)a_0 \leq \left(k_1 - \frac{1}{L}\right)\text{Area}(T).$$

So there is some slope (a, b) on T with a even and b odd such that

$$\ell(a\lambda + b\mu) \leq k_1 \,\text{Area}(T)$$

for some constant k_1 depending on ε.

Let T' be the torus obtained from T by scaling by $\tanh(r)$ in the v direction. As $r > \frac{1}{8}\varepsilon$, we have $\tanh(r) \in \left(\tanh\left(\frac{1}{8}\varepsilon\right), 1\right)$. Since $\tanh(r) < 1$, the shortest slope (p, q) on T' with p even and q odd has length at most $k_1 \,\text{Area}(T)$. The lattice that specifies T' is generated by

$$\lambda' := \tanh(r)\cosh(r)\,\text{Re}(\text{cl}(\gamma)) + \sinh(r)\,\text{Im}(\text{cl}(\gamma))i = \sinh(r)\,\text{cl}(\gamma)$$

and $\mu = 2\pi\sinh(r)i$. So

$$\ell(p\lambda' + q\mu) = |\text{cl}(\gamma)p + 2\pi i q||\sinh(r)|.$$

Hence, by Definition 1.2, the slope $p\lambda' + q\mu$ on T' is the shortest among slopes for which p is even and q is odd. Therefore, its length on T' is at most $k_1 \,\text{Area}(T)$. So

$$\ell(p\lambda + q\mu) \leq \frac{k_1}{\tanh(r)}\,\text{Area}(T) < \frac{k_1}{\tanh\left(\frac{1}{8}\varepsilon\right)}\,\text{Area}(T).$$

So we can set $c_5 := k_1/\tanh\left(\frac{1}{8}\varepsilon\right)$, which concludes the proof of the lemma. □

Proof of Theorem 1.3 We claim that we can build a triangulation $\mathcal{T}$ of $M_{[3\varepsilon/4,\infty)}$ with the following properties:

(1) The number of tetrahedra of $\mathcal{T}$ is at most $c\,\text{vol}(K)$, where c depends on ε.

(2) If n is a closest even integer to slope(K), then some Euclidean geodesic with slope $\lambda - n\mu$ on $\partial N(K)$ is a normal curve in $\partial M_{[3\varepsilon/4,\infty)}$ that intersects each edge of $\mathcal{T}$ at most once.

(3) On the component T of $\partial M_{[3\varepsilon/4,\infty)}$ corresponding to $\partial N(K)$, the edges of $\mathcal{T}$ are Euclidean geodesics with length at most $\frac{1}{15}\varepsilon$.

We follow the construction in the proof of Proposition 3.1, but with different constants. We pick a maximal collection of points in $\partial M_{[3\varepsilon/4,\infty)}$ that are at least $\frac{1}{30}\varepsilon$ apart. We then add points to this collection in the interior of $M_{[3\varepsilon/4,\infty)}$ that have distance at least $\frac{1}{15}\varepsilon$ from each other and from the earlier points. We stop when it is not possible to add any further points, and denote the resulting collection by P. We then form the associated Voronoi diagram, subdivide the 2–cells of this cell structure into triangles without adding

any new vertices, and then triangulate each 3–cell by coning from the relevant point of P. Let $\mathcal{T}$ be the resulting triangulation of $M_{[3\varepsilon/4,\infty)}$.

Exactly the same argument as in the proof of Proposition 3.1 gives that the number of tetrahedra of $\mathcal{T}$ is at most $c\,\mathrm{vol}(K)$, where c depends on ε. The length of each edge in $\partial M_{[3\varepsilon/4,\infty)}$ is now at most $\frac{1}{15}\varepsilon$, because we took points that were at least $\frac{1}{30}\varepsilon$ apart, rather than at least $\frac{1}{8}\varepsilon$ apart. Thus, all that needs to be proved are that the edges of $\mathcal{T}$ in T are Euclidean geodesics and that there is a Euclidean geodesic with slope $\lambda - n\mu$ on $\partial N(K)$ which is a normal curve in $\partial M_{[3\varepsilon/4,\infty)}$ that intersects each edge of $\mathcal{T}$ at most once.

We start by showing that the edges of $\mathcal{T}$ in T are Euclidean geodesics. Following the proof of Proposition 3.1, we need to show that, for each point x on T, its closest points in P all lie in T and have distance at most $\frac{1}{30}\varepsilon$ from x. We also need to show that the shortest geodesic joining x to any of these points remains within the cusp. The first of these statements holds by our choice of P.

Note that T lies within $M_{(0,\varepsilon]}$. By definition of the Margulis constant, $M_{(0,\varepsilon]}$ consists of a cusp and some regular neighbourhoods of geodesics with length at most ε. The Euclidean metrics on T and the cusp component of $\partial M_{(0,\varepsilon]}$ differ by a Euclidean scale factor of $\frac{4}{3}$, and hence are hyperbolic distance $\ln\left(\frac{4}{3}\right) > 0.287$ from each other. On the other hand, the 3–dimensional Margulis constant satisfies $\varepsilon_3 < 0.775$. (See the discussion in [Futer et al. 2022, Section 1.1].) Hence, $\frac{1}{30}\varepsilon < \ln\left(\frac{4}{3}\right)$. We deduce that, for each point x in T, any shortest geodesic to a closest point in P must lie in the cusp. This implies that the restriction to T of the Voronoi diagram for P in M is equal to the Voronoi diagram for $P \cap T$ in T with its Euclidean metric. In particular, the edges of $\mathcal{T}$ in T are Euclidean geodesics, as claimed in (3).

Let $N_{\max}$ be a maximal cusp neighbourhood around K. Then $N_{\max}$ contains T. This torus T is a scaled copy of $\partial N_{\max}$. It is scaled so that, for each point on T, two lifts of this point in $\mathbb{H}^3$ are exactly $\frac{3}{4}\varepsilon$ apart and no two lifts of this point are any closer than this. Say that d is the hyperbolic distance between T and $\partial N_{\max}$. Then the scale factor taking $\partial N_{\max}$ to T is e^{-d}. Now the meridian slope on $\partial N_{\max}$ has length at most 6. Hence, the meridian slope on T has length at most $6e^{-d}$. So any point on T has two lifts to $\mathbb{H}^3$ that are less than $6e^{-d}$ apart, and therefore $\frac{3}{4}\varepsilon \leq 6e^{-d}$. As in the proof of Proposition 3.1, let h be the length in $\partial N_{\max}$ of a Euclidean geodesic that starts and ends on a geodesic with slope $\lambda - n\mu$ and that is orthogonal to this geodesic. It was shown there that $h \geq 0.55$. Hence, the length of the corresponding geodesic on T is at least $0.55e^{-d} \geq \frac{0.55}{6}\left(\frac{3}{4}\varepsilon\right)$. On the other hand, the length of each edge of $\mathcal{T}$ on T is at most $\frac{1}{15}\varepsilon$, and $\frac{1}{15}\varepsilon < \frac{0.55}{6}\left(\frac{3}{4}\varepsilon\right)$. Hence, each such edge can intersect any geodesic with slope $\lambda - n\mu$ at most once. This establishes the claimed properties of $\mathcal{T}$.

Let $T_1, \dots, T_m$ be the components of $\partial M_{[3\varepsilon/4,\infty)}$, where T_i encircles a geodesic $\gamma_i \in \mathrm{OddGeo}\left(\frac{1}{2}\varepsilon\right)$. Let $\mathrm{tw}(\gamma_i) = p\lambda_i + q\mu_i$, where λ_i is the canonical longitude on T_i and μ_i is the meridian, and let C_i be a curve on T_i with this slope. Then

$$\ell(C_i) \leq c_5\,\mathrm{Area}(T_i)$$

by Lemma 6.2. Let

$$C := \bigcup_{i=1}^{m} C_i.$$

Realise each C_i as a Euclidean geodesic in T_i missing the vertices of T_i, and hence as a normal curve in T_i. Since $\ell(C_i) \leq c_5 \operatorname{Area}(T_i)$ and by property (3) of the triangulation $\mathcal{T}$, the normal representative of C_i intersects each edge of $\mathcal{T}$ at most $c_5' \operatorname{Area}(T_i)$ times for a constant c_5' depending only on ε.

We claim that there is a connected normal curve C_i' in T_i for $i \in \{1, \dots, m\}$ with the following properties:

(1) C_i' and C_i are equal in $H_1(T_i; \mathbb{Z}_2)$.

(2) C_i' intersects each edge of $\mathcal{T}$ at most once.

This is constructed as follows. For each edge of $\mathcal{T}$ that intersects C_i an odd number of times, replace this intersection by a single point of intersection. These will be the points of intersection between C_i' and the 1–skeleton of $\mathcal{T}$. Since $|C_i \cap \partial t|$ is even for each triangle t of $\mathcal{T}$, we have $|C_i' \cap \partial t| \in \{0, 2\}$. If $|C_i' \cap \partial t| = 2$, join the two points of $C_i' \cap \partial t$ by a normal arc of C_i'. The result is a collection of simple closed curves in T_i that are mod 2 homologous to C_i. If any of these curves are inessential in T_i, remove them. The resulting curves are essential in T_i. Since they are nontrivial in mod 2 homology, they consist of an odd number of parallel copies of a curve. If this odd number is greater than one, remove all but one of these curves. The result is C_i', and we write

$$C' := \bigcup_{i=1}^{m} C_i'.$$

Let C'' be the union of C' and a normal curve C_K of slope $(1, -n)$ on $\partial N(K)$, where n is a closest even integer to slope(K). We claim that C'' bounds an unoriented surface in $M_{[3\varepsilon/4,\infty)}$. As n is even, there is a compact surface properly embedded in the exterior of K with boundary slope $(1, -n)$. It intersects each geodesic with length at most $\frac{1}{2}\varepsilon$ in a collection of meridians. For a geodesic with odd linking number with K, the number of these meridians is odd. For the others, it is even. As C_i' is homologous to the meridian of T_i over $\mathbb{Z}_2$, we may modify the surface so that its boundary is precisely C''. This proves the claim.

As C'' intersects each edge of $\mathcal{T}$ at most once, we can find a surface F'' in $M_{[\varepsilon/2,\infty)}$ that it bounds such that

$$-\chi(F'') \leq c_6 \operatorname{vol}(K)$$

for some constant c_6, just like in the proof of Theorem 1.7. Now C_i and C_i' are equal in $H_1(T_i; \mathbb{Z}_2)$. Hence, we may insert a compact connected surface F_i into a regular neighbourhood $N(T_i)$ of T_i with $\partial F_i = C_i \cup C_i'$. Since C_i and C_i' intersect each edge of $\mathcal{T}$ at most $c_5' \operatorname{Area}(T_i)$ times, this surface may be chosen so that

$$-\chi(F_i) \leq c_5'' \operatorname{Area}(T_i)$$

for a constant c_5'' depending only on ε. Hence, the surface

$$F := F'' \cup \bigcup_{i=1}^{m} F_i \subset M_{[3\varepsilon/4,\infty)}$$

satisfies $\partial F = C_K \cup C$, and

$$-\chi(F) \leq c_6 \operatorname{vol}(K) + \sum_{i=1}^{m} c_5'' \operatorname{Area}(T_i) \leq c_7 \operatorname{vol}(K) \tag{6.3}$$

for a constant c_7 that depends only on ε. Here the last inequality follows from the observation that $\operatorname{Area}(T_i) \leq c_8 \operatorname{vol}(N(T_i))$ for some constant c_8, where

$$N(T_i) := \{x \in V_i : d(x, T_i) \leq \tfrac{1}{2} r_i\},$$

and V_i is the solid toral component of $M_{(0,3\varepsilon/4]}$ of tube radius r_i that encloses the geodesic $\gamma_i \in \operatorname{OddGeo}(\frac{1}{2}\varepsilon)$.

In each V_i, we construct the surface provided by Lemma 4.6 with boundary $C_i = C \cap V_i$. We attach these surfaces to F to form a surface F_+. We now specify a basis for $H_1(F_+)$. We start by picking a basis for $H_1(V_1 \cap F_+)$. We arrange that all but one of these basis elements have zero winding number around V_1. We then continue to V_2, and so on. We then extend this to a basis for $H_1(F_+)$ by adding some oriented curves in F. We order this basis as follows into $n+1$ blocks. In the first block, we place all the basis elements of $H_1(V_1 \cap F_+)$ that have zero winding number around V_1. In the second block, we do the same for V_2, and so on. In the final block, we place all the remaining basis elements. We saw in the proof of Lemma 6.2 that there is a constant a_0 depending only on ε such that $\operatorname{Area}(T_i) \geq a_0$. As $\sum_{i=1}^{m} \operatorname{Area}(T_i) \leq c_8 \operatorname{vol}(K)$, we have

$$\left|\operatorname{OddGeo}\left(\tfrac{1}{2}\varepsilon\right)\right| \leq \frac{c_8 \operatorname{vol}(K)}{a_0}.$$

This, together with (6.3), implies that the number of elements in this final block is bounded above by a linear function of $\operatorname{vol}(K)$.

Let G be the submatrix of the Goeritz form G_{F_+} consisting of the first n blocks. By Lemma 4.5, $\sigma(G)$ and $\sigma(G_F)$ differ by at most the number of elements in the final block. Note that G is block diagonal. For the block corresponding to V_i, the signature differs from $\sigma(G_{F_+ \cap V_i})$ by at most one by Lemma 4.5. On the other hand,

$$|\sigma(G_{F_+ \cap V_i}) + \kappa(\operatorname{tw}_p(\gamma_i), \operatorname{tw}_q(\gamma_i))| \leq 2$$

by Lemma 4.6. Hence,

$$\left|\sigma(G_{F_+}) + \sum_{\gamma \in \operatorname{OddGeo}(\varepsilon/2)} \kappa(\operatorname{tw}_p(\gamma_i), \operatorname{tw}_q(\gamma_i))\right| \leq c_9 \operatorname{vol}(K)$$

for some constant c_9. By Gordon and Litherland's theorem (Theorem 4.1),

$$\sigma(K) = \sigma(G_{F_+}) + \tfrac{1}{2} e(F_+) = \sigma(G_{F_+}) + \tfrac{1}{2} n.$$

The result follows as n is a closest even integer to $\operatorname{slope}(K)$. □

7 Experimental data and some conjectures about random knots

We set out to find links between hyperbolic and 4–dimensional knot invariants. Initial scatter plots compared some 4–dimensional invariants (the signature and Heegaard Floer invariants τ, ν and ε), the crossing number, and several hyperbolic invariants (volume, meridional and longitudinal translations, and the Chern–Simons invariant). As σ is strongly correlated to τ, ν and ε, we decided to only focus on σ, which is more classical and easier to compute.

The strongest and most surprising correlation was between the signature and the real part of the meridional translation; see Figure 2. There were some more predictable relationships among the hyperbolic invariants.

Figure 10 shows the distribution of

$$c_1(K) := \frac{|2\sigma(K) - \text{slope}(K)|\,\text{inj}(K)^3}{\text{vol}(K)},$$

which indicates that the constant c_1 appearing in Theorem 1.1 is typically quite small. The largest value of this quantity we managed to obtain is less than 0.234, and we conjecture it is always at most 0.3. The left of Figure 11 shows the maximum and the right the mean of $c_1(K)$ by crossing number for the Regina census of knots of at most 16 crossings. See Figure 12 for a scatter plot of injectivity radius versus volume for random hyperbolic knots of 10–80 crossings. This suggests that the injectivity radius is typically not too small as the volume increases.

We will say that a property P holds asymptotically almost surely, or a.a.s. in short, if the probability that P holds for knots of n crossings tends to 1 as $n \to \infty$.

It is known that there is a constant A such that $\text{vol}(K) \leq Ac(K)$, where $c(K)$ is the crossing number of K. From scatter plots, one might conjecture that there is a constant a such that $ac(K) \leq \text{vol}(K)$ a.a.s. Such

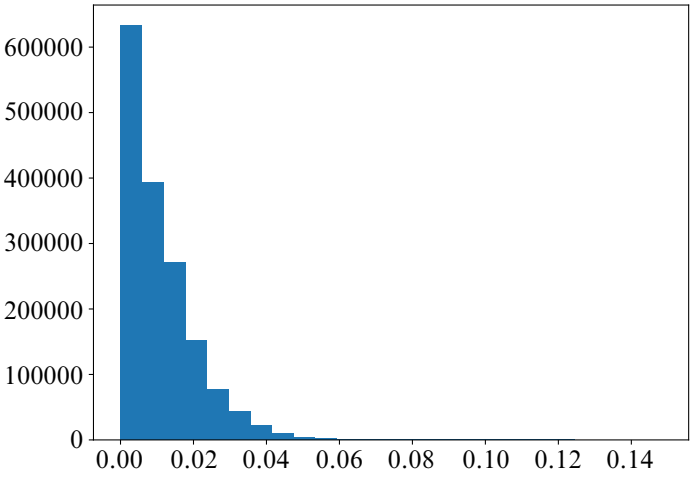


Figure 10: The distribution of $c_1(K) := |2\sigma(K) - \text{slope}(K)|\,\text{inj}(K)^3/\text{vol}(K)$ for knots up to 16 crossings in the Regina census.

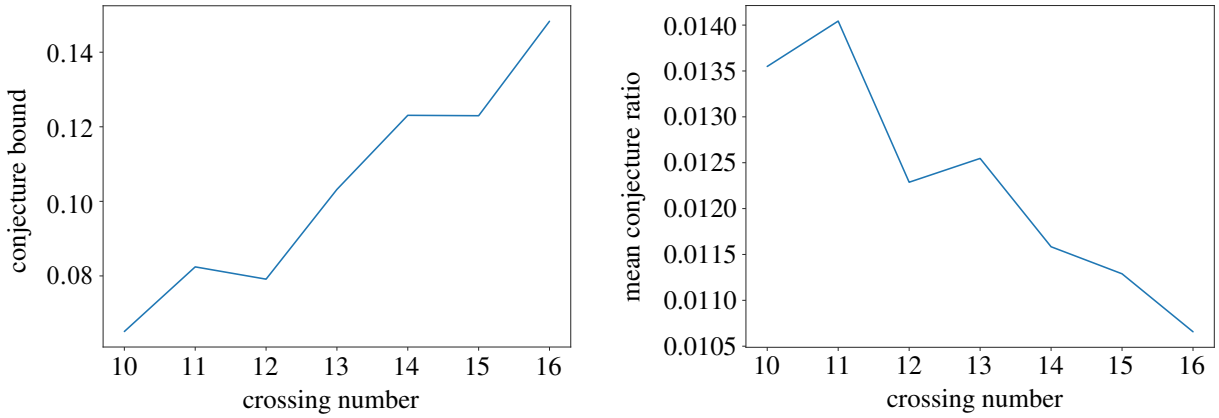


Figure 11: The maximum (left) and the mean (right) of $c_1(K)$ as functions of the crossing number for knots up to 16 crossings in the Regina census.

an inequality cannot hold for all hyperbolic knots K. For example, consider twist knots. More generally, the highly twisted knots considered in Section 5 have bounded volume but unbounded crossing number.

We now consider the behaviour of the signature $\sigma(K)$ for random knots K. By Theorem 4.1, $\sigma(K)$ can be computed from the black surface of a checkerboard colouring of a diagram of K. Hence, it is the signature of a $c(K) \times c(K)$ matrix. If the signs of the eigenvalues of this matrix were independently distributed, then the expected value of $|\sigma(K)|$ would be $C\sqrt{c(K)}$ for some constant C. From computational evidence, it appears the constant is about 2. Based on this heuristic, we introduce the following definition:

Definition 7.1 The *normalised signature* of a hyperbolic knot K is

$$\hat{\sigma}(K) := \frac{\sigma(K)}{\sqrt{\mathrm{vol}(K)}}.$$

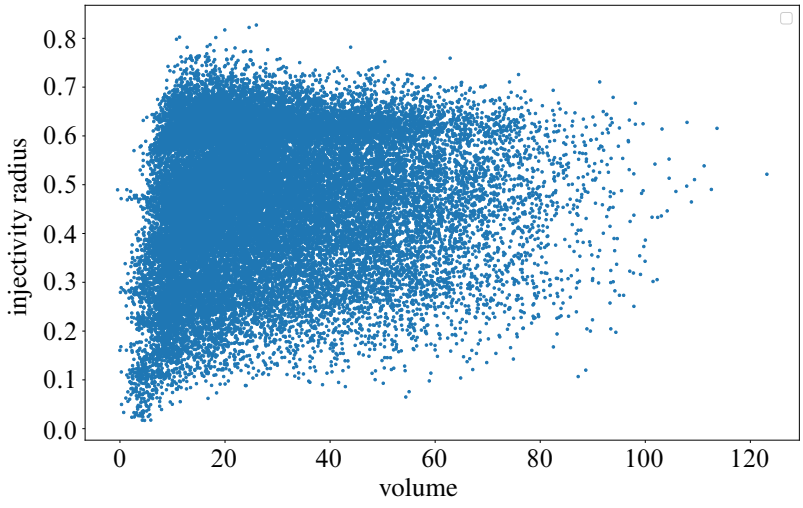


Figure 12: A scatter plot of injectivity radius versus volume for random knots of 10–80 crossings.

We use the volume instead of the crossing number as it is easier to compute using SnapPy and is more regular.

Based on Figure 2, we initially conjectured that, for any hyperbolic knot K in S^3 with $|\widehat{\sigma}(K)| > 1$, the signature $\sigma(K)$ and $\mathrm{Re}(\mu(K))$ have the same sign. However, this turns out not to be true.

Corollary 7.2 *There exists a hyperbolic knot K with $|\widehat{\sigma}(K)| > 1$, but with $\sigma(K)$ and $\mathrm{Re}(\mu(K))$ having opposite signs.*

Proof We start with a hyperbolic link $K \cup C_1 \cup C_2$ in S^3, where C_1 and C_2 bound disjoint embedded discs, and where $\ell_1 = \mathrm{lk}(K, C_1) = 2$ and $\ell_2 = \mathrm{lk}(K, C_2) = 3$. We then build the highly twisted knots $K(q_1, q_2)$ as in Theorem 1.8. Set $q_1 = 17q$ and $q_2 = -8q$, where q is a large positive integer. Then

$$\mathrm{slope}(K(q_1, q_2)) \sim -4 \cdot 17q + 9 \cdot 8q = 4q, \quad \text{whereas} \quad \sigma(K(q_1, q_2)) \sim -2 \cdot 17q + 4 \cdot 8q = -2q.$$

Hence, for q sufficiently large, $\sigma(K(q_1, q_2))$ and $\mathrm{slope}(K(q_1, q_2))$ have opposite signs, and hence $\sigma(K(q_1, q_2))$ and $\mathrm{Re}\big(\mu(K(q_1, q_2))\big)$ also have opposite signs by Lemma 2.4. Note that $\widehat{\sigma}(K(q_1, q_2)) > 1$ if q is sufficiently large, because $|\sigma(K(q_1, q_2))|$ tends to infinity whereas $\mathrm{vol}(K(q_1, q_2))$ is bounded. □

However, we do conjecture the following:

Conjecture 7.3 *If K is a hyperbolic knot in S^3 with $|\widehat{\sigma}(K)| > 1$, then $\sigma(K)$ and $\mathrm{Re}(\mu(K))$ have the same sign asymptotically almost surely.*

We also state the following conjecture, which proposes a more precise relationship between slope and signature:

Conjecture 7.4 *There are constants b and c such that, for any hyperbolic knot K in S^3, we have*

$$|2\sigma(K) - \mathrm{slope}(K)| \le b\sqrt{\mathrm{vol}(K)} + c \tag{7.5}$$

asymptotically almost surely.

By Corollary 5.1, this does not hold for all knots either. In fact, there are families of hyperbolic knots for which $|2\sigma(K) - \mathrm{slope}(K)|$ is not bounded by a linear function of the volume.

The proof of Theorem 1.1 provides some heuristic for Conjecture 7.4. Indeed, if we assume that the signs of the eigenvalues of the Goeritz matrix G_F are independent, then the signature on average is of order $\sqrt{c(K)}$. This justifies the factor $\sqrt{\mathrm{vol}(K)}$ in the upper bound.

If $b < 2$ (and the data supports this; see Figure 13), then Conjecture 7.4 implies Conjecture 7.3 for knots K with sufficiently large volume a.a.s. This is because (7.5) is equivalent to the inequality

$$\left|2\widehat{\sigma}(K) - \frac{\mathrm{slope}(K)}{\sqrt{\mathrm{vol}(K)}}\right| \le b + \frac{c}{\sqrt{\mathrm{vol}(K)}}.$$

If $b < 2$, then $b + c/\sqrt{\mathrm{vol}(K)} < 2$ for all knots with sufficiently large volume. So, if $|\widehat{\sigma}(K)| > 1$, then $\widehat{\sigma}(K)$ and $\mathrm{slope}(K)$ have the same sign.

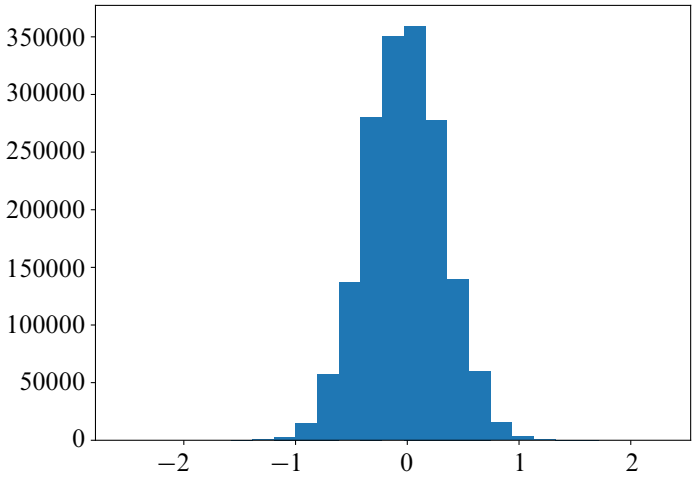


Figure 13: The distribution of the normalised residual $(2\sigma(K) - \text{slope}(K))/\sqrt{\text{vol}(K)}$ for knots up to 16 crossings in the Regina census.

References

[Agol 2000] **I Agol**, *Bounds on exceptional Dehn filling*, Geom. Topol. 4 (2000) 431–449 MR Zbl

[Benard et al. 2021] **L Benard**, **V Florens**, **A Rodau**, *A slope invariant and the A–polynomial of knots*, preprint (2021) arXiv 2103.14151

[Benedetti and Petronio 1992] **R Benedetti**, **C Petronio**, *Lectures on hyperbolic geometry*, Springer (1992) MR Zbl

[Breslin 2009] **W Breslin**, *Thick triangulations of hyperbolic n–manifolds*, Pacific J. Math. 241 (2009) 215–225 MR Zbl

[Burton et al. 1999–2021] **B A Burton**, **R Budney**, **W Pettersson**, et al., *Regina*: *software for low-dimensional topology* (1999–2021) Available at `https://regina-normal.github.io/`

[Cao and Meyerhoff 2001] **C Cao**, **G R Meyerhoff**, *The orientable cusped hyperbolic* 3*–manifolds of minimum volume*, Invent. Math. 146 (2001) 451–478 MR Zbl

[Cooper and Lackenby 1998] **D Cooper**, **M Lackenby**, *Dehn surgery and negatively curved* 3*–manifolds*, J. Differential Geom. 50 (1998) 591–624 MR Zbl

[Culler et al. 2021] **M Culler**, **N M Dunfield**, **M Goerner**, **J R Weeks**, *SnapPy*, *a computer program for studying the geometry and topology of* 3*–manifolds* (2021) Available at `http://snappy.computop.org`

[Degtyarev et al. 2022] **A Degtyarev**, **V Florens**, **A G Lecuona**, *Slopes and signatures of links*, Fund. Math. 258 (2022) 65–114 MR Zbl

[Futer et al. 2019] **D Futer**, **J S Purcell**, **S Schleimer**, *Effective distance between nested Margulis tubes*, Trans. Amer. Math. Soc. 372 (2019) 4211–4237 MR Zbl

[Futer et al. 2022] **D Futer**, **J S Purcell**, **S Schleimer**, *Effective bilipschitz bounds on drilling and filling*, Geom. Topol. 26 (2022) 1077–1188 MR Zbl

[Gordon and Litherland 1978] **C M Gordon**, **R A Litherland**, *On the signature of a link*, Invent. Math. 47 (1978) 53–69 MR Zbl

[Gordon et al. 1981] **C M Gordon**, **R A Litherland**, **K Murasugi**, *Signatures of covering links*, Canad. J. Math. 33 (1981) 381–394 MR Zbl

[Hass et al. 1999] **J Hass**, **J H Rubinstein**, **S Wang**, *Boundary slopes of immersed surfaces in* 3*–manifolds*, J. Differential Geom. 52 (1999) 303–325 MR Zbl

[Kobayashi and Rieck 2011] **T Kobayashi**, **Y Rieck**, *A linear bound on the tetrahedral number of manifolds of bounded volume (after Jørgensen and Thurston)*, from "Topology and geometry in dimension three" (W Li, L Bartolini, J Johnson, F Luo, R Myers, J H Rubinstein, editors), Contemp. Math. 560, Amer. Math. Soc., Providence, RI (2011) 27–42 MR Zbl

[Lackenby 2000] **M Lackenby**, *Word hyperbolic Dehn surgery*, Invent. Math. 140 (2000) 243–282 MR Zbl

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

[Lackenby and Purcell 2016] **M Lackenby**, **J S Purcell**, *Cusp volumes of alternating knots*, Geom. Topol. 20 (2016) 2053–2078 MR Zbl

[Livingston and Moore 2021] **C Livingston**, **A H Moore**, *KnotInfo*: *table of knot invariants*, electronic reference (2021) Available at `https://knotinfo.math.indiana.edu/`

[Morgan 1984] **J W Morgan**, *On Thurston's uniformization theorem for three-dimensional manifolds*, from "The Smith conjecture" (J W Morgan, H Bass, editors), Pure Appl. Math. 112, Academic, Orlando, FL (1984) 37–125 MR Zbl

[Mostow 1968] **G D Mostow**, *Quasi-conformal mappings in n–space and the rigidity of hyperbolic space forms*, Inst. Hautes Études Sci. Publ. Math. 34 (1968) 53–104 MR Zbl

[Thurston 1979] **W P Thurston**, *The geometry and topology of three-manifolds*, lecture notes, Princeton University (1979) Available at `https://url.msp.org/gt3m`

[Voronoi 1908a] **G Voronoi**, *Nouvelles applications des paramètres continus à la théorie des formes quadratiques, deuxième mémoire: Recherches sur les parallélloèdres primitifs*, J. Reine Angew. Math. 134 (1908) 198–287 MR Zbl

[Voronoi 1908b] **G Voronoi**, *Nouvelles applications des paramètres continus à la théorie des formes quadratiques, premier mémoire: sur quelques propriétés des formes quadratiques positives parfaites*, J. Reine Angew. Math. 133 (1908) 97–102 MR Zbl

DeepMind
London, United Kingdom

Mathematical Institute, University of Oxford
Oxford, United Kingdom

Mathematical Institute, University of Oxford
Oxford, United Kingdom

DeepMind
London, United Kingdom

`adavies@google.com, juhasza@maths.ox.ac.uk, lackenby@maths.ox.ac.uk, nenadt@deepmind.com`

Proposed: David Gabai Received: 20 April 2022
Seconded: Cameron Gordon, Stavros Garoufalidis Revised: 26 August 2022

msp

Geometry & Topology **28**:5 (2024) 2345–2398
DOI: 10.2140/gt.2024.28.2345
Published: 24 August 2024

Rigidity and geometricity for surface group actions on the circle

KATHRYN MANN
MAXIME WOLFF

We prove that (topologically) rigid actions of surface groups on the circle by homeomorphisms are necessarily *geometric*, namely, they are semiconjugate to an embedding as a cocompact lattice in a Lie group acting transitively on S^1. This gives the converse to a theorem of the first author; thus characterizing geometric actions as the unique isolated points in the "character space" of surface group actions on S^1.

20H10, 37E10, 37E45, 57S25, 58D29

1 Introduction

Classification results in dynamical systems are often motivated by the study of special examples. Having found a system with interesting (eg stable) behavior, one seeks first to understand its properties and related examples, and then to address the broader problem of classifying all systems with such properties. As a prime example, Anosov observed that hyperbolic linear automorphisms of tori exhibit stability under perturbation, leading to the abstract definition of *Anosov diffeomorphisms*. Smale [32] observed that hyperbolic affine automorphisms of infra-nil manifolds give additional such examples; that this is an exhaustive list of all Anosov diffeomorphisms of closed manifolds up to topological conjugacy is a longstanding open conjecture.

The present work addresses the classification problem for globally rigid actions of surface groups on the circle; equivalently, for rigid, flat topological circle bundles over surfaces. Here, local rigidity, at least in the C^1 setting, already follows from the work of Anosov. A much stronger, global, C^0 rigidity phenomenon was discovered by Matsumoto [28], who proved that all representations $\pi_1\Sigma_g \to \mathrm{Homeo}^+(S^1)$ of equal, extremal Euler class are semiconjugate, in the sense of semiconjugacy for circle actions defined by Ghys [12]. These globally rigid examples are all *geometric* in the sense that they arise from embedding $\pi_1\Sigma_g$ as a cocompact lattice in a Lie subgroup of $\mathrm{Homeo}^+(S^1)$. Matsumoto's result was extended by the first author [23], who showed that, in fact *all* geometric actions of surface groups have this same global rigidity: they are characterized, up to semiconjugacy, by a finite list of rotation numbers which are constant in a neighborhood of each geometric representation. As a consequence, they descend to isolated points in the quotient of the representation space by semiconjugacy. This strong property is the definition of *rigidity* we will use throughout this article; see Section 1.2 for further discussion.

Here we solve the associated classification problem, giving a complete characterization of rigid actions of surface groups on the circle.

Theorem 1.1 *Let Σ_g be a surface of genus $g \geqslant 2$. Then every rigid representation $\pi_1\Sigma_g$ to $\mathrm{Homeo}^+(S^1)$ is geometric: up to semiconjugacy it is obtained by embedding $\pi_1\Sigma_g$ as a lattice in a transitive Lie group in $\mathrm{Homeo}^+(S^1)$.*

The geometric representations referenced in the theorem are easily classified; the Lie groups in question are simply the finite cyclic extensions of $\mathrm{PSL}_2(\mathbb{R})$.

The arc of our proof resembles in spirit the *convergence group theorem* of Tukia [35], Gabai [10] and Casson and Jungries [7]. Both in our case and theirs, one starts with purely dynamical information (in the convergence group case, information on the dynamics of sequences of elements; in ours merely the assumption of rigidity) and from that reconstructs the geometric–topological data of a subgroup of $\mathrm{PSL}_2(\mathbb{R})$ or one of its covers. The key in our case is to show that, under an arbitrary rigid action, elements of $\pi_1\Sigma_g$ which can be represented by nonseparating simple closed curves have the same dynamics as the geometric examples. From there, we again use rigidity to "reconstruct" the topology of the surface, recovering the intersection pattern of these curves on Σ_g.

We note also that, while the statement of Theorem 1.1 resembles Sullivan's "structural stability implies hyperbolicity" for Kleinian groups [33], our methods and conclusion are quite different: for Sullivan, structural stability is a local and C^1 condition, and the groups in consideration are convex-cocompact, acting on their limit set satisfying a hyperbolicity or local hyperbolicity condition.

1.1 Motivation

Our motivation comes from the highly influential work of Milnor, Wood and Goldman. Milnor's contribution to the *Milnor–Wood inequality* is the statement that a principal $\mathrm{PSL}_2(\mathbb{R})$ bundle over a surface admits a flat connection if and only if the *Euler number* of the bundle is bounded in absolute value by the Euler characteristic of the surface. Following this, the natural next question is to what extent the Euler number distinguishes flat bundles. This was answered by Goldman [15], who showed that it is a complete invariant of flat $\mathrm{PSL}_2(\mathbb{R})$ bundles *up to deformation*: the connected components of $\mathrm{Hom}(\pi_1\Sigma_g, \mathrm{PSL}_2(\mathbb{R}))$ are classified by the Euler numbers of the associated bundles.

Here we are interested in these same basic questions in the topological rather than the linear category. Wood [36] showed that Milnor's bound holds in the topological setting as well, demonstrating that topological S^1 bundles over Σ_g which admit a flat connection (or in this case a foliation transverse to the fibers) are precisely those whose Euler numbers are bounded by $\pm(2g-2)$. However, work of the first author [23] showed that Goldman's theorem is no longer true in this setting: there are many connected components of $\mathrm{Hom}(\pi_1\Sigma_g, \mathrm{Homeo}^+(S^1))$ consisting of bundles with the same Euler number.

In fact, the topology of the *space* of flat circle bundles, which can be thought of either as the representation space $\mathrm{Hom}(\pi_1\Sigma_g, \mathrm{Homeo}^+(S^1))$ or the associated *character space* described below, remains very mysterious. For instance, it is an open question whether either space has finitely many or infinitely many connected components. Theorem 1.1 gives the first step towards a global picture, giving a complete classification of the *isolated points* of the character space, and our hope is that the tools we develop should be useful towards the broader program.

1.2 Character spaces and rigidity

As in Goldman's work, the appropriate framing for our work is the study of character spaces. Typically these are defined algebraically, but they generalize naturally to the broad context of groups acting on manifolds.

Let Γ be any discrete group and let G be a topological group such that $G \subset \mathrm{Homeo}(X)$ for some space X. The *representation space* $\mathrm{Hom}(\Gamma, G)$, equipped with the compact-open topology, parametrizes actions of Γ on X with image in G. Typically, G is used to specify the regularity of the action — for instance, taking $G = \mathrm{Diff}(X)$ parametrizes smooth actions, while if G is a Lie group acting transitively on M these are *geometric* actions in the sense of Ehresmann. Since conjugate actions are dynamically equivalent, the appropriate moduli space of actions is the quotient $\mathrm{Hom}(\Gamma, G)/G$ under the natural conjugation action of G. However, this quotient space is typically non-Hausdorff and so in practice difficult to study.

When G is a Lie group and $\mathrm{Hom}(\Gamma, G)$ is an affine variety, algebraic geometers solve this problem by considering the quotient $\mathrm{Hom}(\Gamma, G) /\!\!/ G$ from geometric invariant theory. In the special case where G is a semisimple complex reductive Lie group, this GIT quotient is simply the quotient of $\mathrm{Hom}(\Gamma, G)$ by the equivalence relation $\rho_1 \sim \rho_2$ whenever *the closures* of their conjugacy classes intersect (see Luna [21; 22]); in particular, this relation makes the quotient space Hausdorff. In the well-studied case of $G = \mathrm{SL}(n, \mathbb{C})$, the GIT quotient agrees with the space of *characters* of G–representations, motivating the following terminology.

Definition 1.2 For any discrete group Γ and topological group G, the *character space* $X(\Gamma, G)$ is the largest Hausdorff quotient[1] of $\mathrm{Hom}(\Gamma, G)/G$. Two representations are *weakly conjugate* if they define the same point in $X(\Gamma, G)$.

Loosely speaking, a representation $\rho\colon \Gamma \to G$ is rigid if all deformations of $\rho(\Gamma)$ in G are trivial. This notion can be made precise in the setting of character spaces as follows.

Definition 1.3 A representation $\rho \in \mathrm{Hom}(\Gamma, G)$ is *rigid* if the image of ρ is an isolated point in the character space $X(\Gamma, G)$.

[1] Recall that the largest Hausdorff quotient X_H of a topological space X is a space with the universal property that any continuous map $f : X \to Y$ from X to a Hausdorff topological space factors canonically through the projection $X \to X_H$. One construction of X_H is as the quotient of X by the intersection of all equivalence relations $\sim$ such that $X/\sim$ is Hausdorff.

This is a strong condition on ρ, and less strict forms of rigidity will also be useful. In particular, we say that ρ is *path-rigid* if the path component of ρ in $\mathrm{Hom}(\Gamma, G)$ is contained in a single weak-conjugacy class.

The case of interest in this article is when $G = \mathrm{Homeo}^+(S^1)$, the group of orientation-preserving homeomorphisms of the circle, and $\Gamma = \Gamma_g = \pi_1(\Sigma_g)$ is the fundamental group of an orientable surface of genus $g \geq 2$. As we explain in Section 2.3, in this setting the character space $X(\Gamma, G)$ agrees with the space of *semiconjugacy* classes of actions in the sense of Ghys [12]. In this and related work, semiconjugacy is used to refer to an equivalence relation for group actions on the circle. However, semiconjugacy has a precise and different meaning in topological dynamics. For this reason, we will use the term "weak conjugacy" when referring to the character space $X(\Gamma, G)$, even though this terminology is not yet well established in the literature, and use the term semiconjugacy only when referencing classical results following [12].

1.3 Geometric representations

It is our philosophy that dynamical rigidity often comes from some underlying geometric or algebraic structure. This motivates the following definition.

Definition 1.4 (Mann [24]) Suppose that M is a manifold, and Γ a countable group. A representation $\rho\colon \Gamma \to \mathrm{Homeo}(M)$ is called *geometric* if it is weakly conjugate to a faithful representation with image a cocompact[2] lattice in a transitive, connected Lie group $G \subset \mathrm{Homeo}(M)$.

Indeed, the first known example of a rigid action of a surface group on the circle was a geometric one, due to Matsumoto [28]. Matsumoto's result is that the set of representations with maximal Euler number (equal to $2g-2$ by Milnor–Wood) in $X(\Gamma_g, G)$ consists of a single point — all are weakly conjugate to discrete, faithful representations into $\mathrm{PSL}_2(\mathbb{R}) \subset \mathrm{Homeo}^+(S^1)$. As the Euler number is a continuous function on $\mathrm{Hom}(\Gamma_g, G)$, this implies that representations of maximal Euler number are rigid. The same holds for representations with Euler number $-2g+2$.

While Matsumoto's proof uses maximality of the Euler number in an essential way — a theme that has been taken up in the study of "maximal representations" of surface groups in higher Teichmüller theory, see eg Burger, Iozzi and Wienhard [5] — the idea hints at a separate underlying phenomenon for rigidity, namely *geometricity*.

As hinted above, geometric representations of surface groups in $\mathrm{Homeo}^+(S^1)$ (up to weak conjugacy) all are either discrete, faithful representations into $\mathrm{PSL}_2(\mathbb{R})$, or obtained by lifting such a representation to a finite cyclic extension of $\mathrm{PSL}_2(\mathbb{R})$ (see [24]) and the main result of [23] is their rigidity.

[2] Our choice to require that Γ be cocompact here is motivated by the definition of *model geometries* in the sense of Thurston, where one is interested in compact quotients. It also simplifies the statement of rigidity theorems in low dimensions: noncocompact lattices in $\mathrm{PSL}_2(\mathbb{R})$ and $\mathrm{PSL}_2(\mathbb{C})$ are not rigid, even in the space of representations into $\mathrm{PSL}_2(\mathbb{R})$ or $\mathrm{PSL}_2(\mathbb{C})$.

Theorem 1.5 (Mann [23]) *In the space* $\mathrm{Hom}(\Gamma_g, \mathrm{Homeo}^+(S^1))$, *all geometric representations are rigid.*

Though actually stated there in a slightly weaker form, the proof is carried out on the level of semiconjugacy (or weak-conjugacy, we show in Section 2 these notions coincide) invariants of representations, so actually shows that geometric representations are isolated points in $X(\Gamma_g, \mathrm{Homeo}^+(S^1))$.

1.4 Strategy of proof and outline of the article

The entirety of this work is devoted to the proof of Theorem 1.1, ie the converse of Theorem 1.5. Our main technical result is the following statement, which gives a stronger result for representations of nonzero Euler class.

Theorem 1.6 *Let* $\rho\colon \Gamma_g \to \mathrm{Homeo}^+(S^1)$ *be a path-rigid representation. If* ρ *is not geometric, then its Euler class is zero, and there exists a one-holed, genus* $g-1$ *subsurface* $\Sigma' \subset \Sigma_g$ *such that* $\rho|_{\pi_1\Sigma'}$ *has a finite orbit.*

The surface group representations with Euler class zero are precisely those which can be lifted to actions on the line. It is not entirely surprising that our theorem identifies these as a special case, as more complicated dynamical phenomena sometimes occur for such representations. Notably, Ghys [11] shows that an action of a surface group on S^1 by *real analytic* diffeomorphisms admits a minimal exceptional set only if it has Euler class zero. However, the condition of having a large subsurface with a finite orbit makes it very likely that such a representation could be deformed along a path; giving strong evidence for the fact that all path-rigid representations should in fact be geometric.

The main ingredient in the proof of Theorem 1.6 is the effect of *bending deformations* on the periodic sets of simple closed curves. Bending deformations are classical in (higher) Teichmüller theory (see Section 2.2.2 for a reminder); and we extend their study to representations to $\mathrm{Homeo}^+(S^1)$. While the proof of Theorem 1.6 is quite long, a much simpler argument can be carried out under the additional significant assumption that the relative Euler number on some genus 1 subsurface is equal to 1 (this is the case in particular for representations of Euler class $\geq g$). This much weaker proof is presented in our expository article [25]; the reader may find it helpful to take that work as a starting point or a companion.

We now outline the major steps.

Step 1 (local-to-global) Our proof starts by making a strong additional technical hypothesis on representations that forces them to look "locally" (ie on the level of some pairs of elements) like representations into $\mathrm{PSL}_2^k(\mathbb{R})$. Specifically, we say that the action of two elements $a, b \in \Gamma_g$ representing standard generators of a one-holed torus subsurface of Σ_g *satisfies* $S_k(a,b)$ if $\rho(a)$ and $\rho(b)$ are separately conjugate to hyperbolic elements of $\mathrm{PSL}_2^k(\mathbb{R})$, and their periodic points alternate around the circle. We show the following.

Theorem 1.7 *Let $\rho\colon \Gamma_g \to \mathrm{Homeo}^+(S^1)$ be a path-rigid, minimal representation, and suppose furthermore that there exists $k \geq 1$ such that $S_k(a,b)$ holds for all standard generators of one-holed torus subsurfaces. Then ρ is geometric.*

The proof of Theorem 1.7 uses bending deformations of ρ to move the periodic points of generators of $\pi_1\Sigma_g$; provided ρ is path-rigid, we are able to conclude the periodic points of many simple closed curves are in the same cyclic order as if ρ were geometric. In the toy version we presented in [25] — whose additional hypothesis guarantees that $k = 1$ — this same process was sufficient to demonstrate that ρ has maximal Euler number, hence is geometric. Here in the general case, we need to use a more sophisticated tool, and invoke Matsumoto's theory of *basic partitions*; see Section 3.4.

Step 2 (good and bad tori) We next make extensive use of bending deformations to prove the following result on periodic sets and rotation numbers.

Proposition 1.8 *If a representation $\Gamma_g \to G$ is path-rigid, then all nonseparating simple closed curves have rational rotation number.*

Theorem 1.9 *Suppose ρ is path-rigid and minimal. Then, for all standard generators a, b of one-holed subsurfaces, we have the implication*

$$\mathrm{Per}(\rho(a)) \cap \mathrm{Per}(\rho(b)) = \varnothing \implies S_k(a,b) \text{ for some } k.$$

The upshot of these results is that, if a path-rigid and minimal representation *fails* to be geometric, then many curves are forced to have common periodic points. Common periodic points hint at the existence of a finite orbit for ρ, so we next look for a finite orbit in order to derive a contradiction (indeed, representations with a finite orbit are easily seen to be non-path-rigid). This idea proves difficult to implement, so we search first for curves with rotation number zero, as the dynamics of these are easier to control. This search can be performed separately in every one-holed torus in the surface, where the action of the mapping class group is simple to work with. Accordingly, a one-holed torus in Σ_g is called a *good torus* if it contains a nonseparating simple loop with rotation number zero; otherwise we say it is a *bad torus*. A one-holed torus is called *very good* if its fundamental group has a finite orbit in S^1. We prove:

Proposition 1.10 *Let ρ be path-rigid. Suppose that Σ_g contains a bad torus Σ'. Then its complement $\Sigma_g \setminus \Sigma'$ contains only very good tori.*

Proposition 1.11 *Let ρ be path-rigid, and nongeometric. Then there cannot exist two disjoint good tori that are not very good.*

Theorem 1.12 *Let ρ be a path-rigid representation. Let $\Sigma_{g',1}$ be a subsurface in which all tori are very good. Then $\pi_1\Sigma_{g',1}$ has a finite orbit.*

These three last statements show that if ρ is a path-rigid and nongeometric representation, then it has a subsurface of genus $g-1$ with a finite orbit; the statement about the Euler class in Theorem 1.6 then follows easily.

Conclusion Provided $g \geq 3$, Theorem 1.12 implies that if ρ is path-rigid and nongeometric, then there exist curves a, b, generating a torus subsurface of Σ_g, such that $\rho(a)$ and $\rho(b)$ have a common fixed point. It then follows from a recent theorem of Alonso, Brum and Rivas [1] that ρ cannot be rigid. However, path-rigidity and the genus $g = 2$ case do not follow, so we prove an independent, elementary lemma on rigid representations that shows all torus subsurfaces have only finitely many finite orbits. This applies to all genera, and allows us conclude the proof of Theorem 1.1.

Roadmap The article is organized as follows. Section 2 introduces tools and frameworks that will be frequently used in the proof. We review background and prove new results on *complexes of based curves*; then prove a series of results on the movement of periodic sets under specific bending deformations; and finally discuss character spaces, semiconjugacy, and the Euler class. In Section 3 we prove Theorem 1.7. In Section 4 we prove Proposition 1.8 and Theorem 1.9. The proof of Theorem 1.6 is then completed in Section 5. Finally, in Section 6 we complete the proof of Theorem 1.1 and state some open questions and directions for further work.

Acknowledgements This work was started at MSRI during spring 2015 at a program supported by NSF grant 0932078. Both authors also acknowledge the support of National Science Foundation grants DMS 1107452, 1107263, 1107367 *RNMS*: *geometric structures and representation varieties* (the GEAR network). Mann was partially supported by NSF grant DMS-1606254, and thanks the Institut de Mathématiques de Jussieu and Fondation Sciences Mathématiques de Paris. Parts of this work were written as Wolff was visiting the Institute for Mathematical Sciences, NUS, Singapore, and the Universidad de la República, Montevideo, Uruguay; he wants to thank them for their great hospitality.

2 Preliminaries

2.1 Based curves on surfaces

This subsection should seem familiar to low-dimensional topologists, except that we will give much more attention to *based* curves than is usually present in the literature. As in the introduction, we use the notation $\Gamma_g = \pi_1 \Sigma_g$. While this notation omits mention of a basepoint, all elements of Γ_g are always assumed based. This is crucial — for example, we recall (as a warning) that the set of elements represented by *based* simple closed loops, in Γ_g, is not closed under conjugation. We now set some conventions.

Since we are interested in actions of Γ_g by homeomorphisms on the circle, we will write words in Γ_g (ie products of loops by concatenation) from right to left, in the same order as composition of homeomorphisms. We also fix the commutator notation to be $[a, b] := b^{-1}a^{-1}ba$.

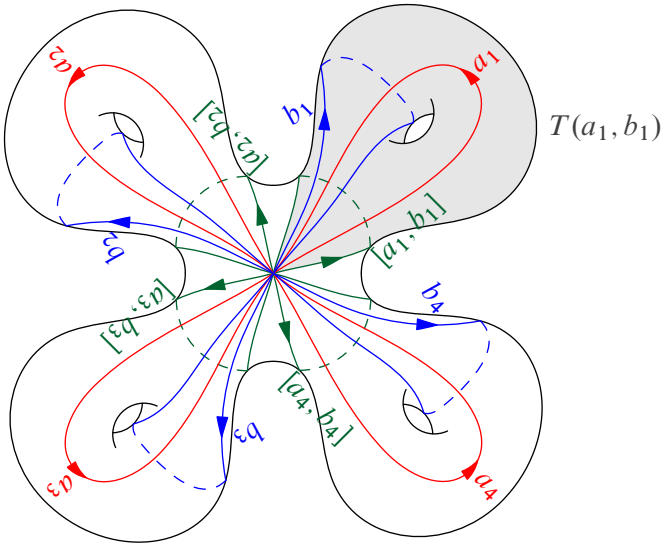


Figure 1: Standard generators on the genus g surface ($g = 4$).

The based curves $(a_1, b_1, \dots, a_g, b_g)$, depicted in Figure 1, are called a *standard system of loops*, and give the following standard presentation of Γ_g:

$$\Gamma_g = \langle a_1, b_1, \dots, a_g, b_g \mid [a_g, b_g] \cdots [a_1, b_1] = 1 \rangle.$$

We will make extensive use of systems of curves that look like those in Figure 2. Accordingly, we will say that a tuple $(\gamma_1, \dots, \gamma_k)$ of elements of Γ_g is an *oriented*, *directed* k*–chain* if these elements of Γ_g can be realized by differentiable based loops, $[0, 1] \to \Sigma_g$, that do not intersect outside the basepoint, and with cyclic order $(\gamma_1'(0), \gamma_2'(0), -\gamma_1'(1), \gamma_3'(0), -\gamma_2'(1), \gamma_4'(0), \dots, -\gamma_k'(1))$. In other words, an oriented, directed k–chain is a k–tuple of loops arising from an orientation-preserving embedding of the graph of Figure 2 (note that we do not require this embedding to be π_1–injective). If we do not insist that the embedding be orientation-preserving, we call it a *directed* k*–chain*, and, similarly, $(\gamma_1, \dots, \gamma_k)$ is simply a k*–chain* if there exist signs $\epsilon_1, \dots, \epsilon_k$ such that $(\gamma_1^{\epsilon_1}, \dots, \gamma_k^{\epsilon_k})$ is a directed k–chain. Also, we will say that a (oriented and/or directed) k–chain is *completable* if it sits in the middle of a (orientable and/or directed) $(k+2)$–chain.

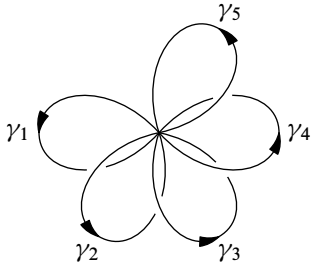


Figure 2: A directed chain of length 5.

For example, $(a_1^{-1}b_1a_1, a_1, b_1^{-1})$ is a noncompletable 3–chain in Σ_g, and the collection

$$(a_1, \delta_1, a_2, \delta_2, \dots, \delta_{g-1}, a_g, b_g^{-1})$$

(as well as its subchains), where we have set $\delta_i = a_{i+1}^{-1}b_{i+1}a_{i+1}b_i^{-1}$, forms a directed chain. Also, the family $(a_1^{-1}b_1a_1, a_1, \delta_1, a_2, b_2^{-1})$ forms a (noncompletable) 5–chain that will be handy in Section 5.3.

If two simple closed loops $a, b \in \Gamma_g$ do not intersect outside of the basepoint, we will write $i(a,b) = 1$ if (a,b) is an oriented, directed 2–chain, and we will write $i(a,b) = -1$ if $i(b,a) = 1$. Otherwise we will write $i(a,b) = 0$; if a and b are nonseparating, to say $i(a,b) = 0$ is equivalent to the existence of a curve c such that (a,c,b) is a 3–chain. Though reminiscent of the algebraic intersection number, $i(a,b)$ is an ad hoc definition, as we do not define $i(a,b)$ for most pairs (a,b) of elements of Γ_g.

Finally, if two curves $a, b \in \Gamma_g$ satisfy $i(a,b) = \pm 1$, we will denote by $T(a,b)$ the genus 1 subsurface of Σ_g defined by a and b; Figure 1 illustrates $T(a_1, b_1)$. While $T(a,b)$ is only defined up to based homotopy, it still makes sense to say, for example, that a curve γ is *disjoint* from $T(a,b)$, if $i(a,\gamma)$, $i(b,\gamma)$, $i([a,b],\gamma)$ are all defined and equal to 0.

We conclude this paragraph with some considerations on complexes of pairs of based curves.

Lemma 2.1 *Let G_0 denote the graph whose vertices are the pairs $(a,b) \in \Gamma_g^2$ with $i(a,b) = \pm 1$, with an edge between two pairs (a,b) and (b,c) whenever (a,b,c) is a 3–chain. Then G_0 is connected.*

The main results of this article do not depend on this lemma, as we will simply need to work on a connected component of this graph; our proof in the companion article [25] follows this strategy. However, the lemma is quite elementary, so here we take the honest approach of giving the proof and using the whole connected graph instead of making reference to a connected component.

The proof of Lemma 2.1 is divided into two main observations. It essentially copies the proof of Proposition 6.7 of [26], but corrects a minor mistake there, where the complex of *based* curves should have been used instead of the standard curve complex.

Observation 2.2 *Let G_1 be the graph whose vertices are the elements of Γ_g represented by simple, nonseparating curves, and with edges between a and b if and only if $i(a,b) = \pm 1$. Then G_1 is connected.*

Proof Let G_2 be the graph with the same vertices, but with edge between a and b whenever $i(a,b)$ is well defined. Let G_3 be the graph with vertex set consisting of the elements of Γ_g represented by simple curves (possibly separating), with an edge between a and b whenever $i(a,b)$ is well defined.

By drilling a puncture in Σ_g at the basepoint, G_3 can be identified with the arc graph of the surface $\Sigma_{g,1}$, which is well known to be connected; see eg [17]. Given a path in G_3 between two vertices of G_2, every time a separating curve appears we may either delete it or replace it by a nonseparating curve, producing a

new path in G_2. Thus, G_2 is connected. Finally, we prove that any path in G_2 can be promoted to a path in G_1. Let $a_1 - a_2$ be an edge of G_2 which is not in G_1, ie we have $i(a_1, a_2) = 0$. Then a neighborhood of the curves a_1 and a_2 in Σ_g is a pair of pants P, with three boundary components, freely homotopic to a_1, a_2 and $a_1 a_2^{\pm 1}$. If Σ, Σ' and Σ'' are, respectively, the connected components of $\Sigma_g \setminus P$ separated from P by a_1, a_2 and $a_1 a_2^{\pm 1}$, then we cannot have $\Sigma' \neq \Sigma''$, for otherwise a_1 or a_2 would be separating. Hence, there exists a curve b such that $a_1 - b - a_2$ is a path in G_1. □

Observation 2.3 *Let a, b and a' be such that $i(a,b) = \pm 1$ and $i(a',b) = \pm 1$. Then (a,b) and (a',b) lie in the same connected component of the graph G_0 from Lemma 2.1.*

Proof Let $\sim$ denote the equivalence relation on vertices of G_0 of being in the same connected component. Let a, b, a' be as in the statement of the observation, and let N be the (geometric) minimum number of disjoint intersections, besides the basepoint, between the based curves a and a'. We will proceed by induction on N, starting with the base case $N = 0$. In this case $i(a,a') \in \{0, \pm 1\}$. If $i(a,a') = 0$, then (a,b,a') is a 3–chain and $(a,b) \sim (b,a')$. If $i(a,a') = \pm 1$, then for some $\epsilon \in \{-1, 1\}$, we have $i(b^\epsilon a, a') = 0$ (this is seen by looking at a neighborhood of the basepoint), hence $(b^\epsilon a, b, a')$ is a 3–chain and $(b^\epsilon a, b) \sim (b, a')$. Now $(b^\epsilon a, b) \sim (a, b)$, because there exists a curve c such that $(b^\epsilon a, b, c)$ and (a, b, c) are both 3–chains. This proves the base case.

Now, suppose $N \geq 1$. Orient the curves a and a' so that their tangent vectors at $t = 0$ are on the same side of b at the basepoint. Let $(x_1, \dots, x_N)$ be the intersection points of a and a', as ordered along the path a. Let a'' be the path obtained from following a' until we hit x_N (actually, any of the x_i would do), and then following the end of the path a. Then we have $i(a,b) = \pm 1$, $i(a',b) = \pm 1$, $i(a'',b) = \pm 1$ and the intersections of a and a' with a'' outside the basepoint are strictly less than N; this concludes our induction. □

Proof of Lemma 2.1 Let (a,b) and (c,d) be such that $i(a,b) = \pm 1$ and $i(c,d) = \pm 1$. There exists a path between b and c in G_1, which can be extended to a path $\gamma_1 - \gamma_2 - \cdots - \gamma_n$ in G_1 with $(a,b,c,d) = (\gamma_1, \gamma_2, \gamma_{n-1}, \gamma_n)$. By Observation 2.3, for all $j \in \{1, \dots, n-2\}$, (γ_j, γ_{j+1}) is connected to $(\gamma_{j+1}, \gamma_{j+2})$ in G_0, hence (a,b) is connected to (c,d). □

Finally, we will also use the following easy variation of Lemma 2.1.

Lemma 2.4 *Let G denote graph whose vertices are the pairs $(a,b) \in \Gamma_g^2$ with $i(a,b) = \pm 1$, with an edge between two pairs (a,b) and (b,c) whenever (a,b,c) is a **completable** 3–chain. Then G is connected.*

Proof First, observe that whenever $T(a,b)$ and $T(c,d)$ are disjoint, (a,b) and (c,d) are in the same connected component of G. Now, observe that if (a,b,c) is a directed 3–chain, then it is completable if and only if ca is nonseparating. (The reader may find it helpful to draw a picture.) It follows that, if (a,b,c) is a noncompletable 3–chain in Σ_g, then there exists a pair (d,e) such that a, b, c do not enter $T(d,e)$. Hence, (a,b) and (b,c) are connected to (d,e) in G, and it follows that G is connected. □

2.2 Actions on the circle

2.2.1 Basic dynamics of circle homeomorphisms We quickly review some definitions for the purpose of setting notation. For more detailed background on this material, the reader may consult [12; 24; 13; 31] for example.

We denote by $\mathrm{Homeo}^{\mathbb{Z}}(\mathbb{R})$ the group of homeomorphisms of $\mathbb{R}$ commuting with translation by 1; we have a natural central extension

$$\mathbb{Z} \to \mathrm{Homeo}^{\mathbb{Z}}(\mathbb{R}) \to \mathrm{Homeo}^{+}(S^1).$$

The *translation number* (or rotation number) of an element $f \in \mathrm{Homeo}^{\mathbb{Z}}(\mathbb{R})$ is defined as

$$\widetilde{\mathrm{rot}}(f) := \lim_{n\to\infty} \frac{f^n(0)}{n} \in \mathbb{R},$$

and the Poincaré *rotation number* of an element $f \in \mathrm{Homeo}^{+}(S^1)$ is defined as $\mathrm{rot}(f) := \widetilde{\mathrm{rot}}(\widetilde{f}) \bmod \mathbb{Z}$, where $\widetilde{f}$ is any lift of f.

We assume the reader is familiar with these invariants, and with their essential properties. Those that we will use most frequently are that rot and $\widetilde{\mathrm{rot}}$ are homomorphisms when restricted to abelian (eg cyclic) subgroups, that $\mathrm{rot}(f) = p/q \in \mathbb{Q} \bmod \mathbb{Z}$ in reduced form if and only if f has a periodic orbit of period q, and that $\widetilde{\mathrm{rot}}$, and hence rot, are invariant under semiconjugacy. (The definition of semiconjugacy is recalled in Section 2.3, where we will be using it.)

We denote by $\mathrm{Per}(f) = \{x \in S^1 \mid f^n(x) = x \text{ for some } n \in \mathbb{Z} - \{0\}\}$ the set of periodic points of f. If $n = 1$, we also denote this by $\mathrm{Fix}(f)$. For $\widetilde{f} \in \mathrm{Homeo}^{\mathbb{Z}}(\mathbb{R})$, we use $\mathrm{Per}(\widetilde{f})$ to denote the set of all lifts of points of $\mathrm{Per}(f)$ to $\mathbb{R}$.

For $f \in \mathrm{Homeo}^{+}(S^1)$ with $\mathrm{Per}(f) \neq \varnothing$, let $q(f)$ denote the smallest positive integer such that $\mathrm{Fix}(f^{q(f)}) \neq \varnothing$, and let $p(f)$ be the least nonnegative integer such that f has rotation number equal to $p(f)/q(f) \bmod \mathbb{Z}$.

Define an *attracting periodic point* for f to be a point $x \in \mathrm{Per}(f)$ with a neighborhood I of x such that $f^{nq(f)}(I) \to x$ as $n \to \infty$. A *repelling* periodic point of f is defined as an attracting periodic point of f^{-1}. The sets of attracting and repelling periodic points will be denoted by $\mathrm{Per}^{+}(f)$ and $\mathrm{Per}^{-}(f)$, respectively.

2.2.2 One-parameter families and bending deformations Let $\gamma \in \Gamma_g$ be a based, simple loop. Cutting Σ_g along γ decomposes Γ_g into an amalgamated product $\Gamma_g = A *_{\langle\gamma\rangle} B$, or an HNN-extension $A*_{\langle\gamma\rangle}$, depending on whether γ is separating.

In both cases, if $\rho\colon \Gamma_g \to \mathrm{Homeo}^{+}(S^1)$ is a representation and if $(\gamma_t)_{t\in\mathbb{R}}$ is a continuous family of homeomorphisms commuting with $\rho(\gamma)$, we may define a deformation of ρ, as follows. If γ is separating and $\Gamma_g = A *_{\langle\gamma\rangle} B$, we define ρ_t to agree with ρ on A, while setting $\rho_t(\delta) = \gamma_t \rho(\delta) \gamma_t^{-1}$ for all $\delta \in B$. If γ is nonseparating, we may write $a_1 = \gamma$ and complete it into a standard generating system $(a_1, \dots, b_g)$, and set ρ_t to agree with ρ on all the generators except b_1, and put $\rho_t(b_1) = \gamma_t \rho(b_1)$.

In both cases, we call this deformation a *bending along* γ. These types of deformations were used by Thurston in order to parametrize *quasi-Fuchsian* representations of surface groups (he actually used more general bendings, as here we bend only along one simple curve). At the level of the representations, this is made explicit for example in [18], and this is the source of our inspiration. Some of our results involving bendings, especially in Section 4, can also be compared to the classical Baumslag's lemma [3, Proposition 1] and its usage in [4] or [20].

Most of the time (but not all) we will use these bendings with a one-parameter group γ_t, ie a morphism $\mathbb{R} \to \mathrm{Homeo}^+(S^1)$, $t \mapsto \gamma_t$, as provided by Lemma 2.7 below. In the special case when $\rho(\gamma) = \gamma_1$, then the deformation defined above, at $t = 1$, is the precomposition of ρ with $\tau_{\gamma *}$, where τ_γ is the Dehn twist along γ. However, for a Dehn twist to make sense as an automorphism of Γ (not up to inner automorphisms), we will use the following convention.

Convention 2.5 Suppose we are given a directed k–chain $(\gamma_1, \dots, \gamma_k)$, and wish to write a Dehn twist along the loop γ_i. Then we will always do so by pushing γ_i outside the basepoint in such a way that it intersects only γ_{i-1} and γ_{i+1} (if these curves exist) in a neighborhood of the chain. Accordingly, if ρ is a given representation and γ_i^t is a one-parameter family commuting with $\rho(\gamma_i)$, then the deformation leaves γ_j unchanged for $|j - i| \geq 2$ and $j = i$, and changes $\rho(\gamma_{i-1})$ into $\gamma_i^{-t}\rho(\gamma_{i-1})$ and $\rho(\gamma_{i+1})$ into $\rho(\gamma_{i+1})\gamma_i^t$.

Not all elements of $\mathrm{Homeo}^+(S^1)$ embed in a one-parameter subgroup. In fact, if $\mathrm{rot}(f)$ is irrational, then f embeds in such a subgroup if and only if the action of f is minimal, in which case f is conjugate to a minimal rotation. However, elements with rational rotation number do have large centralizers, giving us some flexibility in the use of bending deformations. We formalize this in the next lemma. Here, and later on, it will be convenient to fix a section of $\mathrm{Homeo}^+(S^1)$ in $\mathrm{Homeo}^{\mathbb{Z}}(\mathbb{R})$.

Notation 2.6 For $f \in \mathrm{Homeo}^+(S^1)$, let $\widehat{f} \in \mathrm{Homeo}^{\mathbb{Z}}(\mathbb{R})$ be the (unique) lift of f with $\widetilde{\mathrm{rot}}(\widehat{f}) \in [0, 1)$; we will call it the *canonical lift* of f. Later, we will also need to refer to the lift of f with translation number in $(-1, 0]$, this we denote by $\check{f}$. Note that $\widehat{f}^{-1} = \widehat{f^{-1}}$.

Lemma 2.7 (positive one-parameter families) *Let $f \in \mathrm{Homeo}^+(S^1)$ have rational rotation number, and suppose $\mathrm{Per}(f) \neq S^1$. Then there exists a one-parameter group $(f_t)_{t\in\mathbb{R}}$, which commutes with f, such that for all $t \neq 0$, $\mathrm{Fix}(f_t) = \partial\mathrm{Per}(f)$, and for all $t > 0$ and $x \in \mathbb{R} \setminus \partial\mathrm{Per}(\widetilde{f})$, we have $\widehat{f_t}(x) > x$.*

Here and in what follows, ∂X denotes the *frontier* of a subset X of $\mathbb{R}$ or S^1.

Proof The set $S^1 \setminus \partial\mathrm{Per}(f)$ consists of a union of open intervals permuted by f. Choose a single representative interval I_α from each orbit. Note that $f^{q(f)}(I_\alpha) = I_\alpha$ for any such interval, and the restriction of $f^{q(f)}$ to $S^1 \setminus \mathrm{Per}(f)$ is either fixed-point free or the identity. Thus, we may identify each I_α with $\mathbb{R}$ such that $f^{q(f)}$, in coordinates, is $x \mapsto x + C$ for some $C \in \{-1, 0, 1\}$. Define s_t on I_α to be $x \mapsto x + t$. Since these I_α are in different orbits of the action of f on S^1, we may extend s_t equivariantly to a one-parameter family of homeomorphisms of S^1. □

In all the rest of this text, if $f \in \mathrm{Homeo}^+(S^1)$, any family f_t as in Lemma 2.7 will be called a *positive one-parameter family commuting with* f, or simply a *positive one-parameter family* if f is understood.

2.2.3 Periodic sets under deformations We now make some observations on how periodic sets change under bending deformations using positive one-parameter families. The main application of these comes in Section 5.2, but they will also make a few earlier appearances.

Let f and $g \in \mathrm{Homeo}^+(S^1)$ have rational rotation numbers. It follows immediately from the definition of canonical lift that

$$x \in \mathrm{Per}(\widehat{f}) \iff \widehat{f}^{q(f)}(x) = x + p(f).$$

Let f_t be a positive one-parameter family commuting with f. Let $g_t := f_t \circ g$, and let $\widetilde{g}_t = \widehat{f_t} \circ \widehat{g}$. Note that $\widetilde{g}_t = \widehat{g}_t$, provided the rotation number of g_t is constant as t varies.

For all $(x, t_1, \dots, t_{q(g)}) \in S^1 \times \mathbb{R}^{q(g)}$, we set

$$\Delta_{f,g}(x, t_1, \dots, t_{q(g)}) = \widetilde{g}_{t_{q(g)}} \circ \cdots \circ \widetilde{g}_{t_1}(\widetilde{x}) - \widetilde{x} - p(g),$$
$$\delta_{f,g}(x, t) = \Delta_{f,g}(x, t, \dots, t) = (\widetilde{g}_t)^{q(g)}(\widetilde{x}) - \widetilde{x} - p(g).$$

This does not depend on the lift $\widetilde{x} \in \mathbb{R}$ of x, but does depend on the choice of the one-parameter family f_t (so we are somewhat abusing notation). Further, we set

$$P(f, g) = \{x \in S^1 \mid \delta_{f,g}(x, t) = 0 \text{ for all } t \in \mathbb{R}\},$$
$$N(f, g) = \{x \in S^1 \mid \delta_{f,g}(x, t) \neq 0 \text{ for all } t \in \mathbb{R}\},$$
$$U(f, g) = \{x \in S^1 \mid \text{there exists a unique } t \in \mathbb{R} \text{ such that } \delta_{f,g}(x, t) = 0\}.$$

Unlike $\delta_{f,g}$, these sets do not depend on the choice of the positive one-parameter family (provided that it is chosen as in Lemma 2.7).

Assuming $\mathrm{rot}(g_t)$ is constant, then $P(f, g) = \bigcap_{t \in \mathbb{R}} \mathrm{Per}(g_t)$ is the set of *persistent* periodic points; $N(f, g)$ is the set of points that are *never* periodic for any g_t, and $U(f, g)$ is the set of points that lie in $\mathrm{Per}(g_t)$ for a *unique* time t.

Let $T_{f,g} : U(f, g) \to \mathbb{R}$ be the map that assigns to each $x \in U(f, g)$ the unique time $t \in \mathbb{R}$ for which $\delta_{f,g}(x, t) = 0$.

Lemma 2.8 *Suppose g_t has constant rotation number. Then we have the following properties.*

(1) *The set $P(f, g)$ is closed*; *moreover*,

$$P(f, g) = \mathrm{Per}(g) \cap \bigcap_{k=0}^{q(g)-1} g^k(\partial \mathrm{Per}(f));$$

in particular, if $\mathrm{rot}(f) = 0$ then every element of $P(f, g)$ has a finite orbit under the group $\langle f, g \rangle$.

(2) *The sets $P(f, g)$, $N(f, g)$ and $U(f, g)$ partition the circle.*

(3) *The set $U(f,g)$ is open, and the map $T_{f,g}\colon U(f,g)\to\mathbb{R}$ is continuous.*

(4) *For any $\varepsilon>0$, there exists t_0 such that $\mathrm{Per}(f_t\circ g)$ lies in the ε–neighborhood of $P(f,g)\cup\partial N(f,g)$ for all $t>t_0$.*

Proof By construction, the map $\Delta_{f,g}(x,\cdot)$ is (separately, in each variable t_j) constant if $\widetilde{g}_{t_{j-1}}\circ\cdots\circ\widetilde{g}_{t_1}(\widetilde{x})$ is in $\partial\mathrm{Per}(f)$, and is strictly increasing otherwise. Monotonicity implies that the subsets $\Delta_{f,g}(x,\mathbb{R}^{q(g)})$ and $\delta_{f,g}(x,\mathbb{R})$ of $\mathbb{R}$ coincide. The affirmations (1) and (2) are easy consequences of these observations. Let us prove (3). Let $x_0\in U(f,g)$, and write $t_0=T(x_0)$, so $\delta(x_0,t_0)=0$. Fix $\varepsilon>0$. Since $x_0\in U(f,g)$, we have $\delta(x_0,t_0+\varepsilon)>0$, and $\delta(x_0,t_0-\varepsilon)<0$. Since the maps $x\mapsto\delta(x,t_0+\varepsilon)$ and $x\mapsto\delta(x,t_0-\varepsilon)$ are continuous, there exists $\eta>0$ such that for all $x\in(x_0-\eta,x_0+\eta)$, we have $\delta(x,t_0+\varepsilon)>0$ and $\delta(x,t_0-\varepsilon)<0$. Thus, for each $x\in(x_0-\eta,x_0+\eta)$, the map $t\mapsto\delta(x,t)$ takes positive and negative values, hence has a (unique) zero in the interval $(t_0-\varepsilon,t_0+\varepsilon)$. In other words, $(x_0-\eta,x_0+\eta)\subset U(f,g)$, and for all $x\in(x_0-\eta,x_0+\eta)$, we have $|T(x)-T(x_0)|<\varepsilon$.

For statement (4), fix $\varepsilon>0$. Let $I_1,\dots,I_n$ denote the (finitely many) connected components of $U(f,g)$ of length $>\varepsilon$. Let $K\subset U(f,g)$ be the set of points of $U(f,g)$ that are at distance at least ε from $P\cup\partial N$. Then, $K\subset\bigcup_i I_i$, and it follows that K is compact. Since T is continuous, its restriction to K takes values in some segment $[-t_0,t_0]$, this gives the t_0 from the statement. □

The next proposition describes the topology of the sets $P(f,g)$, $N(f,g)$ and $U(f,g)$ in more detail.

Proposition 2.9 *Suppose g_t has constant rotation number. Then all accumulation points of $\partial N(f,g)$ lie in $P(f,g)$.*

The bulk of the proof of this is accomplished by the following lemma.

Lemma 2.10 *Let $x_0\in S^1\setminus\mathrm{Per}(g)$, and suppose there exists $u_k\in U(f,g)$ converging to x_0 from the right. Then there exists $\varepsilon>0$ such that $(x_0,x_0+\varepsilon)\subset U(f,g)$.*

Of course the symmetric statement, with sequences converging to x_0 from the left, holds as well, with a symmetric proof.

Proof Let $x_0\notin\mathrm{Per}(g)$, so we have $d:=d(x_0,g^{q(g)}(x_0))>0$, and suppose some sequence $u_k\in U(f,g)$ converges to x_0 from the right. First, we claim that there exists some $j\in\{1,\dots,q(g)\}$ such that $g^j(x_0)$ is *not* accumulated on the right by points of $\partial\mathrm{Per}(f)$.

To prove the claim, suppose for contradiction that for all $j\in\{1,\dots,q(g)\}$, $g^j(x_0)$ is accumulated on the right by $\partial\mathrm{Per}(f)$. Choose $z_{q(g)}\in\partial\mathrm{Per}(f)\cap\big(g^{q(g)}(x_0),g^{q(g)}(x_0)+\frac{1}{2}d\big)$ and, inductively for $j=q(g)-1,q(g)-2,\dots,1$, define $z_j\in\partial\mathrm{Per}(f)\cap(g^j(x_0),g^{-1}(z_{j+1}))$ for $j\in\{1,\dots,q(g)-1\}$, and set $\delta=g^{-1}(z_1)-x_0$. Then, for all $t>0$, we have $(f_tg)^j(x_0,x_0+\delta)\subset(g^j(x_0),z_j)$, hence

$$(f_tg)^{q(g)}(x_0,x_0+\delta)\subset\big(g^{q(g)}(x_0),g^{q(g)}(x_0)+\tfrac{1}{2}d\big).$$

Now let $k \geq 0$ be such that $u_k \in (x_0, x_0 + \delta)$. Choose $y_1 \in (g(x_0), g(u_k)) \cap \partial\mathrm{Per}(f)$ and, inductively for $j \geq 2$, choose $y_j \in (g^j(x_0), g(y_{j-1})) \cap \partial\mathrm{Per}(f)$. Then we have $(f_t g)^{q(g)}(u_k) \in (y_{q(g)}, z_{q(g)})$ for all $t \in \mathbb{R}$, hence $(f_t g)^{q(g)}(u_k) \in \big(g^{q(g)}(x_0), g^{q(g)}(x_0) + \frac{1}{2}d\big)$; this contradicts that $u_k \in U(f, g)$, and proves the claim.

Let j be the minimum element of $\{1, \dots, q(g)\}$ such that $g^j(x_0)$ is not accumulated on the right by points of $\partial\mathrm{Per}(f)$ (ie satisfying the claim above), and let y be such that $(g^j(x_0), y] \subset S^1 \setminus \partial\mathrm{Per}(f)$. Let k be large enough that $g \circ (f_t \circ g)^{j-1}(u_k) \in (g^j(x_0), y]$ holds for all $t \in \mathbb{R}$. (Such k exists using the argument above, since $g^i(x_0)$ is accumulated on the right by $\partial\mathrm{Per}(f)$ for all $i < j$.) Let $z \in (x_0, u_k)$. We will now show that $z \in U(f, g)$.

Since f_t acts transitively on $(g^j(x_0), y]$, for T sufficiently large we have

$$f_T \circ g \circ (f_T \circ g)^{j-1}(z) > g \circ (f_T \circ g)^{j-1}(u_k).$$

If $T > T(u_k)$, this guarantees that $\delta_{f,g}(z, T) > 0$. Similarly, if T' is small enough, we will have $f_{T'} \circ g \circ (f_{T'} \circ g)^{j-1}(z) < g \circ (f_T \circ g)^{j-1}(u_{k'})$ for any given $u_{k'} \in (x_0, z)$, and choosing $T' < T(u'_k)$ ensures that $\delta_{f,g}(z, T') < 0$. This shows that $z \in U(f, g)$, as desired. □

Proof of Proposition 2.9 Let x_0 be an accumulation point of $\partial N(f, g)$. If $x_0 \notin \mathrm{Per}(g)$, then by Lemma 2.10, on any side of x_0 containing a sequence of points in $\partial N(f, g)$, there is a neighborhood of x_0 containing no points of $U(f, g)$. Since $P(f, g)$, $N(f, g)$ and $U(f, g)$ partition S^1, it follows that there is also a sequence of points in $P(f, g)$ approaching x_0 from this side. Since $P(f, g)$ is closed, $x_0 \in P(f, g) \subset \mathrm{Per}(g)$, a contradiction.

It follows that $x_0 \in \mathrm{Per}(g)$. If also $x_0 \notin P(f, g)$, then $x_0 \in U(f, g)$ since x_0 is a periodic point for $f_0 \circ g = g$. But $U(f, g)$ is open, a contradiction. □

All the discussion above describes the variation of $\mathrm{Per}(g)$ upon deforming g by composition with f_t on the left. However, one may equally well replace g by $g f_t$ and define sets P, N and U with the same properties — indeed, replacing g by $g f_t$ is equivalent to replacing g^{-1} by $f_{-t} g^{-1}$. There is no reason to privilege left-side deformations in the definition of bending, and we will occasionally make use of deformations on the right.

2.3 The character space for $\mathrm{Homeo}^+(S^1)$

Following [12], for a group Γ, two homomorphisms ρ_1 and $\rho_2 \in \mathrm{Hom}(\Gamma, \mathrm{Homeo}^{\mathbb{Z}}(\mathbb{R}))$ are said to be *semiconjugate*[3] if there exists a monotone (possibly noncontinuous or noninjective) map $h\colon \mathbb{R} \to \mathbb{R}$ such that $h(x + 1) = h(x) + 1$ for all $x \in \mathbb{R}$, and $h \circ \rho_1(\gamma) = \rho_2(\gamma) \circ h$ for all $\gamma \in \Gamma$. Similarly, ρ_1 and

[3]Note that this definition is *not* the usual notion of semiconjugacy from topological dynamical systems (eg as in [19]), which is not a symmetric relation.

$\rho_2 \in \mathrm{Hom}(\Gamma, \mathrm{Homeo}^+(S^1))$ are semiconjugate if there is such a map $h\colon \mathbb{R} \to \mathbb{R}$ such that for all γ, there are lifts $\widetilde{\rho_1(\gamma)}$ and $\widetilde{\rho_2(\gamma)} \in \mathrm{Homeo}^{\mathbb{Z}}(\mathbb{R})$ which are semiconjugate by this map h. Ghys [12] proved that, under this definition, semiconjugacy is an equivalence relation. Note that this is particular to actions on S^1 and does not agree with the usual definition of semiconjugacy from topological dynamics.

In [6, Section 1], Calegari and Walker describe an analogy between rotation numbers of elements of $\mathrm{Homeo}^+(S^1)$ and characters of linear representations. Much as characters capture the dynamics of a linear representation; rotation numbers capture representations up to semiconjugacy:

Theorem 2.11 (Ghys [12], Matsumoto [27]) *Let Γ be any group, and let S be a generating set for Γ. For $f, g \in \mathrm{Homeo}^+(S^1)$, define $\tau(f,g) := \widetilde{\mathrm{rot}}(\tilde{f}\tilde{g}) - \widetilde{\mathrm{rot}}(\tilde{f}) - \widetilde{\mathrm{rot}}(\tilde{g})$ for any lifts $\tilde{f}$ and $\tilde{g} \in \mathrm{Homeo}^{\mathbb{Z}}(\mathbb{R})$. With this notation, two representations ρ_1 and ρ_2 in $\mathrm{Hom}(\Gamma, \mathrm{Homeo}^+(S^1))$ are semiconjugate if and only if the following two conditions hold*:

(i) $\mathrm{rot}(\rho_1(s)) = \mathrm{rot}(\rho_2(s))$ *for each* $s \in S$.

(ii) $\tau(\rho_1(a), \rho_1(b)) = \tau(\rho_2(a), \rho_2(b))$ *for all* a *and* b *in* Γ.

We observe here that one can recover Calegari and Walker's analogy from our more general definition of character spaces for arbitrary groups. For a topological group G, recall that $X(\Gamma, G)$ denotes the largest Hausdorff quotient of $\mathrm{Hom}(\Gamma, G)/G$. Let $G /\!\!/ G$ denote the space $X(\mathbb{Z}, G)$; then there is, for each $\gamma \in \Gamma$ a natural, continuous map $X(\Gamma, G) \to G /\!\!/ G$, which sends the class of a representation ρ to the class of $\rho(\gamma)$. For example, when $G = \mathrm{SL}(2, \mathbb{C})$, these are precisely the trace functions. The next proposition says that when $G = \mathrm{Homeo}^+(S^1)$, these are the *rotation numbers*, and the space $X(\Gamma, G)$ is, as a set, exactly the set of semiconjugacy classes of representations.

Proposition 2.12 *Let Γ be a group. Representations $\rho_1, \rho_2 \in \mathrm{Hom}(\Gamma, \mathrm{Homeo}^+(S^1))$ are semiconjugate if and only if they are equivalent in $X(\Gamma, \mathrm{Homeo}^+(S^1))$.*

Following this analogy, the "character variety" for $\mathrm{Homeo}^+(S^1)$ not only comes with its "ring of functions" (the rotation number functions), but with an underlying topological space as well. This gives the most natural setting to speak of rigidity, or to study the global topology of the space of representations.

We defer the proof of Proposition 2.12 in order to make some preliminary observations. The first is the important remark that Proposition 2.12 has no analog in $\mathrm{Homeo}^+(\mathbb{R})$ — a group may have many dynamically distinct actions on the line, but the character space is a single point:

Proposition 2.13 *For any discrete group Γ, the space $X(\Gamma, \mathrm{Homeo}^+(\mathbb{R}))$ consists of a single point.*

Proof Let $\rho \in \mathrm{Hom}(\Gamma, \mathrm{Homeo}^+(\mathbb{R}))$. Let S be a finite, symmetric subset of Γ. Given $\varepsilon > 0$, we will conjugate ρ so that $|\rho(s)(x) - x| < \varepsilon$ holds for all $s \in S$ and $x \in \mathbb{R}$, hence show that conjugates of ρ approach the trivial representation in the compact-open topology.

As a first case, assume also that the subgroup generated by S has no global fixed points in $\mathbb{R}$. Then define $h(0) = 0$, and iteratively, for $n \in \mathbb{Z}$ define $h\big(\frac{1}{2}n\varepsilon\big) = \max_{s\in S} s\big(h\big(\frac{1}{2}(n-1)\varepsilon\big)\big)$ if $n > 0$, and $h\big(\frac{1}{2}n\varepsilon\big) = \min_{s\in S} s\big(h\big(\frac{1}{2}(n+1)\varepsilon\big)\big)$ if $n < 0$. Extend h over the interior of each interval $\big[\frac{1}{2}n\varepsilon, \frac{1}{2}(n+1)\varepsilon\big]$ as an affine map. Since S has no global fixed point, this map h is surjective, hence it is an orientation-preserving homeomorphism. Furthermore, we have $hsh^{-1}\big(\frac{1}{2}n\varepsilon\big) \in \big[\frac{1}{2}(n-1)\varepsilon, \frac{1}{2}(n+1)\varepsilon\big]$ for all $s \in S$. Thus, $|hsh^{-1}(x) - x| < \varepsilon$ holds for all $x \in \mathbb{R}$.

If instead the subgroup generated by S does have a global fixed point, we may define h to be the identity on the set F of global fixed points, and define it as above on each connected component of $\mathbb{R} \setminus F$. □

Recall that the action of any group on S^1 is either minimal, or has a finite orbit, or has a closed, invariant set (called the *exceptional minimal set*) homeomorphic to a Cantor set, on which the restriction of the action is minimal. The following is an easy consequence of the definition of semiconjugacy, which we will use in the proof of Proposition 2.12.

Observation 2.14 *Every action ρ_1 with an exceptional minimal set is semiconjugate to a minimal action ρ_2, by a **continuous** map h satisfying $h \circ \rho_1(\gamma) = \rho_2(\gamma) \circ h$. Furthermore, if ρ_2 is minimal, and ρ_1 arbitrary, then any h satisfying this equation is necessarily continuous. In particular, a semiconjugacy h between two minimal actions is invertible, and hence a **conjugacy**.*

Proof of Proposition 2.12 For one direction, it suffices to prove that the quotient of the space $\mathrm{Hom}(\Gamma, \mathrm{Homeo}^+(S^1))$ by semiconjugacy is Hausdorff. This follows from Theorem 2.11, since the maps rot and τ in the theorem are continuous, well defined on semiconjugacy classes, take values in the (Hausdorff) spaces S^1 and $\mathbb{R}$, and distinguish semiconjugacy classes.

For the converse, if ρ has a finite orbit, then we can employ a similar strategy to the proof of Proposition 2.13 to conjugate it arbitrarily close to an action on the circle by rigid rotations. Hence, there is a unique element of the character space corresponding to the semiconjugacy class of ρ.

Now suppose instead that ρ has an exceptional minimal set. By Observation 2.14 there is a minimal action ρ' and continuous map h such that each $\gamma \in \Gamma$ has lifts satisfying

$$\widetilde{\rho'(\gamma)} \circ h = h \circ \widetilde{\rho(\gamma)}$$

as in the definition of semiconjugacy. Let S be a finite subset of Γ, and fix $\varepsilon > 0$. Let $\delta \in (0, \varepsilon)$ be small enough that for all $s \in S$ and all $x, y \in S^1$, $|x - y| < \delta$ implies $|\rho'(s)(x) - \rho'(s)(y)| < \varepsilon$.

Since h is continuous and commutes with $x \mapsto x + 1$, we can approximate it by a homeomorphism $h' \in \mathrm{Homeo}^{\mathbb{Z}}(\mathbb{R})$ at C^0 distance at most δ from h. Let $s \in S$ and $x \in \mathbb{R}$, and take the lifts $\widetilde{\rho'(s)}$ and $\widetilde{\rho(s)}$ as above. Then we have

$$|\widetilde{\rho'(s)}(x) - \widetilde{\rho'(s)} \circ (h \circ h'^{-1})(x)| < \varepsilon \quad \text{and} \quad |h \circ \widetilde{\rho(s)} \circ h'^{-1}(x) - h' \circ \widetilde{\rho(s)} \circ h'^{-1}(x)| < \varepsilon,$$

hence the definition of semiconjugacy and the triangle inequality gives

$$|\widetilde{\rho'(s)}(x) - h' \circ \widetilde{\rho(s)} \circ h'^{-1}(x)| < 2\varepsilon.$$

This proves that every representation without finite orbit is weakly conjugate to the minimal representation in its semiconjugacy class. □

We conclude this section with two observations and a short lemma that will be useful later on. The observations are simple consequences of Observation 2.14.

Observation 2.15 *Let $\rho_2 \in \mathrm{Hom}(\Gamma, \mathrm{Homeo}^+(S^1))$ be minimal, and let ρ_1 be any action which is semiconjugate to ρ_2* (as in Observation 2.14)*. Then for any $\gamma \in \Gamma$, we have $\mathrm{Per}(\rho_2(\gamma)) = h\,\mathrm{Per}(\rho_1(\gamma))$, and hence $|\mathrm{Per}(\rho_2(\gamma))| \le |\mathrm{Per}(\rho_1(\gamma))|$.*

Observation 2.16 *Suppose that ρ is minimal and path-rigid, and let a and b satisfy $i(a,b) = -1$ and $\mathrm{rot}(\rho(b)) \in \mathbb{Q}$. Since $\rho(b^{q(b)})$ lies in a one-parameter family, there is a bending deformation replacing $\rho(a)$ with $\rho(b^{Nq(b)}a)$ for any $N \in \mathbb{Z}$, which is realized by precomposition with a Dehn twist* (see Section 2.2.2)*. Thus the new representation has the same image as ρ; in particular it is minimal, hence **conjugate** to ρ.*

Lemma 2.17 *Let $f, g \in \mathrm{Homeo}^+(S^1)$ be two homeomorphisms with rational rotation number. The property that f and g share a periodic point depends only on the semiconjugacy class of $\langle f, g\rangle$.*

Proof For $f_1, \dots, f_n \in \mathrm{Homeo}^+(S^1)$, let $\tau(f_1, \dots, f_n) = \widetilde{\mathrm{rot}}(\widetilde{f_n} \circ \cdots \circ \widetilde{f_1}) - \sum_i \widetilde{\mathrm{rot}}(\widetilde{f_i})$, which obviously does not depend on the choices of lifts. Note that

$$\tau(f_1, \dots, f_n) = \tau(f_1, f_n \circ \cdots \circ f_2) - \sum_{j=2}^{n-1} \tau(f_j, f_n \circ \cdots \circ f_{j+1}),$$

so this function can be recovered from the two-variable τ of Theorem 2.11.

To prove the lemma, we prove the stronger statement that f and g sharing a periodic point is equivalent to the following assertion:

For any $\ell \ge 1$ and any integers $n_1, m_1, \dots, n_\ell, m_\ell$, we have

$$\tau(f^{n_1 q(f)}, g^{m_1 q(g)}, \dots, f^{n_\ell q(f)}, g^{m_\ell q(g)}) = 0.$$

Applying Theorem 2.11 gives the desired conclusion.

The assertion is clearly true if f and g share a periodic point. To prove the converse, suppose that $\mathrm{Per}(f) \cap \mathrm{Per}(g) = \varnothing$, so $S^1 \setminus (\mathrm{Per}(f) \cup \mathrm{Per}(g))$ is a union of intervals. As $\mathrm{Per}(f)$ and $\mathrm{Per}(g)$ are closed, disjoint sets, only finitely many of these complementary intervals have one boundary point in each of $\mathrm{Per}(f)$ and $\mathrm{Per}(g)$. Those bounded on the right by a point of $\mathrm{Per}(f)$ and at their left by a point of

$\mathrm{Per}(g)$ alternate with the others (with the roles of right and left reversed), in particular there are an even number of such complementary intervals. Let $I_1, \dots, I_{2\ell}$ denote these intervals, in their cyclic order on the circle, and let $I_j = (x_j, y_j)$. Up to shifting the indices cyclically, we have $x_i, y_{i+1} \in \mathrm{Per}(g)$ and $x_{i+1}, y_i \in \mathrm{Per}(f)$ for all i even.

Choose a point x in I_1. Since the interval (x_1, y_2) contains only points of $\mathrm{Per}(g)$, there exists n_1 such that $f^{n_1 q(f)}(x) \in I_2$. Similarly, there exists a power n_2 of $g^{q(g)}$ which maps $f^{n_1 q(f)}(x)$ into I_3, and so on for n_i, with $i > 2$. The last operation can be done so that the image of x, under a suitable word $g^{n_\ell q(g)} f^{n_\ell q(f)} \dots g^{n_2 q(g)} f^{n_1 q(f)}$, lies to the right of x in I_1. Then, choosing the canonical lifts of $f^{n_i q(f)}$ and $g^{m_i q(g)}$, we observe that $\tau(f^{n_1 q(f)}, g^{m_1 q(g)}, \dots, f^{n_\ell q(f)}, g^{m_\ell q(g)}) \geq 1$. □

Remark 2.18 In the case $\mathrm{Per}(f) \cap \mathrm{Per}(g) = \varnothing$, the integer ℓ in the proof above also only depends on τ; in fact, it is the *minimal* integer such that there exist $m_i, n_i \in \mathbb{Z}$ with

$$\tau(f^{n_1 q(f)}, g^{m_1 q(g)}, \dots, f^{n_\ell q(f)}, g^{m_\ell q(g)}) \geq 1.$$

2.4 The Euler class

Recall that the (*integer*) *Euler class* for circle bundles is a generator e (well defined up to sign) of $H^2(\mathrm{Homeo}^+(S^1); \mathbb{Z}) \cong \mathbb{Z}$; and the *Euler number* of a representation $\rho\colon \Gamma_g \to \mathrm{Homeo}^+(S^1)$ is the integer $\langle \rho^*(e), [\Gamma_g] \rangle$, where $[\Gamma_g]$ denotes the fundamental class, ie a generator of $H_2(\Gamma_g, \mathbb{Z})$. Under the correspondence between second cohomology and central extensions, e is represented by the extension $\mathbb{Z} \to \mathrm{Homeo}^{\mathbb{Z}}(\mathbb{R}) \to \mathrm{Homeo}^+(S^1)$ described in Section 2.2.1 and hence can be seen as the obstruction to lifting a representation to $\mathrm{Homeo}^{\mathbb{Z}}(\mathbb{R})$.

Although this definition only makes sense for fundamental groups of closed surfaces — a surface with boundary has free fundamental group, and $H_2(F_n; \mathbb{Z}) = 0$ — there is a *relative* Euler number for surfaces with boundary, which is additive when such subsurfaces are glued together. This can be made precise in the language of bounded cohomology as explained in [5, Section 4.3]. (Compare also Goldman [14] and Matsumoto [28].) Following [5], we make the following definition.

Definition 2.19 (Euler number for pants) Let $P \subset \Sigma_g$ be a subsurface homeomorphic to a pair of pants; equip it with three based curves a, b and c as in Figure 3. (If P does not contain the basepoint, choose a path in Σ_g from the basepoint to a chosen point in P, and use it to define the curves a, b and c.) Let $\rho\colon \pi_1 \Sigma_g \to \mathrm{Homeo}^+(S^1)$, and let $\widetilde{\rho(a)}$, $\widetilde{\rho(b)}$ be any lifts of $\rho(a)$ and $\rho(b)$ to $\mathrm{Homeo}^{\mathbb{Z}}(\mathbb{R})$, and let $\widetilde{\rho(c)} = \big(\widetilde{\rho(b)}\widetilde{\rho(a)}\big)^{-1}$. Then the contribution of P to the Euler number of ρ is

$$\mathrm{eu}_P(\rho) := \widetilde{\mathrm{rot}}(\widetilde{\rho(a)}) + \widetilde{\mathrm{rot}}(\widetilde{\rho(b)}) + \widetilde{\mathrm{rot}}(\widetilde{\rho(c)}).$$

If the surface Σ_g is cut into pairs of pants, the Euler class of ρ is the sum of the contributions of these pants. See [5, Section 4.3] for a detailed discussion, and [25] for a short exposition and proof that this does not depend on the decomposition. Definition 2.19 extends naturally to one-holed tori: if $T = T(a, b) \subset \Sigma_g$ is a one-holed torus, cutting T along a simple closed curve (say, freely homotopic to a or b) yields the

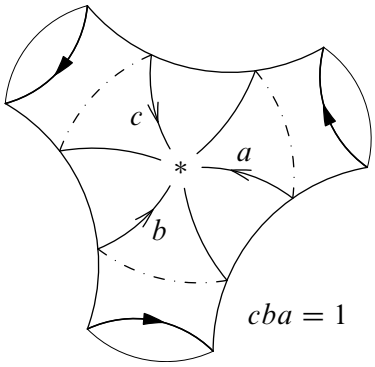


Figure 3: A pair of pants with standard generators of its fundamental group.

formula $\mathrm{eu}_T(\rho) = \widetilde{\mathrm{rot}}(\widetilde{\rho(b)}^{-1}\widetilde{\rho(a)}^{-1}\widetilde{\rho(b)}\widetilde{\rho(a)})$, which, in turn, gives Milnor's classical formula [30], $\mathrm{eu}(\rho) = \prod_{i=1}^{g}[\widetilde{\rho(a_i)}, \widetilde{\rho(b_i)}]$, where $(a_1, \dots, b_g)$ is a standard system of curves, and where the lifts are taken arbitrarily.

3 A first statement

This section proves the main theorem under a strong additional hypothesis. We will show that if ρ is path-rigid and if for every $a, b \in \Gamma_g$ with $i(a,b) = \pm 1$, $\rho(a)$ and $\rho(b)$ resemble, dynamically, a geometric representation, then ρ is in fact geometric. In other words, the local condition that ρ "looks geometric" on pairs a, b with $i(a,b) = \pm 1$ implies global geometricity. To formalize this, we introduce some definitions.

Definition 3.1 Say that an element $f \in \mathrm{PSL}_2^k(\mathbb{R})$ is *hyperbolic* if its projection to $\mathrm{PSL}_2(\mathbb{R})$ is hyperbolic. Equivalently, all its periodic points are hyperbolic in the sense of classical smooth dynamics.

Definition 3.2 Let $a, b \in \Gamma_g$ and $\rho\colon \Gamma_g \to \mathrm{Homeo}^+(S^1)$. Denote by $S_k(a,b)$ (the notation ρ is suppressed) the property that

(i) $i(a,b) = \pm 1$ and $\rho(a)$ and $\rho(b)$ are each separately conjugate to a hyperbolic element of $\mathrm{PSL}_2^k(\mathbb{R})$, and

(ii) their periodic points *alternate* around the circle, meaning that each pair of points of $\mathrm{Per}(a)$ are separated by $\mathrm{Per}(b)$, and vice versa.

If all pairs a, b with $i(a,b) = \pm 1$ have $S_k(a,b)$, then we say that ρ has property S_k.

With this notation we can state the main result of this section.

Theorem 3.3 *Let ρ be a path-rigid, minimal representation, and suppose ρ satisfies S_k for some k. Then ρ is geometric.*

Before embarking on the proof, we discuss some other variations on hyperbolicity to be used later in the section.

Let $f \in \mathrm{Homeo}^+(S^1)$. We say that an open interval $I \subset S^1$ is *attracting for* f if $f(\bar{I}) \subset I$. We say that I is *repelling for* f if it is attracting for f^{-1}. Matsumoto [28] calls homeomorphisms that do not admit attracting intervals *tame*. In line with his terminology, we call those homeomorphisms which do *savage*. More specifically, we have:

Definition 3.4 A homeomorphism $f \in \mathrm{Homeo}^+(S^1)$ is *n–savage* if there exist $2n$ open intervals with pairwise disjoint closures, indexed in cyclic order by $I_1^-, I_1^+, \dots, I_n^-, I_n^+$ such that

$$f\left(S^1 \setminus \left(\bigcup_{j=1}^{n} \overline{I_j^-}\right)\right) = \bigcup_{j=1}^{n} I_j^+.$$

In this sense, savage means 1–savage.

The next observation is an immediate consequence of the definition; we leave the proof to the reader.

Observation 3.5 *If f is n–savage, then f^k is also n–savage for any $k \in \mathbb{Z} \setminus \{0\}$. Furthermore,* $\mathrm{rot}(f^n) = 0$ *and f has at least one periodic point in each interval I_j^+ and I_j^-.*

As a concrete example, note that if f is conjugate to a hyperbolic element in $\mathrm{PSL}_2^k(\mathbb{R})$, then f is n–savage for $n \le k$.

The intervals I_j^+ and I_j^- in the definition of savage are by no way unique, but it will be convenient to use the notation $I^+(f) := \bigcup_{j=1}^n I_j^+$ and $I^-(f) := \bigcup_{j=1}^n I_j^-$, even if these sets depend on choices. We also set $I(f) := I^+(f) \cup I^-(f)$.

Definition 3.6 Two n–savage homeomorphisms $f, g \in \mathrm{Homeo}^+(S^1)$ are *in n–Schottky position* if their respective attracting and repelling intervals $I_j^\pm$ can be chosen so that $I(f)$ and $I(g)$ have disjoint closures.

Note that, if f and g are n–Schottky, then f^{-1} and g are n–Schottky as well. Note also that the condition $S_k(a,b)$ is not equivalent to k–Schottky, although $S_k(a,b)$ does imply that a^N and b^N are k–Schottky for sufficiently large N. We will prove however that hypothesis S_k on a path-rigid representation ρ implies that a and b are indeed k–Schottky whenever $i(a,b) = \pm 1$.

3.1 Outline of proof of Theorem 3.3

We start in Section 3.2 with a series of lemmas that use rigidity and property S_k to show the cyclic order of periodic points of various nonseparating curves agrees with that of a geometric representation, and that certain pairs of curves are k–Schottky. Following this, we show in Section 3.3 that the Euler number of a path-rigid, minimal, S_k representation agrees with a geometric one, ie is equal to $\pm(2g-2)/k$. From there, we need to improve this essentially combinatorial result to the fact that the representation is actually geometric. Our main tool is existing work of Matsumoto on *basic partitions*.

We are now ready to embark on the proof. Throughout, we make the following assumption.

Assumption 3.7 For the rest of this section, ρ denotes a path-rigid minimal representation of Γ_g that satisfies S_k. To simplify notation, we often omit ρ, identifying $a \in \Gamma_g$ with $\rho(a) \in \mathrm{Homeo}^+(S^1)$. Thus, we will speak of $\mathrm{Per}(a)$, denote an attracting point of $\rho(a)$ by a^+, etc.

3.2 Order of periodic points

Property S_k makes it much easier to understand periodic points under deformations. We start with several lemmas to this effect.

Lemma 3.8 *Let $i(a,b) = 1$, let $F \subset S^1$ be a countable set, and let b_t be a positive one-parameter family commuting with $b = \rho(b)$. Then for some $t \in \mathbb{R}$, we have $\mathrm{Per}(b_t\rho(a)) \cap F = \varnothing$.*

Proof We use the notation from Section 2.2.3. Path-rigidity of ρ implies that $\mathrm{rot}(b_t a)$ is constant, and Property S_k and Lemma 2.8 implies that $P(b,a) = \varnothing$, so we need only worry about points in $U = U(b,a)$. Thus, provided $t \notin T_{b,a}(F)$, we have $\mathrm{Per}(b_t a) \cap F = \varnothing$. □

Lemma 3.9 (disjoint curves have disjoint Per) *Let (a,b,c) be a completable directed 3–chain. Then $\mathrm{Per}(a) \cap \mathrm{Per}(c) = \varnothing$. In fact, $\mathrm{Per}(c) \cap b^n(\mathrm{Per}(a)) = \varnothing$ for all $n \in \mathbb{Z}$.*

Proof Fix $n \in \mathbb{N}$. Complete (a,b,c) to a directed 4–chain (a,b,c,d), and apply a bending deformation replacing c with $d_t c$ (leaving the action of a and b unchanged, hence $b^n\,\mathrm{Per}(a)$ unchanged), for a positive family d_t. By Lemma 3.8, there is some t such that $\mathrm{Per}(d_t c) \cap b^n\,\mathrm{Per}(a) = \varnothing$. Now the conclusion follows from path-rigidity of ρ, together with Lemma 2.17. □

Note that, if $i(a,b) = \pm 1$, then for any $n \in \mathbb{Z}$ we also have $i(b^n a, b) = \pm 1$, hence $S_k(b^n a, b)$ holds. The next lemma describes the position of the periodic points of $S_k(b^n a, b)$ for large n. This is particularly useful since there exist bending deformations replacing the pair a, b with $b^n a, b$ provided that $q(b)$ divides n; see Observation 2.16.

Lemma 3.10 (movement of Per by bending) *Suppose $i(a,b) = \pm 1$. Then as $N \to +\infty$, the points of $\mathrm{Per}^+(b^N a)$ approach $\mathrm{Per}^+(b)$, and $\mathrm{Per}^-(b^N a)$ approaches $a^{-1}\,\mathrm{Per}^-(b)$; similarly, as $N \to -\infty$, $\mathrm{Per}^+(b^N a)$ approaches $\mathrm{Per}^- b$ and $\mathrm{Per}^-(b^N a)$ approaches $a^{-1}\,\mathrm{Per}^+(b)$.*

Proof When $a^{-1}\,\mathrm{Per}(b) \cap \mathrm{Per}(b) = \varnothing$, the conclusion of the lemma is an easy exercise. We claim that path-rigidity of ρ implies this extra provision. To see this, suppose for example that $i(a,b) = 1$, and let (c,a,b) be a completable directed 3–chain. By Lemma 3.9, $\mathrm{Per}(c) \cap \mathrm{Per}(b) = \varnothing$. Thus, we can make a positive bending deformation replacing a with ac_t, until $(ac_t)^{-1}\,\mathrm{Per}(b) \cap \mathrm{Per}(b) = \varnothing$. □

Notation 3.11 Let f and g be homeomorphisms of S^1. When talking about cyclic order of periodic points, we use the notation $((f^+, g^+, g^-, f^-))_k$ to mean that, in cyclic order, there is one attracting point for f, followed by an attracting point for g, followed by a repelling point for g, followed by an

attracting point for f, with this pattern repeating k times. The notation $f^{\pm}$ means any point from $\mathrm{Per}(f)$. We also use other obvious variations, such as $((f^{\pm}, g^{-}, f^{\pm}, g^{+}))_k$, and extend this naturally to periodic points of three or more homeomorphisms.

When such a cyclic order is given, we call an interval $I \subset S^1$ of *type* (f^{+}, g^{-}) if it is bounded on the left (proceeding anticlockwise, using the natural orientation of S^1) by a point of $\mathrm{Per}^{+}(f)$ and on the right by a point of $\mathrm{Per}^{-}(g)$, and if it does not contain a proper subinterval with this property. We also use other obvious variations.

Lemma 3.12 (periodic points of 3–chains) *Let (a, b, c) be a completable directed 3–chain. Then, up to reversing the orientation of the circle, the periodic points of a, b and c come in the cyclic order*

$$((a^{-}, b^{-}, a^{+}, c^{\pm}, b^{+}, c^{\pm}))_k.$$

Proof Up to reversing orientation of S^1, we may suppose that the cyclic order of points in $\mathrm{Per}(a) \cup \mathrm{Per}(b)$ is $((a^{-}, b^{-}, a^{+}, b^{+}))_k$. Choose two consecutive points of $\mathrm{Per}(b)$ (in cyclic order), and denote these by b^{-} and b^{+}. Let a^{+} be the point of $\mathrm{Per}(a)$ between b^{-} and b^{+}, and let $c^{\pm}$ be the periodic point of c in this interval (there is exactly one by hypothesis S_k). The points of $\mathrm{Per}(a)$ in the interval (b^{-}, b^{+}) are in cyclic order $(b^{-}, a^{+}, b^{q(b)}(a^{+}), b^{+})$.

By Lemma 3.9, $c^{\pm}$ cannot be equal to a^{+} or $b^{q(b)}(a^{+})$. Suppose for contradiction that $c^{\pm}$ lies in the interval (b^{-}, a^{+}), or in the interval $(b^{q(b)}(a^{+}), b^{+})$. Then the closed segment $[a^{+}, b^{q(b)}(a^{+})]$ does not contain any periodic point of c. Let $(c_t)_{t\in\mathbb{R}}$ be a positive one-parameter family commuting with c, and use this to perform a bending along c as in Section 2.2.3. Using the notation from this section, we have $\delta_{c,b}(a^{+}, 0) > 0$, but for t sufficiently negative, we have $\Delta_{c,b}(a^{+}, 0, \dots, 0, t) < 0$. Thus, for some $t_0 < 0$, we have $\delta_{c,b}(a^{+}, t_0) = 0$, ie $a^{+} \in \mathrm{Per}(c_{t_0} b) \cap \mathrm{Per}(a)$. This, together with Lemma 2.17 and the path-rigidity of ρ, yields a contradiction.

The same argument applies to an interval of the form (b^{+}, b^{-}), where b^{+} and b^{-} denote two other consecutive points of $\mathrm{Per}(b)$. In that case, the argument shows that the (unique) periodic point of c in this interval lies between points of the form $b^{q(b)}(a^{-})$ and a^{-}, proving the lemma. □

In particular, for all pairs $a, c \in \Gamma_g$ such that there exists a completable 3–chain (a, b, c), Lemma 3.12 provides information about the periodic sets of a and c.

Corollary 3.13 *Let a and c be two nonseparating curves with $i(a, c) = 0$, and suppose c is not conjugate to a or a^{-1}. Then their periodic points are in cyclic order $((a^{\pm}, a^{\pm}, c^{\pm}, c^{\pm}))_k$.*

Proposition 3.14 *Since c is not conjugate to $a^{\pm 1}$, we may find b such that (a, b, c) is a completable directed 3–chain. Then, up to reversing the orientation of the circle, the periodic points of a, b and c and the b–preimages of $\mathrm{Per}(c)$ are in cyclic order*

$$((a^{-}, b^{-1}(c^{\pm}), b^{-}, b^{-1}(c^{\pm}), a^{+}, c^{\pm}, b^{+}, c^{\pm}))_k.$$

Proof of Proposition 3.14 Apply a bending deformation of ρ replacing b with $c^{Nq(c)}b$, and leaving the action of c and a unchanged. By Lemma 3.10, for N sufficiently large, $\mathrm{Per}^-(c^{Nq(c)}b)$ approaches $b^{-1}\,\mathrm{Per}^-(c)$, and $\mathrm{Per}^-(c^{-Nq(c)}b)$ approaches $b^{-1}\,\mathrm{Per}^+(c)$. Since ρ is path-rigid, the cyclic order of periodic points is invariant under these deformations, hence the points $b^{-1}(c^{\pm})$ all must lie in intervals of type (a^-, a^+).

Now up to replacing c with c^{-1} (its orientation is unimportant in this proof) we may assume that the order of periodic points given by Lemma 3.12 is $((a^-, b^-, a^+, c^+, b^+, c^-))_k$. Then $b^{-1}\,\mathrm{Per}^-(c)$ lies in the intervals of type (b^+, b^-), as b preserves these intervals. Thus, points of $b^{-1}\,\mathrm{Per}^-(c)$ are between consecutive points of $\mathrm{Per}^-(a)$ and $\mathrm{Per}^-(b)$. Similarly, the points $b^{-1}(c^+)$ are between consecutive points of the form b^- and a^+. □

The following variation is proved using the same style of argument.

Lemma 3.15 *Let $a, b, c \in \Gamma_g$ be three nonseparating curves such that $i(a,b) = -1$ and c is disjoint from $T(a,b)$. Up to reversing the orientation of S^1, we may suppose that the periodic points of a and b are in the order $((a^-, b^+, a^+, b^-))_k$. Then the periodic points of c all lie in intervals of type (b^-, a^-).*

Note that the order in which we prefer to take the periodic points of a and b is different here than in the two preceding statements, because here $i(a,b) = -1$.

Proof Similar to the proof of Proposition 3.14, we perform bending deformations. Since ρ is path-rigid, the cyclic order of periodic points does not change after the bending deformation replacing b with $a^{Nq(a)}b$ (leaving a and c unchanged). The effect of these deformations is to push $\mathrm{Per}^+(b)$ as close as we want to either $\mathrm{Per}^+(a)$ or $\mathrm{Per}^{-1}(a)$. Applying Lemma 3.10 as in the proof of Proposition 3.14 shows that periodic points of c cannot be in the intervals of type (a^-, b^+) or (b^+, a^+); as the argument is entirely analogous, we omit the details. The same argument again using the deformation replacing a by $b^{Nq(b)}a$ shows that the periodic points of c cannot be in the intervals of type (a^+, b^-), either. □

Proposition 3.16 *Let a and c be two nonseparating curves with $i(a,c) = 0$, and suppose c is not conjugate to a or a^{-1}. Then $\rho(a)$ and $\rho(c)$ are in k–Schottky position.*

Proof Up to changing the orientation of c, we may choose nonseparating curves b and d such that (a, b, c, d) is the beginning of a standard basis of $\pi_1\Sigma_g$.

Using a deformation as in Lemma 3.9, path-rigidity of ρ implies that the points of $\mathrm{Per}^-(d)$, $c^{-1}\,\mathrm{Per}^+(d)$, $\mathrm{Per}^-(b)$ and $a^{-1}\,\mathrm{Per}^+(b)$ are all distinct. Fix small disjoint neighborhoods U^+ of $\mathrm{Per}^-(d)$, U^- of $c^{-1}\,\mathrm{Per}^+(d)$, and also V^+ of $\mathrm{Per}^-(b)$, and V^- of $a^{-1}\,\mathrm{Per}^+(b)$.

By Lemma 3.10, $d^{-nq(d)}c(S^1 \setminus U^-) \subset U^+$ and $b^{-nq(b)}a(S^1 \setminus V^-) \subset V^+$ if n is large enough, so we may find $2k$ disjoint attracting and repelling intervals for $d^{-nq(d)}c$ and $b^{-nq(b)}a$ as in the definition of

k–Schottky. Now there exists a bending deformation that replaces c with $d^{-nq(d)}c$ and a with $b^{-nq(b)}a$, and it follows from Observation 2.16 that this deformation is conjugate to the original action. Thus, a and c are k–Schottky. □

Proposition 3.17 *Let a and c be two nonseparating curves with $i(a,c) = \pm 1$. Then $\rho(a)$ and $\rho(c)$ are in k–Schottky position.*

Proof Choose b and d so that (b,a,c,d) is a 4–chain. Now follow the proof above. □

From Proposition 3.16 we deduce an enhanced version of Lemma 3.12.

Proposition 3.18 *Let (a,b,c) be a completable directed 3–chain. Then, up to reversing the orientation of the circle, the periodic points of a, b and c are in cyclic order $((a^-,b^-,a^+,c^-,b^+,c^+))_k$.*

Proof By Lemma 3.15, we need only discard the possibility that the order is $((a^-,b^-,a^+,c^+,b^+,c^-))_k$. Suppose for contradiction that this order does hold. By Proposition 3.16, we know that a and c each have $2k$ intervals as in Definition 3.4, with pairwise disjoint closures. As $|\mathrm{Per}(a)| = |\mathrm{Per}(c)| = 2k$, each of these intervals contains exactly one periodic point, so their cyclic order is specified by the order of periodic points given above.

Note that ca is nonseparating, as the 3–chain (a,b,c) is completable. Also, $\rho(ca)$ is k–savage, and we may take $I^-(ca) \subset I^-(a)$ and $I^+(ca) \subset I^+(c)$. With the same argument as above, $\rho(ca)$ has exactly one repelling periodic point in each interval of $I^-(ca)$, and one attracting periodic point in each interval of $I^+(ca)$.

If $\mathrm{Per}(b)$ is disjoint from $I^-(a) \cup I^+(c)$, then this is enough to imply that the periodic points of ca and b alternate, contradicting Lemma 3.12, since $i(ca,b) = 0$. Thus, it only remains to prove that $\mathrm{Per}(b)$ can be made disjoint from $I^-(a) \cup I^+(c)$ to finish the proof. This can be done in the same manner as that of Proposition 3.16. First, complete (a,b,c) into a directed 5–chain (α,a,b,c,γ). Then, consider a bending deformation of ρ, where b is unchanged but the action of a is replaced by that of $a\alpha^{Nq(\alpha)}$ and the action of c by $\gamma^{Nq(\gamma)}c$ for N large. By Observation 2.16 this new action is conjugate to ρ. Now, provided N is large enough, we can choose our Schottky intervals to be as narrow as we want, around the points α^-, $a(\alpha^+)$, γ^+ and $c^{-1}(\gamma^-)$ which, using Lemma 3.9, are disjoint from $\mathrm{Per}(b)$. □

3.3 Euler number

As a consequence of the work in the previous section, we show that the Euler number of ρ agrees with a geometric representation.

Theorem 3.19 *Let ρ be path-rigid, minimal and satisfy S_k. Then $|\mathrm{eu}(\rho)| = (2g-2)/k$.*

In fact, we will show the following stronger statement, which implies Theorem 3.19 by additivity of the Euler number on subsurfaces.

Theorem 3.20 *Up to changing the orientation of the circle, for every pair-of-pants subsurface $P \subset \Sigma_g$, the relative Euler class of ρ on P is $-1/k$.*

Definition 3.21 Let $i(a,b) = 1$. We say that the ordered pair (a,b) is of type $+$ if the periodic points of a and b are in the cyclic order $((a^-, b^-, a^+, b^+))_k$. Otherwise, we say that (a,b) is of type $-$.

As a consequence of Proposition 3.18, for every oriented, completable directed 3–chain (a,b,c), the pairs (a,b) and (b,c) have the same type. Thus, Lemma 2.4 implies that all one-holed tori have the same type. Thus, up to conjugating ρ by an orientation-reversing homeomorphism, we may suppose the type is always $+$.

Proof of Theorem 3.20 We begin by proving the claim for a pair of pants P such that at least two boundary components of P are nonseparating. Denote by a^{-1}, c^{-1} and ac the three boundary components of P, with the convention of Figure 3, and suppose that a and c are nonseparating. With these choices of orientations, the Euler number of ρ on P will be equal to $\widetilde{\mathrm{rot}}(\widehat{ac}) - \widetilde{\mathrm{rot}}(\widehat{a}) - \widetilde{\mathrm{rot}}(\widehat{c})$, and there exists a curve b such that (a,b,c) is an oriented, completable, directed 3–chain — the end of the proof of Observation 2.2 justifies the existence of such a curve b.

Since (a,b) is of type $+$, it follows from Proposition 3.18 that the periodic points of a and c are in cyclic order $((a^-, a^+, c^-, c^+))_k$; and by Proposition 3.16, they are in k–Schottky position, with Schottky intervals $I_j^{\pm}(a)$ and $I_j^{\pm}(c)$. Lift these to intervals $\widetilde{I}_j^{\pm}(a)$ and $\widetilde{I}_j^{\pm}(c) \subset \mathbb{R}$, indexed by integers, and in order

$$\dots \widetilde{I}_j^-(a), \widetilde{I}_j^+(a), \widetilde{I}_j^-(c), \widetilde{I}_j^+(c), \widetilde{I}_{j+1}^-(a), \dots$$

such that the projection to S^1 is given by taking indices mod k. It follows easily from the definition of Savage (see also Observation 3.5) that $\widehat{a}(\widetilde{I}_j^+(a)) \subset \widetilde{I}_{j+\ell}^+(a)$ for some ℓ (which depends on a) and in this case $\ell/k = \widetilde{\mathrm{rot}}(\widehat{a})$. An analogous statement holds also for c; let m/k denote its translation number.

Since a and c are in k–Schottky position, their product ac is k–savage, and we can take $I^-(ac) = I^-(c)$ and $I^+(ac) \subset I^+(a)$. Note that each of the k intervals of $I^+(ac)$ is contained in a different interval of $I^+(a)$. We now track images of intervals to compare translation numbers. Set the indexing of the intervals $\widetilde{I}^{\pm}(ac)$ so that $\widetilde{I}_1^+(a) = \widetilde{I}_1^+(ac)$. This lies between $\widetilde{I}_0^+(c)$ and $\widetilde{I}_1^-(c)$, so we have

$$c(\widetilde{I}_1^+(ac)) \subset \widetilde{I}_m^+(c),$$

and similarly, since $\widetilde{I}_m^+(c)$ lies between $\widetilde{I}_m^+(a)$ and $\widetilde{I}_{m+1}^-(a)$, we have

$$ac(\widetilde{I}_1^+(ac)) \subset a(\widetilde{I}_m^+(c)) \subset \widetilde{I}_{m+\ell}^+(a) = \widetilde{I}_{m+\ell}^+(ac).$$

Thus, $k \cdot \widetilde{\mathrm{rot}}(\widehat{ac}) = m + \ell - 1 = k \cdot \widetilde{\mathrm{rot}}(\widehat{a}) + k \cdot \widetilde{\mathrm{rot}}(\widehat{c}) - 1$ and hence $k(\widetilde{\mathrm{rot}}(\widehat{ac}) - \widetilde{\mathrm{rot}}(\widehat{a}) - \widetilde{\mathrm{rot}}(\widehat{c})) = -1$, as desired.

This implies Theorem 3.19, as we can cut the surface Σ_g into pairs of pants whose boundary components are all nonseparating.

Now, if P is a pair of pants with possibly more than one separating boundary component, then $\Sigma_g \setminus P$ admits a pants decomposition whose pants all have at most one separating boundary component. The fact that the contribution of P to the Euler class of ρ is $-1/k$ is then a consequence of Theorem 3.19 and the additivity of the Euler class. □

3.4 Basic partitions and combinations

Fix disjoint, nonseparating curves $C_1, \dots, C_{3g-3}$ so that $\Sigma_g \setminus \left(\bigcup_i C_i\right)$ is a disjoint union of pairs of pants $P_1, \dots, P_{2g-2}$. For concreteness, the reader may use the decomposition suggested in Figure 4.

We briefly part from the convention for the presentation of $\pi_1\Sigma_g$ that was given in Section 2.1, and instead present $\pi_1\Sigma_g$ as the fundamental group of a graph of groups. Choose basepoints $x_i \in P_i$ and $y_j \in C_j$, identifying x_1 with the basepoint of $\pi_1\Sigma_g$. Also, choose paths in P_i from x_i to each basepoint of each boundary component of P_i. This collects all the basepoints of the pants and curves as the vertices of a graph G embedded in Σ_g; fix an orientation for each of its edges, and a spanning tree $T \subset G$. This data gives a *graph of groups*: the vertex (resp. edge) groups are the fundamental groups of the based pairs of pants (resp. curves), and for each edge C_j, the chosen paths define monomorphisms ϕ_j and ψ_j from $\pi_1 C_j \simeq \mathbb{Z}$ to the fundamental groups of the two adjacent (initial and final endpoints of the edge, respectively) pairs of pants. The Seifert–Van Kampen theorem then identifies $\pi_1\Sigma_g$ with the fundamental group of this graph of groups; this is the group generated by the union of the $\pi_1 P_i$, as well as one extra generator e_j for each edge that is not in T, subject to the relations that for each edge C_j (in T or not), and each $\gamma \in \pi_1 C_j$, we have $\phi_j(\gamma) = e_j^{-1}\psi_j(\gamma)e_j$ (taking $e_j = 1$ for the edges in T).

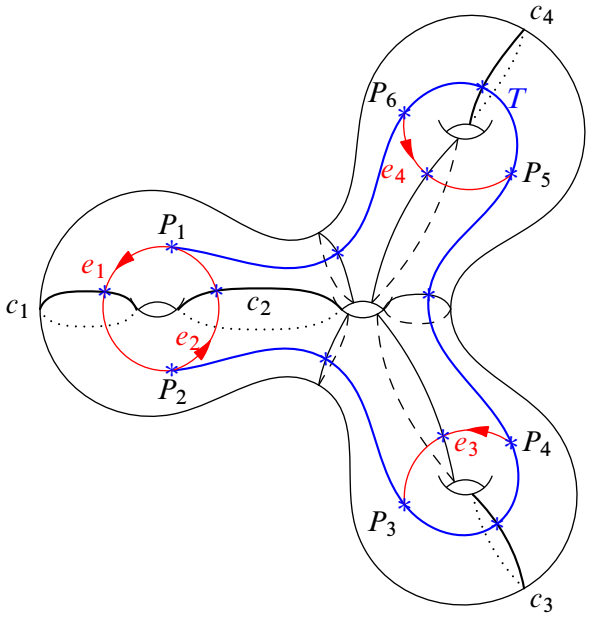


Figure 4: A decomposition of $\pi_1\Sigma_4$ into a graph of groups.

Our representation ρ gives rise to a representation of each $\pi_1(P_i)$, by using the spanning tree T to identify based curves in P_i with based curves in Σ_g. Similarly, each additional edge generator e_j can be identified with a closed, based loop in Σ_g, hence to an element $\rho(e_j)$.

We now define a geometric representation that will be our candidate for a representation semiconjugate to ρ. As a consequence of Theorem 3.20, $(2g-2)/k$ is an integer, hence a Fuchsian representation of Euler class $2g-2$ can be lifted to $\mathrm{PSL}_2^k(\mathbb{R})$. The choice of such a lift amounts to the choice of rotation numbers (in $(1/k)\mathbb{Z} \bmod \mathbb{Z}$) for the elements of a homology basis of $\pi_1\Sigma_g$. Let $c_1,\dots,c_g,e_1,\dots,e_g$ be the homology basis depicted in Figure 4, with c_j a generator of $\pi_1(C_j)$. Thus, as just observed, there exists a geometric representation ρ_0 with the same Euler class as ρ, and with $\mathrm{rot}(\rho_0(\gamma)) = \mathrm{rot}(\rho(\gamma))$ for each γ in $\{c_1,\dots,c_g,e_1,\dots,e_g\}$. This also holds for each $\gamma \in \{c_{g+1},\dots,c_{3g-3}\}$. Indeed, the contribution of the Euler class of ρ and ρ_0 on each pairs of pants are equal, and they are sums of rotation numbers, so we can propagate these equalities to the whole family of cutting curves.

To show ρ and ρ_0 are semiconjugate, thereby concluding the proof, we use (an adaptation of) Matsumoto's theory of basic partitions and combinations.

Definition 3.22 (Matsumoto [29]) Let Γ be a group generated by a finite symmetric set S, and let $\rho\colon \Gamma \to \mathrm{Homeo}^+(S^1)$. A *basic partition* (*BP*) for $\rho(\Gamma)$ is a collection P of disjoint closed intervals of S^1 such that

(i) for each $I \in P$, there is a unique $s_I \in S$ such that $\rho(s_I)(I)$ is a union of $m = m(I)$ elements of P and $m-1$ complementary intervals to P,

(ii) for any $s \neq s_I$ in S, the image $\rho(s)(I)$ is a proper subset of an element of P, and

(iii) for any complementary interval J to P and $s \in S$, either $\rho(s)(I)$ is contained in the interior of P, or is a complementary interval to P.

Following the last condition, we may put the complementary intervals to P into a directed graph, with an edge from J_1 to J_2 if there is a generator sending J_1 to J_2. A basic partition is called *pure* if this graph consists of disjoint nontrivial cycles.

Applying this to our context, for each pair of pants P_i, choose two "preferred" boundary components as generators for $\pi_1 P_i$ (identified with a subgroup of Γ_g via T). Let a_i^{-1} and c_i^{-1} denote these elements, and consider their images under ρ. The proof of Theorem 3.20 shows that the periodic points of a_i, c_i, $c_i a_i$ and $a_i c_i$ are in the cyclic order

$$((a_i^-, a_i^+, (a_i c_i)^+, (a_i c_i)^-, c_i^-, c_i^+, (c_i a_i)^+, (c_i a_i)^-))_k$$

and that the $4k$ intervals of types $(a_i^+, (a_i c_i)^+)$, $((a_i c_i)^-, c_i^-)$, $(c_i^+, (c_i a_i)^+)$ and $((c_i a_i)^-, a_i^-)$ form a *pure basic partition* for the action of $\pi_1 P_i$ on the circle with respect to the symmetric generating set $(a_i, c_i, a_i^{-1}, c_i^{-1})$. This conclusion rested only upon rigidity and the hypothesis S_k, and the combinatorics of the BP (the images of intervals and complementary intervals following conditions (i)–(iii) of the definition) depends only on the rotation numbers of the generators. Thus, ρ_0 admits a basic partition with

the same combinatorics as ρ, ie there exists a cyclic-order-preserving map sending the basic partition of one to the other, which intertwines the two actions. In this case, [29, Theorem 4.7] states that the restrictions of ρ and ρ_0 to $\pi_1 P_i$ are semiconjugate.

It remains to improve this to a global semiconjugacy between ρ and ρ_0. With the notation above, in a pair of pants P_i, let J_a (resp. J_c, resp. J_{ac}) denote the union of all intervals of type (a_-, a_+) (resp. (c_-, c_+), resp. $((ac)_+, (ac)_-)$. These are called the *entries* of the basic partition described above; their stabilizers in $\pi_1 P_i$ are the cyclic groups generated by a, c and ac, respectively.

Now consider an edge e_j of G (in T or not). It serves to conjugate one generator of $\pi_1 P_i$ for some i, a^{-1}, c^{-1} or ac, into *the inverse* of the corresponding generator of this boundary component on the adjacent pair of pants. It follows that if, say, a_i and $a_{i'}$ are the generators of $\pi_1 P_i$ and $\pi_1 P_{i'}$ on each side of an edge e_j, then the sets J_{a_i} and $\rho(e_j)(J_{a_{i'}})$ form a partition of S^1, up to the finitely many periodic points of a_i. In this situation, Matsumoto says that the two entrances J_{a_i} and $J_{a_{i'}}$ are *combinable*. More generally, given a graph of groups decomposition of a group Γ as ours, and pure basic partitions for each vertex group that have combinable entrances for every edge, Matsumoto says the collection of all basic partitions for the vertex groups form a *basic configuration* for the action $\rho(\Gamma)$ on the circle. (Matsumoto works with trees of groups; but this definition generalizes immediately to the graph setting.)

As we already argued for the $\pi_1 P_i$, the equalities between rotation numbers of ρ and ρ_0 on the curves C_i and on the edge elements e_j imply that they admit basic configurations with the same combinatorics; in other words there exists a cyclic-order-preserving bijection which maps the basic partitions of ρ to those of ρ_0, intertwining the actions.

Matsumoto's main result [29, Theorem 6.7] is that a cyclic-order-preserving bijection between basic configurations can be promoted to a semiconjugacy between ρ and ρ_0. We comment briefly on the proof. To produce a semiconjugacy, it suffices to show that some orbit of ρ and some orbit of ρ_0 are in the same cyclic order. Matsumoto's proof strategy begins by showing this property holds for elements of vertex groups (ie of some $\pi_1 P_i$) — this is the content of [29, Theorem 4.7] cited above. He then proceeds with elements of the form $\gamma_i e_j \gamma_{i'}$ (where $\gamma_i \in P_i$ and $\gamma_{i'} \in P_{i'}$ belong to adjacent pairs of pants), then of the form $\gamma_{i_3} e_{j_2} \gamma_{i_2} e_{j_1} \gamma_{i_1}$, and so on, inductively. While his proof is not carried out in the language of Bass–Serre theory, and the context is specialized to a tree of groups decompositions of $\pi_1 \Sigma_g$, the arguments adapt without modification.

4 Periodic considerations

The content of this section is the proof of the following two statements.

Proposition 4.1 *If a representation $\Gamma_g \to G$ is path-rigid, then all nonseparating simple closed curves have rational rotation number.*

Theorem 4.2 *Suppose ρ is path-rigid and minimal. Then, for all a, b with $i(a,b) = \pm 1$, we have the implication*

$$\mathrm{Per}(a) \cap \mathrm{Per}(b) = \varnothing \implies S_k(a,b) \text{ for some } k.$$

Proof of Proposition 4.1 Suppose for contradiction that there exists a nonseparating simple curve a with $\rho(a) \notin \mathbb{Q}$. After semiconjugacy, we may assume that ρ is minimal. If $\rho(a)$ is conjugate into SO(2), then it lies in a one-parameter subgroup a_t of rotations, and for any b with $i(a,b) = 1$, the bending deformation $a_t\rho(b)$ has nonconstant rotation number, contradicting rigidity. Thus, $\rho(a)$ has an invariant minimal Cantor set, which we denote by K. We next show that K is $\rho(b)$–invariant, for any curve b with $i(a,b) = 1$. This suffices to prove the proposition since Γ_g is generated by $\{a\} \cup \{b \mid i(a,b) = 1\}$, whence K is $\rho(\Gamma_g)$–invariant, contradicting minimality of ρ.

To show invariance, suppose for contradiction that $\rho(b)(K) \not\subset K$; the case where $\rho(b^{-1})(K) \not\subset K$ is analogous. Let $K' \subset K$ be the set of two-sided accumulation points of K. Since $\overline{K'} = K$, there exists $x \in K'$ such that $\rho(b)(x) \notin K$. Let I be the connected component of $S^1 \setminus K$ containing $\rho(b)(x)$. Minimality of the action of $\rho(a)$ on K implies there exists $N \in \mathbb{Z}$ such that $\rho(a)^N(I) \subset \rho(b)^{-1}(I)$, and in particular $\mathrm{rot}(\rho(a^N b)) = 0$. We work now with the pair $(a, a^N b)$ with intersection number ± 1. Let β_t be a positive one-parameter family commuting with $\rho(a^N b)$. Since $\rho(a^N b)$ does not preserve K, we can find a connected component J of $S^1 \setminus \mathrm{Fix}(\rho(a^N b))$ such that $J \cap K' \neq \varnothing$, and then find $M \in \mathbb{Z}$ such that $\rho(a)^M(J) \cap J \neq \varnothing$.

Let $\widetilde{x} \in \mathbb{R}$ be a lift of a point in $\rho(a)^M(J) \cap J$. Adapting the notation from Section 2.2.3, set

$$\Delta(\widetilde{x}, t_1, \dots, t_M) = \widehat{\beta_{t_M}}\,\widehat{\rho(a)} \circ \cdots \circ \widehat{\beta_{t_1}}\,\widehat{\rho(a)}(\widetilde{x}) - \widetilde{x} - k,$$

where k is chosen so that $\widehat{\rho(a)}^M(\widetilde{J}) \cap (\widetilde{J} + k) \neq \varnothing$ for any lift of J, and we set $\delta(\widetilde{x}, t) = \Delta(\widetilde{x}, t, \dots, t)$. Up to reversing orientation, we can suppose that $\delta(\widetilde{x}, 0) > 0$. Since $\widetilde{J}$ contains both $\widetilde{x}$ and $\widehat{\rho(a)}^M(\widetilde{x})$, there exists $t < 0$ such that $\Delta(\widetilde{x}, 0, \dots, 0, t) < 0$, hence $\delta(\widetilde{x}, t) < 0$. Thus, there exists t_0 such that $\delta(\widetilde{x}, t_0) = 0$, hence $\mathrm{rot}(\rho_{t_0}(a)) = k/M \in \mathbb{Q}$, contradicting rigidity. □

4.1 Proof of Theorem 4.2

For this subsection, we assume ρ is path-rigid, $i(a,b) = \pm 1$, and $\mathrm{Per}(a) \cap \mathrm{Per}(b) = \varnothing$. Recall from Proposition 4.1 that $\mathrm{Per}(a)$ and $\mathrm{Per}(b)$ are nonempty. We will first establish some properties that do not use minimality, so are robust under deformations of ρ. We add the hypothesis that ρ is minimal only at the end of the proof.

Borrowing notation from the previous section, say that a connected component of $S^1 \setminus (\mathrm{Per}(a) \cup \mathrm{Per}(b))$ is of type (x, y) if it is bounded to the left by a point of $\mathrm{Per}(x)$ and to the right by a point of $\mathrm{Per}(y)$, for $x, y \in \{a, b\}$.

Definition 4.3 Let X_a denote the set of connected components of $S^1 \setminus \mathrm{Per}(a)$ that contain points of $\mathrm{Per}(b)$. We say an element I of X_a is *positive* if $a^{q(a)}$ is increasing on the interval I, and *negative* otherwise.

The set X_b and its positive and negative elements are defined by reversing the roles of a and b above. Since each (a,b) interval in $S^1 \setminus (\mathrm{Per}(a) \cap \mathrm{Per}(b))$ is followed by a collection — perhaps empty — of (b,b) intervals, and then a (b,a) interval, and $\mathrm{Per}(a)$ and $\mathrm{Per}(b)$ are disjoint closed sets, there exists an integer $m = m(\rho) \geq 1$ such that S^1 contains exactly m intervals of type (a,b) and m intervals of type (b,a), alternating around the circle, and thus $|X_a| = |X_b| = m(\rho)$. By Remark 2.18, m depends only on the semiconjugacy class of ρ.

Lemma 4.4 *The set X_a is $\rho(a)$–invariant, and the subset of positive (resp. negative) intervals in X_a is also $\rho(a)$–invariant.*

Proof Let $I \in X_a$ be a positive interval; we show that its image under a is another positive interval in X_a. The negative case is analogous. Since $a(I)$ is an interval between two consecutive points of $\mathrm{Per}(a)$ on which $a^{q(a)}$ is increasing, we need only show that $a(I) \cap \mathrm{Per}(b) \neq \varnothing$.

Suppose for contradiction that $a(I) \cap \mathrm{Per}(b) = \varnothing$. Then $a(\bar{I}) \subset J$ for some $J \in X_b$. Let b_t be a positive one-parameter family commuting with b, let $x \in I \cap \mathrm{Per}(b)$, and take lifts $\widetilde{x} \in \widetilde{I}$ of x and I to $\mathbb{R}$. Positivity implies $\delta_{b,a}(x,0) > 0$. If $t < 0$ is negative enough that $b_t(a(I)) \cap a(I) = \varnothing$, then we have $\widehat{b_t}(\widehat{a}(\widetilde{x})) < \widehat{a}(\widetilde{I})$; it follows that $\delta_{b,a}(x,t) < 0$. Therefore, there exists $t_0 \in \mathbb{R}$ such that $\delta_{b,a}(x,t_0) = 0$, ie $x \in \mathrm{Per}(b_{t_0}a) \cap \mathrm{Per}(b)$. This contradicts path-rigidity via Lemma 2.17. □

Obviously, reversing the roles of a and b above shows the positive and negative intervals of X_b are b–invariant. The next lemma shows X_a and X_b are invariant under particular bending deformations.

Lemma 4.5 *Let b_t be a positive one-parameter family commuting with b. For $t \in \mathbb{R}$, let $X_b(t)$ denote the set of connected components I of $S^1 \setminus \mathrm{Per}(b)$ such that $I \cap \mathrm{Per}(b_t a) \neq \varnothing$. Then $X_b(t) = X_b(0)$ for all t.*

Proof Let $X_b(t)$ be as in the statement of the lemma and let $X_a(t)$ denote the set of connected components of $S^1 \setminus \mathrm{Per}(b_t a)$ containing points of $\mathrm{Per}(b)$. By our discussion above, path-rigidity of ρ implies that the cardinality of $X_b(t)$ is constant. Let $K_a = \{(x,t) \in S^1 \times \mathbb{R} \mid x \in \mathrm{Per}(b_t a)\}$, and $K_b = \mathrm{Per}(b) \times \mathbb{R}$. These are closed, disjoint sets, and their intersections with each horizontal slice $S^1 \times \{t\}$ are the periodic sets of $b_t a$ and b, respectively.

For each connected component $I \subset S^1 \setminus \mathrm{Per}(b)$, we set

$$T_I = \{t \in \mathbb{R} \mid I \in X_b(t)\} = \{t \in \mathbb{R} \mid I \cap \mathrm{Per}(b_t a) \neq \varnothing\}.$$

Note that T_I is the projection of $K_a \cap (\bar{I} \times \mathbb{R})$ onto the $\mathbb{R}$–factor, so in particular is closed. We claim T_I is also open. To see this, let $t_0 \in T_I$, and let $I_2, \ldots, I_m$ be the other components of $S^1 \setminus \mathrm{Per}(b)$ such that $t_0 \in T_{I_j}$. If $d > 0$ is the distance (for the product metric) between the disjoint compact sets $(S^1 \times [t_0 - 1, t_0 + 1]) \cap K_a$ and $(S^1 \times [t_0 - 1, t_0 + 1]) \cap K_b$, let $I_{m+1}, \ldots, I_N$ be the remaining connected

components of $S^1 \setminus \mathrm{Per}(b)$ of length $\geq d$. Any component J of shorter length tautologically satisfies $T_J \cap [t_0-1, t_0+1] = \varnothing$. Since the sets T_{I_j} are closed, there exists $\varepsilon > 0$ such that $(t_0-\varepsilon, t_0+\varepsilon) \cap T_{I_j} = \varnothing$ for all $j \geq m+1$, hence $(t_0-\varepsilon, t_0+\varepsilon) \subset T_I$, for otherwise $|X_b(t)|$ would fail to be constant. This proves that T_I is open, hence equal to $\varnothing$ or $\mathbb{R}$, and the intervals in $X_b(t)$ do not depend on t. □

The next two lemmas establish some properties of a and b which are, in particular, held by pairs of homeomorphisms semiconjugate to hyperbolic elements of $\mathrm{PSL}_2^k(\mathbb{R})$ satisfying $S_k(a,b)$. Of course, both lemmas also hold with the roles of a and b exchanged.

Lemma 4.6 *Any two consecutive intervals of X_a have opposite sign. In particular, $m(\rho) = 2k$ for some $k \geq 1$.*

Proof Let b_t be a positive one-parameter family commuting with $\rho(b)$. Suppose for contradiction that X_a has two successive positive intervals I_1 and I_2 (the negative case is analogous). Let $I \in X_b$ be the interval such that $I_1 \cap I \neq \varnothing$ and $I_2 \cap I \neq \varnothing$. Take $x \in I_1 \setminus I$ such that $a^{q(a)}(x) \in I$. For t large enough, we have $a^{q(a)} b_t a^{q(a)}(x) \in I_2 \setminus I$. Since b_t has positive dynamics, it follows that $(b_t a^{q(a)})^2$ moves every point of I to the right; thus, $\Delta_{b,a}(y, 0, \dots, 0, t) > 0$ for all $y \in I$, and $\mathrm{Per}(b_t a) \cap I = \varnothing$ for t large enough. But this contradicts Lemma 4.5. □

Lemma 4.7 *Let $I \in X_b$ have left endpoint in a positive interval of X_a. Then $a(I) \subset J$ for some $J \in X_b$. If, instead, $I \in X_b$ has left endpoint in a negative interval of X_a, then $a^{-1}(I) \subset J$ for some $J \in X_b$.*

Note that Lemma 4.6 implies that, in both cases, J is a positive interval of X_b if and only if I is.

Proof Let $x_1, x_2, \dots, x_6$ be points in cyclic order such that (x_1, x_3) and (x_4, x_6) are consecutive (positive and negative, respectively) intervals in X_a, and $I = (x_2, x_5) \in X_b$. Let $y_i = a(x_i)$ for $i = 1, 3, 4, 6$. Then (y_1, y_3) and (y_4, y_6) are in X_a, and both intersect some interval of X_b, say (y_2, y_5). The statement of the lemma is that $a(x_5) \leq y_5$ and $a(x_2) \geq y_2$.

Similar to the proof of Lemma 4.4, we assume the contrary and find a deformation with a common periodic point for a and b. Suppose $a(x_5) > y_5$ (the proof of the other inequality is symmetric), and choose a positive one-parameter family b_t commuting with b. Since $a^{-1}(y_5) \in (x_2, x_5)$, there is $t \in \mathbb{R}$ with $b_t a^{-1}(y_5) \in (x_1, x_3)$. As (y_1, y_3) is $a^{q(a)}$–invariant, it follows that $a^{-q(a)+1} b_t a^{-1}(y_5) < y_5$, ie $\Delta_{b,a}(y_5, 0, \dots, 0, t, 0) > 0$. On the other hand, as (y_4, y_6) is a negative interval of X_a, we have $\delta_{b,a}(y_5, 0) < 0$. Thus, there exists $t_0 \in \mathbb{R}$, such that $y_5 \in \mathrm{Per}(b_{t_0} a)$. Since $y_5 \in \mathrm{Per}(b)$, this contradicts path-rigidity by Lemma 2.17. The statement concerning $\rho(a)^{-1}$ is symmetric, and proved in the same manner. □

Now we state a lemma of purely technical nature, that will allow us to compress the periodic sets in each interval of X_a or of X_b to singletons. In the statement and proof, we let $\tau_t : \mathbb{R} \to \mathbb{R}$ denote the translation $x \mapsto x + t$.

Lemma 4.8 *Let $n \geq 1$, and for all $i = 1, \dots, n$, let f_i be an increasing homeomorphism from $\mathbb{R}$ to some interval $(a_i, b_i) \subset \mathbb{R}$. Assume that $a_i > -\infty$ for at least one i, and that $b_j < +\infty$ for at least one j. For all $t \in \mathbb{R}$, we set $F_t = \tau_t \circ f_n \circ \cdots \circ \tau_t \circ f_1$. Then there exists a subset $N \subset \mathbb{R}$ of finite Lebesgue measure and consisting of a countable union of segments, such that for all $t \notin N$, the map F_t admits a unique fixed point in $\mathbb{R}$.*

The statement of this lemma came from our attempt to better understand the argument in the first four lines of [9, page 644]. In particular, the case $n = 1$ gives an alternative end to the proof of [9, Lemma 2.7]. We defer the proof of Lemma 4.8 to the next paragraph, and use it now to finish the proof of Theorem 4.2.

Proof of Theorem 4.2 Assume now that ρ is minimal. Let b_t be a positive one-parameter family commuting with b. We will first find t such that $b_t a$ has exactly $2k$ periodic points; the conclusion will then follow easily.

Let X_a^+ denote the set of positive intervals of X_a. As observed in Lemma 4.4, $\rho(a)$ induces a permutation of X_a^+; which has, say, ℓ orbits, all of cardinality $n = k/\ell$. Fix an interval $I_0 \in X_b$ whose left endpoint lies in an element of X_a^+. Successive applications of Lemma 4.7, for $j = 1, 2, \dots n-1$, gives $\rho(a)^j(I_0) \subset I_j$ for some $I_j \in X_b$. Also, $\rho(a)^n(I_0) \subset I_0$ because $\rho(a)^n$ fixes X_a^+. Note that there exists some j such that $\rho(a)(I_{j-1}) \subset I_j$ is a strict inclusion at the left endpoint (and similarly, another for the right endpoint) as otherwise some endpoint of I_0 would lie in $\mathrm{Per}(a) \cap \mathrm{Per}(b)$.

For each j, let $\phi_j : I_j \to \mathbb{R}$ be a homeomorphism such that $\phi_j \circ b_t \circ \phi_j^{-1} = \tau_t$, and for $j \in \{1, \dots, n\}$ set $f_j = \phi_{j+1} \circ a \circ \phi_j^{-1}$, using cyclic notation for the last index. Then Lemma 4.8 applies, giving a set $N_{I_0} \subset \mathbb{R}$ of finite Lebesgue measure, such that for all $t \notin N_{I_0}$, $(b_t a)^n = \phi_1^{-1} \circ F_t \circ \phi_1$ has a unique fixed point in I_0.

We repeat this procedure for each element I of X_b, using a^{-1}, instead of a for the intervals of X_b whose left endpoint lies in an element of X_a^-. The resulting, finitely many, sets N_I, each of finite Lebesgue measure, cannot cover $\mathbb{R}$, hence there exists $t \in \mathbb{R}$ such that each element of X_b intersects $\mathrm{Per}(b_t a)$ as a singleton. By Lemma 4.5, $\mathrm{Per}(b_t a) \subset X_b$, hence $b_t a$ has exactly $2k$ periodic points. As $b_t a$ is obtained by a bending deformation that does not change the dynamics of a, by Lemma 4.6 these $2k$ periodic points have alternating attracting and repelling dynamics. One may now repeat the same procedure reversing the roles of a and b, to obtain a further deformation where b has exactly $2k$ periodic points, alternately attracting and repelling. Minimality of ρ and Observation 2.15 implies the original action of $\rho(a)$ and $\rho(b)$ also had this dynamics. □

Proof of Lemma 4.8 We suggest the reader take $n = 1$ at first reading, as the argument is less technical in that case. We will show that there exists a countable union of segments, $N_+ \subset \mathbb{R}_+$, of finite Lebesgue measure, such that F_t has a unique fixed point for all $t \in \mathbb{R}_+ \setminus N_+$. The case for $t < 0$ is symmetric and left to the reader.

Let j be an index such that $b_j < +\infty$. Let $A_t = \tau_t \circ f_j \circ \cdots \circ \tau_t \circ f_1$, and let $B_t = \tau_t \circ f_n \circ \cdots \circ \tau_t \circ f_{j+1}$. For fixed t, both maps A_t and B_t are homeomorphisms to their images so $F_t = B_t \circ A_t$ has a unique fixed point x if and only if $A_t \circ B_t$ has a unique fixed point (namely, $B_t(x)$). In other words, we may suppose without loss of generality that $j = n$.

Let $G(t, x) = F_t(x) - t$. Then G is strictly increasing in x, and increasing (strictly, if $n \geq 2$) in t. The monotonicity of G, and the assumptions $\sup(a_j) > -\infty$ and $b_n < +\infty$, imply that the range of the map $G : \mathbb{R}_{\geq 0} \times \mathbb{R} \to \mathbb{R}$ is a bounded interval, say (a_0, b_0), where $b_0 = b_n$.

If $x \geq b_0$, the map $t \mapsto F_t(x)$ is a homeomorphism between $\mathbb{R}_{\geq 0}$ and $[F_0(x), +\infty)$, and

$$F_0(x) = G(0, x) < b_0.$$

Hence, there is a unique $t = T(x)$ such that $F_t(x) = x$. This defines a function $T : [b_0, +\infty) \to (0, +\infty)$.

Sublemma 4.9 *The map T satisfies the following inequalities*:

(T1) *For every $x \in [b_0, +\infty)$, we have $a_0 < x - T(x) < b_0$.*

(T2) *For all $x_1, x_2 \in [b_0, +\infty)$ such that $x_1 < x_2$, we have*

$$f_1(x_1) - f_1(x_2) < T(x_2) - T(x_1) < x_2 - x_1.$$

In particular, T is continuous, at bounded distance from the identity, and its rate of increase is bounded above by 1.

The proof of Sublemma 4.9 is a straightforward consequence of the definition of T, the defining identity $F_{T(x)}(x) = x$, and monotonicity of G. We leave it as an exercise, noting for (T2) that the first inequality is trivially satisfied if $T(x_2) \geq T(x_1)$, and the second if $T(x_2) \leq T(x_1)$.

For the next step, define a map $H : \mathbb{R}_{\geq b_0} \to [T(b_0), +\infty)$ by

$$H(x) = \sup\{T(x') \mid x' \leq x\}.$$

The reader may verify that H is continuous, surjective, and for all $A \geq T(b_0)$, the set $H^{-1}(A)$ is a segment of the form $[a, b]$ (possibly $a = b$), with $T(a) = T(b) = A$.

Now let $W = \{w \in [T(b_0), +\infty) \mid H^{-1}(w) \text{ is not a singleton}\}$, and for all $w \in W$ denote $H^{-1}(w)$ by $[a_w, b_w]$. Since these segments are disjoint and of positive length, W is countable. By definition of H, we have $F_w(a_w) = a_w$, ie $G(w, a_w) + w = a_w$; and the same holds for b_w in place of a_w. Thus, the segment $[G(w, a_w), G(w, b_w)]$ has the same length $b_w - a_w$. The reader may now easily deduce from monotonicity of G that these segments are disjoint; as they are contained in $[a_0, b_0]$, this implies $\sum_{w \in W} b_w - a_w \leq b_0 - a_0$.

Finally, for all $w \in W$, define $N_w := [w - (b_w - a_w), w]$, and define

$$N_+ = [0, b_0 - a_0] \cup \bigcup_{w \in W} N_w.$$

This may not be a disjoint union, but the remarks above imply this countable union of segments has finite Lebesgue measure. Hence, the proof of Lemma 4.8 amounts to the following sublemma.

Sublemma 4.10 *For all $t \in \mathbb{R}_{\geq 0} \setminus N_+$, the map F_t has a **unique** fixed point.*

Proof Let $t > b_0 - a_0$ be such that F_t has at least two distinct fixed points, say x_1, x_2 with $x_1 < x_2$. By definition, these satisfy $G(t, x_i) + t = x_i$. Since $G(t, x) > a_0$ for all x, and $t > b_0 - a_0$, this implies $x_1, x_2 \in [b_0, +\infty)$. By definition of T, we have $T(x_1) = T(x_2) = t$. Let $x_0 = \min\{x \leq x_2 \mid T(x) = H(x_2)\}$. Then $x_0 < x_2$. Indeed, if $H(x_2) = t$ then $x_0 \leq x_1$, and if $H(x_2) > t$ then the maximum $H(x_2)$ is reached at some point to the left of x_2. Thus, $x_0 = a_w$ for some $w \in W$, and we also have $b_w \geq x_2$.

We claim now that $t \in N_w$. Since $x_2 \leq b_w$, by definition of H we have $w = H(b_w) \geq t = T(x_2)$. Applying inequality (T2) to x_2 and b_w now gives $w - t \leq b_w - x_2$, so $w - t \leq b_w - a_w$, hence $t \geq w - (b_w - a_w)$. Thus we indeed have $t \in N_w$. □

This concludes the proof of Lemma 4.8. □

5 Proof of Theorem 1.6

In this section we finish the proof of the main result for path-rigid representations, showing that a path-rigid representation ρ of Γ_g is either geometric, or has Euler class zero and a genus $g-1$ subsurface whose fundamental group has finite orbit under ρ. (We believe the latter case cannot actually occur.) As in Section 3, we will frequently drop the notation ρ when the context is clear, using a to denote $\rho(a)$.

Recall from the introduction that, if ρ is a given representation and $T \subset \Sigma_g$ is a one-holed torus, we say that T is a *good torus* if it contains a nonseparating simple closed curve a with $\mathrm{rot}(a) = 0$, and that T is *bad* otherwise. We say T is *very good* if $\pi_1(T)$ has a finite orbit in S^1.

Note that very good implies good: if $T(a, b)$ is very good, then $\mathrm{rot}\colon \pi_1(T) \to \mathbb{R}/\mathbb{Z}$ is a homomorphism onto a finite subgroup, so if $0 \neq |\mathrm{rot}(a)| \leq |\mathrm{rot}(b)| < 1$, one may find n such that $|\mathrm{rot}(a^n b)| < |\mathrm{rot}(a)|$. Iterating this process produces a simple closed curve with rotation number zero.

Assumption 5.1 For the remainder of this section, we assume $\rho\colon \Gamma_g \to \mathrm{Homeo}^+(S^1)$ is path-rigid.

5.1 Bad tori

This subsection contains the proof of Proposition 1.10: under Assumption 5.1 we show that if Σ_g contains a bad torus T, then $\Sigma_g \setminus T$ contains only very good tori.

Definition 5.2 Let $f, g \in \mathrm{Homeo}^{\mathbb{Z}}(\mathbb{R})$. We say that g *dominates* f, and write $f < g$, if $f(x) < g(x)$ for all $x \in \mathbb{R}$.

Note that $<$ is a left- and right-invariant partial order on $\mathrm{Homeo}^{\mathbb{Z}}(\mathbb{R})$, and satisfies the following obvious properties:

(1) For all $f, g \in \mathrm{Homeo}^{\mathbb{Z}}(\mathbb{R})$, $f > g \iff f^{-1} < g^{-1}$.

(2) For all $f \in \mathrm{Homeo}^{+}(S^1)$, $\widehat{f} > \mathrm{Id} \iff \mathrm{rot}(f) \neq 0$.

(3) For all $f, g \in \mathrm{Homeo}^{\mathbb{Z}}(\mathbb{R})$,

$$f < g \implies \widetilde{\mathrm{rot}}(f) \leq \widetilde{\mathrm{rot}}(g) \quad \text{and} \quad (f < g \text{ or } g < f) \iff \widetilde{\mathrm{rot}}(f^{-1}g) \neq 0.$$

Property (2) uses the notation $\widehat{f}$ from Notation 2.6, which is also adopted throughout this section. The following easy observation will be handy; it follows directly from property (3) above.

Observation 5.3 *Let $f, g \in \mathrm{Homeo}^{\mathbb{Z}}(\mathbb{R})$. Suppose that $\widetilde{\mathrm{rot}}(f) < \widetilde{\mathrm{rot}}(g)$ and also that $\widetilde{\mathrm{rot}}(g^{-1}f) \neq 0$. Then $f < g$.*

Building on this observation, we have the following.

Lemma 5.4 *Let (a, b) be standard generators of a bad torus T. Then there exist integers m and n, unique and well defined modulo $q(a)$, with $(n-m)p(a) = 1 \bmod q(a)$, and such that for all j not divisible by $q(a)$, we have $\widehat{a^n b} < \widehat{a^j}$, and $\widetilde{a^j} < \widetilde{a^m b}$. Moreover, if $p(a) = 1$, then we have $\widehat{a^n b^2} < \widehat{a}$, or $\widetilde{a^{n-1} b^{-2}} < \widehat{a}$, or both.*

Assumption 5.1 is used in the proof only to guarantee that all nonseparating simple closed curves have rational rotation number (Proposition 4.1).

Proof Let F be a finite orbit of a. If there exists some point $x \in F \cap b^{-1}(F)$, then there exists $N \geq 0$ such that $\rho(a)^N \rho(b)(x) = x$, thus $\mathrm{rot}(a^N b) = 0$, contradicting the fact that T was bad. Thus, $F \cap b^{-1}(F) = \varnothing$.

Now we claim that F and $b^{-1}(F)$ alternate. Suppose for contradiction that some connected component $I = (x_1, x_2)$ of $S^1 \setminus F$ contains at least two points of $b^{-1}(F)$. Let $y_1 \in b^{-1}(F)$ be the leftmost point of $b^{-1}(F)$ in I, and y_2 be the second leftmost such point. Then there exists $N > 0$ such that $a^N b(y_1) = x_1$. It follows that $a^N b(y_2) = x_2$ and $(a^N b)^{-1}(I) = (y_1, y_2) \subset I$, so $\mathrm{rot}(a^N b) = 0$, giving the desired contradiction.

Now that we know these sets alternate, choose $x \in b^{-1}(F)$, and let y_ℓ, $y_r \in F$ be the left and right endpoints of the component of $S^1 \setminus F$ containing x. Then there exists a unique pair $(n, m) \in \{0, \dots, q(a)-1\}^2$ such that $a^n b(x) = y_r$ and $a^m b(x) = y_\ell$. In particular, $(n-m)p(a) = 1 \bmod q(a)$. These m, n are obviously the only candidates, modulo $q(a)$, for the dominations $\widehat{a^n b} < \widehat{a^j}$ and $\widetilde{a^m b} > \widetilde{a^{-j}}$, for an integer j such that $a^j(y_\ell) = y_r$. (This shows m and n do not depend on F.) We claim that this pair (n, m) satisfies the statement of the lemma.

To see this, lift F to $\widetilde{F} \subset \mathbb{R}$ and let $x_1 < x_2 < \cdots < x_{q(a)}$ be consecutive points of $\widetilde{F}$. Then $\widehat{a^n b}(x_i) \leq x_{i+1}$ for all i, hence $\widetilde{\mathrm{rot}}(\widehat{a^n b}^{q(a)}) \leq 1$ and $\widetilde{\mathrm{rot}}(\widehat{a^n b}) \leq 1/q(a)$. Also, for any integer j not divisible by $q(a)$ we

have $\widetilde{\mathrm{rot}}(\widehat{a^n b}) \le \widetilde{\mathrm{rot}}(\widehat{a^j})$. Since T is bad, $\widetilde{\mathrm{rot}}(\widehat{a^j}^{-1}\widehat{a^n b}) \neq 0$, so we must have $\widehat{a^n b} < \widehat{a^j}$ by Observation 5.3. An essentially identical argument shows that $\widetilde{a^m b} > \widetilde{a^j}$.

It remains only to prove the statement regarding the case $p(a) = 1$, where $n - 1 = m \bmod q(a)$. We know that $\widehat{a} > \widehat{a^n b}$ and $\widehat{a} > \widehat{b^{-1}a^{1-n}} = \widetilde{a^{n-1}b^{-1}}$, and this immediately implies $\widehat{a} = \widehat{a^n b} \cdot \widehat{b^{-1}a^{1-n}}$. As $(a, a^n b)$ and hence $(b^{-1}a^{1-n}, a^n b)$ are also standard generating sets of $\pi_1(T)$, we must either have $\widehat{b^{-1}a^{1-n}} > \widehat{a^n b}$, or $\widehat{b^{-1}a^{1-n}} < \widehat{a^n b}$, otherwise the nonseparating simple closed curve $a^{n-1}ba^n b$ would have rotation number zero. The statement follows. □

As a consequence, we have the following.

Proposition 5.5 *Let (a, b) be a standard generating set for a bad torus. Let $(a_k, b_k)_{k\ge 0}$ be the sequence of standard generating sets, defined inductively as follows.*

- *Define $(a_0, b_0) = (a, b)$.*
- *If k is even, let $a_{k+1} = a_k$ and $b_{k+1} = a_k^{n(k)} b_k$, where $0 \le n(k) \le q(a_k) - 1$ is the integer given by Lemma 5.4 applied to the generators (a_k, b_k).*
- *If k is odd, let $b_{k+1} = b_k$ and $a_{k+1} = b_k^{n(k)} a_k$, where $0 \le n(k) \le q(a_k) - 1$ is obtained, similarly, by inputting (b_k, a_k) into Lemma 5.4.*

Then for all $k \ge 0$ even, we have $\widehat{a_{k+1}} > \widehat{b_{k+1}}$, and for $k \ge 0$ odd, we have $\widehat{a_{k+1}} < \widehat{b_{k+1}}$.

Moreover, for all $k \ge 0$, we have $\widehat{a_k} > \widehat{a_{k+2}}^2$, and $\widehat{b_k} > \widehat{b_{k+2}}^2$. In particular, both sequences $(\mathrm{rot}(a_k))_{k\ge 0}$ and $(\mathrm{rot}(b_k))_{k\ge 0}$ converge to zero.

Note that the sequence (a_k, b_k) is built so that both $\mathrm{rot}(a_k)$ and $\mathrm{rot}(b_k)$ converge to zero from above. This choice is arbitrary.

Proof The first consideration follows immediately from the first statement of Lemma 5.4. Let us prove the second. Let $k \ge 0$ be even. If $p(a_k) \ge 2$, let $n = n(k) \ge 0$ be such that $np(a_k) = 1 \bmod q(a_k)$. Then $\mathrm{rot}(a_k^n) = 1/q(a_k)$, and $\widehat{a_k^n}^{p(a_k)} = \widehat{a_k}$. By a direct application of Lemma 5.4 we conclude that $\widehat{b_{k+1}} < \widehat{a_k^n}$, hence $\widehat{b_{k+1}}^{p(a_k)} < \widehat{a_k}$, and $\widehat{a_{k+2}}^2 < \widehat{a_k}$.

Otherwise, $p(a_k) = 1$, and again we take $n(k)$ as in Lemma 5.4. If $\widehat{a_k^{n(k)} b_k}^2 < \widehat{a_k}$, then we may conclude as above. Otherwise, $\widehat{b_k^{-1} a_k^{1-n}}^2 < \widehat{a_k}$, ie $\widehat{b_{k+1}^{-1} a_{k+1}}^2 < \widehat{a_k}$. Thus, either $n(k+1)$ is equal to -1 modulo $q(b_{k+1})$, or not; in which case we have

$$\widetilde{\mathrm{rot}}\big(\widehat{b_{k+1}^{n(k+1)} a_{k+1}}\big) < \widetilde{\mathrm{rot}}\big(\widehat{b_{k+1}^{-1} a_{k+1}}\big),$$

and then $\widehat{b_{k+1}^{n(k+1)} a_{k+1}} < \widehat{b_{k+1}^{-1} a_{k+1}}$. In either case we conclude that $\widehat{a_{k+2}}^2 < \widehat{a_k}$.

The argument is symmetric for k odd, and for b_k instead of a_k. In particular, $\widehat{a_{k+2}}^2 < \widehat{a_k}$ implies that $0 < \widetilde{\mathrm{rot}}(\widehat{a_{k+2}}) < \frac{1}{2}\widetilde{\mathrm{rot}}(\widehat{a_k})$, hence the sequences $(\widetilde{\mathrm{rot}}(\widehat{a_k}))$ and $(\widetilde{\mathrm{rot}}(\widehat{b_k}))$ converge to zero from above. □

Let $T = T(a,b)$ be a bad torus, and let (a_k, b_k) be the sequence furnished by Proposition 5.5. Let $x \in S^1$, and let $\tilde{x} \in \mathbb{R}$ be a lift of x. Then, by Proposition 5.5, the sequence $(\widehat{a_k}(\tilde{x}))_k$ is decreasing, bounded below by $\tilde{x}$, hence it converges to some real number that we denote by $\tilde{x} + j_T(x)$. Note that $j_T(x)$ does not depend on the choice of the lift of x. We define

$$\mathcal{A}_T := \{x \in S^1 \mid j_T(x) = 0\}.$$

The reader should interpret this as the set of points that are moved arbitrarily small distances by elements of $\{a_k\}$. Although the notation (a,b) is suppressed, $\mathcal{A}_T$ as defined is dependent on the generating set we started with. (But see Step 1 of the proof of Proposition 5.7 below.) As usual, we let $\widetilde{\mathcal{A}}_T$ denote the preimage of $\mathcal{A}_T$ in $\mathbb{R}$. The following proposition may be viewed as an algorithmic proof (as it runs essentially on the Euclidean algorithm as introduced in Proposition 5.5) of Hölder's classical result that any group acting freely on the circle is abelian.

Proposition 5.6 (properties of $\mathcal{A}_T$) (1) *$\mathcal{A}_T$ is a nonempty, proper subset of S^1, with no isolated points, hence is infinite.*

(2) *For every $x \in S^1$, we have $\min\{\widetilde{\mathcal{A}}_T \cap [\tilde{x}, \infty)\} = \tilde{x} + j_T(x)$. In particular, $x + j_T(x) \in \mathcal{A}_T$ for all x.*

(3) *The commutator $[a,b]$ fixes $\mathcal{A}_T$ pointwise.*

Proof Let $x \in \mathbb{R}$. For all $k \geq 0$ we have $\widehat{a_k}(x) > x + j_T(x)$, from which follows $\widehat{a_k}^2(x) > x + j_T(x) + j_T(x + j_T(x))$. But $\widehat{a_{k-2}}(x) > \widehat{a_k}^2(x)$, and, by definition, $\widehat{a_{k-2}}(x)$ converges to $x + j_T(x)$. This proves that $x + j_T(x) \in \mathcal{A}_T$ and thus $\mathcal{A}_T$ is nonempty. Further, if the open interval $(x, x + j_T(x))$ contained a point $y \in \widetilde{\mathcal{A}}_T$, then for large k we would have $x + j_T(x) > \widehat{a_k}(y) > y > x$, contradicting that a_k preserves orientation. This proves property (2).

To prove property (3), let $x \in \widetilde{\mathcal{A}}_T$ and observe, as above, that the sequence $\widehat{a_k}^4(x)$ also converges to x. Fix $\varepsilon > 0$, and let k be even, and large enough that $x_1 = x$, $x_2 = \widehat{a_k}(x)$, $x_3 = \widehat{a_k}^2(x)$ and $x_4 = \widehat{a_k}^3(x)$ all lie in the interval $[x, x + \varepsilon]$. By Lemma 5.4, $a_{k+1} = a_k$ and $\widehat{b_{k+1}}$ is dominated by $\widehat{a_{k+1}}$. Thus, $\widehat{b_{k+1}}(x_3) \in (x_3, x_4)$, and $\widehat{b_{k+1}}^{-1}(x_2, x_3) \subset (x_1, x_3)$. It follows that $[a_{k+1}, b_{k+1}] = [a,b]$ maps the point x_2 into the interval (x_1, x_3), hence, for all $\varepsilon > 0$, $[a,b]$ maps a point of $[x, x + \varepsilon]$ in $[x, x + \varepsilon]$, whence $[a,b](x) = x$.

It remains to prove that $\mathcal{A}_T \neq S^1$, and $\mathcal{A}_T$ has no isolated point. If $\mathcal{A}_T = S^1$, then $[a,b] = \mathrm{id}$ and the restriction of ρ to $\langle a, b\rangle$ would have abelian image; this contradicts the fact that T is bad. Finally if x were an isolated point of $\mathcal{A}_T$, we could take $x_0 \in S^1$ such that $[x_0, x) \cap \mathcal{A}_T = \varnothing$. Let x_1 be the next point of $\mathcal{A}_T$ to the right of x. Then $x_0 + j_T(x_0) = x$, so for all $k \geq 0$, we have $\widehat{a_k}(x_0) > x$. But then x_1 is the next point of $\mathcal{A}_T$ to the right of $\widehat{a_k}(x_0)$, so $\widehat{a_k}^2(x_0) > x_1$ holds, and hence, also, $\widehat{a_{k-2}}(x_0) > x_1$. As this is true for all k, it contradicts the fact that $\widehat{a_{k-2}}(x_0)$ converges to x as $k \to \infty$. □

Using j_T, we now prove the following major step towards Proposition 1.10.

Proposition 5.7 *There cannot exist two disjoint bad tori in Σ_g.*

Proof By contradiction, let $T = T(a,b)$ and $T' = T(a',b')$ be two disjoint bad tori. Up to re-indexing and reversing some of these curves, we may suppose that (a,b,a',b') is the beginning of a standard basis of $\pi_1\Sigma_g$.

Step 1 *We have $j_T = j_{T'}$.*

We proceed by contradiction. Suppose for some $x_0 \in S^1$ we have $j_T(x_0) \neq j_{T'}(x_0)$; without loss of generality assume $j_T(x_0) < j_{T'}(x_0)$. Let $(a_k,b_k)_{k\geq 0}$ and $(a'_k,b'_k)_{k\geq 0}$ be the sequences of generators of T and T' furnished by Proposition 5.5. For k large enough, we have $\widehat{a_k}(x_0) < x_0 + j_{T'}(x_0)$. Let m be as in Lemma 5.4 applied to (a_k,b_k), and put $\alpha = a_k$, and $\beta = a_k^m b_k$. Then (α,β) is a standard generating set for T, and $\widehat{\alpha} > \widehat{\beta^{-1}}$. Since $\mathrm{rot}(b'_\ell) \to 0$, for $\ell \geq 0$ large enough we have $\widetilde{\mathrm{rot}}(\widehat{b'_\ell}) < \widetilde{\mathrm{rot}}(\widehat{\beta^{-1}})$. But $\widehat{b'_\ell}(x_0) > x_0 + j_{T'}(x_0)$ (indeed, $\widehat{b'_\ell}$ dominates $\widehat{a'_{\ell+1}}$, by construction of the sequences in Proposition 5.5); hence $\widehat{a_k}$ does not dominate $\widehat{b'_\ell}$. We now prove a sublemma to derive a contradiction; this will conclude the proof of Step 1.

Sublemma 5.8 *Let $T(a,b)$ be a bad torus, and let b' be a nonseparating simple curve outside $T(a,b)$ such that $b'^{-1}a$ and bb' are simple. Suppose that $\widehat{a} > \widehat{b^{-1}}$ and $\widetilde{\mathrm{rot}}(\widehat{b^{-1}}) > \widetilde{\mathrm{rot}}(\widehat{b'})$. Then $\widehat{a}$ dominates $\widehat{b'}$.*

Proof Suppose that $\widehat{a}$ does not dominate $\widehat{b'}$. Then $\widehat{b^{-1}}$ does not dominate $\widehat{b'}$ either. Observation 5.3 then asserts that $\mathrm{rot}(b'^{-1}a) = \mathrm{rot}(bb') = 0$. Now $i(b'^{-1}a, bb') = \pm 1$, and $b'^{-1}a$ lies in a one-parameter family, so, as in Observation 2.16, there is a path-deformation of ρ replacing the action of bb' with $b'^{-1}a \cdot bb'$. Hence,

$$\mathrm{rot}(bb') = 0 = \mathrm{rot}(b'^{-1}a \cdot bb') = \mathrm{rot}(ab).$$

This contradicts that $T(a,b)$ is bad. □

Step 2 *We can deform the representation so that $j_T \neq j_{T'}$.*

As shown in the proof of Proposition 5.6, $[a,b] \neq \mathrm{id}$, but $\mathcal{A}_T \subset \mathrm{Fix}([a,b])$. Let $x \in S^1 \setminus \mathrm{Fix}([a,b])$, so then $j_T(x) > 0$. Let $y = x + j_T(x)$, let I be the connected component of $S^1 \setminus \mathrm{Fix}([a,b])$ containing x, and let c_t be a one-parameter family of homeomorphisms commuting with $[a,b]$, and with support equal to $\bar{I}$.

Then the distance between $c_t(x)$ and $c_t(y)$ varies, in a nonconstant way, with t: it goes to zero as $t \to \infty$ if $y \in I$, and simply changes if $y \notin I$. Now, consider a bending deformation of ρ defined by $\rho_t(\gamma) = \rho(\gamma)$ for all curves outside T, and $\rho_t(\gamma) = c_t \rho(\gamma) c_{-t}$ for $\gamma \in \langle a,b\rangle$. This deformation changes the value of $j_T(x)$, without changing the value of $j_{T'}(x)$. In particular, after this path-deformation, Step 1 no longer holds! This gives a contradiction. □

Supposing again that $T(a,b)$ is a bad torus, it remains to show that any torus in $\Sigma_g \setminus T(a,b)$ is not only good, but *very good*. The next lemma will allow us to easily achieve this goal.

Lemma 5.9 *Let $T = T(a,b)$ be a bad torus, and let γ be a nonseparating simple closed curve outside of T, with $\mathrm{rot}(\gamma) = 0$. Then $\mathscr{A}_T \subset \mathrm{Fix}(\gamma)$.*

Proof Let $(a_k, b_k)_{k\geq 0}$ be the sequence given by Proposition 5.5, and orient γ so that $\gamma^{-1}a_k$ is also a (nonseparating) simple curve. Fix $k \geq 0$, and let $\alpha = a_k$ and $\beta = a_k^m b_k$, as in Lemma 5.4. Then, by Sublemma 5.8, we have $\widehat{a_k} > \hat{\gamma}$. This holds for all $k \geq 0$; hence, for all $x \in \mathbb{R}$, we have $\hat{\gamma}(x) \leq x + j_T(x)$. In particular, if $x \in \widetilde{\mathscr{A}_T}$, we have $\hat{\gamma}(x) \leq x$.

For the reverse inequality, first note the conditions $\check{\alpha} < \widetilde{b^{-1}}$ and $\widetilde{\mathrm{rot}}(\widetilde{b^{-1}}) < \widetilde{\mathrm{rot}}(\check{\gamma})$ imply the domination $\check{\alpha} < \check{\gamma}$ (this is exactly the statement of Sublemma 5.8 after reversing the orientation of $\mathbb{R}$), and $\check{\gamma} = \hat{\gamma}$ since $\mathrm{rot}(\gamma) = 0$. Let $x \in \widetilde{\mathscr{A}_T}$, and fix $\varepsilon > 0$. For k large enough, the sequence (a_k, b_k) from Proposition 5.5 satisfies $\widehat{a_k}(x) < x + \varepsilon$. Let $(a', b') = (a_k, b_k)$ for such a large k, and define $b'' = b'$ and $a'' = (b')^m a'$ and then $\alpha = a''$ and $\beta = (a'')^n b''$, where m, and then n, are given by Lemma 5.4 with these two successive pairs. Then, we have $\widetilde{\mathrm{rot}}(\check{\alpha}) < \widetilde{\mathrm{rot}}(\widetilde{\beta^{-1}}) < \widetilde{\mathrm{rot}}(\check{\gamma})$, hence, $\check{\alpha} < \check{\gamma}$, ie $\widehat{\alpha^{-1}}$ dominates $\hat{\gamma}^{-1}$. It follows that $x \leq \hat{\gamma}(x + \varepsilon)$. This shows $\check{\gamma}(x) \geq x$, as desired. □

End of the proof of Proposition 1.10 Suppose that $T = T(a,b)$ is a bad torus, and let T' be a torus disjoint from T. By Sublemma 5.8, T' is good and we may take $T' = T(a', b')$, where $\mathrm{rot}(a') = 0$. Then we have $\mathrm{Fix}(a') \supset \mathscr{A}_T$ by Lemma 5.9. This is also true after replacing a' with a deformation $b'_t a'$, so $\mathrm{Per}(b') \supset \mathscr{A}_T$, or equivalently, $\mathrm{Fix}((b')^{q(b')}) \supset \mathscr{A}_T$. Since this is also true after replacing b' with any deformation $a'_t b'$, we conclude $\mathscr{A}_T \subset P(a', b')$. By Lemma 2.8(1), this means that $\langle a', b' \rangle$ has a finite orbit in S^1. □

5.2 Good tori

In this section, we prove Proposition 1.11: if ρ is path-rigid and nongeometric, then there cannot exist two disjoint good tori which are both not very good. In the course of the proof, we will develop some tools that will be used again in Section 6 for the proof of Theorem 1.1.

To motivate the first step, observe that if ρ has two disjoint good tori $T(a,b)$ and $T(d,e)$ with $\mathrm{rot}(a) = \mathrm{rot}(e) = 0$, and if neither of these tori are very good, then $P(a,b) = P(e,d) = \varnothing$. We can also find c so that (a,b,c,d,e) is a 5–chain. This is the set-up of the next proposition.

Proposition 5.10 *Let ρ be path-rigid minimal and let (a,b,c,d,e) be a 5–chain. Suppose that both $P(a,b)$ and $P(e,d)$ are empty. Then we have $S_k(b,c)$ for some $k \geq 1$.*

Proof After changing orientations of these curves, we may suppose that (a,b,c,d,e) is a directed 5–chain. By Theorem 4.2, it suffices to show that $\mathrm{Per}(b) \cap \mathrm{Per}(c) = \varnothing$. Since $P(a,b) = \varnothing$, Proposition 2.9 says that $\partial N(a,b)$ is finite. Choose a positive one-parameter family $(e_t)_{t\in\mathbb{R}}$, commuting with $\rho(e)$. Since $P(e,d) = \varnothing$, we have $\mathrm{Per}(e_t d) \subset U(e,d)$ for all t, so the sets $\mathrm{Per}(e_t d)$, for varying t, are pairwise disjoint and we can choose t_0 so that $\mathrm{Per}(e_{t_0} d) \cap \partial N(a,b) = \varnothing$. Abusing notation, we now replace d

with $e_{t_0}d$ (we will not further use e). With this change in notation, we now have $\partial N(a,b)\cap P(d,c)=\varnothing$. The remaining step will be a useful tool later in Section 6, so we split it off to a separate statement (Lemma 5.11), proved below. □

Lemma 5.11 *Let ρ be path-rigid, and let (a,b,c,d) be a 4–chain. Suppose that $P(a,b)=\varnothing$ and $\partial N(a,b)\cap P(d,c)=\varnothing$. Then $\mathrm{Per}(b)\cap\mathrm{Per}(c)=\varnothing$.*

Proof Orient the curves so that (a,b^{-1},c,d) is a directed 4–chain. Let a_t and d_t be positive one-parameter families commuting with a and d, respectively. By Lemma 2.17, it suffices to find t and s such that $\mathrm{Per}(a_tb)\cap\mathrm{Per}(d_sc)=\varnothing$.

Let $F_0=\partial N(a,b)\cap\partial N(d,c)$. Since $P(a,b)=\varnothing$, Proposition 2.9 says $\partial N(a,b)$ is finite. Hence, F_0 is finite. Let $F_1=\partial N(a,b)\setminus F_0$ and $F_2=(P(d,c)\cup\partial N(d,c))\setminus F_0$. By construction, the F_i are disjoint closed sets; let $\varepsilon>0$ be smaller than the minimum distance between any two of them. Fix t large, so that (by Lemma 2.8), $\mathrm{Per}(a_tb)$ is contained in the ε–neighborhood of $F_0\cup F_1$, hence disjoint from F_2. Since $F_0\subset N(a,b)$, it is also disjoint from $\mathrm{Per}(a_tb)$, ie $\mathrm{Per}(a_tb)\cap(F_0\cup F_2)=\varnothing$. Now let $\eta>0$ be smaller than the distance between $F_0\cup F_2$ and $\mathrm{Per}(a_tb)$. By Lemma 2.8 again, for s large enough, the set $\mathrm{Per}(d_sc)$ is in the η–neighborhood of $F_0\cup F_2$. Hence, $\mathrm{Per}(a_tb)$ and $\mathrm{Per}(d_sc)$ are disjoint, as desired. □

Our next goal is to propagate $S_k(\cdot,\cdot)$ to other curves. For this, we define two stronger properties.

Definition 5.12 (strengthenings of S_k) Say that two curves a and b satisfy $S_k^+(a,b)$ if they satisfy $S_k(a,b)$ and if additionally $a(\mathrm{Per}(b))\cap\mathrm{Per}(b)=\varnothing$. Say that a and b satisfy $S_k^{++}(a,b)$ if they satisfy both $S_k^+(a,b)$ and $S_k^+(b,a)$.

Property $S_k^+(\cdot,\cdot)$ allows one to move families of periodic points continuously by twist deformations, as described in the following lemma.

Lemma 5.13 *Let a and b be any curves with $i(a,b)=-1$ satisfying $S_k^+(a,b)$. There exists a continuous family a_t commuting with a such that $\mathrm{Per}(a_tb)\cap\mathrm{Per}(a_sb)=\varnothing$ for all $s\neq t$, and $|\mathrm{Per}(a_tb)|=2k$ for all t.*

Since property $S_k(a,b)$ immediately implies that $\mathrm{Per}(b)\subset U(a,b)$, the nontrivial part of this lemma is controlling the cardinality of $\mathrm{Per}(a_tb)$. This requires a special construction of one-parameter family a_t, which is, for once, not a one-parameter group.

Proof With Lemma 4.7, the assumption $a\,\mathrm{Per}(b)\cap\mathrm{Per}(b)=\varnothing$ completely prescribes the cyclic order on the set $\bigcup_n a^n(\mathrm{Per}(b))$; it follows that we may choose a neighborhood V of $\mathrm{Per}(b)$, consisting of $2k$ open intervals, such that $a^n(V)\cap a^m(V)=\varnothing$ for all $n,m\in\mathbb{Z}$. We now construct a continuous family of homeomorphisms a_t commuting with a, supported on $\bigcup_{n\in\mathbb{Z}}a^nV$.

Choose one point in each of the periodic orbits of b; let $x_1, x_2, \dots, x_m$ denote these points. Parametrize S^1 so that, for each x_i, b agrees with a rigid rotation by $p(b)/q(b)$ on a small neighborhood of $b^k(x_i)$ for $k = 0, 1, \dots, q(b)-2$ and so that b maps a neighborhood of $b^{q(b)-1}(x_i)$ to a neighborhood of $x_i = b^{q(b)}(x_i)$ by the map $x \mapsto 2x$ or $x \mapsto \frac{1}{2}x$, in coordinates, depending on whether the orbit of x_i is repelling or attracting.

Let $V_{i,k}$ denote the connected component of V containing $b^k(x_i)$. Define a_t to be the identity on $V_{i,k}$ for $k = 0, 1, \dots, q(b)-2$ and all i. On $V_{i,q(b)-1}$, using the local coordinates in which b is linear, define a_t to agree in a neighborhood of 0 with the translation $x \mapsto x + t$, and extend a_t equivariantly (with respect to a) over S^1. This all can be done continuously in t. After shrinking the $V_{i,k}$ if needed, by construction, each $(a_t b)^{q(b)}$ has a unique fixed point in each $V_{i,k}$, and these vary continuously. Additionally, for t sufficiently small, no new fixed points will be introduced; this proves the lemma. □

The next lemma and proposition allow one to propagate S_k^{++} along chains.

Lemma 5.14 *Let (a, b, c) be a completable 3–chain. Then $S_k^+(a, b)$ implies $S_k(b, c)$.*

Proposition 5.15 *Let (a, b, c) be a completable 3–chain. Suppose that $S_k^{++}(a, b)$ holds. Then $S_k^{++}(b, c)$ holds as well.*

To prove these two statements, we will need a quick sublemma.

Sublemma 5.16 (Per has empty interior) *Let a and b be any curves with $i(a, b) = \pm 1$, and let b_t be a positive one-parameter family commuting with b. Then, for all but countably many t, the set $\mathrm{Per}(b_t a)$ has empty interior.*

Proof Let $X = S^1 \setminus P(b, a)$. Then for $t \neq s$, we have $\mathrm{Per}(b_t a) \cap \mathrm{Per}(b_s a) \cap X = \varnothing$. In particular, the set $T = \{t : \mathrm{Per}(b_t a) \cap X \text{ contains a nonempty open set}\}$ is countable. Also if $U \subset \mathrm{Per}(b_t a)$ is nonempty and open, then $U \cap X = U \setminus P(b, a)$ is nonempty and open since $P(b, a)$ is closed with empty interior, hence $t \in T$. It follows that for all $t \notin T$, $\mathrm{Per}(b_t a)$ has empty interior. □

Proof of Lemma 5.14 Complete (a, b, c) to a 4–chain (a, b, c, d), and let $(d_t)_{t \in \mathbb{R}}$ be a positive one-parameter family commuting with d. By Sublemma 5.16, $\mathrm{Per}(d_{t_0} c)$ has empty interior for some $t_0 \in \mathbb{R}$. Now, by Lemma 5.13, there exists a continuous family $(a_s)_{s \in \mathbb{R}}$, an interval $I \subset \mathbb{R}$ and $2k$ maps, $\phi_j : I \to S^1$, each a homeomorphism to its image, such that for all $s \in I$, the $2k$ periodic points of $\mathrm{Per}(a_s b)$ are precisely $\phi_1(s), \dots, \phi_{2k}(s)$. The set $\bigcap \phi_j^{-1}(\mathrm{Per}(d_{t_0} c))$ then has empty interior in I, hence there exists $s_0 \in I$ such that $\mathrm{Per}(a_{s_0} b) \cap \mathrm{Per}(d_{t_0} c) = \varnothing$, and $\mathrm{Per}(b) \cap \mathrm{Per}(c) = \varnothing$ by Lemma 2.17. We conclude by using Theorem 4.2. □

Proof of Proposition 5.15 Complete the 3–chain into a 5–chain, (e, a, b, c, d), and apply Lemma 5.14 to the 3–chains (a, b, c) and (e, a, b) to conclude $S_k(b, c)$ and $S_k(a, e)$. By Lemma 3.8, we may then use a bending deformation of a along e to move the periodic set of a disjoint from any finite set,

so in particular $\mathrm{Per}(a) \cap \mathrm{Per}(c) = \varnothing$. Let a_t be a positive one-parameter family, commuting with a. Then $\mathrm{Per}(a) \cap \mathrm{Per}(c) = \varnothing$, and $a_{-t}\,\mathrm{Per}(c)$ moves continuously in t, so there exists some t such that $b\,\mathrm{Per}(c) \cap a_{-t}\,\mathrm{Per}(c) = \varnothing$. Thus, $a_t b\,\mathrm{Per}(c) \cap \mathrm{Per}(c) = \varnothing$; hence, by Lemma 2.17, $b\,\mathrm{Per}(c) \cap \mathrm{Per}(c) = \varnothing$. Thus, we conclude that $S_k^+(b,c)$ holds. By Lemma 5.14, this implies that $S_k(c,d)$ holds as well. In particular, $\mathrm{Per}(d)$ is finite. We can now apply Lemma 3.8 and use a bending deformation so that $\mathrm{Per}(a_t b) \cap \mathrm{Per}(d) = \varnothing$, which implies that $\mathrm{Per}(b) \cap \mathrm{Per}(d) = \varnothing$, and repeat the argument above (with d and c playing the roles of a and b) to conclude $S_k^+(c,b)$ holds as well. □

Proposition 5.15, Theorem 3.3, and the connectedness of the graph in Lemma 2.4 immediately gives:

Corollary 5.17 *Let ρ be a path-rigid, minimal representation, and suppose there exists (a,b) such that $S_k^{++}(a,b)$ holds. Then ρ is geometric.*

This consequence is strong enough to imply the main result of the companion article [25]. We explain this now, as it will be used again in Section 6.

Corollary 5.18 *Let ρ be a path-rigid, minimal representation, and suppose that there is some torus $T(a,b)$ such that the relative Euler number of $T(a,b)$ is ± 1. Then ρ is semiconjugate to a Fuchsian representation.*

Proof Since $T(a,b)$ has Euler number 1, it follows from [29] that the restriction of ρ to $\langle a,b\rangle$ is semiconjugate to a geometric representation in $\mathrm{PSL}_2(\mathbb{R})$. (This is not difficult: that $\widetilde{\mathrm{rot}}([\widehat{\rho(a)}, \widehat{\rho(b)}]) = \pm 1$ easily implies that $\rho(a)$ and $\rho(b)$ are 1–Schottky, hence are semiconjugate to a geometric representation in $\mathrm{PSL}_2(\mathbb{R})$. See the beginning of [29, Section 3].) In particular, property $S_1^{++}(a,b)$ holds, and Corollary 5.17 implies that ρ is geometric. □

Given Corollary 5.17, the main goal of this section reduces to the following.

Proposition 5.19 *Let (a,b,c,d,e) be a 5–chain, and suppose that $P(a,b) = P(e,d) = \varnothing$. Then we have $S_k^{++}(b,c)$.*

Proof Suppose $P(a,b) = P(e,d) = \varnothing$. By Proposition 5.10, we have $S_k(b,c)$ and $S_k(c,d)$ for some $k \geq 1$. Since $P(e,d) = \varnothing$ and $\mathrm{Per}(b)$ is finite, we have a bending deformation $e_t d$ such that $\mathrm{Per}(b) \cap \mathrm{Per}(e_t d) = \varnothing$; hence $\mathrm{Per}(b) \cap \mathrm{Per}(d) = \varnothing$. Hence, $\mathrm{Per}(b) \cap d_t c\,\mathrm{Per}(b) = \varnothing$ for some t, so we have $\mathrm{Per}(b) \cap c\,\mathrm{Per}(b) = \varnothing$, ie $S_k^+(c,b)$ holds. By Lemma 5.14, this gives $S_k(a,b)$. In particular, $\mathrm{Per}(a)$ is finite, and so there exists a bending deformation replacing c with $d_t c$ such that $\mathrm{Per}(a) \cap \mathrm{Per}(d_t c) = \varnothing$, and hence $\mathrm{Per}(a) \cap \mathrm{Per}(c) = \varnothing$. Repeating the argument above, we conclude $S_k^+(b,c)$ holds. □

The main result of this section is now a quick corollary. We restate it here for convenience and to summarize our work.

Corollary 5.20 *Let ρ be a path-rigid, minimal representation. Suppose ρ admits two disjoint good tori that are not very good. Then ρ is geometric.*

Proof Let $T(a,b)$ and $T(d,e)$ be two disjoint good tori. Since they are good, we may suppose $\mathrm{rot}(a) = \mathrm{rot}(e) = 0$. Since they are not very good, we have $P(a,b) = \varnothing$ and $P(e,d) = \varnothing$. We may find a curve c such that (a,b,c,d,e) is a 5–chain, and then Proposition 5.19 and Corollary 5.17 imply that ρ is geometric. □

5.3 Finite orbits

The goal of this section is the proof of the following proposition.

Proposition 5.21 *Let $\rho\colon \Gamma_g \to \mathrm{Homeo}^+(S^1)$ be a path-rigid representation, and let $\Sigma = \Sigma_{g-1,1}$ be a subsurface containing only very good tori. Then $\rho_{|\pi_1\Sigma}$ has a finite orbit.*

If $T(a,b)$ is very good, then a and b act with a finite orbit, so $\mathrm{rot}(ab) = \mathrm{rot}(a) + \mathrm{rot}(b)$. Thus, in a subsurface where all tori are very good, rotation number is additive on any pair of curves with intersection number ± 1. This motivates the following proposition, which gives our first step.

Proposition 5.22 *Let Σ be a one-holed surface of genus ≥ 2. Suppose that $\pi_1\Sigma$ acts on the circle in such a way that all nonseparating simple curves have rational rotation number, and that for all γ_1, γ_2 with $i(\gamma_1,\gamma_2) = \pm 1$, we have $\mathrm{rot}(\gamma_1\gamma_2) = \mathrm{rot}(\gamma_1) + \mathrm{rot}(\gamma_2)$.*

Then, there exist two curves γ_1, γ_2 with $i(\gamma_1,\gamma_2) = \pm 1$ and $\mathrm{rot}(\gamma_1) = \mathrm{rot}(\gamma_2) = 0$.

Proof Let $(a_1,\dots,b_g)$ be a standard generating set of $\pi_1\Sigma$, and consider the noncompletable directed 5–chain $(\gamma_1,\gamma_2,\gamma_3,\gamma_4,\gamma_5) = (a_1^{-1}b_1a_1, a_1, \delta_1, a_2, b_2^{-1})$, with the notation of Section 2.1.

Let $r_i = \mathrm{rot}(\gamma_i)$ and let τ_i denote the map on rotation numbers induced by the Dehn twist along γ_i. Then $\tau_i(r_1,r_2,r_3,r_4,r_5) = (r_1',\dots,r_5')$, where $r_{i-1}' = r_{i-1} - r_i$ and $r_{i+1}' = r_{i+1} + r_i$, and $r_j' = r_j$. As Dehn twists preserve chains, the proof of the proposition is reduced to showing that the operations τ_i can be iterated to transform any vector in $(\mathbb{Q}/\mathbb{Z})^5$ to a vector of the form $(0,0,r_3,r_4,r_5)$. This is a straightforward exercise (and should be familiar to anyone familiar with the symplectic group $\mathrm{Sp}(2g,\mathbb{Z})$); we leave the details to the reader. □

Proposition 5.22 is useful because it is much easier to control the dynamics of two curves if their rotation numbers are zero, as in the next proposition.

Proposition 5.23 *Suppose $\mathrm{rot}(a) = \mathrm{rot}(b) = 0$. Then for every $\varepsilon > 0$, there exists a one-parameter family $(a_t)_{t\in\mathbb{R}}$ commuting with a, an interval $J \subset \mathbb{R}$, and a finite collection of homeomorphisms $\phi_i : J \to S^1$ with disjoint images, such that for all $t \in J$,*

$$\mathrm{Fix}(a_t b) \cap \big(S^1 \setminus V_\varepsilon(P(a,b))\big) = \{\phi_1(t),\dots,\phi_n(t)\}.$$

In other words, for all $t \in J$, the fixed points of $a_t b$ at distance $\geq \varepsilon$ to $P(a,b)$ are finite in number and move continuously in t. Compare with Lemma 5.13. Note that we do not require a_t to be a *positive* family.

Proof Fix a positive one-parameter family α_t commuting with a. We will modify α_t to obtain the desired family a_t.

When $\mathrm{rot}(a) = \mathrm{rot}(b) = 0$, we have $P(a,b) = \mathrm{Fix}(b) \cap \partial\mathrm{Fix}(a)$, and the set $U(a,b)$ has a very simple description: $x \in U(a,b)$ if and only if x and $b(x)$ are in the same connected component of $S^1 \setminus \partial\mathrm{Fix}(a)$. Thus, $U(a,b) = \bigcup_I (I \cap b^{-1}(I))$, where I ranges over the connected components of $S^1 \setminus \partial\mathrm{Fix}(a)$. As each connected component I is a–invariant, we may define a_t separately on each connected component, affecting only $\mathrm{Fix}(a_t b) \cap I$.

For every connected component I of $S^1 \setminus \partial\mathrm{Fix}(a)$, let $U(I)$ denote $I \cap b^{-1}(I)$. By definition, each endpoint of $U(I)$ lies in $\partial N(a,b) \cup P(a,b)$. Thus, by Proposition 2.9, all but finitely many intervals $U(I)$ lie in $V_\varepsilon(P(a,b))$. On all the corresponding connected components I of $S^1 \setminus \partial\mathrm{Fix}(a)$ we set $a_t = \alpha_t$.

Now we treat the remaining (finitely many) intervals I of $S^1 \setminus \mathrm{Fix}(a)$ such that $U(I)$ is nonempty, considering the configuration of I and $b^{-1}(I)$. As a first case, suppose that I and $b^{-1}(I)$ share an endpoint, ie a point in $P(a,b)$. If this is the right endpoint, define $a_t = \alpha_t$ on I. If the left endpoint is shared, take instead $a_t = \alpha_{-t}$. If $I = b(I)$, either choice will work. In each case, for all s sufficiently large, we have

$$\mathrm{Fix}(a_s b) \cap I \subset V_\varepsilon(P(a,b)). \tag{5-1}$$

As a second case, suppose b shifts I. If the shift is to the right, ie $I = (x_1, x_3)$ and $b(I) = (x_2, x_4)$ with x_1, x_2, x_3, x_4 in cyclic order, define $a_t = \alpha_t$ on I, and if the shift is to the left, set $a_t = \alpha_{-t}$. In either case, for all s sufficiently large, we have

$$\mathrm{Fix}(a_s b) \cap I = \varnothing. \tag{5-2}$$

We are left with the case where either $b(\bar{I}) \subset I$ or $\bar{I} \subset b(I)$. Suppose the first holds, as the second can be dealt with by a symmetric argument. Note that (using α_t and b) we are in the case $n = 1$ of Lemma 4.8 of the preceding section. Thus, there exists $s \in \mathbb{R}$ such that $\alpha_s b$ has a unique fixed point in I. Moreover, $b(\bar{I}) \subset I$ implies that this unique fixed point is an attracting point, ie we may take local coordinates so that the map $\alpha_s b$ agrees with $x \mapsto \frac{1}{2}x$ at the origin. After reparametrization of α_t on I, we may assume that this time s is sufficiently large to satisfy (5-1) and (5-2) above. Working in coordinates, let $(-\delta, \delta)$ be a neighborhood of 0 contained in a fundamental domain for a. Let τ_t be a smooth family of bump functions supported on $(-\delta, \delta)$ and agreeing with $x \mapsto x + t$ on an even smaller (fixed) neighborhood of 0, for all $t < \delta' < \delta$. Extend τ_t a–equivariantly to a homeomorphism of I. Now define a_t on I to agree with α_t for $t < s$, to agree with $\tau_{t-s}\alpha_s$ for $s \leq t \leq s + \delta'$, and arbitrarily (for example, constant in t) for $t \geq s + \delta'$. Varying t in $J := (s, s + \delta')$, the homeomorphism $a_t b$ has a unique fixed point in I that moves continuously with t, as desired. Of course, we can choose parametrizations of a_t on each of these (finitely many) intervals so that J does not depend on I. This proves the lemma. □

Using this tool, we can propagate finite orbits over chains.

Proposition 5.24 *Let $a, \gamma_1, \gamma_2, \gamma_3, \dots \gamma_k$ be a chain. Suppose that $\mathrm{Per}(a)$ has empty interior, $\mathrm{rot}(\gamma_i) = 0$ for all i, the subgroup $\langle a, \gamma_1\rangle$ has a finite orbit and $\langle \gamma_i, \gamma_{i+1}\rangle$ has a global fixed point. Then $\langle a, \gamma_i, \dots, \gamma_k\rangle$ has a finite orbit.*

Proof Inductively, suppose the statement holds for chains of length k and take a chain of length $k+1$ of the form $a, \gamma_1, \dots, \gamma_k$. By inductive hypothesis the group generated by the first k elements $\langle a, \gamma_1, \dots, \gamma_{k-1}\rangle$ has a finite orbit, ie there is a periodic orbit of a contained in $\bigcap_{i=1}^{k-1} \mathrm{Fix}(\gamma_i)$.

Since $\mathrm{Per}(a)$ has empty interior, for any $n \in \mathbb{N}$, we can use Proposition 5.23 to produce a homeomorphism $c(n)$ lying in a one-parameter family commuting with γ_k such that

$$\mathrm{Fix}(c(n)\gamma_{k-1}) \cap \mathrm{Per}(a) \subset V_{1/n}(P(\gamma_{k-1}, \gamma_k)).$$

Indeed, with the notation of that proposition, there exists $t \in J$ such that $\phi_j(t) \notin \mathrm{Per}(a)$ for all j, because $\bigcap_j \phi_j^{-1}(\mathrm{Per}(a))$ has empty interior in J. Do this for each $n \in \mathbb{N}$; we do not require that the $c(n)$ all belong to a common one-parameter family, all that is important is that they are each obtainable by a bending deformation, hence give a semiconjugate representation.

The result is a sequence of bending deformations $c(n)\gamma_{k-1}$ of γ_{k-1} such that

$$\mathrm{Fix}(c(n)\gamma_{k-1}) \cap \mathrm{Per}(a) \subset V_{1/n}(\mathrm{Fix}(\gamma_{k-1}) \cap \mathrm{Fix}(\gamma_k)).$$

Since $\langle a, \gamma_1, \dots, \gamma_{k-1}\rangle$ has a finite orbit, and this property is stable under semiconjugacy, it follows that, for every n, $\bigcap_{i=1}^{k-2} \mathrm{Fix}(\gamma_i) \cap \mathrm{Fix}(c(n)\gamma_{k-1})$ contains a full orbit of a. For each n, choose one such full orbit, and denote it by $\mathbb{O}_n$. After passing to a subsequence, the sets $\mathbb{O}_n$ converge pointwise to a finite subset of $\bigcap_{i=1}^{k-2} \mathrm{Fix}(\gamma_i) \cap \mathrm{Per}(a)$ that is invariant under a (as these are both closed conditions) so the limit is a full orbit. Moreover, this orbit is contained in every open neighborhood of $\mathrm{Fix}(\gamma_{k-1}) \cap \mathrm{Fix}(\gamma_k)$, so also lies in $\mathrm{Fix}(\gamma_{k-1}) \cap \mathrm{Fix}(\gamma_k)$. This gives a periodic orbit of a in $\bigcap_{i=1}^{k} \mathrm{Fix}(\gamma_i)$, as desired. □

We now prove the main result advertised at the beginning of this section.

Proof of Proposition 5.21 Let $\Sigma_{g-1,1}$ be a surface with one boundary component, in which all tori are very good. Recall that our goal is to show that ρ has a finite orbit. Since all tori are very good, we may use Proposition 5.22 to find a standard system of generators $a_1, b_1, \dots, a_{g-1}, b_{g-1}$ where $\mathrm{rot}(a_i) = \mathrm{rot}(b_i) = 0$ for all $i = 2, 3, \dots, g-1$. Since $T(a_1, b_1)$ is good, we may also assume that $\mathrm{rot}(b_1) = 0$.

Let $\delta_i = a_{i+1}^{-1} b_{i+1} a_{i+1} b_i^{-1}$ as in Section 2.1, so that $(a_1, \delta_1, a_2, \delta_2, \dots, \delta_{g-2}, a_{g-1}, b_{g-1})$ forms a chain. For each i, we can use Sublemma 5.16 in order to assume without loss of generality that $\mathrm{Per}(\delta_i)$ has empty interior, and then apply Proposition 5.24 to the chain (δ_i, a_i, b_i). It follows that $\langle \delta_i, b_i\rangle$ has a finite orbit, hence

$$\mathrm{rot}(\delta_i) + \mathrm{rot}(b_i) = \mathrm{rot}(a_{i+1}^{-1} b_{i+1} a_{i+1}) = \mathrm{rot}(b_{i+1}).$$

Thus, $\mathrm{rot}(\delta_i) = 0$ for all i.

Sublemma 5.16 implies that, after a deformation, we may assume that $\mathrm{Per}(a_1)$ has empty interior. We can apply Proposition 5.24 to the chain $(a_1, \delta_1, a_2, \delta_2, \dots \delta_{g-2}, a_{g-1}, b_{g-1})$ to conclude that the subgroup generated by these elements has a finite orbit. As this subgroup is equal to $\pi_1(\Sigma_{g-1,1})$, this proves the proposition. □

5.4 Proof of Theorem 1.6

Theorem 1.6 is now a quick consequence of Proposition 5.21 and Corollary 5.18.

Proof of Theorem 1.6 Let $\rho\colon \pi_1(\Sigma_g) \to \mathrm{Homeo}_+(S^1)$ be a path-rigid representation, and suppose that ρ is not geometric. If Σ contains a bad torus T, then by Proposition 1.11, $\Sigma \setminus T$ contains only very good tori. If Σ contains no bad torus, but some torus T' that is not very good, then Proposition 1.11 implies that $\Sigma \setminus T'$ contains only very good tori. In either case, there is a genus $g-1$ subsurface $\Sigma_{g-1,1}$ containing only very good tori, hence by Proposition 5.21 the restriction of ρ to $\Sigma_{g-1,1}$ has a finite orbit. In particular, the boundary curve of this subsurface has zero rotation number, and the restriction of ρ to this subsurface has relative Euler number zero.

It follows that the Euler number of the remaining (not very good) torus is either 0 or ± 1. By Corollary 5.18, if it is ± 1, then ρ is geometric. Thus, the remaining torus has Euler number 0, and by additivity the Euler number of ρ is zero. □

The proof of Proposition 5.21 also shows the following, which will be useful to us in the next section of this work.

Corollary 5.25 *Suppose ρ is a path-rigid representation such that Σ has only very good tori. Then ρ has a finite orbit.*

Proof To show this, one simply runs the proof of Proposition 5.21 for a genus g surface (rather than a genus $g-1$ surface with boundary), finding a standard system of generators $a_1, b_1, \dots, a_g, b_g$ and ignoring the extra relation. The remainder of the proof applies verbatim, with g replacing $g-1$. □

6 Proof of Theorem 1.1 and last comments

6.1 Proof of Theorem 1.1

Here is where we use the stronger hypothesis of rigidity. Our proof relies on the following observation, inspired by work in the recent article [1].

Lemma 6.1 *Let ρ be a rigid, minimal representation. Let $T = T(a,b)$ be a very good torus. Then only finitely many points of S^1 have a finite orbit under $\langle a, b\rangle$. In particular, if $\mathrm{rot}(a) = 0$, then $P(a,b)$ is a finite set.*

This lemma is the *only* place where we use rigidity instead of path-rigidity.

Proof Let $F(a,b)$ denote the set of points whose orbit under $\langle a,b\rangle$ is finite. To simplify the exposition of the proof, fix a metric on S^1 so that a and b act on $F(a,b)$ by rigid rotations. Given any $\varepsilon>0$, let $J_1, J_2, \dots$ denote the (finitely many) connected components of $S^1 \setminus F(a,b)$ consisting of intervals of length greater than ε — by our choice of metric, this is a $\langle a,b\rangle$–invariant set. If $F(a,b)$ is finite, and ε small enough, then $\bigcup_i \bar{J}_i = S^1$. Otherwise (even in the case where $\bigcup_i \bar{J}_i = \varnothing$), we may divide $S^1 \setminus \bigcup_i \bar{J}_i$ into finitely many disjoint open intervals $I_1, I_2, \dots$ each of length at most ε and with endpoints in $F(a,b)$, such that these intervals are permuted by $\langle a,b\rangle$, and such that $S^1 = \big(\bigcup_i \bar{J}_i\big) \cup \big(\bigcup_i \bar{I}_i\big)$.

Since T is very good, we can suppose without loss of generality that $\mathrm{rot}(a)=0$. We claim that there exist $a', b' \in \mathrm{Homeo}^+(S^1)$, agreeing with a and b on $S^1 \setminus \bigcup_i I_i$, such that $[a',b'] = [a,b]$ holds globally, and such that $\mathrm{Per}(b') \cap \bigcup I_i = \varnothing$.

Let $c = [a,b]$. As $\bigcup_i I_i$ is a,b–invariant, constructing a' and b' amounts to solving the equation $b'c = a'^{-1}ba'$ on $\bigcup_i J_i$. That this can be solved is shown in [9, Lemma 2.7]; as their notation and context is slightly different, we explain the strategy. Take coordinates identifying each J_i with $\mathbb{R}$. If b' is defined on some J_i (with image in J_j) to increase sufficiently quickly (as a homomorphism $\mathbb{R} \to \mathbb{R}$), then $b'c$ will also be strictly increasing, hence conjugate to b'. One then defines a' to be this conjugacy.

Let ρ' be the representation obtained from ρ by replacing (a,b) by (a',b'). As $\varepsilon>0$ is arbitrary, this ρ' can be taken arbitrarily close to ρ in $\mathrm{Hom}(\Gamma_g, \mathrm{Homeo}^+(S^1))$. Rigidity implies that, for small enough ε, ρ' is semiconjugate to ρ. Minimality implies that there is a *continuous* semiconjugacy $h\colon S^1 \to S^1$ such that $h \circ \rho' = \rho \circ h$. Let

$$F' := \{x \in S^1 \mid x \text{ has finite orbit under } \langle \rho'(a), \rho'(b)\rangle\}.$$

By construction of ρ', this set is finite. However, $h(F') = F(a,b)$. It follows that $F(a,b)$ was finite as well. □

To apply this to the proof of Theorem 1.1, let ρ be a rigid, minimal representation, and assume for contradiction that ρ is nongeometric. If ρ has a bad torus T, then by Theorem 1.6 any torus $T(a,b)$ disjoint from T is very good. In particular, we can take such a torus where $\mathrm{rot}(a)=0$. Lemma 5.9 implies then that $A_T \subset \mathrm{Fix}(a)$. Since the same holds after replacing a with a deformation $b_t a$, we conclude that $A_T \subset P(a,b)$. However, Proposition 5.6 states that A_T is infinite, contradicting Lemma 6.1. We conclude that ρ has no bad tori.

In order to derive a contradiction, we will show that all good tori are actually very good. We pursue this with an argument in the spirit of Proposition 5.10.

Lemma 6.2 *Suppose $P(a,b)=\varnothing$. Then $\partial N(a,b) \subset \partial\mathrm{Per}(a) \cup b^{-1}(\partial\mathrm{Per}(a))$.*

Proof Assume $P(a,b)=\varnothing$ and let $x \in \partial N(a,b)$. Since $P(a,b)=\varnothing$, the set $N(a,b)$ is closed, hence $x \in N(a,b) \cap \overline{U(a,b)}$.

Suppose that $x \notin (\partial\mathrm{Per}(a) \cup b^{-1}(\partial\mathrm{Per}(a))$. Then there exists two intervals I, J, neighborhoods of x, with $I \subset S^1 \setminus \partial\mathrm{Per}(a)$ and $J \subset S^1 \setminus b^{-1}(\partial\mathrm{Per}(a))$. As $x \in \overline{U(a,b)}$, there exists $u \in U(a,b) \cap I \cap J$. Let a_t be a positive one-parameter family commuting with a. Since $b(J)$ contains $b(x)$ and $b(u)$ and $b(J) \cap \partial\mathrm{Per}(a) = \varnothing$, there exists $t_0 \in \mathbb{R}$ such that $a_{t_0} b(x) = b(u)$. Similarly, there exists $t_1 \in \mathbb{R}$ such that $a_{t_1}(u) = x$. Thus, $\Delta_{a,b}(x, t_1 + T(u), T(u), \dots, T(u), T(u) + t_0) = 0$, and it now follows easily that $x \in U(a,b)$. This proves the lemma. □

Lemma 6.3 *Suppose* $\mathrm{rot}(a) = 0$ *and that* $\langle a, b\rangle$ *has no finite orbit. Choose a positive one-parameter group* b_t *that commutes with* b. *Then for all* $x \in S^1$, *there exist at most two values of* t *such that* $x \in \partial N(b_t a, b)$.

Proof Since $\langle a, b\rangle$ has no finite orbit, $P(a,b) = \varnothing$ and hence $P(b_t a, b) = \varnothing$ for all t. Let $x \in S^1$; we will apply Lemma 6.2 to the pairs $(b_t a, b)$. If $x \in \mathrm{Per}(b)$, then $x \notin N(b_t a, b)$, and in particular $x \notin \partial N(b_t a, b)$ for all $t \in \mathbb{R}$. Thus, suppose $x \notin \mathrm{Per}(b)$.

By Lemma 6.2, if $x \in \partial N(b_t a, b)$, then $x \in \partial\mathrm{Per}(b_t a) \cup b^{-1}(\partial\mathrm{Per}(b_t a))$. Note that x cannot be in $P(b,a)$, as $x \notin \mathrm{Per}(b)$. Hence, if there exists some $t \in \mathbb{R}$ such that $x \in \mathrm{Per}(b_t a)$, then $x \in U(b,a)$, and this t is unique. Similarly, if there exists some $t \in \mathbb{R}$ such that $b(x) \in \mathrm{Per}(b_t a)$, then $b(x) \in U(b,a)$, and this t is unique. This concludes the proof. □

Using these tools, we will now show that ρ (always assumed rigid and minimal) satisfies hypothesis S_k.

Lemma 6.4 *Let* (a,b,c,d) *be a* 4*–chain, and suppose* $\mathrm{rot}(a) = \mathrm{rot}(d) = 0$ *holds. Suppose that* $T(a,b)$ *is good but not very good. Then we have* $S_k(b,c)$.

Proof If $T(d,c)$ is good but not very good, then $P(d,c)$ is empty. Otherwise, it is very good and so by Lemma 6.1, the set $P(d,c)$ is finite. In either case, using Lemma 6.3, we can first deform a to some $b_t a$, so that $\partial N(a,b)$ does not intersect $P(d,c)$. Then by Lemma 5.11, we have $\mathrm{Per}(b) \cap \mathrm{Per}(c) = \varnothing$, and so Theorem 4.2 says that $S_k(b,c)$ holds. □

Lemma 6.5 *Let* (a,b,c,d) *be a* 4*–chain, and suppose* $S_k(a,b)$ *and* $\mathrm{rot}(d) = 0$ *hold. Then we have* $S_k(b,c)$.

Proof Similarly to the previous lemma, in this case we may again use Lemma 6.1 to conclude that the set $P(d,c)$ is finite. By Lemma 3.8 in the torus $T(a,b)$, the set $\mathrm{Per}(b)$ is disjoint from $P(d,c)$.

Hence, $\mathrm{Per}(b) \subset U(d,c) \cup N(d,c)$, and $\mathrm{Per}(b)$ is finite. Thus, for all but finitely many t, we have $\mathrm{Per}(b) \cap \mathrm{Per}(d_t c) = \varnothing$. Hence $\mathrm{Per}(b) \cap \mathrm{Per}(c) = \varnothing$ by Lemma 2.17. □

Now we can complete the proof of the theorem.

Proof of Theorem 1.1 Let ρ be a rigid, minimal representation. As remarked above, ρ has no bad torus. If all tori are very good, then by Corollary 5.25, we know that ρ admits a finite orbit, a contradiction.

Thus, ρ admits a good torus, $T(a,b)$, which is not very good. We may suppose $\mathrm{rot}(a)=0$. As all tori are good, we may choose a curve d outside $T(a,b)$ with $\mathrm{rot}(d)=0$, and we may form a 4–chain (a,b,c,d). By Lemma 6.4, we have $S_k(b,c)$ for some k.

Now rename (b,c) into (a,b), and forget about the other curves, remembering only that we have two curves a,b with $S_k(a,b)$. Since all tori are good, we may choose a curve d outside $T(a,b)$ such that $\mathrm{rot}(d)=0$, and such that there exists a standard generating system beginning with (a,b,d,γ). Define $u=\gamma a^{-1}b^{-1}a$ and $v=\gamma a^{-1}$. Then (u,a,b,v), (d,u,a,b) and (a,b,v,d) are 4–chains; we encourage the reader to refer to Figure 1 and draw these curves u and v for him/herself. Apply Lemma 6.5 to the 4–chain (a,b,v,d). This proves that $S_k(b,v)$ holds. The same lemma applied to the 4–chain (d,u,a,b) implies $S_k(u,a)$. Hence, the 4–chain (u,a,b,v) satisfies $S_k(u,a)$, $S_k(a,b)$ and $S_k(b,v)$. We can deform a along u, thanks to Lemma 3.8, in such a way that $\mathrm{Per}(a)\cap\mathrm{Per}(v)=\varnothing$, hence we have $S_k^+(b,a)$, and we can deform b along v, in such a way that $\mathrm{Per}(b)\cap\mathrm{Per}(u)=\varnothing$, hence we have $S_k^+(a,b)$. Finally, this proves $S_k^{++}(a,b)$, and thus ρ is geometric by Corollary 5.17. □

6.2 Comments and further questions

We conclude this paper by discussing some natural questions and directions for further work.

6.2.1 Path-rigidity Given Theorem 1.6, we expect that path-rigidity should suffice to imply that a representation is geometric. The most obvious route to this result would be through an improvement of Lemma 6.1, as it is the only place where we use the stronger hypothesis of rigidity.

Question 6.6 *Does Lemma 6.1 hold when "rigid" is replaced by "path-rigid"?*

This question also arises naturally out of the work of Alonso, Brum and Rivas in [1]. Their main result is the following.

Theorem 6.7 (Alonso–Brum–Rivas) *Let ρ be in $\mathrm{Hom}(\Gamma_g,\mathrm{Homeo}^+(S^1))$ or $\mathrm{Hom}(\Gamma_g,\mathrm{Homeo}^+(\mathbb{R}))$. In any neighborhood U of ρ, there exists a representation ρ' without global fixed points.*

Since it is unknown whether these representation spaces are locally connected, their result does not imply that there is a *path-deformation* of ρ without global fixed points. Thus, the obvious problem arising out of their work is to upgrade this result to path-deformations. A first step in this direction would be to attempt to reprove [1, Lemma 3.9, 3.10]. These lemmas show that, in any neighborhood of ρ, there exists a representation ρ' whose fixed points are isolated and either attracting or repelling points. Can ρ' be attained by deforming along a path? If so, can this be generalized to finite orbits, rather than fixed points, for actions on S^1?

6.2.2 The commutator equation More general than Question 6.6 above, the following basic problem appears to be essential in understanding the topology of $\mathrm{Hom}(\Gamma_g, \mathrm{Homeo}^+(S^1))$.

Probem 6.8 *For fixed $h \in \mathrm{Homeo}^+(S^1)$, describe the topology of the set*

$$\nu_h := \{f, g \in \mathrm{Homeo}^+(S^1) \times \mathrm{Homeo}^+(S^1) \mid [f, g] = h\}.$$

As it stands, remarkably little is known about this space. If $\mathrm{rot}(h) \in \mathbb{Q} \setminus \{0\}$, then it is known that ν_h is not connected; however, we do not know the number of connected components, nor do we know in any circumstances whether ν_h is locally connected or not.

Probem 6.8 is strongly related to the following major problem.

Probem 6.9 *Classify the connected components of $X(\Gamma_g, \mathrm{Homeo}^+(S^1))$.*

As was mentioned in the introduction, it is still unknown whether $X(\Gamma_g, \mathrm{Homeo}^+(S^1))$ (or equivalently, $\mathrm{Hom}(\Gamma_g, \mathrm{Homeo}^+(S^1))$) has finitely many or infinitely many connected components. The relationship with Probem 6.8 comes through the analogy with Goldman's work on $\mathrm{Hom}(\Gamma_g, \mathrm{PSL}_2(\mathbb{R}))$. Indeed, Goldman's classification of connected components of $\mathrm{Hom}(\Gamma_g, \mathrm{PSL}_2(\mathbb{R}))$ given in [15] is built upon a complete understanding of the space $\nu_h \cap (\mathrm{PSL}_2(\mathbb{R}) \times \mathrm{PSL}_2(\mathbb{R}))$. This is of course a much easier problem, as $\mathrm{PSL}_2(\mathbb{R})$ is a finite-dimensional Lie group, and the commutator map is smooth. The result of the first author in [23] — that Euler number does not classify connected components of $\mathrm{Hom}(\Gamma_g, \mathrm{Homeo}^+(S^1))$, unlike the $\mathrm{PSL}_2(\mathbb{R})$ case — may also serve as warning that the topology of ν_h should be more complicated than its intersection with $\mathrm{PSL}_2(\mathbb{R}) \times \mathrm{PSL}_2(\mathbb{R})$.

Throughout this paper, we navigated within ν_h by making bending deformations. This raises a few obvious questions, such as the following.

Question 6.10 *Let $h \in \mathrm{Homeo}^+(S^1)$, and let (f, g) and (f', g') be in the same path-component of ν_h. Identifying f, g with the image of generators of a one-holed torus, is there a path from (f, g) to (f', g') consisting of a sequence of bending deformations? More generally, given ρ and ρ' in the same path-component of $\mathrm{Hom}(\Gamma_g, \mathrm{Homeo}^+(S^1))$, is there a path from ρ to ρ' using bending deformations in simple closed curves on Σ_g?*

This question is reminiscent of Thurston's earthquake theorem [34] for Teichmüller space. It also calls to mind work of Goldman and Xia [16], who use the analogous (positive) result for bending deformations in connected components of classical character varieties in order to studying the action of the mapping class group on these varieties. As well as justifying our use of bending deformations alone, a positive answer to Question 6.10 would give another analogy between classical character varieties and $\chi(\Gamma_g, \mathrm{Homeo}^+(S^1))$.

6.2.3 Bad tori In Section 5, we used a long series of lemmas to prove that a path-rigid representation cannot contain two disjoint bad tori. However, we do not know any example of a path-rigid representation

with even a single bad torus. Besides being an interesting question in itself, the question of existence of bad tori could provide a means of showing path-rigid representations are geometric: if one showed that path-rigid representations of Γ_g have no bad tori, an enhanced version of Lemma 5.11 would complete the proof.

However, we were somewhat surprised to be unable to tackle the following even more basic question.

Question 6.11 *Let $T(a,b)$ be a one-holed torus. Does there exist a representation*

$$\rho\colon \pi_1(T) \to \mathrm{Homeo}^+(S^1)$$

such that the rotation number of every nonseparating simple closed curve is rational, but nonzero?

This is obviously related to understanding mapping class group actions on character varieties, as we are asking for a nonseparating simple closed curve.

By contrast, relaxing the condition that curves be simple gives a problem already solved by a classical result of Antonov. See [31, Exercise 2.3.24] for an outline of the proof. An equivalent statement can be found in [8, Proposition 5.2].

Theorem 6.12 (Antonov [2]) *Let $\rho\colon \langle a,b\rangle \to \mathrm{Homeo}^+(S^1)$ be a minimal action. Either ρ has abelian image and is conjugate to an action by rotations, or — up to taking a quotient of S^1 by a finite-order rotation commuting with ρ — the probability that the rotation number of the image of a random word of length N in $\{a,b,a^{-1},b^{-1}\}$ (with respect to some nondegenerate distribution on the set) is zero tends to 1 as N tends to ∞.*

In the case where ρ commutes with a finite order rotation, say of order n, but does not have image conjugate into $\mathrm{SO}(2)$, the rotation numbers of random words equidistribute in $\{0, 1/n, \dots, (n-1)/n\}$. Thus, for any such action, almost all words have rational rotation number.

6.2.4 Local versus global rigidity Thus far, we have discussed rigidity and path-rigidity of representations; rigidity being the natural notion to study from our interest in character spaces, and path-deformations being easier to work with in practice. However, from a dynamical perspective, it is also interesting to study *local rigidity* or *stability* of actions.

Definition 6.13 ([24, Definition 3.1]; see also [1]) A representation ρ is *locally rigid* if it has a neighborhood *in the representation space* $\mathrm{Hom}(\Gamma_g, \mathrm{Homeo}^+(S^1))$ containing only representations semiconjugate to ρ.

In many circumstances, this condition is much easier to satisfy than rigidity or path-rigidity. For example, a savage element $g \in \mathrm{Homeo}^+(S^1)$ (as in Definition 3.4 above), thought of as a representation of $\mathbb{Z}$, is easily seen to be locally rigid, but it is semiconjugate to the identity. We do not know if this phenomenon generalizes to representations of Γ_g.

Question 6.14 *Is there a representation $\rho \in \mathrm{Hom}(\Gamma_g, \mathrm{Homeo}^+(S^1))$ that is locally rigid, but not rigid?*

Again, a natural first step to this question could be to study the local topology of the sets v_h defined above.

References

[1] **J Alonso**, **J Brum**, **C Rivas**, *Orderings and flexibility of some subgroups of* $\mathrm{Homeo}_+(\mathbb{R})$, J. Lond. Math. Soc. 95 (2017) 919–941 MR Zbl

[2] **V A Antonov**, *Modeling of processes of cyclic evolution type*: *synchronization by a random signal*, Vestnik Leningrad. Univ. Mat. Mekh. Astronom. (1984) 67–76 MR Zbl In Russian

[3] **G Baumslag**, *On generalised free products*, Math. Z. 78 (1962) 423–438 MR Zbl

[4] **E Breuillard**, **T Gelander**, **J Souto**, **P Storm**, *Dense embeddings of surface groups*, Geom. Topol. 10 (2006) 1373–1389 MR Zbl

[5] **M Burger**, **A Iozzi**, **A Wienhard**, *Higher Teichmüller spaces: from* $\mathrm{SL}(2,\mathbb{R})$ *to other Lie groups*, from "Handbook of Teichmüller theory, IV" (A Papadopoulos, editor), IRMA Lect. Math. Theor. Phys. 19, Eur. Math. Soc., Zürich (2014) 539–618 MR Zbl

[6] **D Calegari**, **A Walker**, *Ziggurats and rotation numbers*, J. Mod. Dyn. 5 (2011) 711–746 MR Zbl

[7] **A Casson**, **D Jungreis**, *Convergence groups and Seifert fibered* 3*–manifolds*, Invent. Math. 118 (1994) 441–456 MR Zbl

[8] **B Deroin**, **V Kleptsyn**, **A Navas**, *Sur la dynamique unidimensionnelle en régularité intermédiaire*, Acta Math. 199 (2007) 199–262 MR Zbl

[9] **D Eisenbud**, **U Hirsch**, **W Neumann**, *Transverse foliations of Seifert bundles and self-homeomorphism of the circle*, Comment. Math. Helv. 56 (1981) 638–660 MR Zbl

[10] **D Gabai**, *Convergence groups are Fuchsian groups*, Ann. of Math. 136 (1992) 447–510 MR Zbl

[11] **É Ghys**, *Classe d'Euler et minimal exceptionnel*, Topology 26 (1987) 93–105 MR Zbl

[12] **É Ghys**, *Groupes d'homéomorphismes du cercle et cohomologie bornée*, from "The Lefschetz centennial conference, III" (A Verjovsky, editor), Contemp. Math. 58, Amer. Math. Soc., Providence, RI (1987) 81–106 MR Zbl

[13] **É Ghys**, *Groups acting on the circle*, Enseign. Math. 47 (2001) 329–407 MR Zbl

[14] **W M Goldman**, *The symplectic nature of fundamental groups of surfaces*, Adv. Math. 54 (1984) 200–225 MR Zbl

[15] **W M Goldman**, *Topological components of spaces of representations*, Invent. Math. 93 (1988) 557–607 MR Zbl

[16] **W M Goldman**, **E Z Xia**, *Ergodicity of mapping class group actions on* $\mathrm{SU}(2)$*–character varieties*, from "Geometry, rigidity, and group actions" (B Farb, D Fisher, editors), Univ. Chicago Press (2011) 591–608 MR Zbl

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

[18] **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, MA (1987) 48–106 MR Zbl

[19] **A Katok**, **B Hasselblatt**, *Introduction to the modern theory of dynamical systems*, Encycl. Math. Appl. 54, Cambridge Univ. Press (1995) MR Zbl

[20] **S-h Kim**, **T Koberda**, **M Mj**, *Flexibility of group actions on the circle*, Lecture Notes in Math. 2231, Springer (2019) MR Zbl

[21] **D Luna**, *Sur certaines opérations différentiables des groupes de Lie*, Amer. J. Math. 97 (1975) 172–181 MR Zbl

[22] **D Luna**, *Fonctions différentiables invariantes sous l'opération d'un groupe réductif*, Ann. Inst. Fourier (Grenoble) 26 (1976) 33–49 MR Zbl

[23] **K Mann**, *Spaces of surface group representations*, Invent. Math. 201 (2015) 669–710 MR Zbl

[24] **K Mann**, *Rigidity and flexibility of group actions on the circle*, from "Handbook of group actions, IV", Adv. Lect. Math. 41, International, Somerville, MA (2018) 705–752 MR Zbl

[25] **K Mann**, **M Wolff**, *A characterization of Fuchsian actions by topological rigidity*, Pacific J. Math. 302 (2019) 181–200 MR Zbl

[26] **J Marché**, **M Wolff**, *The modular action on* $\mathrm{PSL}_2(\mathbb{R})$*–characters in genus* 2, Duke Math. J. 165 (2016) 371–412 MR Zbl

[27] **S Matsumoto**, *Numerical invariants for semiconjugacy of homeomorphisms of the circle*, Proc. Amer. Math. Soc. 98 (1986) 163–168 MR Zbl

[28] **S Matsumoto**, *Some remarks on foliated* S^1 *bundles*, Invent. Math. 90 (1987) 343–358 MR Zbl

[29] **S Matsumoto**, *Basic partitions and combinations of group actions on the circle: a new approach to a theorem of Kathryn Mann*, Enseign. Math. 62 (2016) 15–47 MR Zbl

[30] **J Milnor**, *On the existence of a connection with curvature zero*, Comment. Math. Helv. 32 (1958) 215–223 MR Zbl

[31] **A Navas**, *Groups of circle diffeomorphisms*, Univ. Chicago Press (2011) MR Zbl

[32] **S Smale**, *Differentiable dynamical systems*, Bull. Amer. Math. Soc. 73 (1967) 747–817 MR Zbl

[33] **D Sullivan**, *Quasiconformal homeomorphisms and dynamics, II: Structural stability implies hyperbolicity for Kleinian groups*, Acta Math. 155 (1985) 243–260 MR Zbl

[34] **W P Thurston**, *Earthquakes in two-dimensional hyperbolic geometry*, from "Low-dimensional topology and Kleinian groups" (D B A Epstein, editor), Lond. Math. Soc. Lect. Note Ser. 112, Cambridge Univ. Press (1986) 91–112 MR Zbl

[35] **P Tukia**, *Homeomorphic conjugates of Fuchsian groups*, J. Reine Angew. Math. 391 (1988) 1–54 MR Zbl

[36] **J W Wood**, *Bundles with totally disconnected structure group*, Comment. Math. Helv. 46 (1971) 257–273 MR Zbl

Department of Mathematics, Cornell University
Ithaca, NY, United States
Sorbonne Universités, UPMC Univ. Paris 06, Institut de Mathématiques de Jussieu-Paris Rive Gauche, UMR 7586, CNRS, Univ. Paris Diderot, Sorbonne Paris Cité
Paris, France
k.mann@cornell.edu, maxime.wolff@math.univ-toulouse.fr

Proposed: David Fisher
Seconded: Leonid Polterovich, Dan Abramovich
Received: 20 September 2022
Revised: 28 November 2022

msp

Geometry & Topology **28**:5 (2024) 2399–2482
DOI: 10.2140/gt.2024.28.2399
Published: 24 August 2024

Embedding surfaces in 4–manifolds

DANIEL KASPROWSKI
MARK POWELL
ARUNIMA RAY
PETER TEICHNER

We prove a surface embedding theorem for 4–manifolds with good fundamental group in the presence of dual spheres, with no restriction on the normal bundles. The new obstruction is a Kervaire–Milnor invariant for surfaces and we give a combinatorial formula for its computation. For this we introduce the notion of band characteristic surfaces.

57K40, 57N35

1 Introduction

We study whether a given map of a surface to a topological 4–manifold is homotopic to an embedding. Here and throughout the article, embeddings and immersions in the topological category are by definition *locally flat*, meaning they are locally modelled on linear inclusions $\mathbb{R}^2 \hookrightarrow \mathbb{R}^4$ or $\mathbb{R}^2_+ \hookrightarrow \mathbb{R}^4$.

Even for maps of 2–spheres, this question has only been completely addressed in a handful of simple manifolds, such as S^4, $\mathbb{C}\mathbb{P}^2$ [Tristram 1969, page 264] and $S^2 \times S^2$ [Tristram 1969, Theorem 4.5; Kervaire and Milnor 1961, Corollary 1; [Freedman 1982, Corollary 1.1]]. Lee and Wilczyński [1990; 1997] and

Hambleton and Kreck [1993] described the minimal genus of an embedded surface in a fixed homology class, in any given simply connected, closed 4–manifold, assuming that the fundamental group of the complement is abelian. This was recently extended by [Feller et al. 2021] to knot traces. In the relative setting even the simplest case is open: which knots in S^3 bound a (locally flat) embedded disc in D^4?

The main available tool for proving positive results is Freedman's embedding theorem (Theorem 4.3), which shows that maps of discs and spheres to a 4–manifold with *good* fundamental group, with vanishing intersection and self-intersection numbers, and with framed algebraically dual spheres, are regularly homotopic to embeddings [Freedman 1982; Freedman and Quinn 1990, Corollary 5.1B]; see also [Powell et al. 2020; Behrens et al. 2021]. Surgery and the s–cobordism theorem for topological 4–manifolds with good fundamental group are consequences of this theorem [Quinn 1982; Freedman and Quinn 1990]. Our aim, realised in Theorems 1.2 and 1.6 below, is to extend Freedman's theorem to all compact surfaces with algebraically dual spheres, in any connected 4–manifold with good fundamental group. In Section 1.4, we explain how to apply this to the question from the opening paragraph of whether a given homotopy class contains an embedding. In Section 1.5, we give some applications to knot theory. In particular, we show that every knot bounds an embedded surface of genus one in $M \setminus \mathring{D}^4$ for every simply connected closed 4–manifold M not homeomorphic to S^4. Recall that, for $M = S^4$, this does not hold because the slice genera of knots can be arbitrary large.

Throughout the paper, we will work in the following setting unless otherwise specified.

Convention 1.1 We assume that M is a connected, topological 4–manifold and that Σ is a nonempty compact surface with connected components $\{\Sigma_i\}_{i=1}^m$. The notation $F = \{f_i\}_{i=1}^m : (\Sigma, \partial\Sigma) \looparrowright (M, \partial M)$ represents a generic immersion (Definition 2.3) with components $f_i : (\Sigma_i, \partial\Sigma_i) \looparrowright (M, \partial M)$.

By assumption, the map F restricts to an embedding on $\partial\Sigma$ and $F^{-1}(\partial M) = \partial\Sigma$, where $\partial\Sigma$ and ∂M are permitted to be empty. There is no requirement for Σ or M to be orientable, and M could be noncompact. Weakening the hypotheses of Freedman's theorem to allow for the algebraically dual spheres to be unframed introduces an additional obstruction, the *Kervaire–Milnor invariant* $\mathrm{km}(F) \in \mathbb{Z}/2$ (Definition 1.4), which vanishes in the presence of framed algebraically dual spheres.

Theorem 1.2 (Surface embedding theorem) *Let $F = \{f_i\}_{i=1}^m : (\Sigma, \partial\Sigma) \looparrowright (M, \partial M)$ be as in Convention 1.1. Suppose that $\pi_1(M)$ is good and that F has algebraically dual spheres $G = \{[g_i]\}_{i=1}^m \subseteq \pi_2(M)$. Then the following statements are equivalent*:

(i) *The intersection numbers $\lambda(f_i, f_j)$ for all $i < j$, the self-intersection numbers $\mu(f_i)$ for all i, and the Kervaire–Milnor invariant $\mathrm{km}(F) \in \mathbb{Z}/2$ all vanish.*

(ii) *There is an embedding $\overline{F} = \{\bar{f}_i\}_{i=1}^m : (\Sigma, \partial\Sigma) \hookrightarrow (M, \partial M)$, regularly homotopic to F relative to $\partial\Sigma$, with geometrically dual spheres $\overline{G} = \{\bar{g}_i : S^2 \looparrowright M\}_{i=1}^m$ such that $[\bar{g}_i] = [g_i] \in \pi_2(M)$ for all i.*

Equivariant intersection and self-intersection numbers of immersed discs and spheres have a long history (see eg [Wall 1970]). In the theorem above, we consider generalised versions for arbitrary compact surfaces,

lying in quotients of the group ring $\mathbb{Z}[\pi_1(M)]$, which we denote by $\Gamma_{f_i,f_j} \ni \lambda(f_i, f_j)$ for the intersection numbers and $\Gamma_{f_i} \ni \mu(f_i)$ for the self-intersection numbers. We describe these quotients in detail in Sections 2.2 and 2.3. As in the simply connected case, these invariants require based maps (Definition 2.11) but their vanishing as in Theorem 1.2(i) does not depend on the choice of basing. Vanishing of all the $\lambda(f_i, f_j)$ for $i < j$ and all the $\mu(f_i)$ is equivalent to the vanishing of the self-intersection number $\mu(F)$, which is defined as follows.

Definition 1.3 Let $F = \{f_i\}_{i=1}^m : (\Sigma, \partial\Sigma) \looparrowright (M, \partial M)$ be as in Convention 1.1. Assume in addition that M and Σ are based and that F is a based map. The *self-intersection number* of this possibly disconnected immersed surface is given by counting all double points of F, as follows:

$$\mu(F) := \sum_{i<j} \lambda(f_i, f_j) + \sum_i \mu(f_i) \in \bigoplus_{i<j} \Gamma_{f_i,f_j} \oplus \bigoplus_i \Gamma_{f_i}.$$

The self-intersection number $\mu(F)$ is a regular homotopy invariant that vanishes if and only if there is a collection of Whitney discs that pair all double points of F (Corollary 2.30), just like for connected surfaces. The Whitney discs may be chosen to form a *convenient* collection, meaning that all Whitney discs have disjointly embedded boundaries, are framed and have interiors transverse to F (Definition 2.31).

Definition 1.4 Let $F\colon (\Sigma, \partial\Sigma) \looparrowright (M, \partial M)$ be as in Convention 1.1. By definition, $\mathrm{km}(F) \in \mathbb{Z}/2$ vanishes if and only if, after finitely many finger moves taking F to some F', there is a convenient collection of Whitney discs pairing all the double points of F' and whose interiors are disjoint from F'.

We think of $\mu(F)$ as the primary embedding obstruction, and $\mathrm{km}(F)$ as the secondary embedding obstruction. Note that $\mathrm{km}(F) = 0$ implies $\mu(F) = 0$ but $\mathrm{km}(F)$ is always defined even if $\mu(F) \neq 0$. In Section 1.1, we will give a combinatorial description of $\mathrm{km}(F)$ in the case that $\mu(F) = 0$.

The Kervaire–Milnor invariant is named in homage to [Kervaire and Milnor 1961], in which the authors defined an embedding obstruction and used it to give the first proof that the Whitney trick fails in dimension 4. Section 3 gives details on the connection of our Kervaire–Milnor invariant to the original obstruction and other secondary embedding obstructions in the literature.

A group is said to be *good* if it satisfies the π_1*–null disc property* [Freedman and Teichner 1995] (see also [Kim et al. 2021]); we shall not repeat the definition. In practice, it suffices to know that virtually solvable groups and groups of subexponential growth are good, and that the class of good groups is closed under taking subgroups, quotients, extensions and colimits [Freedman and Teichner 1995; Krushkal and Quinn 2000].

If Σ is a union of discs or spheres, Theorem 1.2 follows from [Freedman and Quinn 1990, Theorem 10.5(1)]. The latter theorem contained an error found by Stong [1994] (see Theorem 5.7), but this is not relevant to Theorem 1.2 because of the way we defined the Kervaire–Milnor invariant. It is, however, very relevant to how to compute the Kervaire–Milnor invariant, and Stong's correction will be one of the ingredients in our results (see Section 1.1).

For an arbitrary Σ, one could try to prove Theorem 1.2 by using general position to embed the 0– and 1–handles of Σ (relative to $\partial\Sigma$) and removing a small open neighbourhood thereof from M. This gives a new connected 4–manifold M_0 with the same fundamental group as M, and only the 2–handles $\{h_i\colon (D^2, S^1) \looparrowright (M_0, \partial M_0)\}$, one for each component Σ_i of Σ, remain to be embedded. One then hopes to apply [Freedman and Quinn 1990, Theorem 10.5(1)] (ie Theorem 1.2 for Σ a union of discs) to these maps of 2–handles to produce the desired embedded surface. The original algebraically dual spheres $\{g_i\}$ for the $\{f_i\}$ perform the same role for the $\{h_i\}$ in M_0. The intersection and self-intersection numbers λ and μ remain unchanged; hence, they also vanish for $\{h_i\}$. However, the Kervaire–Milnor invariant may behave differently. That is, it may become nonzero for the embedding problem for the discs $\{h_i\}$, whereas it was trivial for the original F. We show that this phenomenon can occur in Example 9.3. This difference stems from the fact that, in applying [Freedman and Quinn 1990, Theorem 10.5(1)], we fix an embedding of the 1–skeleton of Σ and try to extend it across the 2–handles. As usual in obstruction theory, it might be advantageous to go back and change the solution of the problem on the 1–skeleton.

1.1 Computing the Kervaire–Milnor invariant

The strength of Theorem 1.2 versus the above strategy using [Freedman and Quinn 1990, Theorem 10.5(1)] lies in our computation of the Kervaire–Milnor invariant for arbitrary compact surfaces. In the case of discs and spheres, Stong showed that the Kervaire–Milnor invariant vanishes in more situations than claimed by Freedman and Quinn, due to the ambiguity arising from sheet choices when pairing up double points by Whitney discs, when the associated fundamental group element has order two. As we recall in Theorem 5.7, Stong [1994] introduced the notion of an *r–characteristic* surface, short for $\mathbb{RP}^2$*–characteristic* surface (Definition 5.5), to give a criterion, in terms of copies of $\mathbb{RP}^2$ immersed in the ambient manifold M, to decide whether the sheet changing move is viable. Combined with the work of Freedman and Quinn, this enabled the computation of the Kervaire–Milnor invariant, and therefore answered the embedding problem for every finite union of discs or spheres with algebraically dual spheres, in an ambient 4–manifold with good fundamental group (see Remark 5.8 for more details).

In order to compute the Kervaire–Milnor invariant $\mathrm{km}(F)$ for general surfaces, we extend the notion of r–characteristic surfaces to a notion of *b–characteristic* surfaces, short for *band characteristic* (Definition 5.17), defined using bands (annuli and Möbius bands) immersed in M. The next theorem is a generalisation of Stong's computation of $\mathrm{km}(F)$ to arbitrary compact surfaces.

Definition 1.5 Let $F = \{f_i\}_{i=1}^m\colon (\Sigma, \partial\Sigma) \looparrowright (M, \partial M)$ be as in Convention 1.1 with $\mu(F) = 0$. Choose a convenient collection $\mathcal{W} = \{W_l\}$ of Whitney discs that pair all double points of F and define

$$t(F, \mathcal{W}) := \sum_{l,i} |\mathrm{Int}\, W_l \pitchfork f_i| \mod 2.$$

In other words, $t(F, \mathcal{W})$ is the mod 2 count of transverse intersections between F and the interiors of the Whitney discs in $\mathcal{W}$.

We will often apply this definition to the restriction F^ω of $F = \{f_1, \dots, f_m\}$ to the subsurface $\Sigma^\omega \subseteq \Sigma$, which includes a component Σ_i of Σ precisely if its image does not admit a framed immersion $g_i : S^2 \looparrowright M$ with $\lambda(f_j, g_i) = \delta_{ij}$ for all $j = 1, \dots, m$ (Definition 5.1). The main result of the article is as follows.

Theorem 1.6 *Let $F\colon (\Sigma, \partial\Sigma) \looparrowright (M, \partial M)$ be as in Convention 1.1. Suppose that $\mu(F) = 0$ and that F has algebraically dual spheres. If F^ω is not b–characteristic then* $\mathrm{km}(F) = 0$. *If F^ω is b–characteristic, then the secondary embedding obstruction satisfies*

$$\mathrm{km}(F) = t(F^\omega, \mathcal{W}^\omega) \in \mathbb{Z}/2$$

for every convenient collection of Whitney discs $\mathcal{W}^\omega$ pairing all the double points of F^ω.

The main novelty in this theorem lies in finding the right condition on F that makes the combinatorial formula $t(F^\omega, \mathcal{W}^\omega)$ independent of the choice of Whitney discs, namely that F^ω is b–characteristic. Note that, if $\pi_1(M)$ is good, then, for $\mathrm{km}(F) = 0$ in the previous theorem, Theorem 1.2 gives an embedding regularly homotopic to F. In practice, it can often be easy to determine if a given surface is b–characteristic, as demonstrated by the following corollaries, derived in Section 9 as consequences of the more general Proposition 9.1.

Corollary 1.7 *If M is a simply connected 4–manifold and Σ is a connected, oriented surface with positive genus, then any generic immersion $F\colon (\Sigma, \partial\Sigma) \looparrowright (M, \partial M)$ with vanishing self-intersection number is not b–characteristic. Thus, if F has an algebraically dual sphere, then* $\mathrm{km}(F) = 0$, *and, since $\pi_1(M)$ is good, the map F is regularly homotopic, relative to $\partial\Sigma$, to an embedding.*

This corollary in particular implies that, for every simply connected 4–manifold M with boundary a disjoint union of homology spheres, every primitive class in $H_2(M; \mathbb{Z})$ can be represented by an embedded torus. This recovers [Lee and Wilczyński 1997, Theorem 1.1] in the case of divisibility $d = 1$. We also have the following extension to the case of arbitrary 4–manifolds.

Corollary 1.8 *Let $F\colon (\Sigma, \partial\Sigma) \looparrowright (M, \partial M)$ be as in Convention 1.1, with $\mu(F) = 0$ and Σ connected. If F' is obtained from F by an ambient connected sum with an embedding $S^1 \times S^1 \hookrightarrow S^4$, then F' is not b–characteristic. Thus, if F has an algebraically dual sphere, then* $\mathrm{km}(F') = 0$, *and if $\pi_1(M)$ is good, then F' is regularly homotopic, relative to $\partial\Sigma$, to an embedding.*

See Corollaries 1.13 and 1.14 for the nonorientable analogues of these two results. In particular, the latter concerns the case where we replace the embedded torus in Corollary 1.8 by an embedded $\mathbb{RP}^2$.

1.2 Band characteristic maps

We briefly explain how the notion of a map being b–characteristic arises in the context of embedding general surfaces. Given $F\colon \Sigma \looparrowright M$ as in Convention 1.1, assume that its double points are paired by a convenient collection of Whitney discs $\mathcal{W}$. Then the interiors of the discs in $\mathcal{W}$ could be tubed into spheres

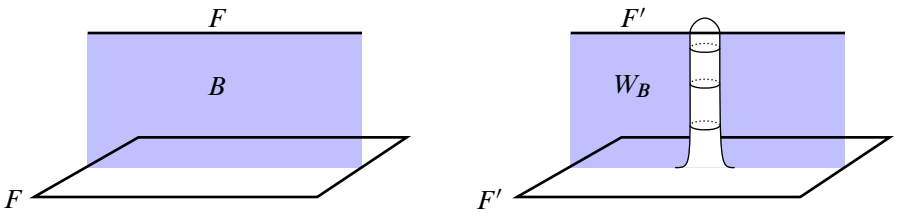


Figure 1: Two portions of the immersion F and part of a band B are shown on the left. A finger move produces F' with two new double points, paired by W_B.

in M, potentially changing the count t from Theorem 1.6. The condition that F is *s–characteristic*, short for *spherically characteristic* (Definition 5.2), precisely ensures that the count is preserved under this move.

Similarly, consider a band, ie an annulus or Möbius band, immersed in M with boundary lying on $F(\Sigma)$ minus the double points, as in Figure 1. Then, as shown in the figure, we may perform a finger move on F along a fibre of the band, creating F' with two new intersections, paired by a new Whitney disc W_B arising from the band B (see Figure 1). We call this move the *band fibre finger move* and give further details in Construction 7.2. Adding W_B to $\mathcal{W}$ might in principle change the count t, but the requirement that F^ω is b–characteristic ensures it does not. In the case that Σ has only simply connected components, the boundary of the band is null-homotopic in Σ, and therefore the band can be closed off by discs to produce either a sphere (from an annulus) or an $\mathbb{RP}^2$ (from a Möbius band). Thus in this case it suffices to consider r–characteristic maps.

However, for general Σ there may exist a band in M with a boundary curve that is nontrivial in $\pi_1(\Sigma)$. This necessitates the new notion of b–characteristic maps, which by definition requires that a function $\Theta\colon \mathcal{B}(F) \to \mathbb{Z}/2$ vanishes (Definitions 5.12 and 5.14), where $\mathcal{B}(F)$ consists of the homology classes in $H_2(M, \Sigma; \mathbb{Z}/2)$ that can be represented by certain immersed bands in M with boundary on Σ (Definition 5.9). These additional conditions on the bands have to do with the first Stiefel–Whitney classes of M and Σ; when both are oriented, $\mathcal{B}(F)$ consists precisely of the classes in $H_2(M, \Sigma; \mathbb{Z}/2)$ that are represented by maps of annuli and Möbius bands. Roughly speaking, the vanishing of Θ means that every band with boundary on Σ intersects Σ evenly many times in its interior. Intersections among the boundary components of the bands and a relative Euler number also play a role; see Sections 5 and 3.7 for details. If $\Theta \equiv 0$, then, for every band B, adding W_B does not change the t–count, and in fact t is well defined if and only if F is b–characteristic (Lemma 7.3). See Example 9.3 for a map which is r–characteristic but not b–characteristic.

The first step for deciding whether F is b–characteristic is to determine the subset $\mathcal{B}(F)$. In general this could be difficult, but in practice it is often soluble. If this can be done, then, as shown in Figure 2, by computing $\lambda_\Sigma|_{\partial\mathcal{B}(F)}$ and $\Theta\colon \mathcal{B}(F) \to \mathbb{Z}/2$, we can decide whether F is b–characteristic. Both of these are functions on a finite group, so in principle these computations are manageable.

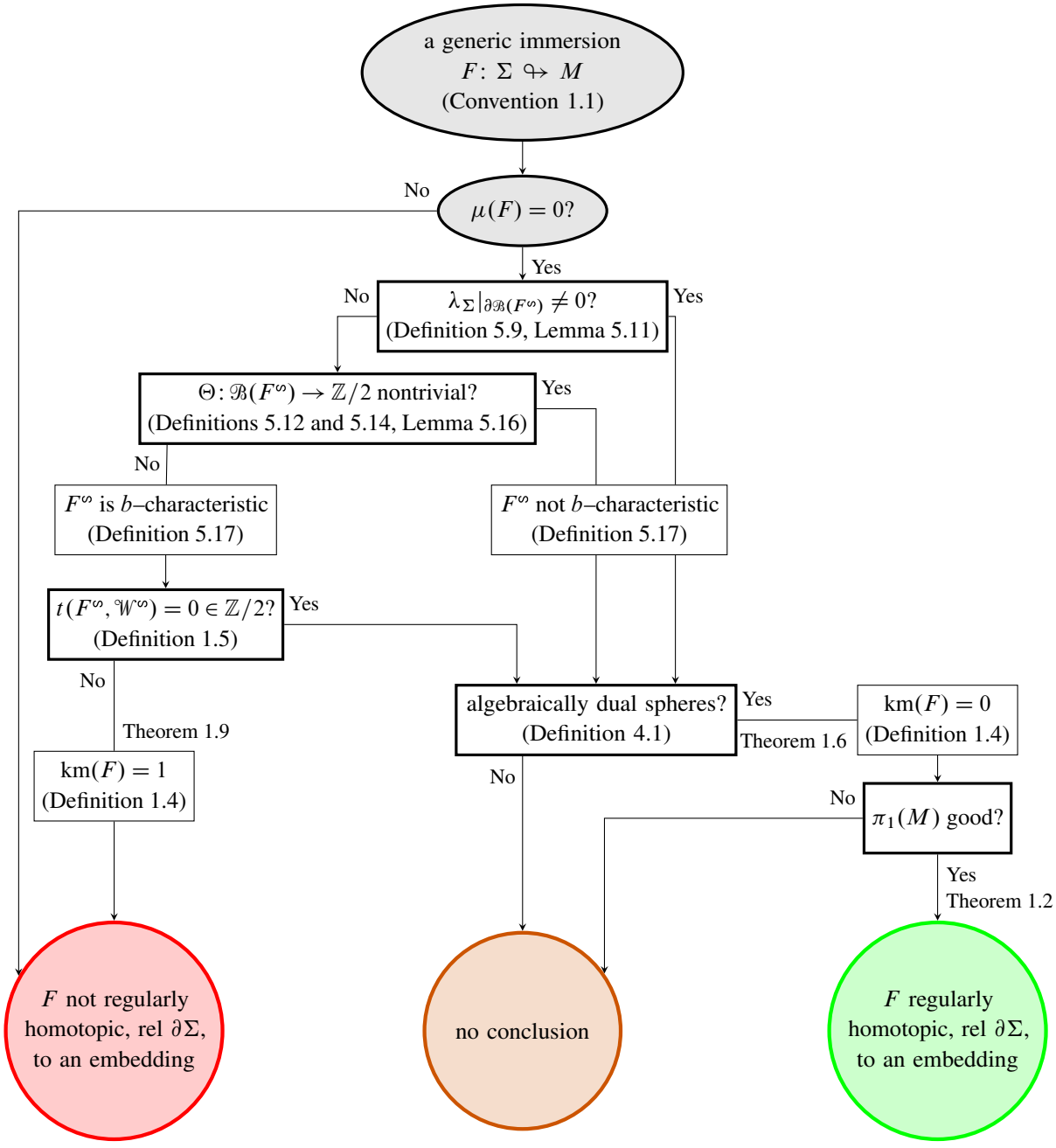


Figure 2: A flowchart deciding whether a generic immersion F is regularly homotopic, relative to the boundary, to an embedding.

1.3 An embedding obstruction without dual spheres

Irrespective of whether F has algebraically dual spheres, we obtain a secondary embedding obstruction in the b–characteristic case.

Theorem 1.9 *Let $F\colon (\Sigma,\partial\Sigma)\looparrowright(M,\partial M)$ be as in Convention 1.1 with $\mu(F)=0$. Let $\mathcal{W}$ be a convenient collection of Whitney discs for the double points of F. Then F is b–characteristic if and only if, for every F' regularly homotopic to F and convenient collection $\mathcal{W}'$ for the double points of F', we have $t(F,\mathcal{W})=t(F',\mathcal{W}')$.*

For b–characteristic F, we denote the resulting regular homotopy invariant by $t(F)\in\mathbb{Z}/2$. Then, if $\mathrm{km}(F)=0$ — for instance, if F is an embedding — then $t(F)=0$.

Note that, in particular, if F is b–characteristic and a map H is regularly homotopic to F, then H is b–characteristic (Lemma 5.19). If F is not b–characteristic, it is still possible that some restriction F' of F to a union of connected components $\Sigma'\subseteq\Sigma$ is b–characteristic. Then we obtain an obstruction to embedding F', and as a consequence to embedding F. A frequent example of this phenomenon is $F'=F^{\omega}$ from Theorem 1.6. Note that, by Lemma 5.3, if F is b–characteristic then $F=F^{\omega}$.

As part of our analysis of the obstruction t, in Section 9 we shall prove the following additivity properties.

Proposition 1.10 *Let M_1 and M_2 be oriented 4–manifolds. Let $F_1\colon(\Sigma_1,\partial\Sigma_1)\looparrowright(M_1,\partial M_1)$ and $F_2\colon(\Sigma_2,\partial\Sigma_2)\looparrowright(M_2,\partial M_2)$ be generic immersions of connected, compact, oriented surfaces, each with vanishing self-intersection number. If F_i is b–characteristic for each i, then both the disjoint union*

$$F_1\sqcup F_2\colon\Sigma_1\sqcup\Sigma_2\looparrowright M_1\#M_2$$

and any interior connected sum

$$F_1\#F_2\colon\Sigma_1\#\Sigma_2\looparrowright M_1\#M_2$$

are b–characteristic, and satisfy

$$t(F_1\sqcup F_2)=t(F_1\#F_2)=t(F_1)+t(F_2).$$

Theorem 1.9 and Proposition 1.10 imply the following corollary.

Corollary 1.11 *For any g, there exists a smooth, closed 4–manifold M_g, a closed, connected, oriented surface Σ_g of genus g, and a smooth, b–characteristic, generic immersion $F\colon\Sigma_g\looparrowright M_g$ with $t(F)\neq0$, and therefore $\mathrm{km}(F)\neq0$.*

By contrast, we show in Example 9.5 that every map of a closed surface to $S^1\times S^3$ is homotopic to an embedding. One could ask whether there exists a 4–manifold M and immersions $\Sigma_g\looparrowright M$ with nontrivial Kervaire–Milnor invariant, for every g. However, as a partial negative answer we will show in Proposition 9.7 that a b–characteristic generic immersion $F\colon\Sigma\looparrowright M$ from a closed surface Σ to a compact 4–manifold M with abelian fundamental group with n generators must satisfy $\chi(\Sigma)\geq-2n$.

1.4 Homotopy versus regular homotopy

Theorems 1.2, 1.9 and 1.6 together give a framework for deciding whether or not an immersed surface is regularly homotopic to an embedding, as illustrated by the flowchart in Figure 2. However, in the first

sentence of the article, we began by considering whether a given continuous map is homotopic to an embedding. We explain now how to extend the framework of Figure 2 to decide this, for maps of surfaces that admit algebraically dual spheres.

For a map f from a connected surface to a 4–manifold, we will show in Theorem 2.32 that there are either infinitely many or precisely two regular homotopy classes in the homotopy class of f, according to whether $f^*(w_1(M))$ is trivial or nontrivial, respectively. Our strategy is to make a judicious choice of regular homotopy class to which we apply our previous theory.

Begin with a continuous map $F\colon \Sigma \to M$ that restricts to an embedding on $\partial\Sigma$ and satisfies $F^{-1}(\partial M) = \partial\Sigma$, where Σ and M are as in Convention 1.1. Denote the components of F by $f_i\colon (\Sigma_i, \partial\Sigma_i) \to (M, \partial M)$, and suppose that F has algebraically dual spheres. Note that homotopies preserve the intersection numbers $\lambda(f_i, f_j)$, but might not preserve the self-intersection number $\mu(f_i)$, since adding a local cusp in f_i changes $\mu(f_i)_1$, the coefficient of $1 \in \pi_1(M)$, by ± 1. Depending on the behaviour of the orientation characters of M and Σ, the coefficient $\mu(f_i)_1$ lies in either $\mathbb{Z}$ or $\mathbb{Z}/2$, and is preserved under regular homotopy (see Lemma 2.24 and Proposition 2.25).

Now, in order to decide whether F is homotopic to an embedding, we will either find a generic immersion in the homotopy class of F which is regularly homotopic to an embedding, or show that this is impossible. First, by performing a homotopy we may assume without loss of generality that F is a generic immersion such that $\mu(f_i)_1 = 0$ for every component f_i of F. If $\mu(F) \neq 0$, then F is not homotopic to an embedding. On the other hand, if $\mu(F) = 0$, we have the two following cases. Below, $(f_i)_\bullet$ is the map induced on fundamental groups by f_i using some choice of path connecting $f_i(\Sigma_i)$ to the basepoint of M.

Case 1 $w_1(\Sigma)|_{\ker(f_i)_\bullet}$ is trivial for every $f_i \in F^\omega$.

By Theorem 2.32, the regular homotopy class of F^ω is uniquely determined by the condition that $\mu(f_i)_1 = 0$ for each i with $f_i \in F^\omega$. Run the analysis in Figure 2 on F to determine whether it is regularly homotopic to an embedding. Note that the outcome of this analysis only depends on the regular homotopy class of F^ω, rather than all of F. In particular, if $\pi_1(M)$ is good, then F is homotopic to an embedding if and only if F is regularly homotopic to an embedding.

Case 2 There exists $f_i \in F^\omega$ with $w_1(\Sigma)|_{\ker(f_i)_\bullet}$ nontrivial.

In this case we use the following theorem, which we prove in Section 6.

Theorem 1.12 *Let $F = \{f_i\}_{i=1}^m\colon (\Sigma, \partial\Sigma) \looparrowright (M, \partial M)$ be as in Convention 1.1 with $\mu(F) = 0$. Suppose that there is at least one $f_i \in F^\omega$ with $w_1(\Sigma)|_{\ker(f_i)_\bullet}$ nontrivial. Then there exists a generic immersion F' homotopic to F with $\mu(F') = 0$, and a convenient collection of Whitney discs $\mathcal{W}'$ such that $t((F')^\omega, (\mathcal{W}')^\omega) = 0$. Thus, if F' has algebraically dual spheres, then $\mathrm{km}(F') = 0$, and if moreover $\pi_1(M)$ is good, then F' is regularly homotopic, relative to $\partial\Sigma$, to an embedding.*

Using this theorem, we can immediately conclude that our F as in Case 2 is homotopic to an embedding, as long as $\pi_1(M)$ is good. Notably, it is not relevant in this case whether F^ω is b–characteristic. This completes the analysis of whether a given continuous map of a surface into a 4–manifold is homotopic to an embedding.

We now sketch the proof of Theorem 1.12. By the vanishing of $\mu(F)$, there is a convenient collection of Whitney discs $\mathcal{W}$ for F and therefore for F^ω. When $t(F^\omega, \mathcal{W}^\omega) = 0$, the proof is completed by setting $F' = F$. When $t(F^\omega, \mathcal{W}^\omega) = 1$, we use Construction 6.1 to find another generic immersion F' homotopic to F. Briefly, Construction 6.1 involves creating four new double points in the component f_i with nontrivial $w_1(\Sigma)|_{\ker(f_i)_\bullet}$ using local cusps, and then cancelling them using a suitable choice of Whitney arcs and discs. For further details on the proof, see Section 6.

Theorem 1.12 also has the following immediate corollaries. These are the nonorientable analogues of Corollaries 1.7 and 1.8. They provide homotopies to embeddings rather than regular homotopies, and again it is not relevant whether F^ω is b–characteristic.

Corollary 1.13 *If M is a simply connected 4–manifold and Σ is a connected, nonorientable surface, then a generic immersion $F\colon (\Sigma, \partial\Sigma) \looparrowright (M, \partial M)$ with vanishing self-intersection number and an algebraically dual sphere is homotopic, relative to $\partial\Sigma$, to an embedding.*

Corollary 1.14 *Let $F\colon (\Sigma, \partial\Sigma) \looparrowright (M, \partial M)$ be as in Convention 1.1, with Σ connected and $\pi_1(M)$ good. Suppose that F has vanishing self-intersection number and an algebraically dual sphere. If F' is obtained from F by an ambient connected sum with any embedding $\mathbb{RP}^2 \hookrightarrow S^4$, then F' is homotopic, relative to $\partial\Sigma$, to an embedding.*

As in our analysis for the embedding problem up to regular homotopy, our techniques are primarily applicable in the presence of algebraically dual spheres and good fundamental group of the ambient space. It is however sometimes possible to conclude that a map without algebraically dual spheres is homotopic to an embedding. For example, we show in Example 9.5 that every map of a closed surface to $S^1 \times S^3$ is homotopic to an embedding.

1.5 Applications to knot theory

Theorem 1.2 can be applied to the problem of finding embedded surfaces in general 4–manifolds bounded by knots in their boundary. Given a closed 4–manifold M, let M° denote the punctured manifold $M \setminus \mathring{D}^4$. The *$M$–genus* of a knot $K \subseteq S^3 = \partial M^\circ$, denoted by $g_M(K)$, is the minimal genus of an embedded orientable surface bounding K in M°. If M is smooth, we also consider the *smooth M–genus*, denoted by $g_M^{\mathrm{Diff}}(K)$, the minimal genus of a smoothly embedded orientable surface with boundary K. The quantities g_{S^4} and $g_{S^4}^{\mathrm{Diff}}$ coincide with the topological and smooth slice genus of knots in S^3, respectively. Note that $g_{\overline{M}}(K) = g_M(\overline{K})$, so (2) and (3) below imply the corresponding results for $\overline{\mathbb{CP}^2}$ and $*\overline{\mathbb{CP}^2}$, respectively.

Corollary 1.15 *For every knot $K \subseteq S^3$,*

(1) $g_M(K) = 0$ *for every simply connected 4–manifold M not homeomorphic to one of S^4, $\mathbb{CP}^2$ or $*\mathbb{CP}^2$;*

(2) $g_{\mathbb{CP}^2}(K) \leq 1$ *and* $g_{\mathbb{CP}^2}\big(\#^3 T(2,3)\big) = 1$; *and*

(3) $g_{*\mathbb{CP}^2}(K) \leq 1$ *and* $g_{*\mathbb{CP}^2}\big(\#^2 T(2,3)\big) = 1$.

See Section 9 for the proof. The smooth $\mathbb{CP}^2$–genus has been studied by [Yasuhara 1991; 1992; Aït Nouh 2009; Pichelmeyer 2020; Marengon et al. 2024], and differs dramatically from the topological result in Corollary 1.15(2); in particular, it can be arbitrarily high [Marengon et al. 2024].

Corollary 1.15(1) is straightforward to prove if M topologically splits as a connected sum with $S^2 \times S^2$ or $S^2 \tilde{\times} S^2$, because $g_{S^2\times S^2}(K) = g_{S^2\tilde{\times}S^2}(K) = 0$ for all K by the Norman trick [1969, Corollary 3, Remark]. For the $K3$ surface, this implies that $g_{K3}(K) = 0$ for all knots K. On the other hand, it is an open question whether there exists a K with $g^{\mathrm{Diff}}_{K3}(K) \neq 0$ [Manolescu et al. 2024, Question 6.1].

Given a knot $K \subseteq S^3$ and an integer $n \in \mathbb{Z}$, we build the *n–trace* $X_n(K)$ by attaching a 2–handle D^4 along K with framing n. The minimal genus of an embedded surface representing a generator of $H_2(X_n(K);\mathbb{Z})$ is called the *n–shake genus* of K, denoted by $g^{\mathrm{sh}}_n(K)$. Similarly, the *smooth n–shake genus* of K is denoted by $g^{\mathrm{sh,Diff}}_n(K)$. We recover the following result of [Feller et al. 2021].

Corollary 1.16 [Feller et al. 2021, Proposition 8.7] *For any knot $K \subseteq S^3$, $g^{\mathrm{sh}}_{\pm1}(K) = \mathrm{Arf}(K) \in \{0,1\}$.*

By contrast, the smooth ±1–shake genus of a knot can be arbitrarily high. For example, for $q \geq 5$ we have that $g^{\mathrm{sh,Diff}}_{\pm1}(T(2,q)) \geq \frac{1}{2}(q+1)$, by the slice-Bennequin inequality [Lisca and Matić 1998; Cochran and Ray 2016, Corollary 5.2].

Outline of the paper

In Section 2, we describe the primary embedding obstructions arising from the theory of equivariant intersection numbers for surfaces in 4–manifolds. In Section 3, we define the Kervaire–Milnor invariant carefully. We prove Theorem 1.2 in Section 4. In Section 5, we explain our combinatorial method for computing the Kervaire–Milnor invariant, and define b–characteristic surfaces, postponing almost all proofs to Section 7. In Section 6, we give the proof of Theorem 1.12. In Section 8, we prove Theorems 1.6 and 1.9. Finally, in Section 9, we prove Corollaries 1.7, 1.8, 1.11, 1.15 and 1.16 and Proposition 1.10, and we give further applications and examples.

Acknowledgements

We are grateful to Rob Schneiderman and the referee for several insightful comments on a previous version, and to Allison N Miller and Andrew Nicas for helpful conversations. Much of this research was conducted at the Max Planck Institute for Mathematics. Kasprowski was supported by the Deutsche

Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy — GZ 2047/1, Projekt-ID 390685813. Powell was partially supported by EPSRC New Investigator grant EP/T028335/2 and EPSRC New Horizons grant EP/V04821X/2.

2 Generic immersions and intersection numbers

In Section 2.1, we carefully define and study generic immersions of surfaces in 4–manifolds in the topological category. We show they admit well-behaved normal bundles, and introduce generic homotopies and ambient isotopies between them.

In Sections 2.2 and 2.3, we study equivariant intersection and self-intersection numbers of generically immersed surfaces. In the case of immersions of spheres and discs, these have a long history, in particular in surgery theory (see eg [Wall 1970]). For the first time in the literature, as far as we are aware, we give a careful account of intersection and self-intersection numbers in full generality for compact surfaces and for any possible combination of orientation characters. The specific groups in which these numbers live depend on the input surfaces, and it is somewhat subtle to describe them. A preliminary version for orientable surfaces was considered in eg [Cochran et al. 2003, Section 7], and the self-intersection number for annuli was considered in [Schneiderman 2003].

In Section 2.4, we discuss Whitney discs, which arise if the intersection and self-intersection numbers vanish. We define the important notion of a convenient collection of Whitney discs. In Section 2.5, Theorem 2.32 explains the difference between homotopy and regular homotopy of generic immersions of surfaces, in terms of the Euler number of the normal bundle or the self-intersection number.

2.1 Topological generic immersions

We start with the definition of an immersion of manifolds in the topological setting. For $m \geq 0$, let $\mathbb{R}^m_+ := \{(x_1, \dots, x_m) \in \mathbb{R}^m \mid x_1 \geq 0\}$. For $k \leq n$, we consider the standard inclusions

$$\begin{aligned} \iota &: \mathbb{R}^k = \mathbb{R}^k \times \{0\} \hookrightarrow \mathbb{R}^k \times \mathbb{R}^{n-k} = \mathbb{R}^n, \\ \iota_+ &: \mathbb{R}^k_+ = \mathbb{R}^k_+ \times \{0\} \hookrightarrow \mathbb{R}^k \times \mathbb{R}^{n-k} = \mathbb{R}^n, \\ \iota_{++} &: \mathbb{R}^k_+ = \mathbb{R}^k_+ \times \{0\} \hookrightarrow \mathbb{R}^k_+ \times \mathbb{R}^{n-k} = \mathbb{R}^n_+. \end{aligned}$$

Definition 2.1 A continuous map $F\colon \Sigma^k \to M^n$ between topological manifolds of dimensions $k \leq n$ is an *immersion* if locally it is a flat embedding, that is, if, for each point $p \in \Sigma$, there is a chart φ around p and a chart Ψ around $F(p)$ fitting into one of the commutative diagrams

(2-1)
$$\begin{array}{ccc} \mathbb{R}^k & \xrightarrow{\iota} & \mathbb{R}^n \\ \downarrow{\scriptstyle\varphi} & & \downarrow{\scriptstyle\Psi} \\ \Sigma & \xrightarrow{F} & M \end{array} \qquad \begin{array}{ccc} \mathbb{R}^k_+ & \xrightarrow{\iota_+} & \mathbb{R}^n \\ \downarrow{\scriptstyle\varphi} & & \downarrow{\scriptstyle\Psi} \\ \Sigma & \xrightarrow{F} & M \end{array} \qquad \begin{array}{ccc} \mathbb{R}^k_+ & \xrightarrow{\iota_{++}} & \mathbb{R}^n_+ \\ \downarrow{\scriptstyle\varphi} & & \downarrow{\scriptstyle\Psi} \\ \Sigma & \xrightarrow{F} & M \end{array}$$

The first diagram is for $p \in \operatorname{Int}\Sigma$ and $F(p) \in \operatorname{Int} M$, the second diagram is for $p \in \partial\Sigma$ and $F(p) \in \operatorname{Int} M$, and the third is for $p \in \partial\Sigma$ and $F(p) \in \partial M$. In particular, F is required to map interior points of Σ to interior points of M.

Some authors prefer to call this notion a *locally flat immersion*.

Definition 2.2 A (linear) *normal bundle* for an immersion $F\colon \Sigma^k \to M^n$ is an $(n-k)$–dimensional real vector bundle $\pi\colon \nu_F \to \Sigma$, together with an immersion $\widetilde{F}\colon \nu_F \to M$ that restricts to F on the zero section s_0, ie $\widetilde{F} \circ s_0 = F$, and such that each point $p \in \Sigma$ has a neighbourhood U such that $\widetilde{F}|_{\pi^{-1}(U)}$ is an embedding.

We now restrict to the relevant dimensions for this paper, $k = 2$ and $n = 4$, and take M to be a connected topological 4–manifold as in Convention 1.1. The *singular set* of an immersion $F\colon \Sigma \to M$ is the set

$$\mathcal{S}(F) := \{m \in M : |F^{-1}(m)| \geq 2\}.$$

Recall that a continuous map is said to be *proper* if the inverse image of every compact set in the codomain is compact.

Definition 2.3 Let Σ be a surface, possibly noncompact. A continuous, proper map $F\colon \Sigma \to M$ is said to be a (topological) *generic immersion*, denoted by $F\colon \Sigma \looparrowright M$, if it is an immersion and the singular set is a closed, discrete subset of M consisting only of transverse double points, each of whose preimages lies in the interior of Σ. In particular, whenever $m \in \mathcal{S}(F)$, there are exactly two points $p_1, p_2 \in \Sigma$ with $F(p_i) = m$, and there are disjoint charts φ_i around p_i for $i = 1, 2$, where φ_1 is as in the leftmost diagram of (2-1) and φ_2 is the same, with respect to the same chart Ψ around m, but with ι replaced by

$$\iota'\colon \mathbb{R}^2 = \{0\} \times \mathbb{R}^2 \hookrightarrow \mathbb{R}^2 \times \mathbb{R}^2 = \mathbb{R}^4.$$

Theorem 2.4 *A generic immersion $F\colon \Sigma \looparrowright M$, for possibly noncompact Σ, has a normal bundle as in Definition 2.2 with the additional property that $\widetilde{F}$ is an embedding outside a neighbourhood of $F^{-1}(\mathcal{S}(F))$, and near the double points $\widetilde{F}$ plumbs two coordinate regions $\pi^{-1}(\varphi_i(\mathbb{R}^2)) \cong \varphi_i(\mathbb{R}^2) \times \mathbb{R}^2$ for $i = 1, 2$ together, ie $\widetilde{F} \circ (\varphi_1(x), y) = \widetilde{F} \circ (\varphi_2(y), x)$.*

Proof Let $\partial_1\Sigma \subseteq \partial\Sigma$ denote the union of the components of $\partial\Sigma$ mapped to ∂M. Then, since $F|_{\partial_1\Sigma}$ is an embedding of a 1–manifold in a 3–manifold, it has a normal bundle. We extend this to a collar neighbourhood of $F(\partial_1\Sigma)$ contained in a collar neighbourhood of ∂M. To do this, first note that, since ∂M is closed in M, $\mathcal{S}(F)$ is closed and contained in $\operatorname{Int}(M)$, and manifolds are normal, it follows that there is an open neighbourhood of ∂M disjoint from $\mathcal{S}(F)$. Then argue as in Connelly's proof [1971] that boundaries of manifolds have collars, to obtain a homeomorphism of pairs

$$G\colon (M, F(\Sigma)) \xrightarrow{\cong} \big(M \cup (\partial M \times [0,1]),\, F(\Sigma) \cup (F(\partial_1\Sigma) \times [0,1])\big).$$

The normal bundle over the boundary extends to a collar in the codomain; hence, its pullback extends to a collar in the domain.

Next, let $\partial_2\Sigma \subseteq \partial\Sigma$ denote the union of the components of $\partial\Sigma$ mapped to $\operatorname{Int} M$. We see that $F(\partial_2\Sigma)$ has a normal bundle by [Freedman and Quinn 1990, Theorem 9.3], as a submanifold of M. Let $\widetilde{F}$ be the embedding of the total space, as in Definition 2.2. By using the inward-pointing normal for $\partial_2\Sigma$ in Σ, we obtain an orthogonal decomposition of each fibre as $\nu_{\partial_2\Sigma\hookrightarrow\Sigma} \oplus V$, where V is a 2–dimensional subspace. Then translates of V in the direction of the inward-pointing normal give rise to a normal bundle on the intersection of a collar of $\partial_2\Sigma$ with the image of the normal bundle of $\partial_2\Sigma$ under $\widetilde{F}$.

Now we want to extend the normal bundle that we have just constructed on a neighbourhood of $\partial\Sigma$ to the rest of Σ. First we will produce a normal bundle in a neighbourhood of both preimages of each double point, and then finally we will extend the normal bundle to the rest of the interior of Σ.

Let $m \in \mathcal{S}(F)$ be a double point of F, so that there exist $p_1, p_2 \in \Sigma$ with $F(p_i) = m$ for $i = 1, 2$. By the definition of a generic immersion, there is a chart Ψ for M at m, and charts φ_i around p_i, such that $F\circ\varphi_1(x) = \Psi(x, 0)$ and $F\circ\varphi_2(y) = \Psi(0, y)$. We assume that $F(\Sigma)\cap\Psi(\mathbb{R}^4) = F(\varphi_1(\mathbb{R}^2))\cup F(\varphi_2(\mathbb{R}^2))$, and moreover that the images of the charts for different elements of $\mathcal{S}(F)$ do not overlap one another, and also are disjoint from the images of the normal bundles already constructed close to $\partial\Sigma$. Then we take a trivial $\mathbb{R}^2$–bundle over each $\varphi_i(\mathbb{R}^2)$, and we define the map $\widetilde{F}$ on $\varphi_1(\mathbb{R}^2)\times\mathbb{R}^2$ and $\varphi_2(\mathbb{R}^2)\times\mathbb{R}^2$ by setting $\widetilde{F}(\varphi_1(x), y) = \Psi(x, y)$ and $\widetilde{F}(\varphi_2(x), y) = \Psi(y, x)$. Then $\widetilde{F}\circ(\varphi_2(y), x) = \Psi(x, y) = \widetilde{F}\circ(\varphi_1(x), y)$, as needed.

Let U_1^m and U_2^m be open neighbourhoods in Σ of p_1 and p_2, contained within the images of φ_1 and φ_2 above, respectively. Define $\Sigma' := \Sigma\setminus\bigcup_{m\in\mathcal{S}(F)}(U_1^m\cup U_2^m)$. Then the restriction of F gives an embedding of Σ' in M. We already have a normal bundle defined on a neighbourhood of $\partial\Sigma'$. Apply [Freedman and Quinn 1990, Theorem 9.3A] to extend the given normal bundle on $\overline{U_1^m\cup U_2^m}$ and $\partial\Sigma$ to all of Σ' and therefore we have a normal bundle on all of Σ. □

Remark 2.5 Freedman and Quinn [1990, Theorem 9.3] produce an *extendable* normal bundle for every submanifold of a 4–manifold. The extendibility condition is technical with an important consequence: extendable normal bundles are unique up to isotopy. One can always find an extendable normal bundle embedded in the total space of any given normal bundle.

The proof of Theorem 2.4 also applies in the more general setting where $\partial_2\Sigma$ is not embedded in M, but $F|_{\partial_2\Sigma}$ factors as a composition of generic immersions $\partial_2\Sigma \looparrowright S \looparrowright M$ for some surface S. We will use this case in the definition of b–characteristic maps in Section 5, so we introduce nomenclature.

Definition 2.6 Let $g\colon S \looparrowright M$ be a generic immersion of a surface in a 4–manifold M. Let (B, Z) be a pair consisting of a surface B and a collection $Z \subseteq \partial B$ of connected components of its boundary. A map $H\colon B \to M$ is called a *generic immersion of pairs* if $H(Z) \subseteq g(S)$ and

(i) $H|_{B\setminus Z}$ is a generic immersion that is transverse to g and has image disjoint from $H(Z)$;

(ii) $H(B)$ is disjoint from the double points of g, which implies there is a unique map $h\colon Z \to S$ with $g\circ h = H|_Z$;

(iii) the map h is a generic immersion; and

(iv) there is a collar N of Z in B with $H(N \setminus Z) \subseteq M \setminus g(S)$.

We denote such maps by $H\colon (B, Z) \looparrowright (M, S)$, and sometimes identify h with $H|_Z$.

Corollary 2.7 *Let $g\colon S \looparrowright M$ be a generic immersion of a surface in a 4–manifold M and let $H\colon (B, Z) \looparrowright (M, S)$ be a generic immersion of pairs. Then H admits a normal bundle, ie a normal bundle for B in M such that the restriction to Z contains a normal bundle for Z in S.*

Proof Note that Z has a normal bundle in S, and then the sum of this with the normal bundle of S in M guaranteed by Theorem 2.4 gives rise to a normal bundle for Z in M. The rest of the proof proceeds as before. □

Observe that smooth generic immersions are topological immersions. Next we show that when both notions make sense they coincide, which justifies the terminology.

Theorem 2.8 *Consider a smooth compact surface Σ and a (topological) generic immersion $F\colon \Sigma \looparrowright M$. If M is noncompact then let $M' := M$, and if M is compact then choose $p \in M \setminus F(\Sigma)$ and set $M' := M \setminus \{p\}$. Then F is a smooth generic immersion in some smooth structure on M'.*

We know that M' has a smooth structure by [Freedman and Quinn 1990, Theorem 8.2; Quinn 1982, Corollary 2.2.3].

Proof Fix a smooth structure on ∂M such that the generic immersion F restricted to those connected components of $\partial\Sigma$ that map to ∂M is a smooth embedding. To find such a smooth structure, first use the standard smooth structure on the normal bundle of $F|_{\partial\Sigma}$, and then extend this to a smooth structure on all of ∂M. Since any two smooth structures on a 3–manifold are isotopic, this could also be arranged by an isotopy of $\partial\Sigma$, but our aim is to use the given map without isotoping it.

By Theorem 2.4, there is a normal bundle $(\nu_F, \widetilde{F})$ for F. Let $D(\nu_F) \to \Sigma$ be the (closed) disc bundle. This yields a regular neighbourhood $N(F) := \widetilde{F}(D(\nu_F))$ of $F(\Sigma)$, a codimension zero submanifold of M'. The regular neighbourhood $N(F)$ can be identified with a smooth manifold obtained from $D(\nu_F)$ after the requisite plumbing operations and smoothing corners. Use such an identification to fix a smooth structure on $N(F) \subseteq M$. With respect to this smooth structure, the map $\Sigma \to N(F)$ is a smooth generic immersion.

Now, the boundary of $N(F)$ inherits a smooth structure. The complement of $\operatorname{Int} N(F) \cup (N(F) \cap \partial M')$ in M' is a connected, noncompact 4–manifold with a prescribed smooth structure on its boundary. Then the interior has a compatible smooth structure by [Freedman and Quinn 1990, Theorem 8.2; Quinn 1982, Corollary 2.2.3], giving rise to a smooth structure on all of M'. Since the smooth structure on $N(F)$ is unaltered, F become a smooth generic immersion, as desired. □

Recall that an *isotopy* of homeomorphisms of a manifold M is a map $H\colon M\times[0,1]\to M$ such that the track $M\times[0,1]\to M\times[0,1]$ given by $(m,t)\mapsto(H(m,t),t)$ is a homeomorphism.

Definition 2.9 An *ambient isotopy* between generic immersions $F,G\colon \Sigma \looparrowright M$ consists of two isotopies $H_\Sigma\colon \Sigma\times[0,1]\to\Sigma$ and $H_M\colon M\times[0,1]\to M$ such that

(1) $H_\Sigma(-,0)$ and $H_M(-,0)$ are both the identity; and

(2) $G(x)=H_M\big(F(H_\Sigma(x,1)),1\big)$ for all $x\in\Sigma$.

This is motivated by the smooth result which states that two generic immersions are ambiently isotopic (in the sense of Definition 2.9 but with homeomorphism replaced by diffeomorphism in the definition of an isotopy) if and only if they are connected by a path in the space of generic immersions [Golubitsky and Guillemin 1973, Chapter III, Theorem 3.11]. Note that for embeddings one does not need the isotopy H_Σ.

Mirroring the smooth notion, a *generic homotopy* between generically immersed surfaces in a 4–manifold is by definition a sequence of ambient isotopies, finger moves, Whitney moves and cusp homotopies. The moves in question are defined in local coordinates exactly as in the smooth setting. A *regular homotopy* between generically immersed surfaces in a 4–manifold is by definition a sequence of ambient isotopies, finger moves and Whitney moves. The following proposition explains that maps of surfaces in a 4–manifold can be assumed to be generic immersions, and homotopies between generic immersions may be assumed to be generic as well.

Proposition 2.10 [Powell et al. 2020, Proposition 3.1] *Let Σ be a compact surface and let M be a topological 4–manifold.*

(1) *Every map $(\Sigma,\partial\Sigma)\to(M,\partial M)$ is homotopic (relative to the embedded boundary) to a generic immersion.*

(2) *Every homotopy $(\Sigma,\partial\Sigma)\times[0,1]\to(M,\partial M)$ that restricts to a generic immersion on $\Sigma\times\{0,1\}$ is homotopic (relative to the boundary) to a generic homotopy.*

Briefly, the proposition is proven as follows. Homotope the maps away from a point of M using cellular approximation, remove that point, choose a smooth structure on the complement of the point, and then apply the smooth theory of generic immersions, combining [Hirsch 1976, Theorems 2.2.6 and 2.2.12] with [Golubitsky and Guillemin 1973, Chapter III, Corollary 3.3].

2.2 Intersection numbers

We define intersection numbers between compact, connected surfaces in 4–manifolds. In order to accommodate the fundamental group in equivariant intersection numbers, we need to use basings.

Definition 2.11 We call a manifold X *based* if it is equipped with basepoints $p_i \in X_i$ for each connected component $X_i \subseteq X$, together with a local orientation at each p_i. A generic immersion $F\colon X \to Y$ between based manifolds with Y connected is said to be *based* if it is equipped with *whiskers*, ie paths in Y from the basepoint of Y to $F(p_i)$ for each basepoint p_i of X.

For the remainder of this section, let M be a connected, based 4–manifold and let Σ and Σ' be based, compact, connected surfaces, unless specified otherwise.

Let $f\colon \Sigma \to M$ and $g\colon \Sigma' \to M$ be based maps that are *transverse*, ie around each intersection point $f(s) = g(s')$ with $s \in \Sigma$ and $s' \in \Sigma'$, there are coordinates that make f and g (in a neighbourhood of s in Σ and a neighbourhood of s' in Σ') resemble the standard inclusions $\mathbb{R}^2 \times \{0\}$ and $\{0\} \times \mathbb{R}^2$ into $\mathbb{R}^4$, respectively, as in (2-1). We assume that these intersections are the only singularities between f and g and that $f(\partial\Sigma)$ and $g(\partial\Sigma')$ are disjoint.

Let v_f and v_g be whiskers for f and g. The intersection number $\lambda(f, g)$ is the sum of signed fundamental group elements

$$\lambda(f, g) := \sum_{p \in f \pitchfork g} \varepsilon(p) \cdot \eta(p)$$

as follows. A priori this is the formal sum of a list of elements of the set $\{\pm 1\} \times \pi_1(M)$. It will ultimately give rise to an element of a quotient of $\mathbb{Z}[\pi_1(M)]$, given in Definition 2.12, after we factor out the effect of finger and Whitney moves and the effect of the choice of the paths γ_f^p and γ_g^p in the first bullet point below.

Fix $p \in f \pitchfork g$. Next we define $\varepsilon(p) \in \{\pm 1\}$ and $\eta(p) \in \pi_1(M)$. We use $*$ to denote concatenation of paths.

- Let γ_f^p be a path in Σ from the basepoint to $f^{-1}(p)$ and let γ_g^p be a path in Σ' from the basepoint to $g^{-1}(p)$.
- The sign $\varepsilon(p) \in \{\pm 1\}$ is determined as follows. Transport the local orientation of Σ at the basepoint to $f^{-1}(p)$ along γ_f^p, and the local orientation of Σ' at the basepoint to $g^{-1}(p)$ along γ_g^p. This induces a local orientation at p, by ordering f before g. Another local orientation is obtained by transporting the local orientation at the basepoint of M to p along the concatenated path $v_g * (g \circ \gamma_g^p)$. We define $\varepsilon(p) = +1$ when the two local orientations match at p, and -1 otherwise.
- The element $\eta(p) \in \pi_1(M)$ is by definition the concatenation $v_f * (f \circ \gamma_f^p) * (g \circ \gamma_g^p)^{-1} * v_g^{-1}$.

For a generic immersion $f\colon \Sigma \looparrowright M$, we define $\lambda(f, f) := \lambda(f, f^+)$, where f^+ is a push-off of f along a section of its normal bundle transverse to the zero section. If the embedding $f|_{\partial\Sigma}$ is equipped with a specified framing for its normal bundle, then f^+ is defined to be a push-off of f along a section restricting to the first vector of that framing on $\partial\Sigma$.

If f_t is a homotopy of f that is transverse to g for all t then $\lambda(f_t, g)$ is independent of t as a set of signed fundamental group elements, assuming the above choices of $\gamma^p_{f_t}$ are made carefully. However, if f_t describes a finger move of f into g, there is a single time t_0 at which f_{t_0} and g are not transverse, because there is a tangency. After the tangency, two new intersection points p and q arise. These have the same group element $\eta(p) = \eta(q)$ and opposite signs $\varepsilon(p) = -\varepsilon(q)$, with appropriate choices of $\gamma^p_{f_t}$ and $\gamma^q_{f_t}$. Similarly, a Whitney move reduces the intersections between f and g by such a pair. To get a regular homotopy invariant notion, it is thus important to specify the home of $\lambda(f, g)$ carefully.

For Σ and Σ' simply connected, the sum $\lambda(f, g)$ is usually considered as an element of $\mathbb{Z}[\pi_1(M)]$ and is independent of the choice of $\{\gamma^p_f\}_p$ and $\{\gamma^p_g\}_p$. In the abelian group $\mathbb{Z}[\pi_1(M)]$ the relations $-a + a = 0$ for each $a \in \pi_1(M)$ are built in, and if one identifies the sign $\varepsilon(p)$ with the inverse in this abelian group then finger moves and Whitney moves do not change $\lambda(f, g)$ as an element in the group ring.

For nonsimply connected Σ and Σ', the homotopy class of γ^p_f and γ^p_g may be changed by wrapping around nontrivial elements in $\pi_1(\Sigma)$ or $\pi_1(\Sigma')$. This wrapping may also change the induced local orientations at the intersection points of f and g. We describe this in more detail next. Let $w^M : \pi_1(M) \to \{\pm 1\}$, $w^\Sigma : \pi_1(\Sigma) \to \{\pm 1\}$ and $w^{\Sigma'} : \pi_1(\Sigma') \to \{\pm 1\}$ denote the orientation characters.

Definition 2.12 Let $\Gamma_{f,g}$ be the abelian group generated by the elements of $\pi_1(M)$ and with relators

$$\gamma - w^\Sigma(\alpha) w^{\Sigma'}(\beta) w^M(g_\bullet(\beta)) \cdot f_\bullet(\alpha) * \gamma * g_\bullet(\beta), \tag{2-2}$$

for all $\alpha \in \pi_1(\Sigma), \beta \in \pi_1(\Sigma'), \gamma \in \pi_1(M)$. Here $f_\bullet(\alpha) := v_f * (f \circ \alpha) * v_f^{-1}$ and $g_\bullet(\beta) := v_g * (g \circ \beta) * v_g^{-1}$ are elements of $\pi_1(M)$.

For transverse $f : \Sigma \to M$ and $g : \Sigma' \to M$, the intersection number $\lambda(f, g) \in \Gamma_{f,g}$ is well defined. The relations precisely account for wrapping around elements of $\pi_1(\Sigma)$ or $\pi_1(\Sigma')$ as described above. We will show in Proposition 2.18 that this target also makes $\lambda(f, g)$ a homotopy invariant.

Remark 2.13 In the case that M, Σ and Σ' are all oriented,

$$\Gamma_{f,g} \cong \mathbb{Z}[f_\bullet(\pi_1(\Sigma)) \backslash \pi_1(M) / g_\bullet(\pi_1(\Sigma'))],$$

the free abelian group generated by the double coset quotient of $\pi_1(M)$ by left and right multiplication by the images of loops in Σ and Σ', respectively.

In general, due to the signs introduced by the orientation characters, there may be torsion in $\Gamma_{f,g}$. For example, consider $f : \mathbb{RP}^2 \looparrowright M$ with $f_\bullet(\mathbb{RP}^1) = 1$, without any assumption on M. Then, for every $g : \Sigma' \to M$, the group

$$\Gamma_{f,g} \cong (\mathbb{Z}/2)[\pi_1(M)/g_\bullet(\pi_1(\Sigma'))]$$

is 2–torsion, due to the relations $\gamma = w^{\mathbb{RP}^2}(\mathbb{RP}^1) \cdot f_\bullet(\mathbb{RP}^1) * \gamma = -\gamma$ for every $\gamma \in \pi_1(M)$, arising from setting $\alpha := \mathbb{RP}^1$ and $\beta := 1$ in (2-2).

To understand $\Gamma_{f,g}$ better, we introduce some notation. Write $\pm\pi_1(M) := \{\pm 1\} \times \pi_1(M)$. There is a natural inclusion $\pm\pi_1(M) \to \mathbb{Z}[\pi_1(M)]$ given by $(\pm 1, \gamma) \mapsto \pm\gamma$. Write $[a] \in \Gamma_{f,g}$ for the equivalence class of $a \in \mathbb{Z}[\pi_1(M)]$, and let $\sim$ denote the equivalence relation on $\pm\pi_1(M)$ induced by the composition $\pm\pi_1(M) \hookrightarrow \mathbb{Z}[\pi_1(M)] \twoheadrightarrow \Gamma_{f,g}$, ie for $a, b \in \pm\pi_1(M)$, $a \sim b$ if and only if the images of a and b in $\Gamma_{f,g}$ coincide. The following lemma is immediate from the definitions.

Lemma 2.14 *Let $\gamma_1, \gamma_2 \in \pi_1(M)$. One of the relations $[\gamma_1] = \pm[\gamma_2] \in \Gamma_{f,g}$ holds if and only if γ_1 and γ_2 represent the same element in the double coset $f_\bullet(\pi_1(\Sigma))\backslash\pi_1(M)/g_\bullet(\pi_1(\Sigma'))$.*

Let $\mathrm{pm}\colon \pm\pi_1(M)/\sim \twoheadrightarrow f_\bullet\pi_1(\Sigma)\backslash\pi_1(M)/g_\bullet\pi_1(\Sigma')$ be the map sending $\pm\gamma$ to the class of γ. We write $|\gamma| := \mathrm{pm}(\gamma)$. Here one should think that pm stands for dividing out "plus–minus". Note that pm has fibres of order 1 or 2 and we can decompose the double coset as a disjoint union $B_1 \sqcup B_2$ according to this distinction, where pm gives a bijection $\mathrm{pm}^{-1}(B_1) \leftrightarrow B_1$ while $\mathrm{pm}^{-1}(B_2) \to B_2$ is two-to-one.

Remark 2.15 We give examples in the cases from Remark 2.13. In the case that M, Σ and Σ' are all oriented, $B_1 = \varnothing$ and $B_2 = f_\bullet(\pi_1(\Sigma))\backslash\pi_1(M)/g_\bullet(\pi_1(\Sigma'))$. If we have $f\colon \mathbb{RP}^2 \looparrowright M$ with $f_\bullet(\mathbb{RP}^1) = 1$, and $g\colon \Sigma' \to M$ is arbitrary, then $B_1 = \pi_1(M)/g_\bullet(\pi_1(\Sigma'))$ and $B_2 = \varnothing$.

Choose a section s of pm. For each $s(b) \in \mathrm{pm}^{-1}(B_2)$, we denote the other element of $\mathrm{pm}^{-1}(b)$ by $-s(b)$. Their images in $\Gamma_{f,g}$ are indeed inverse to one another, which motivates the notation.

Lemma 2.16 *Fix a section s for* pm *as above. The abelian group $\Gamma_{f,g}$ is a direct sum $\Gamma_{f,g} = FA \oplus V$ of a free abelian group FA on the set $s(B_2) \subseteq \pm\pi_1(M)/\sim \subseteq \Gamma_{f,g}$ and a $\mathbb{Z}/2$–vector space V with basis $s(B_1) = \mathrm{pm}^{-1}(B_1) \subseteq \pm\pi_1(M)/\sim$.*

Reading off the coefficients in this decomposition gives homomorphisms $c_{s(b)}\colon \Gamma_{f,g} \to \mathbb{Z}$ for each $b \in B_2$, and $c_b\colon \Gamma_{f,g} \to \mathbb{Z}/2$ for each $b \in B_1$, yielding a decomposition of $\Gamma_{f,g}$ as a direct sum of copies of $\mathbb{Z}$ and $\mathbb{Z}/2$.

In particular, the homomorphisms $c_{s(b)}$ and c_b determine the isomorphisms displayed in Remark 2.13.

Proof Starting with the free abelian group with basis $\pi_1(M)$, a relator in (2-2) does one of the following three things:

(1) It identifies two distinct basis elements γ_1 and γ_2 if and only if $\gamma_2 = f_\bullet(\alpha) * \gamma_1 * g_\bullet(\beta) \in \pi_1(M)$ and $w^\Sigma(\alpha)w^{\Sigma'}(\beta)w^M(g_\bullet(\beta)) = 1$ for some $\alpha \in \pi_1(\Sigma)$ and $\beta \in \pi_1(\Sigma')$.

(2) It identifies a basis element γ_1 with the inverse $-\gamma_2$ of another basis element $\gamma_2 \neq \gamma_1$ if and only if $\gamma_2 = f_\bullet(\alpha) * \gamma_1 * g_\bullet(\beta)$ and $w^\Sigma(\alpha)w^{\Sigma'}(\beta)w^M(g_\bullet(\beta)) = -1$ for some $\alpha \in \pi_1(\Sigma)$ and $\beta \in \pi_1(\Sigma')$.

(3) It identifies a basis element γ with its inverse $-\gamma$ if and only if

$$\gamma = f_\bullet(\alpha) * \gamma * g_\bullet(\beta) \quad \text{and} \quad w^\Sigma(\alpha)w^{\Sigma'}(\beta)w^M(g_\bullet(\beta)) = -1$$

for some $\alpha \in \pi_1(\Sigma)$ and $\beta \in \pi_1(\Sigma')$.

The first two types of relators reduce the basis to the double coset

$$f_\bullet\pi_1(\Sigma)\backslash\pi_1(M)/g_\bullet\pi_1(\Sigma').$$

The third type adds the relations $2[\gamma]=0$ to the fundamental group elements γ in question, which then generate V, because these are exactly the γ such that $-[\gamma]=[\gamma]$, ie where $\#\mathrm{pm}^{-1}(|\gamma|)=1$. Those γ where $\#\mathrm{pm}^{-1}(|\gamma|)=2$ remain of infinite order and generate FA. Note that the second type of relator forces us to choose the section s in order to write down a consistent basis for FA. □

The subgroup V and its basis clearly do not depend on our choice of section, but the basis of FA depends on this choice. If we change the section s at a point $b\in B_2$ to s' so that $s'(b)=-s(b)$, the associated basis element changes to its inverse. It follows that the subgroup FA of $\Gamma_{f,g}$ does not depend on the choice of s. It also follows that the coefficient maps $c_{s(b)}$ only depend on s up to sign and satisfy $c_{-s(b)}=-c_{s(b)}$.

For a given $a\in\pm\pi_1(M)/\sim$, we can choose a section s as above with $s(\mathrm{pm}(a)):=a$ and hence we get a coefficient map c_a that is independent of the other values of s. For example, we can take $a:=[\gamma]\in\pm\pi_1(M)/\sim$ with $\gamma\in\pi_1(M)$ to get c_γ.

Definition 2.17 For $\gamma\in\pi_1(M)$, write $\lambda(f,g)_\gamma:=c_\gamma(\lambda(f,g))$. This quantity lies in $\mathbb{Z}$ (respectively $\mathbb{Z}/2$) when $|\gamma|$ lies in B_2 (respectively B_1), or equivalently when $[\gamma]$ has infinite order (respectively order 2) in $\Gamma_{f,g}$. The values do not depend on the choice of s and satisfy $c_{\gamma_1}=-c_{\gamma_2}$ whenever $[\gamma_1]=-[\gamma_2]$.

The following can be proven using Proposition 2.10 (see eg [Freedman and Quinn 1990, Section 1.7; Powell and Ray 2021b] for the case of discs and spheres).

Proposition 2.18 *Let $f\colon\Sigma\to M$ and $g\colon\Sigma'\to M$ be based maps that are transverse to one another. The intersection number $\lambda(f,g)$ is preserved by homotopies that are ambient isotopies near $\partial\Sigma\sqcup\partial\Sigma'$.*

Remark 2.19 The geometric definition of λ given above has a well-known algebraic version in the case that f and g correspond to classes in $H_2(M,\partial M;\mathbb{Z}[\pi_1(M)])$. This extends to the case of positive genus, as we now sketch. We restrict ourselves to the case that M, Σ and Σ' are closed and oriented for convenience.

Choose a basepoint in the universal cover of M, lifting the basepoint of M. The maps f and g lift uniquely (with respect to this choice of basepoint) to covers $\widehat{M}$ and $\widehat{M}'$, corresponding to the subgroups $f_\bullet(\pi_1(\Sigma))$ and $g_\bullet(\pi_1(\Sigma'))$, respectively. These lifts represent classes

$$[f]\in H_2(\widehat{M};\mathbb{Z})\cong H_2\big(M;\mathbb{Z}[\pi_1(M)/f_\bullet\pi_1(\Sigma)]\big)$$

and

$$[g]\in H_2(\widehat{M}';\mathbb{Z})\cong H_2\big(M;\mathbb{Z}[\pi_1(M)/g_\bullet\pi_1(\Sigma')]\big).$$

Then we have

$$\mathrm{PD}^{-1}([f])\smile\mathrm{PD}^{-1}([g])\in H^4\big(M;\mathbb{Z}[\pi_1(M)/f_\bullet\pi_1(\Sigma)]\otimes_{\mathbb{Z}}\mathbb{Z}[\pi_1(M)/g_\bullet\pi_1(\Sigma')]\big).$$

By Poincaré duality, this yields an element in

$$H_0\big(M;\mathbb{Z}[\pi_1(M)/f_\bullet\pi_1(\Sigma)]\otimes_{\mathbb{Z}}\mathbb{Z}[\pi_1(M)/g_\bullet\pi_1(\Sigma')]\big),$$

which is isomorphic as an abelian group to

$$\mathbb{Z}[f_\bullet\pi_1(\Sigma)\backslash\pi_1(M)]\otimes_{\mathbb{Z}[\pi_1(M)]}\mathbb{Z}[\pi_1(M)/g_\bullet\pi_1(\Sigma')].$$

Here $\mathbb{Z}[f_\bullet(\pi_1(\Sigma))\backslash\pi_1(M)]$ denotes $\mathbb{Z}[\pi_1(M)/f_\bullet\pi_1(\Sigma)]$ considered as a right $\mathbb{Z}[\pi_1(M)]$–module.

Finally, we have the isomorphism

$$\begin{aligned}\mathbb{Z}[f_\bullet\pi_1(\Sigma)\backslash\pi_1(M)]\otimes_{\mathbb{Z}[\pi_1(M)]}\mathbb{Z}[\pi_1(M)/g_\bullet\pi_1(\Sigma')]&\to\mathbb{Z}[f_\bullet\pi_1(\Sigma)\backslash\pi_1(M)/g_\bullet\pi_1(\Sigma')],\\ [a]\otimes[b]&\mapsto[ab]\quad\text{ for } a,b\in\pi_1(M).\end{aligned}$$

We shall not prove that this formulation agrees with the geometric definition.

2.3 Self-intersection numbers

Next we turn to the *self-intersection number* for a based generic immersion $f:\Sigma\looparrowright M$ of a connected surface Σ, with whisker v_f. The definition of μ, given below, is similar to that of λ in the previous subsection, except that there is no longer a clear choice of which sheet to consider first at a given double point. Consequently, the values of μ lie in a further quotient of the group $\Gamma_{f,f}$ from Definition 2.12.

We write $f\pitchfork f\subseteq M$ for the set of double points of f. We record the self-intersections of f by the sum of signed group elements

$$\mu(f):=\sum_{p\in f\pitchfork f}\varepsilon(p)\cdot\eta(p)$$

as follows:

- For $p=f(x_1)=f(x_2)$ for $x_1\neq x_2\in\Sigma$, let γ_1^p and γ_2^p be paths in Σ from the basepoint to x_1 and x_2, respectively.
- The sign $\varepsilon(p)\in\{\pm1\}$ is defined as follows. Transport the local orientation of Σ at the basepoint to x_1 along γ_1^p, and along γ_2^p to x_2. This induces a local orientation at p. Another local orientation is obtained by transporting the local orientation at the basepoint of M to p along the concatenated path $v_f*(f\circ\gamma_2^p)$. We define $\varepsilon(p)=1$ when the two local orientations match at p, and -1 otherwise.
- The element $\eta(p)\in\pi_1(M)$ is given by the concatenation $v_f*(f\circ\gamma_1^p)*(f\circ\gamma_2^p)^{-1}*v_f^{-1}$.

There is a similar discussion about homotopy invariance of $\mu(f)$ as for $\lambda(f,g)$ earlier: homotopies f_t that are generic immersions for all t preserve the formal sum of signed elements but finger moves and Whitney moves (of f with itself) create pairs $-\eta(p)+\eta(p)$, so it is convenient to use abelian groups. This takes care of regular homotopies of f but there is an additional subtlety for cusp homotopies f_t, where there is exactly one time t_0 for which f_{t_0} is not an immersion. These issues will be discussed carefully below.

For simply connected Σ, the self-intersection invariant $\mu(f)$ is well defined in the quotient (as an abelian group) of $\mathbb{Z}[\pi_1(M)]$ obtained by introducing the relators

$$\gamma - w^M(\gamma)\cdot\gamma^{-1}$$

for all $\gamma \in \pi_1(M)$. For general Σ, the quantity $\mu(f)$ is well defined in the abelian group

$$\Gamma_f := \Gamma_{f,f}/\langle \gamma - w^M(\gamma)\cdot\gamma^{-1}\rangle, \tag{2-3}$$

ie in this quotient of $\Gamma_{f,f}$ from Definition 2.12. Here, as above, $w^M\colon \pi_1(M) \to \{\pm 1\}$ is the orientation character.

We now change our notation slightly from the discussion of $\lambda(f,g)$ in order to work in this further quotient. Let $\sim$ denote the equivalence relation on $\pm\pi_1(M)$ induced by the composition

$$\pm\pi_1(M) \hookrightarrow \mathbb{Z}[\pi_1(M)] \twoheadrightarrow \Gamma_f$$

sending $a \mapsto [a]$ and let $|\pi_1(M)|$ be the quotient of $\pm\pi_1(M)$ obtained by identifying γ_1 and γ_2 whenever $[\gamma_1] = \pm[\gamma_2] \in \Gamma_f$. Then we obtain the following analogues of Lemmas 2.14 and 2.16.

Lemma 2.20 *Let $\gamma_1, \gamma_2 \in \pi_1(M)$. One of the relations $[\gamma_1] = \pm[\gamma_2]$ holds if and only if γ_1 and γ_2 represent the same element in the quotient of the double coset by inversion. In other words, the identity map induces a bijection*

$$|\pi_1(M)| \leftrightarrow (f_\bullet\pi_1(\Sigma)\backslash\pi_1(M)/f_\bullet\pi_1(\Sigma))/\approx,$$

where $\approx$ is the equivalence relation identifying γ and γ^{-1} for all $\gamma \in \pi_1(M)$.

We write $|\gamma| := \mathrm{pm}(\gamma)$ for the quotient map $\mathrm{pm}\colon \pm\pi_1(M)/\sim \twoheadrightarrow |\pi_1(M)|$. Again, pm has fibres of order 1 or 2 and we decompose $|\pi_1(M)|$ as a disjoint union $B_1 \sqcup B_2$ according to this distinction as before. Choose a section $s\colon |\pi_1(M)| \to \pm\pi_1(M)/\sim$ of pm. As before, for $b \in B_2$ we denote the elements of the fibre by $\mathrm{pm}^{-1}(b) = \{s(b), -s(b)\}$.

Lemma 2.21 *The abelian group Γ_f is a direct sum $\Gamma_f = FA \oplus V$ of a free abelian group FA on the set $s(B_2) \subseteq \pm\pi_1(M)/\sim \subseteq \Gamma_f$ and a $\mathbb{Z}/2$–vector space V with basis $s(B_1) = \mathrm{pm}^{-1}(B_1) \subseteq \pm\pi_1(M)/\sim$.*

Reading off the coefficients in this decomposition gives homomorphisms $c_{s(b)}\colon \Gamma_f \to \mathbb{Z}$ for $b \in B_2$, and $c_b\colon \Gamma_f \to \mathbb{Z}/2$ for $b \in B_1$, leading to a decomposition of Γ_f as a direct sum of copies of $\mathbb{Z}$ and $\mathbb{Z}/2$.

Proof The proof is analogous to that of Lemma 2.16. □

Remark 2.22 If M and Σ are oriented, then

$$\Gamma_f \cong \mathbb{Z}[f_\bullet\pi_1(\Sigma)\backslash\pi_1(M)/f_\bullet\pi_1(\Sigma)]/\langle\gamma - \gamma^{-1}\rangle = \mathbb{Z}[(f_\bullet\pi_1(\Sigma)\backslash\pi_1(M)/f_\bullet\pi_1(\Sigma))/\approx]$$

is free abelian. In this case, $B_1 = \varnothing$ and $B_2 = |\pi_1(M)| = (f_\bullet\pi_1(\Sigma)\backslash\pi_1(M)/f_\bullet\pi_1(\Sigma))/\approx$ by Lemma 2.20.

Consider instead $f\colon \mathbb{RP}^2 \looparrowright M$ with $f_\bullet(\mathbb{RP}^1) = 1$, without any assumption on M. Then

$$\Gamma_f \cong (\mathbb{Z}/2)[\pi_1(M)]/\langle \gamma - w^M(\gamma)\cdot\gamma^{-1}\rangle = (\mathbb{Z}/2)[\pi_1(M)/{\approx}].$$

As in Remark 2.13, this is 2–torsion due to the relations $\gamma = w^{\mathbb{RP}^2}(\mathbb{RP}^1)\cdot f_\bullet(\mathbb{RP}^1) * \gamma = -\gamma$. In this case, $B_1 = |\pi_1(M)| = \pi_1(M)/{\approx}$ and $B_2 = \varnothing$.

As before, the subgroups FA and V of Γ_f do not depend on the choice of s; only the basis of FA does. As a consequence, the coefficient maps $c_{s(b)}$ only depend on s up to sign and satisfy $c_{-s(b)} = -c_{s(b)}$. Given $a \in \pm\pi_1(M)/{\sim}$, we may again take $s(\mathrm{pm}(a)) := a$ to get c_a and, in particular, c_γ for $\gamma \in \pi_1(M)$ independent of the choice of s at other points. This gives the following definition.

Definition 2.23 For $\gamma \in \pi_1(M)$, we write $\mu(f)_\gamma := c_\gamma(\mu(f))$. This quantity lies in $\mathbb{Z}$ (respectively $\mathbb{Z}/2$) when $|\gamma|$ lies in B_2 (respectively B_1), or equivalently when $[\gamma]$ has infinite order (respectively order 2) in Γ_f. The values do not depend on the choice of s and satisfy $c_{\gamma_1} = -c_{\gamma_2}$ whenever $[\gamma_1] = -[\gamma_2]$.

We focus on $\mu(f)_1$, which plays an important role in the distinction between the homotopy class and regular homotopy class of f, as we will discuss in the next subsection. In the usual case, where Σ is simply connected, $\mu(f)_1 \in \mathbb{Z}$. However, in general, $\mu(f)_1$ may lie in either $\mathbb{Z}$ or $\mathbb{Z}/2$. The following lemma gives the precise conditions determining the home of $\mu(f)_1$.

Lemma 2.24 *Let $f\colon \Sigma \looparrowright M$ be a based, generic immersion, with whisker v. Recall that the map $f_\bullet\colon \pi_1(\Sigma) \to \pi_1(M)$ is given by $\alpha \mapsto v * (f\circ\alpha) * v^{-1}$.*

If w^Σ is trivial on $\ker(f_\bullet)$ and w^M is trivial on $\mathrm{Im}(f_\bullet)$, then $[1] \in \Gamma_f$ has infinite order and thus $\mu(f)_1 \in \mathbb{Z}$. Otherwise, $[1]$ has order 2 and $\mu(f)_1 \in \mathbb{Z}/2$.

Proof By definition, for $1 \in \pi_1(M)$, we know that $[1] \in \Gamma_{f,f}$ has order 2 precisely if

(i) there exists $\alpha, \beta \in \pi_1(\Sigma)$ such that $f_\bullet(\alpha) * f_\bullet(\beta) = f_\bullet(\alpha * \beta) = 1$ and $w^\Sigma(\alpha)w^\Sigma(\beta)w^M(f_\bullet(\beta)) = w^\Sigma(\alpha*\beta)w^M(f_\bullet(\beta)) = -1$, or

(ii) there exists $\delta \sim 1$ where δ has order two in $\pi_1(M)$ and $w^M(\delta) = -1$.

Suppose that w^Σ is trivial on $\ker(f_\bullet)$ and w^M is trivial on $\mathrm{Im}(f_\bullet)$. Then the first case (i) cannot happen since, if $f_\bullet(\alpha*\beta) = 1$, then $\alpha*\beta \in \ker(f_\bullet)$ so $w^\Sigma(\alpha*\beta)w^M(f_\bullet(\beta)) = 1\cdot 1 = 1$. Similarly, (ii) cannot happen: suppose δ has order two in $\pi_1(M)$ and $\delta \sim 1$. Then, by definition, $\delta = f_\bullet(\alpha)*1*f_\bullet(\beta) = f_\bullet(\alpha*\beta)$ in $\pi_1(M)$, for some $\alpha, \beta \in \pi_1(M)$. In particular, $\delta \in \mathrm{Im}(f_\bullet)$, and so again $w^M(\delta) = 1$ by hypothesis, contradicting (ii). Therefore, $[1]$ has infinite order, as claimed.

Now suppose there is some $\alpha \in \pi_1(\Sigma)$ with $w^M(f_\bullet(\alpha)) = -1$. Then we have $f_\bullet(\alpha^{-1}) * 1 * f_\bullet(\alpha) = 1$ and $w^\Sigma(\alpha^{-1})w^\Sigma(\alpha)w^M(f_\bullet(\alpha)) = -1$, so $[1] \sim -[1]$ and $[1]$ has order two.

Finally, suppose that there is some $\alpha \in \ker(f_\bullet)$ with $w^\Sigma(\alpha) = -1$. Then we have $f_\bullet(\alpha) * 1 * f_\bullet(1_\Sigma) = 1$ and $w^\Sigma(\alpha)w^\Sigma(1_\Sigma)w^M(f_\bullet(1_\Sigma)) = -1$, where 1_Σ denotes the trivial element in $\pi_1(\Sigma)$. Then, again, we have $[1] \sim -[1]$ and $[1]$ has order two. □

As with Proposition 2.18, the proof of the following proposition is virtually identical to the case of discs and spheres using Proposition 2.10 (see eg [Powell and Ray 2021b]), and we leave it for the interested reader.

Proposition 2.25 *Let $f\colon \Sigma \looparrowright M$ be a based generic immersion. The self-intersection number $\mu(f)$ is preserved under regular homotopies that are ambient isotopies near $\partial\Sigma$.*

In this and the previous subsection, we have considered intersection and self-intersection numbers of connected surfaces. By combining these invariants, we can define the conglomerate notion of self-intersection number for disconnected surfaces $F = \{f_i\}_{i=1}^m\colon \Sigma \to M$, as considered in Definition 1.3:

$$\mu(F) := \sum_{i<j} \lambda(f_i, f_j) + \sum_i \mu(f_i) \in \bigoplus_{i<j} \Gamma_{f_i,f_j} \oplus \bigoplus_i \Gamma_{f_i}.$$

Propositions 2.18 and 2.25 imply that $\mu(F)$ is preserved under regular homotopies of F that are ambient isotopies near $\partial\Sigma$.

2.4 Whitney discs

A Whitney move cancels a pair of double points of a generic immersion $F\colon \Sigma \looparrowright M$ as in Convention 1.1, provided all the assumptions on the guiding Whitney disc are satisfied. In our setting, where Σ and M need be neither simply connected nor orientable, this requires some care. We start with the notion of arcs A and A' pairing double points p and q, and the corresponding notion of (p, q, A, A') having *opposite sign*.

Definition 2.26 Let $f\colon \Sigma \to M$ and $g\colon \Sigma' \to M$ be based maps that either intersect transversely, or $f = g$ and f is a generic immersion. We say that two points $p, q \in f \pitchfork g \subseteq M$ are *paired by arcs* if we equip them with the extra data of an arc $A\colon [0,1] \to \Sigma$ from $f^{-1}(p)$ to $f^{-1}(q)$ and an arc A' in Σ' from $g^{-1}(q)$ to $g^{-1}(p)$. In the case that $f = g$, we require that each point in $f^{-1}(p)$ and in $f^{-1}(q)$ is the endpoint of precisely one of the arcs A and A', ie A and A' lie in distinct sheets at both p and q.

With the extra data of the arcs A and A', we can make sense of whether two intersection points that are paired by arcs have opposite sign.

Definition 2.27 Let $f\colon \Sigma \to M$ and $g\colon \Sigma' \to M$ be based maps that either intersect transversely, or $f = g$ and f is a generic immersion. Two intersection points $p, q \in f \pitchfork g \subseteq M$ paired by arcs A in Σ and A' in Σ' have *opposite sign* if the following holds. Fix local orientations of Σ at $f^{-1}(p)$ and of Σ' at $g^{-1}(p)$. This choice induces a local orientation of M at p. Transport the local orientation of Σ from $f^{-1}(p)$ to $f^{-1}(q)$ along A, and the local orientation of Σ' from $g^{-1}(p)$ to $g^{-1}(q)$ along A'. This gives a local orientation of M at the point q. Compare this with the local orientation on M at q induced by transporting the local orientation from p to q along the arc $f \circ A$. If these orientations disagree, then the points p and q are said to have opposite sign (with respect to A and A'), and otherwise they are said to have the same sign. The dependence on the choice of arcs A and A' is sometimes neglected.

Note that double points having the same sign could be "paired" by an embedded disc, but this does not mean that a Whitney move using this disc is possible, because the required section of the normal bundle of the disc is not available; in this case, any rank one subbundle of the normal bundle of the disc, restricted to the boundary, that is tangent to one sheet of Σ and normal to the other sheet turns out to be a Möbius bundle. So one does not study such discs and assumes that a Whitney disc always pairs two double points of opposite sign.

In the setting of based transverse maps $f \pitchfork g$, with Σ and Σ' connected, recall from Section 2.2 that $\lambda(f,g)$ is a sum of terms $\varepsilon(p)\cdot\eta(p)$, one for each double point $p \in f \pitchfork g$, with $\eta(p) \in \pi_1(M)$ and $\varepsilon(p) \in \{\pm 1\}$. This sum is well defined in the abelian group $\Gamma_{f,g}$ and each signed group element $a \in \pm\pi_1(M)$ represents a unique element $[a] \in \Gamma_{f,g}$. The same proof as in the case of simply connected surfaces [Powell and Ray 2021b, Proposition 11.10] yields the following result.

Lemma 2.28 *Let Σ and Σ' be compact connected surfaces and let $f\colon \Sigma \to M$ and $g\colon \Sigma' \to M$ be based maps with transverse double points $p, q \in f \pitchfork g \subseteq M$. Then $[\varepsilon(p)\cdot\eta(p) + \varepsilon(q)\cdot\eta(q)] = 0 \in \Gamma_{f,g}$ if and only if p and q can be paired by arcs $A \subseteq \Sigma$ and $A' \subseteq \Sigma'$ such that*

(i) *the closed loop $f\circ A \cup_{p,q} g\circ A'$ is null-homotopic in M, and*

(ii) *the points p and q have opposite sign with respect to the arcs A and A'.*

If (i) and (ii) are satisfied for p and q, we say that $W\colon D^2 \to M$ is a (map of a) *Whitney disc pairing p and q* if its boundary is the closed loop in (i), the union of its two *Whitney arcs* $f\circ A$ and $g\circ A'$. We leave it to the reader to formulate the analogous notion for a pair of transverse self-intersection points of $f\colon \Sigma \to M$. This gives rise to the following corollary to Lemma 2.28.

Corollary 2.29 *Let Σ and Σ' be compact connected surfaces and let $f\colon \Sigma \to M$ and $g\colon \Sigma' \to M$ be transverse based maps. Then $\lambda(f,g) = 0$ if and only if all intersection points between f and g can be paired by maps of Whitney discs.*

Moreover, a based generic immersion $f\colon \Sigma \looparrowright M$ satisfies $\mu(f) = 0$ if and only if all self-intersection points of f can be paired by maps of Whitney discs.

Note that, by the geometric Casson lemma (Lemma 4.2), the vanishing of $\lambda(f,g)$ is equivalent to the existence of a regular homotopy of f and g that makes their images disjoint, at the cost of introducing self-intersections in f and g. There is no analogue of this argument if $\mu(f) = 0$ by the failure of the Whitney trick in dimension 4, as for example exhibited by the secondary embedding obstruction $\mathrm{km}(f)$ (see Section 3).

By the geometric characterisation in Corollary 2.29, it is meaningful to refer to f and g having trivial intersection number, and to f having trivial self-intersection number, without using a basing.

The analogue of the characterisation in the second part of Corollary 2.29 holds for generic immersions $F = \{f_i\}_{i=1}^m : \Sigma \looparrowright M$ from Convention 1.1, ie for compact but possibly disconnected domains, if we use Definition 1.3 from the introduction for the self-intersection number $\mu(F)$.

Corollary 2.30 *Let $F\colon (\Sigma, \partial\Sigma) \looparrowright (M, \partial M)$ be as in Convention 1.1. Then $\mu(F) = 0$ if and only if the double points of F can be paired by maps of Whitney discs.*

Proof This is a direct consequence of Propositions 2.18 and 2.25 and Corollary 2.29 because every double point of F is either a self-intersection point of a component f_i or an intersection point between distinct components f_i and f_j (where we can assume that $i < j$). Note that both cases represent self-intersection points of F. □

Collections of Whitney discs as above may be assumed to be *convenient* in the following sense (see eg [Freedman and Quinn 1990, Section 1.4; Powell and Ray 2021b]).

Definition 2.31 Let $F\colon (\Sigma, \partial\Sigma) \looparrowright (M, \partial M)$ be as in Convention 1.1. A *convenient* collection of Whitney discs for F is a collection of framed, generically immersed Whitney discs pairing all the double points of F, with interiors transverse to F and with disjointly embedded boundaries. A collection of arcs in $F(\Sigma)$ is called a collection of *Whitney arcs* if the union of the arcs is the boundary of a convenient collection of Whitney discs.

By pushing double points of a convenient collection across the boundaries of Whitney discs [Powell and Ray 2021b, Figure 11.4], we may further assume that all Whitney discs are pairwise disjoint and embedded. However, the (resulting and preexisting) intersections between the original surface F and the Whitney discs can in general not be removed, as detected by the secondary invariant $\mathrm{km}(F)$.

2.5 Homotopy versus regular homotopy of generic immersions

Let $f\colon \Sigma \looparrowright M$ be a generic immersion. Local orientations of M and Σ determine a local orientation of νf. Hence, given a framing of $f|_{\partial\Sigma}$, one can define a relative Euler class of the normal bundle νf in $H^2(\Sigma, \partial\Sigma; \mathbb{Z}^{w_1(\nu f)})$. If $f^*(w_1(M)) = w_1(\nu f) + w_1(T\Sigma) = 0$ then the local orientation of Σ determines a Poincaré duality isomorphism from this twisted cohomology group to $\mathbb{Z}$, and we denote the resulting integer by $e(\nu f)$. Note that $e(\nu f)$ does not depend on the local orientation of Σ but only on the local orientation of M. If $f^*(w_1(M)) \neq 0$ then there is still a mod 2 normal Euler number, which we also denote by $e(\nu f) \in \mathbb{Z}/2$.

A useful interpretation of $e(\nu f)$ is as follows. A vector in $\mathbb{R}^2$ together with the framing of $f|_{\partial\Sigma}$ determines a nonvanishing section of νf on $f(\partial\Sigma)$. Extend this to a section of νf over all of $f(\Sigma)$, transverse to the zero section. Then $e(\nu f)$ counts, with sign, the number of zeros of the section, in $\mathbb{Z}$ or $\mathbb{Z}/2$ as appropriate.

Next we give an extension of [Powell et al. 2020, Theorem 1.2] from the simply connected to the general setting, restricting ourselves to the case of connected Σ for convenience. We note that [Powell et al. 2020, Theorem 1.2] was based on [Freedman and Quinn 1990, Lemma 1.2 and Proposition 1.6], but that the latter proposition was not proven in [Freedman and Quinn 1990].

By the following theorem, in some cases, for example when M is orientable, $e(\nu f) \in \mathbb{Z}$ is an additional invariant of regular homotopy classes of immersions. It changes by ± 2 during a cusp homotopy (see eg [Conant et al. 2012b, Figure 19]) and hence there can be infinitely many regular homotopy classes of immersions, all of which are homotopic as continuous maps.

In Theorem 2.32, in the case that Σ has nontrivial boundary, we fix a framing on the embedding $f|_{\partial\Sigma}$, in order to define the relative Euler number $e(\nu \tilde{f})$ for $\tilde{f}$ any generic immersion homotopic to f.

Theorem 2.32 *Let Σ be a compact, connected surface and let M be a 4–manifold. Then the inclusion of the subspace of generic immersions $\mathrm{Imm}(\Sigma, M)$ in the space of all continuous maps induces a map*

$$\frac{\mathrm{Imm}(\Sigma, M)}{\{\text{regular homotopy}\}} \xrightarrow{i} [\Sigma, M]_\partial,$$

where $[\Sigma, M]_\partial$ denotes the set of homotopy classes of continuous maps that restrict on $\partial\Sigma$ to embeddings disjoint from the image of the interior of Σ.

(1) *i is surjective.*

(2) *The fibres of i are related by cusp homotopies. More precisely, suppose that f and g are homotopic generic immersions. Then we can add cusps to f and g, to obtain f' and g', respectively, such that f' and g' are regularly homotopic.*

(3) *For every $f \in [\Sigma, M]_\partial$, there is a bijection*

$$i^{-1}(f) \cong \begin{cases} 2\mathbb{Z} & \text{if } f^*(w_1(M)) = 0 \text{ and } w_2(\nu \tilde{f}) = 0, \\ 2\mathbb{Z}+1 & \text{if } f^*(w_1(M)) = 0 \text{ and } w_2(\nu \tilde{f}) = 1, \\ \mathbb{Z}/2 & \text{otherwise}, \end{cases}$$

where $\nu \tilde{f}$ is a normal bundle for $\tilde{f}$, a generic immersion in $i^{-1}(f)$. When $f^*(w_1(M)) = 0$, the bijection is given by

$$\tilde{f} \mapsto e(\nu \tilde{f}).$$

Otherwise, the bijection is given by

$$\tilde{f} \mapsto \mu(\tilde{f})_1 \in \mathbb{Z}/2.$$

(4) *If $f^*(w_1(M)) = 0$ and $w_1(\Sigma)|_{\ker(f_\bullet)} = 0$ for $\tilde{f}$ a generic immersion in $i^{-1}(f)$, the quantities $\mu(\tilde{f})_1$ and $e(\nu \tilde{f})$ are related by the formula*

$$\lambda(\tilde{f}, \tilde{f})_1 = 2\mu(\tilde{f})_1 + e(\nu \tilde{f}) \in \mathbb{Z}$$

and so $\mu(\tilde{f})_1 \in \mathbb{Z}$ also detects the regular homotopy class of $\tilde{f} \in i^{-1}(f)$.

While we prefer the upcoming direct argument analysing singularities, Theorem 2.32 could in principle also be proven via Smale–Hirsch immersion theory, which has a version in the topological category. The main novelty of the theorem is that we give precise conditions in terms of the Stiefel–Whitney classes to control how large the fibres of i are, and which invariants detect them.

Proof By Proposition 2.10(1), the map is surjective. That is, every homotopy class contains a generic immersion. This proves (1).

For (2), note that, if f and g are homotopic generic immersions, then, by Proposition 2.10(2), there exists a generic homotopy H between them, which by definition is a sequence of ambient isotopies, finger moves, Whitney moves and cusp homotopies. We can modify H so that there are real numbers $t_1 < t_2 \in [0,1]$ such that the singularities of H in $[0,t_1]$ only consist of cusp homotopies that create double points, the singularities in $[t_1,t_2]$ only consist of finger moves and Whitney moves, and those in $[t_2,1]$ only consist of cusp homotopies that remove double points. The statement then follows by taking $f' := H_{t_1}$ and $g' := H_{t_2}$.

To achieve this modification, note that we can bring all the creating cusp singularities forward, so that they occur earlier, and we can delay all the removing cusps. To arrange for a creating cusp to be rearranged earlier than a finger or Whitney move, choose an arc in the image of H starting from $H_t(\Sigma)$ for some $t \in (0,t_1)$, and ending at the cusp, which intersects each level in a point and is disjoint from all Whitney arcs and double points. The homotopy can then be altered in a neighbourhood of this arc so that the cusp singularity occurs at time t. Delaying a removing cusp is the same procedure but with the direction of time reversed. This completes the proof of (2).

The proof of (3) splits naturally into two cases.

Case 1 $f^*(w_1(M)) = 0$.

As noted in Section 2, the sign of an intersection point is not always well defined. Nevertheless, in the case that $f^*(w_1(M)) = 0$, the sign of a cusp homotopy is well defined. The key point is that a cusp not only specifies a double point p but also an arc between the preimages of p. In the case that $f^*(w_1(M)) = 0$, using this path, the sign of the double point p is well defined, independent of the choice of path transporting the local orientation at the basepoint to the double point. Thus in this setting we define the *sign* of a cusp to be the sign of the double point it creates or removes. We will use the terminology of *creating cusps* for cusps that create a double point and *removing cusps* for those that remove a double point.

Since $f^*(w_1(M)) = 0$, $e(\nu\tilde{f})$ is defined in $\mathbb{Z}$ for any generic immersion $\tilde{f}$ homotopic to f. Recall that $w_2(\nu\tilde{f}) \equiv e(\nu\tilde{f}) \bmod 2$. Since regularly homotopic generic immersions have equal Euler numbers, the map in the theorem statement is well defined on equivalence classes in the domain of i. Note that a cusp homotopy changes $e(\nu\tilde{f})$ by 2 or -2, depending on the sign of the cusp and whether it is a creating or a removing cusp. So every element of $2\mathbb{Z}$ or $2\mathbb{Z}+1$, depending on $w_2(\nu\tilde{f})$, can be realised as the Euler number of a generic immersion in $i^{-1}(f)$.

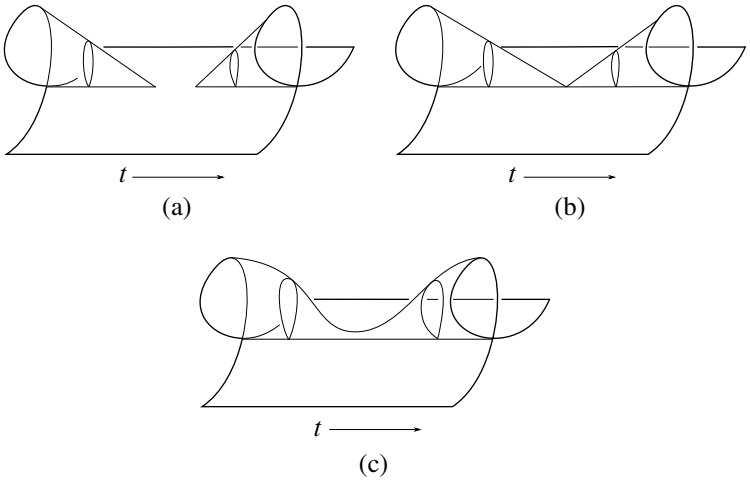


Figure 3: A schematic picture showing how a removing cusp singularity and creating cusp singularity with the same sign can be cancelled. In each of (a), (b) and (c), a homotopy is traced out in the direction of t. At every time t, except the times of the cusp singularities, we depict an arc of a generic immersion homotopic to f. (a) Two cusp singularities are shown: a removing cusp occurring first, followed by a creating cusp of the same sign. (b) Modify the homotopy, delaying the removing cusp until it coincides with the creating cusp. This involves choosing an arc in Σ joining the two cusp points. (c) A further local modification removes the two cusps.

To complete the proof when $f^*(w_1(M)) = 0$, it remains to show injectivity. We will show that, given a generic homotopy between generic immersions with equal Euler numbers, we can modify the homotopy to cancel cusps, until we are left with a regular homotopy.

First note that, when we have a removing cusp, and later in the homotopy we have a creating cusp with the same sign, we can cancel these two cusps along a level-preserving path in the homotopy, as indicated in Figure 3.

However, this is not sufficient. We also have to show that we can also cancel cusps given

(i) two creating cusps of opposite sign, or two removing cusps of opposite sign; or

(ii) a creating cusp paired with a later removing cusp, both of the same sign.

Suppose that we have a generic homotopy H between generic immersions with equal Euler numbers consisting of two creating cusp homotopies of opposite sign, as in (i). Create a self-homotopy H_0 of the starting immersion, ie the immersion at $t = 0$, consisting of a trivial finger move together with two removing cusps for the double points created by the finger move, as shown in Figure 4. Then concatenate H_0 with the original homotopy H. The new homotopy can be modified as in Figure 3 to cancel the removing cusps in H_0 and the creating cusps in H, leaving only the finger move behind. An analogous argument shows how to cancel two removing cusps of opposite sign, this time concatenating at the end of H.

Similarly, for the situation in (ii), suppose that we have a generic homotopy H between generic immersions with equal Euler numbers consisting of a creating cusp and a later removing cusp of the same sign. Again

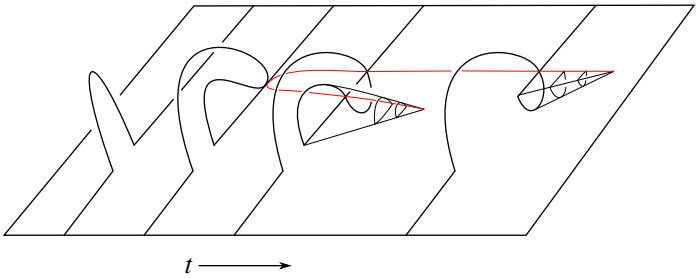


Figure 4: A schematic picture showing a self-homotopy consisting of a trivial self-finger move followed by two removing cusps. The homotopy is traced out in the direction of t. At every time t, except at the times of the cusp and finger move singularities, we depict an arc of a generic immersion homotopic to f. In red we show the arc of self-intersections of f; note that it starts at one cusp singularity and ends at the other.

we construct the self-homotopy H_0 and concatenate with H. In the result, we use the procedure from Figure 3 to cancel the creating cusp in H with one of the removing cusps in H_0. This entire operation has so far replaced a cusp with a cusp of opposite sign and direction. As before we can repeat the operation at the end of the homotopy to replace the removing cusp with a creating one, also with the opposite sign. Thus when we have a creating cusp with a later removing cusp of the same sign, we can replace both by cusps of opposite sign and direction. Since now the removing cusp happens before the creating cusp, the two can be cancelled and we are done with case (ii). This completes the proof of (3) in the case that $f^*(w_1(M)) = 0$.

Case 2 $f^*(w_1(M)) \neq 0$.

Note that a cusp homotopy changes $\mu(\tilde{f})_1 \in \mathbb{Z}/2$ by one. So both values of $\mathbb{Z}/2$ can be realised within the homotopy class. To show injectivity in this case, we have to show that we can cancel cusps in a homotopy in arbitrary pairs. First use the trading argument above to get all the removing cusps before the creating cusps in the homotopy. Then, for any pair of cusps, one removing and one creating, choose some level-preserving path in the homotopy between the first and the second cusp, and restrict to a small disc containing the path. If they have the same sign with respect to this disc, cancel the two cusps as before.

If they have opposite signs, change the choice of the arc to arrange that the union of the new arc and the old arc maps nontrivially under $w_1(M)$. Such an arc exists since $f^*(w_1(M))$ is nontrivial and Σ is connected. With this new choice, the signs of the cusps in the disc become the same and we can again cancel the cusps. This completes the proof of both halves of (3).

Finally, for (4), note that, if $f^*(w_1(M)) = 0$ and $w_1(\Sigma)|_{\ker(f_\bullet)} = 0$, then, by Lemma 2.24, $\mu(\tilde{f})_1$ is well defined in $\mathbb{Z}$. By the discussion above the statement of the theorem, $e(\nu\tilde{f})$ is also well defined in $\mathbb{Z}$. In this case, the formula

$$\lambda(\tilde{f}, \tilde{f})_1 = 2\mu(\tilde{f})_1 + e(\nu\tilde{f}) \in \mathbb{Z}$$

holds by the proof of the corresponding fact for discs and spheres (see eg [Powell and Ray 2021b, Proposition 11.8]). Any cusp homotopy leaves $\lambda(\tilde{f}, \tilde{f})_1$ unchanged, while it changes $\mu(\tilde{f})_1$ by ± 1. By

the formula, it changes $e(\nu\tilde{f})$ by ∓ 2. Thus, if $\tilde{f}$ and $\tilde{f}'$ are generic immersions homotopic to f, then $e(\nu\tilde{f}) = e(\nu\tilde{f}')$ if and only if $\mu(\tilde{f})_1 = \mu(\tilde{f}')_1$. Hence, (4) follows from (3). □

3 Secondary embedding obstructions

The Whitney trick implies that every map $F\colon S^n \to M^{2n}$ is homotopic to an embedding whenever M is a simply connected $2n$–dimensional manifold and $n > 2$. In order to prove the failure of the Whitney trick in dimension 4, Kervaire and Milnor [1961] devised an obstruction that gave counterexamples to the above statement for $n = 2$. They showed that the homotopy class of $3\cdot\mathbb{CP}^1$ is not represented by an embedded sphere in $\mathbb{CP}^2$. In a smooth, oriented, closed 4–manifold M, consider the formula

$$\theta(c) := \tfrac{1}{8}(c\cdot c - \sigma(M)) \mod 2, \tag{3-1}$$

where the $\mathbb{Z}/2$–reduction of $c \in H_2(M;\mathbb{Z})$ is Poincaré dual to $w_2(M)$ and $\sigma(M)$ is the signature of the intersection form $(x, y) \mapsto x\cdot y$ on $H_2(M;\mathbb{Z})$. In this setting, if c is represented by an embedded sphere, then $\theta(c) = 0$. Recall that, for a unimodular form ℓ and a *characteristic* element c, ie one satisfying $\ell(c, x) \equiv \ell(x, x) \bmod 2$, the difference $\ell(c, c) - \sigma(\ell)$ is always divisible by 8. The condition on c being dual to $w_2(M)$ is stronger than being characteristic for the intersection form since the mod 2 intersection condition holds for all $x \in H_2(M;\mathbb{Z}/2)$, not just for integral homology classes. For example, if M is the Enriques surface (double covered by the $K3$ surface), then $\theta(0) \neq 0$, so 0 cannot be dual to $w_2(M)$, even though the intersection form on $H_2(M;\mathbb{Z})$ is even.

For the proof that θ is an embedding obstruction, Kervaire and Milnor added $1 - [F]\cdot[F]$ copies of $(\mathbb{CP}^2, \mathbb{CP}^1)$ to a proposed characteristic pair $(M, F\colon S^2 \hookrightarrow M)$, with F assumed to be an embedding, to obtain an embedded sphere with self-intersection number 1. Then they blow down that characteristic sphere to arrive at a spin manifold M' with $\sigma(M') = \sigma(M) + (1 - [F]\cdot[F]) - 1 = \sigma(M) - [F]\cdot[F]$. Rokhlin's theorem [1952] — that the signature of a smooth, closed, spin 4–manifold is divisible by 16 — is equivalent to the original condition $\theta([F]) = 0$ in M.

The Kervaire–Milnor result also has consequences for spin manifolds, where it says that any (characteristic) homology class $c = 2b$ that is represented by an embedded sphere must satisfy $b\cdot b \equiv 0 \bmod 4$. For example, $2\Delta \in \pi_2(S^2\times S^2)$ is not represented by an embedding for Δ the diagonal 2–sphere.

For about a decade, it remained an open problem to find a combinatorial formula for $\theta(c)$ in terms of geometric representatives for c.

3.1 Combinatorial formulas: Rokhlin's Arf invariant

Rokhlin [1972] picked an embedded representative $F\colon \Sigma \hookrightarrow M$ for $c \in H_2(M;\mathbb{Z})$ as above and assumed that $H_1(M;\mathbb{Z}/2)$ vanishes. Any simple closed curve r in (the image of) F then bounds a compact surface R in M. The reader should think of R as an "unoriented cap" and check that it has a relative Euler

number, just like a Whitney disc or an ordinary cap. Rokhlin then asserted that setting $q_F(r) := |\mathrm{Int}\, R \pitchfork F|$ for R with vanishing relative Euler number defines a quadratic enhancement

$$q_F \colon H_1(\Sigma; \mathbb{Z}/2) \to \mathbb{Z}/2$$

that refines the mod 2 intersection form on Σ. Independence from the choice of R follows from F being dual to $w_2(M)$, in this setting using intersections of F with all classes of the form $[R \cup R'] \in H_2(M; \mathbb{Z}/2)$. Rokhlin stated that the Arf invariant $\mathrm{Arf}(q_F)$ is equal to $\theta(c) = \theta([F])$. A nice consequence of this equality is that $\mathrm{Arf}(q_F) = \theta(c)$ vanishes whenever c can be represented by an embedded sphere, because q_F is then defined on the zero vector space.

3.2 Combinatorial formulas: Freedman and Kirby's characteristic bordism

Using the same definitions, Freedman and Kirby [1978] proved Rokhlin's claims from above, on their way to a geometric proof of Rokhlin's original theorem. They worked with an arbitrary smooth, closed, oriented 4–manifold M, but before computing q_F they performed surgery on circles in M to arrange that $H_1(M; \mathbb{Z}/2) = 0$; alternatively, they could have made M simply connected and used discs for R, ie ordinary caps. They showed that $\mathrm{Arf}(q_F)$ is invariant under "characteristic bordism", implying independence from the choice of surgeries, as well as establishing the equality $\mathrm{Arf}(q_F) = \theta([F])$ by checking it on the generators of Ω_4^{char}. A different proof of $\mathrm{Arf}(q_F) = \theta([F])$ was given in [Matsumoto 1986].

On a historical note, Freedman and Kirby wrote that they learnt these results from Casson and that they only heard of Rokhlin's results after finishing their paper. The Rokhlin method was extended to nonorientable characteristic surfaces in closed 4–manifolds in [Guillou and Marin 1980; Kirby and Taylor 2001].

3.3 Combinatorial formulas: Matsumoto's t–invariant

Matsumoto [1978], in the same proceedings as [Freedman and Kirby 1978], started with a spherical class $c \in \pi_2(M)$ and represented it by a generic immersion $F \colon S^2 \looparrowright M$ with $2g$ algebraically cancelling double points. He assumed that $H_1(M; \mathbb{Z}) = 0$, using this condition to find "Whitney surfaces", ie oriented surfaces $R_1, \dots, R_g$ bounded by pairs of Whitney arcs in F. Again there is a relative Euler number and we may assume that every R_i has vanishing relative Euler number. Matsumoto proved that, if $[F] \in H_2(M; \mathbb{Z})$ is characteristic, then

$$\mathrm{Arf}(q_F) = \sum_{i=1}^{g} |\mathrm{Int}\, R_i \pitchfork F| =: t(F) \in \mathbb{Z}/2 \tag{3-2}$$

by adding g tubes based at pairs of double points of F to turn it into an embedding of a surface Σ of genus g, where q_F is the quadratic enhancement defined above. The new surface has pairs of framed caps (D_i, R_i), where D_i is a meridional disc of the i^{th} tube and hence has one interior intersection with

F, so $q_F(\partial D_i) = 1$. Since the boundaries of these caps form a hyperbolic basis of $H_1(\Sigma; \mathbb{Z}/2)$, the result follows from the usual formula

$$\operatorname{Arf}(q_F) = \sum_{i=1}^{g} q_F(\partial D_i) \cdot q_F(\partial R_i) = \sum_{i=1}^{g} q_F(\partial R_i) = \sum_{i=1}^{g} |\operatorname{Int} R_i \pitchfork F|.$$

3.4 Summary of the secondary embedding obstructions from the 1970s

Given $[F] \in \pi_2(M)$ such that its Hurewicz image in $H_2(M; \mathbb{Z}/2)$ is Poincaré dual to $w_2(M)$, the above results show that

$$\theta([F]) = \operatorname{Arf}(q_F) = t(F) \in \mathbb{Z}/2$$

is an obstruction to representing $[F]$ by an embedding $F\colon S^2 \hookrightarrow M$. Note that θ only depends on the homology class $h([F]) \in H_2(M; \mathbb{Z})$ by definition, whereas that is not clear for the other two invariants.

An attractive aspect of Matsumoto's $t(F)$ is that it can be computed combinatorially from a generic immersion $F\colon S^2 \looparrowright M$. One argues directly that $t(F)$ is an obstruction to representing $[F] \in \pi_2(M)$ by an embedded sphere and independence of the choice of R_i comes from $[F]$ being characteristic.

Matsumoto's formula was extended in a number of ways. For example, in recent work of Kasprowski, Land, Powell and Teichner [Kasprowski et al. 2017; 2021b; 2020] on the stable diffeomorphism classification of spin 4–manifolds, a version of Matsumoto's t–invariant was used to compute the relevant Arf invariant. We describe further extensions presently.

It follows from topological transversality [Freedman and Quinn 1990, Section 9.5] that, in a smooth, closed, oriented 4–manifold M, the quantity $\theta(c)$ is also an obstruction to representing an element c as before by a topological — ie locally flat — embedding $F\colon S^2 \hookrightarrow M$. If M is not smooth, one adds the Kirby–Siebenmann invariant and then the formula

$$\theta_{\mathrm{TOP}}(c) := \theta(c) + \operatorname{ks}(M)$$

defines such an obstruction; see [Conant et al. 2012a, Introduction] for details. For example, it follows that the generator of $\pi_2(*\mathbb{C}\mathbb{P}^2)$ is not represented by an embedding. Historically speaking, these applications were not known at the time of publication of [Rokhlin 1972; Freedman and Kirby 1978; Matsumoto 1978].

In the following, we will return to considering topological manifolds and obstructions to topological embeddings.

3.5 Secondary obstructions to embedding genus zero surfaces with dual spheres

If Σ is a union of discs or spheres and $F\colon (\Sigma, \partial\Sigma) \looparrowright (M, \partial M)$ has algebraically dual spheres, then Freedman and Quinn [1990, Definition 10.8A] gave a version of Matsumoto's t–invariant, calling it the Kervaire–Milnor invariant. Rather than restricting $H_1(M; \mathbb{Z})$ as in the discussions above, they assumed that $\mu(F) = 0$, ie that all double points of F can be paired by Whitney discs. They used the same

formula as in (3-2), but counted intersections with F^ω, restricting to the Whitney discs in a convenient collection $\mathcal{W}^\omega$ that pair double points of F^ω. They claimed that this mod 2 count, $t(F^\omega, \mathcal{W}^\omega)$ from Definition 1.5, is a secondary obstruction to representing F by an embedding. However, this is only true if F^ω is r–characteristic (Definition 5.5), as Stong's correction [1994] showed. Stong noticed that the choice of sheets for double points whose group elements have order 2 is related to immersed $\mathbb{RP}^2$s in M. If F is dual to $w_2(M)$ then F is also r–characteristic, but not vice versa, so this obstruction is more generally defined than $\theta([F])$.

The embedding theorem for unions of discs and spheres [Freedman and Quinn 1990, Theorem 10.5], as corrected by Stong, says, in our notation, that, for good fundamental group $\pi_1(M)$, such an F is homotopic to a topological embedding if and only if there exists a convenient collection of Whitney discs $\mathcal{W}^\omega$ for the double points of F^ω such that $t(F^\omega, \mathcal{W}^\omega) = 0$. We give more details about the Freedman–Quinn–Stong embedding result in Section 5.

3.6 Secondary obstructions to embedding unions of spheres

Matsumoto's invariant $t(F)$ from (3-2) was extended to a secondary embedding obstruction in [Schneiderman and Teichner 2001] for $F = \{f_i\}_{i=1}^m$, not assuming dual spheres, where each $f_i : S^2 \looparrowright M$ is a generic immersion and assuming $\mu(F) = 0$ and that M is oriented. By counting interior intersections of F with a convenient collection $\mathcal{W}$ of Whitney discs pairing the double points of F, and remembering group elements, signs and components of F, the authors defined an intersection count $\tau(F, \mathcal{W}) \in \mathcal{T}(\pi_1(M), m)$. Here $\mathcal{T}(\pi_1(M), m)$ is the abelian group given by the direct sum of $m + \binom{m}{2} + \binom{m}{3}$ copies of $\mathbb{Z}[\pi_1(M) \times \pi_1(M)]$. To obtain a secondary embedding obstruction, Schneiderman and Teichner [2001, Section 8] gave a list of relations such that the subgroup $\mathcal{R}(M, F) \leq \mathcal{T}(\pi_1(M), m)$ generated by these relations has the property that

$$\tau(f_1, \dots, f_m) = \tau(F) := [\tau(F, \mathcal{W})] \in \mathcal{T}(\pi_1(M), m)/\mathcal{R}(M, F)$$

does not depend on the choice of convenient collection $\mathcal{W}$. In our current language, the main result of that paper is that $\tau(F) = 0$ if and only if $\mathrm{km}(F) = 0$ as in Definition 1.4. In the absence of dual spheres, $\tau(F) = 0$ does not imply that F is homotopic to an embedding. For example, there are obstructions from higher-order Whitney towers.

If F is r–characteristic then the augmentation map $\mathcal{E} \colon \mathcal{T}(\pi_1(M), m) \to \mathbb{Z}/2$, summing all possible coefficients, takes $\mathcal{R}(M, F)$ to zero and $\tau(F)$ to Matsumoto's $t(f_1 \# \cdots \# f_m)$, for an arbitrary choice of interior connected sum of the $\{f_i\}_{i=1}^m$. Moreover, if F has algebraic dual spheres then $\mathcal{E}$ induces an isomorphism of $\mathcal{T}(\pi_1(M), m)/\mathcal{R}(M, F)$ with either $\mathbb{Z}/2$ or 0, depending on whether F^ω is r–characteristic or not. This gives the relationship to Section 3.5.

3.7 Secondary embedding obstructions for arbitrary compact surfaces

It is likely possible to extend the invariant τ from Section 3.6 to arbitrary immersed compact surfaces, not just spheres. However, determining the analogue of $\mathcal{R}(M, F)$ would be a formidable task. In this

paper we take the first step, namely by defining the right notion of b–characteristic surfaces for which Matsumoto's invariant extends from spheres to a secondary embedding obstruction for arbitrary compact surfaces. We also generalise the work of Freedman, Quinn and Stong to all compact surfaces in the presence of algebraically dual spheres.

Recall from Definition 1.4 that, for $F\colon (\Sigma, \partial\Sigma) \looparrowright (M, \partial M)$ as in Convention 1.1, by definition the *Kervaire–Milnor invariant* $\mathrm{km}(F) \in \mathbb{Z}/2$ vanishes if and only if, after finitely many finger moves on the interior of F, taking F to some F', there is a convenient collection of Whitney discs, with interiors disjoint from F', pairing all the double points of F'.

The finger moves in this definition are relevant because finger moves can add relations to the fundamental group $\pi_1(M \smallsetminus F)$, making it easier to find (Whitney) discs in the complement of F.

We could have allowed arbitrary regular homotopies, from F to F', in the definition of km. However, this is not needed, as the following result shows. Note that a nonregular homotopy can change $\mathrm{km}(F)$; see Corollary 6.3.

Proposition 3.1 *Let Σ and M be as in Convention 1.1. If $F_1, F_2\colon \Sigma \looparrowright M$ are regularly homotopic generic immersions, then $\mathrm{km}(F_1) = \mathrm{km}(F_2) \in \mathbb{Z}/2$.*

Proof To show that $\mathrm{km}(F_1) = \mathrm{km}(F_2)$, by symmetry it suffices to show that $\mathrm{km}(F_1) = 0$ implies $\mathrm{km}(F_2) = 0$. Suppose that $\mathrm{km}(F_1) = 0$, and let F_1' be obtained from F_1 by finger moves such that the intersections of F_1' can be paired up by Whitney discs $\{W_i\}$ as in Definition 1.4. Since F_1' and F_2 are regularly homotopic, there is a generic immersion F_3 such that F_3 can be obtained from both F_1' and F_2 by finger moves and ambient isotopies. Since F_3 is obtained from F_1' by finger moves and ambient isotopies and the finger moves can be assumed to be disjoint from $\{W_i\}$, all the double points of F_3 can also be paired up by Whitney discs with interiors disjoint from F_3, as in Definition 1.4. Since F_3 is obtained from F_2 by finger moves, by taking $F_2' := F_3$ it follows that $\mathrm{km}(F_2) = 0$. □

Definition 1.4 is optimised for the proof of Theorem 1.2, as we will see shortly, but is difficult to use in practice. In particular, while one may fortuitously detect specific finger moves and Whitney discs to show $\mathrm{km}(F) = 0$, without a combinatorial description it appears, for a given F, to be hard to prove that the required finger moves from F to some F', together with Whitney discs for F', do not exist. We provide precisely such a combinatorial reformulation in Theorem 1.9, generalising Matsumoto's invariant to our formula for $t(F)$ for b–characteristic F. In the proof of Theorem 1.6 we will show that in the presence of dual spheres this agrees with Definition 1.4.

4 The proof of the surface embedding theorem

The surface embedding theorem (Theorem 1.2) can be deduced using the proof of [Freedman and Quinn 1990, Theorem 10.5(1)], combined with an observation in [Powell et al. 2020, Theorem A and Lemma 6.5] for the condition on the homotopy class of $\overline{G}$, using our definition of the Kervaire–Milnor

invariant (Definition 1.4). Since the surface embedding theorem does not follow directly from the statement of [Freedman and Quinn 1990, Theorem 10.5(1)], as previously discussed, and also since the latter requires a correction by Stong [1994], it can be hard for the uninitiated to piece together a correct proof. Therefore we provide one in this section.

Further, the statement of [Freedman and Quinn 1990, Theorem 10.5(1)] is itself quite complicated, and our version, focussed on the surface embedding problem, may be useful for those looking to apply the technology of Freedman–Quinn without delving into the details.

4.1 Ingredients

The statement of the surface embedding theorem uses the notions of algebraically and geometrically dual spheres. We recall the definitions.

Definition 4.1 Let $F\colon (\Sigma, \partial\Sigma) \looparrowright (M, \partial M)$ be as in Convention 1.1, with components $\{f_i\}_{i=1}^m$.

(1) A collection $G = \{g_i\colon S^2 \looparrowright M\}_{i=1}^m$ of generic immersions is said to be *algebraically dual* to F if $F \sqcup G$ is a generic immersion and $\lambda(f_i, g_j) = [\delta_{ij}] \in \Gamma_{f_i, g_j}$ for all i and j for some choice of basings for F and G.

(2) A collection $\overline{G} = \{\bar{g}_i\colon S^2 \looparrowright M\}_{i=1}^m$ of generic immersions is *geometrically dual* to F if $F \sqcup G$ is a generic immersion and the geometric count of intersections satisfies $|f_i \pitchfork \bar{g}_j| = \delta_{ij}$ for all i and j.

We will need the following lemma, the idea behind which is due to Casson [1986]; see also [Freedman 1982, Section 3]. The formulation we give here is from [Powell et al. 2020, Lemma 5.1].

Lemma 4.2 (Geometric Casson lemma) *Let F and G be transversely intersecting generic immersions of compact surfaces in a connected 4–manifold M. Assume that the intersection points $\{p, q\} \subseteq F \pitchfork G$ are paired by a Whitney disc W. Then there is a regular homotopy from $F \cup G$ to $\overline{F} \cup \overline{G}$ such that $\overline{F} \pitchfork \overline{G} = (F \pitchfork G) \setminus \{p, q\}$. That is, the two paired intersections have been removed. The regular homotopy may create many new self-intersections of F and G; however, these are algebraically cancelling.*

The proof of the surface embedding theorem also relies on Freedman's disc embedding theorem, whose statement we recall.

Theorem 4.3 (Disc embedding theorem [Freedman 1982; Freedman and Quinn 1990; Powell et al. 2020]; see also [Behrens et al. 2021]) *Let M be a connected topological 4–manifold with good fundamental group, and let*

$$F = \{f_i\}_{i=1}^m\colon (D^2 \sqcup \cdots \sqcup D^2, S^1 \sqcup \cdots \sqcup S^1) \looparrowright (M, \partial M)$$

be a generic immersion of finitely many discs. Assume that F has framed algebraically dual spheres $G = \{[g_i]\}_{i=1}^m \subseteq \pi_2(M)$ such that $\lambda(g_i, g_j) = 0 = \mu(g_i)$ for all $i \neq j$. Then there is a flat embedding

$\overline{F} = \{\bar{f}_i\}_{i=1}^m \colon \left(\bigsqcup D^2, \bigsqcup S^1\right) \hookrightarrow (M, \partial M)$, *which is equipped with geometrically dual spheres* $\overline{G} = \{\bar{g}_i\}_{i=1}^m$, *such that* $\overline{F}$ *and* F *have the same framed boundary and* $[\bar{g}_i] = [g_i] \in \pi_2(M)$ *for all* i.

We will also freely use standard constructions such as symmetric contraction, boundary twisting and interior twisting (ie adding local cusps). See [Freedman and Quinn 1990, Chapters 1–2; Powell and Ray 2021a] for further details.

4.2 Proof of the surface embedding theorem

We recall the statement for the convenience of the reader.

Theorem 1.2 (Surface embedding theorem) *Let* $F = \{f_i\}_{i=1}^m \colon (\Sigma, \partial\Sigma) \looparrowright (M, \partial M)$ *be as in Convention 1.1. Suppose that* $\pi_1(M)$ *is good and that* F *has algebraically dual spheres* $G = \{[g_i]\}_{i=1}^m \subseteq \pi_2(M)$. *Then the following statements are equivalent*:

(i) *The self-intersection number* $\mu(F)$ *and the Kervaire–Milnor invariant* $\mathrm{km}(F) \in \mathbb{Z}/2$ *vanish.*

(ii) *There is an embedding* $\overline{F} = \{\bar{f}_i\}_{i=1}^m \colon (\Sigma, \partial\Sigma) \hookrightarrow (M, \partial M)$, *regularly homotopic to* F *relative to* $\partial\Sigma$, *with geometrically dual spheres* $\overline{G} = \{\bar{g}_i \colon S^2 \looparrowright M\}_{i=1}^m$ *such that* $[\bar{g}_i] = [g_i] \in \pi_2(M)$ *for all* i.

Proof The direction (ii) $\Longrightarrow$ (i) follows from the fact that the intersection and self-intersection numbers, as well as the Kervaire–Milnor invariant, are invariant under regular homotopy (relative to the boundary) by Propositions 2.18, 2.25 and 3.1, respectively.

The proof of the direction (i) $\Longrightarrow$ (ii) reduces to the disc embedding theorem (Theorem 4.3) as follows. The argument is similar to the proof of [Freedman and Quinn 1990, Corollary 5.1B] (see also the proof of [Powell et al. 2020, Theorem 8.1]).

Apply the geometric Casson lemma (Lemma 4.2) to upgrade $G = \{g_i\}$ from algebraically to geometrically dual spheres $G' = \{g_i'\}$, changing F to F' by a regular homotopy in the process. The intersection and self-intersection numbers, and the Kervaire–Milnor invariant, vanish for F. So they also vanish for F', since all three quantities are preserved under regular homotopy relative to the boundary by Propositions 2.18, 2.25 and 3.1.

Then, by the definition of the Kervaire–Milnor invariant (Definition 1.4), after further finger moves changing F' to some F'', we can find a convenient collection of Whitney discs $\mathcal{W} = \{W_l\}$ for F'' whose interiors are disjoint from F''. Moreover, F'' and G' are still geometrically dual, since the finger moves may be assumed to miss G'.

We shall apply the disc embedding theorem (Theorem 4.3) to the collection of generically immersed discs $\mathcal{W}$ in the 4–manifold $M \setminus \nu F''$, so we verify that the hypotheses are satisfied. The Whitney discs $\mathcal{W}$ have framed algebraically dual spheres as follows. The Clifford tori at the double points of F'' are geometrically dual to $\mathcal{W}$. Symmetrically contract half of these tori, one per Whitney disc, using meridional discs for F'' tubed into the geometrically dual spheres $G' = \{g_i'\}$. The resulting spheres

are only algebraically dual to $\mathcal{W}$ since the components of $\mathcal{W}$ and G' may intersect arbitrarily; however, they have vanishing intersection and self-intersection numbers since they were produced by symmetric contraction. They are also framed, as we argue briefly now. If a sphere g_i in G' is not framed, then the symmetric contraction uses incorrectly framed caps. However, in the symmetric contraction process, each cap is used twice, with opposite orientations, and so any framing discrepancies cancel out. Since F'' has geometrically dual spheres, $\pi_1(M \setminus \nu F'') \cong \pi_1(M)$ and is thus good. This verifies the hypotheses of the disc embedding theorem (Theorem 4.3) for $\mathcal{W}$, as desired.

Apply the disc embedding theorem to the Whitney discs $\mathcal{W}$ in $M \setminus \nu F''$ to obtain disjointly embedded, flat, framed Whitney discs $\{\overline{W}_l\}$ for the double points of F'', with interiors still disjoint from F'', along with a collection of geometrically dual spheres for the $\{\overline{W}_l\}$ in $M \setminus \nu F''$.

Tube any intersections of G' with $\{\overline{W}_l\}$ into the geometrically dual spheres for $\{\overline{W}_l\}$, giving a new collection of spheres $\overline{G} = \{\bar{g}_i\}$ disjoint from $\{\overline{W}_l\}$. Now we have that the interiors of $\{\overline{W}_l\}$ lie in the complement of $F'' \cup \overline{G}$, and moreover F'' and $\overline{G}$ are geometrically dual. Perform Whitney moves on F'' along $\{\overline{W}_l\}$ to arrive at an embedding $\overline{F}$, as claimed. By construction, $\overline{F}$ and $\overline{G}$ are geometrically dual. That $[\bar{g}_i] = [g_i] \in \pi_2(M)$ for each i follows from [Powell et al. 2020, Lemma 6.5]. □

4.3 The π_1–negligible surface embedding theorem

Recall that a map $F\colon X \to Y$ is called *π_1–negligible* if the inclusion $Y \setminus F(X) \subseteq Y$ induces an isomorphism on π_1 for all basepoints. Here is a reformulation of the surface embedding theorem.

Corollary 4.4 (The π_1–negligible surface embedding theorem) *Let $F\colon (\Sigma, \partial\Sigma) \looparrowright (M, \partial M)$ be as in Convention 1.1. Suppose that F is π_1–negligible and that $\pi_1(M)$ is good. Then $\mu(F)$ and the Kervaire–Milnor invariant of F both vanish if and only if there exists a π_1–negligible embedding $\overline{F}\colon (\Sigma, \partial\Sigma) \hookrightarrow (M, \partial M)$ regularly homotopic to F, relative to the boundary.*

This corollary follows from the surface embedding theorem and the fact that a generic immersion $F\colon \Sigma \looparrowright M$ is π_1–negligible if and only if F admits geometrically dual spheres, which can be seen as follows. For the forward direction, the meridional circles are null-homotopic in M, so by π_1–negligibility they are null-homotopic in $M \setminus F(\Sigma)$. The union of null-homotopies with meridional discs gives geometrically dual spheres. For the reverse direction, first note that, by general position, the homomorphism $\pi_1(M \setminus F(\Sigma)) \to \pi_1(M)$ is surjective. The kernel is normally generated by a collection consisting of one meridional circle for each connected component of Σ. Since geometrically dual spheres provide null-homotopies for these meridians, the assertion follows. By the geometric Casson lemma (Lemma 4.2), the map F in the statement of the surface embedding theorem (Theorem 1.2) is regularly homotopic to a π_1–negligible map, due to the existence of the algebraically dual spheres G. Indeed, this is the first step of the proof of the surface embedding theorem. Note that Theorem 1.2 also controls the homotopy class of the dual spheres, and so is slightly stronger than Corollary 4.4.

5 Band characteristic maps and the combinatorial formula

In this section we define b–characteristic surfaces (Definition 5.17) and motivate the combinatorial formula for the Kervaire–Milnor invariant (Definition 1.5). We postpone many of the proofs to Section 7. We hope this will help the reader to assimilate the overall structure more easily. We work towards the definition of b–characteristic surfaces by first defining the related notions of s–characteristic and r–characteristic surfaces, mirroring the historical development. These latter definitions are simpler to state and serve to motivate the more complicated definition of b–characteristic surfaces.

A sphere $g\colon S^2 \looparrowright M$ in a topological 4–manifold M is said to be *twisted* if the Euler number of the normal bundle is odd. We say g is *framed* if the normal bundle is trivial. To coincide with the usual meaning of framed, one can also implicitly choose a trivialisation, although we will not make use of such a choice. Observe that, if the normal bundle of g has even Euler number, then it is homotopic to a generically immersed sphere with trivial normal bundle, via adding local cusps.

Definition 5.1 Let $F = \{f_1, \dots, f_m\}\colon (\Sigma, \partial\Sigma) \looparrowright (M, \partial M)$ be as in Convention 1.1. We define $\Sigma^\omega \subseteq \Sigma$ so that, for $i \in \{1, \dots, m\}$, the component $\Sigma_i \subseteq \Sigma^\omega$ if and only if there is no framed immersed sphere g_i with $\lambda(f_j, g_i) = \delta_{ij}$ for all $j = 1, \dots, m$. Then we use

$$F^\omega = \{f_i^\omega\colon \Sigma_i^\omega \to M\}$$

to denote the restriction of F to Σ^ω. Note that, if an f_i does not admit an algebraically dual sphere at all, then it belongs to F^ω.

Recall that $x \in H_2(M, \partial M; \mathbb{Z}/2)$ is said to be *characteristic* if $x\cdot a = a\cdot a \in \mathbb{Z}/2$ for every $a \in H_2(M; \mathbb{Z}/2)$, where $-\cdot-$ denotes the intersection pairing $H_2(M; \mathbb{Z}/2) \times H_2(M, \partial M; \mathbb{Z}/2) \to \mathbb{Z}/2$. The next definition gives a weaker notion.

Definition 5.2 Let $F\colon (\Sigma, \partial\Sigma) \looparrowright (M, \partial M)$ be as in Convention 1.1. The map F is called *spherically characteristic* (or *s–characteristic* for short) if $F \cdot a = a \cdot a \in \mathbb{Z}/2$ for all $a \in \pi_2(M)$, considered as an element of $H_2(M; \mathbb{Z}/2)$.

We will show in Lemma 5.18 that b–characteristic maps are s–characteristic.

Lemma 5.3 *Let $F\colon (\Sigma, \partial\Sigma) \looparrowright (M, \partial M)$ be as in Convention 1.1.*

(i) *If F is s–characteristic, then $F = F^\omega$.*

(ii) *If F has algebraically dual spheres, then F^ω is s–characteristic or empty.*

Proof To prove (i), suppose F is not equal to F^ω; then there exists a component Σ_i of Σ with a framed dual sphere g_i, ie with $\lambda(f_j, g_i) = \delta_{ij}$ for all $j \neq i$. This leads to the contradiction

$$1 = f_i \cdot g_i = f_i \cdot g_i + \sum_{j \neq i} f_j \cdot g_i = F \cdot g_i = g_i \cdot g_i = 0 \in \mathbb{Z}/2,$$

where the second-to-last equality follows from F being s–characteristic.

To prove (ii), suppose that F has algebraically dual spheres. Note that the dual spheres for $F^{\omega} \subseteq F$ are necessarily twisted. Assume that F^{ω} neither s–characteristic nor empty. Then there exists $a \in \pi_2(M)$ such that $F^{\omega} \cdot a \neq a \cdot a \in \mathbb{Z}/2$. By tubing into a dual sphere to a component of F^{ω} if necessary, we can assume that a is untwisted, ie $a \cdot a$ is zero and that $F^{\omega} \cdot a = 1$. Choose some component f_j of F^{ω} such that $a \pitchfork f_j$ is nonempty. Except for one of the intersections between f_j and a, tube all the intersections of a and F^{ω} into the corresponding dual spheres to F^{ω}. Call the resulting sphere a'_j. Since $F^{\omega} \cdot a = 1$, we tubed into an even number of dual spheres, so $a'_j \cdot a'_j = a \cdot a = 0 \in \mathbb{Z}/2$. Via adding local cusps, we may assume that a'_j is framed. We also have that $\lambda(f_i, a'_j) = \delta_{ij}$ for all i and j. This contradicts the definition of F^{ω}. □

Recall that a *convenient* collection of Whitney discs $\mathcal{W}$ for the intersections within F consists of framed, generically immersed Whitney discs with interiors transverse to F, and with disjointly embedded boundaries. Recall the invariant t from Definition 1.5 appearing in Theorem 1.6, where $t(F, \mathcal{W})$ is the mod 2 count of transverse intersections between F and the interiors of the Whitney discs in $\mathcal{W}$.

If F is not s–characteristic, then we can change $t(F, \mathcal{W})$ as follows. Given $a \in \pi_2(M)$ with $F \cdot a$ odd but $a \cdot a$ even, one can tube a framed Whitney disc W for F into a framed representative $\tilde{a}\colon S^2 \looparrowright M$, keeping the new Whitney disc framed but adding an odd number of interior intersections with F. If $F \cdot a$ is even and $a \cdot a$ is odd, one can tube W into a representative $\tilde{a}$ and also add an odd number of boundary twists to keep the new Whitney disc framed but again adding an odd number of interior intersections with F. Using Lemma 5.3, this is one reason for the appearance of F^{ω} in the following statements.

The following lemma is also used in the proof of Theorem 1.6, and shows that the vanishing of t for a given collection of Whitney discs implies the vanishing of km.

Lemma 5.4 *Let $F\colon (\Sigma, \partial\Sigma) \looparrowright (M, \partial M)$ be as in Convention 1.1. Suppose that F admits algebraically dual spheres, and that all double points of F are paired by a convenient collection $\mathcal{W}$ of Whitney discs. Let $\mathcal{W}^{\omega} \subseteq \mathcal{W}$ denote the subcollection of Whitney discs for the intersections within F^{ω}, where F^{ω} is as in Definition 5.1. If $t(F^{\omega}, \mathcal{W}^{\omega}) = 0$, then $\mathrm{km}(F) = 0$.*

We wish to find practically verifiable conditions on F that guarantee that $t(F^{\omega}, \mathcal{W}^{\omega})$ is independent of the collection of Whitney discs $\mathcal{W}^{\omega}$. More precisely, the value of $t(F^{\omega}, \mathcal{W}^{\omega})$ should be independent of the pairing of double points, the Whitney arcs joining the paired double points (which includes the choice of sheets at each double point), and finally the Whitney discs. In the case that each Σ_i is simply connected, $t(F^{\omega}, \mathcal{W}^{\omega})$ agrees with [Freedman and Quinn 1990, Definition 10.8A]. However, Freedman and Quinn claim in their Lemma 10.8B that, for simply connected Σ, the quantity $t(F^{\omega}, \mathcal{W}^{\omega})$ only depends on F^{ω}, and not on the Whitney discs, as long as F^{ω} is s–characteristic. This is not true in general, as pointed out and corrected by Stong [1994]. Further, again with $\pi_1(\Sigma_i) = 1$ for all i, Stong established that the value of t does not depend on the choice of $\mathcal{W}^{\omega}$ using the notion of *r–characteristic* discs and spheres. Here is our generalisation of his notion.

Definition 5.5 Let Σ and M be as in Convention 1.1. A map $F\colon(\Sigma,\partial\Sigma)\to(M,\partial M)$ is called $\mathbb{RP}^2$–*characteristic* (or simply *r–characteristic*) if $F\cdot R = R\cdot R\in\mathbb{Z}/2$ for every map $R\colon\mathbb{RP}^2\to M$ satisfying $R^*w_1(M)=0$.

Remark 5.6 A map $c\colon\mathbb{RP}^2\to S^2$ of odd degree (eg a collapse map) composed with elements of $\pi_2(M)$ can be used to show that r–characteristic maps are s–characteristic. Indeed, given F and $a\in\pi_2(M)$, we obtain $a\circ c\colon\mathbb{RP}^2\to M$, and we have $0=F\cdot(a\circ c)+(a\circ c)\cdot(a\circ c)=F\cdot a+a\cdot a\in\mathbb{Z}/2$, where the second equality uses that c has odd degree.

Stong's key observation [1994] was that, in some instances involving elements of order two in $\pi_1(M)$, one can change the choice of sheets at two double points of F^ω, and hence the Whitney arcs and corresponding Whitney disc, with a resulting change in the value of $t(F^\omega,\mathcal{W}^\omega)$ by one. The restriction to r–characteristic maps removes this source of indeterminacy. To summarise, Stong showed the following theorem.

Theorem 5.7 [Stong 1994] *Let $F\colon(\Sigma,\partial\Sigma)\looparrowright(M,\partial M)$ be as in Convention 1.1, with Σ a union of discs or spheres. Suppose $\mu(F)=0$ and that F admits algebraically dual spheres. If F^ω is not r–characteristic, then* $\mathrm{km}(F)=0$, *and if F^ω is r–characteristic, then* $\mathrm{km}(F)=t(F^\omega,\mathcal{W}^\omega)$ *for any choice of Whitney discs $\mathcal{W}^\omega$ for the intersections within F^ω.*

Remark 5.8 Combining Theorems 2.32 and 5.7 with Theorem 1.2 gives the complete answer to the embedding problem for spheres and discs with algebraically dual spheres for good fundamental groups, due to Freedman, Quinn and Stong. Theorem 2.32 allows one to fix the regular homotopy class of generic immersions within the homotopy class to be that with $\mu(-)_1=0$, Theorem 5.7 computes the Kervaire–Milnor invariant, and then one applies Theorem 1.2 to conclude whether or not there is a regular homotopy to an embedding.

Our contribution in the present paper extends this solution to the case that the components of Σ are not all simply connected. In this case there is a further source of indeterminacy coming from the choice of Whitney arcs on Σ. For this reason we need a stronger restriction on F^ω.

Definition 5.9 A *band* refers to either of the two D^1–bundles over S^1, ie a band is either an annulus or a Möbius band. Let $F\colon(\Sigma,\partial\Sigma)\looparrowright(M,\partial M)$ be as in Convention 1.1. Let $\mathcal{M}_F$ be the mapping cylinder of F. Write $\mathcal{B}(F)\subseteq H_2(M,\Sigma;\mathbb{Z}/2):=H_2(\mathcal{M}_F,\Sigma;\mathbb{Z}/2)$ for the subset of elements of the relative homology group that can be represented by a square

$$\begin{array}{ccc}\partial B & \xrightarrow{\ h\ } & \Sigma\\ \big\downarrow{\scriptstyle \iota} & & \big\downarrow{\scriptstyle F}\\ B & \xrightarrow{\ g\ } & M\end{array}$$

where B is a band and $\iota\colon\partial B\hookrightarrow B$ is the inclusion such that

$$\langle w_1(M),g(C)\rangle+\langle w_1(\Sigma),h(\partial B)\rangle=0\in\mathbb{Z}/2, \tag{5-1}$$

where C is the core curve of B.

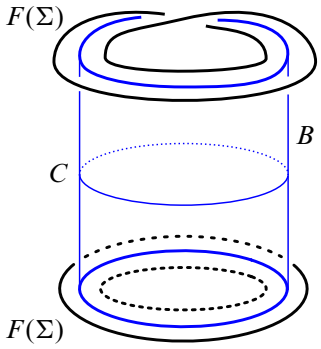


Figure 5: An annular band B (blue) is shown with boundary on F (black). One of the boundary components of B is nonorientable on F and one is orientable, so $\langle w_1(\Sigma), h(\partial B)\rangle = 1$. Therefore, in order for this to be an element of $\mathcal{B}(F)$, we must have $\langle w_1(M), g(C)\rangle = 1$, where M is the ambient 4–manifold and C is the core curve of the annulus, shown in blue.

See Figure 5 for an example of a band. Note that every element of $\mathcal{B}(F)$ can be represented by a generic immersion of pairs $(B, \partial B) \looparrowright (M, \Sigma)$ (Definition 2.6). Writing $H_2(M, \Sigma; \mathbb{Z}/2)$ in place of $H_2(\mathcal{M}_F, \Sigma; \mathbb{Z}/2)$ is a slight but standard abuse of notation. The pair (g, h) induces a relative homology class since the map

$$g \sqcup (h \times \mathrm{Id}_{[0,1]})\colon B \sqcup (\partial B \times [0,1]) \to M \sqcup (\Sigma \times [0,1])$$

descends to a map $\mathcal{M}_\iota \to \mathcal{M}_F$. The mapping cylinder $\mathcal{M}_\iota$ is homeomorphic to B, and so we obtain a map $(B, \partial B) \to (\mathcal{M}_F, \Sigma)$. The image of the relative fundamental class $[B, \partial B]$ in $H_2(\mathcal{M}_F, \Sigma; \mathbb{Z}/2)$ is an element of $\mathcal{B}(F)$. From now on, since $\mathcal{M}_F \simeq M$, to simplify the notation we will not mention the mapping cylinder and refer to $\mathcal{B}(F) \subseteq H_2(M, \Sigma; \mathbb{Z}/2)$.

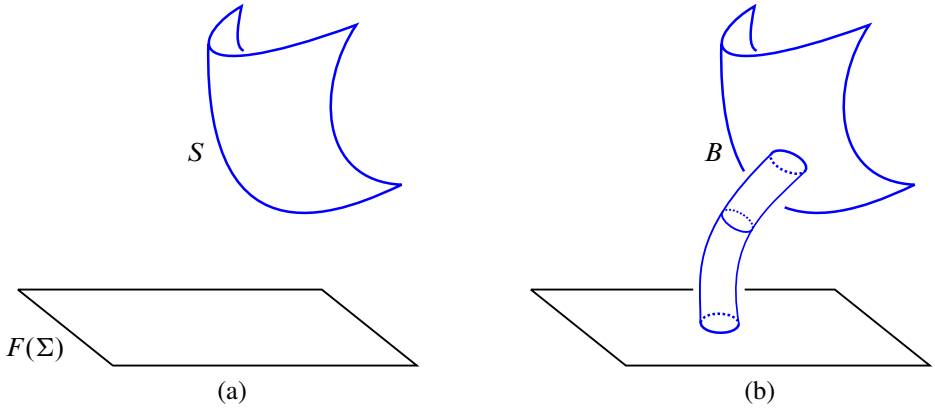


Figure 6: (a) An immersed surface S (blue) in the ambient manifold M, and the image (black) of a generic immersion $F\colon (\Sigma, \partial\Sigma) \looparrowright (M, \partial M)$. (b) A thin tube is added, with one boundary component on Σ and one on S. The surface B (blue) is obtained by cutting out a disc on S and gluing in the tube. Note that, compared to S, the surface B has a new boundary component lying on Σ.

We will see in Lemma 7.7 that, given a generic immersion $F\colon(\Sigma,\partial\Sigma)\looparrowright(M,\partial M)$ as in Convention 1.1, every element of $H_2(M,\Sigma;\mathbb{Z}/2)$ can be represented by an immersion of some compact surface S into M, with interior transverse to F, and with boundary generically immersed in $F(\Sigma)$ away from the double points. The subset $\mathcal{B}(F)$ consists of those homology classes for which S can be chosen to be a band satisfying condition (5-1).

We use the notation

$$\partial\colon H_2(M,\Sigma;\mathbb{Z}/2)\to H_1(\Sigma;\mathbb{Z}/2)$$

for the connecting homomorphism from the long exact sequence of the pair. A class represented by a compact surface S is mapped to its boundary ∂S under the map ∂. Then $\partial\mathcal{B}(F)\subseteq H_1(\Sigma;\mathbb{Z}/2)$ consists of (the homology classes of) those closed 1–manifolds immersed in Σ whose images under F bound bands in M satisfying (5-1).

Construction 5.10 *Given a generic immersion $F\colon(\Sigma,\partial\Sigma)\looparrowright(M,\partial M)$ as in Convention 1.1, suppose we have a generically immersed surface S in M with boundary on Σ, ie admitting maps satisfying*

$$\begin{array}{ccc}\partial S & \xrightarrow{h} & \Sigma\\ \big\downarrow{\scriptstyle \iota} & & \big\downarrow{\scriptstyle F}\\ S & \xrightarrow{g} & M\end{array}$$

where possibly ∂S is empty. Then the tubing procedure shown in Figure 6 can be used to create a band, as follows. If S is a disc, the procedure gives an annulus B with boundary lying on Σ. This annulus satisfies (5-1)*, and therefore lies in $\mathcal{B}(F)$, if and only if $\langle w_1(\Sigma),h(\partial S)\rangle=0\in\mathbb{Z}/2$, since the core of B is null-homotopic in M and the newly created boundary component of B is null-homotopic in Σ.*

In the case that S is a sphere, we can perform the tubing procedure of Figure 6 to S twice. In this case, both boundary components of the annulus created are null-homotopic on Σ, so we always produce an element of $\mathcal{B}(F)$.

Finally, if S is an $\mathbb{RP}^2$, the tubing procedure creates a Möbius band with boundary on Σ, which lies in $\mathcal{B}(F)$ if and only if $\langle w_1(M),g(\mathbb{RP}^1)\rangle=0\in\mathbb{Z}/2$, where $\mathbb{RP}^1\subseteq\mathbb{RP}^2$.

When defining b–characteristic surfaces, we will restrict to the case that the $\mathbb{Z}/2$–valued intersection form λ_Σ is trivial on $\partial\mathcal{B}(F)$. We can restrict in this way because of the following lemma.

Lemma 5.11 *Let $F\colon(\Sigma,\partial\Sigma)\looparrowright(M,\partial M)$ be as in Convention 1.1, with $\mu(F)=0$. If the $\mathbb{Z}/2$–valued intersection form λ_Σ on $H_1(\Sigma;\mathbb{Z}/2)$ is nontrivial on $\partial\mathcal{B}(F)$, then we can change F by a regular homotopy to F' such that there are convenient collections of Whitney discs $\mathcal{W}$ and $\mathcal{W}'$ for the double points of F and F', respectively, such that $t(F,\mathcal{W})\neq t(F',\mathcal{W}')$.*

Moreover, if F has dual spheres and the $\mathbb{Z}/2$–valued intersection form λ_{Σ^o} on $H_1(\Sigma^o;\mathbb{Z}/2)$ is nontrivial on $\partial\mathcal{B}(F^o)$ then $\operatorname{km}(F)=0$.

In the case that $\lambda_\Sigma|_{\partial\mathcal{B}(F)}$ is trivial, we define an invariant Θ on the set $\mathcal{B}(F)$. We will need the following notions:

(1) For Z a closed 1–manifold generically immersed in Σ, the self-intersection number $\mu_\Sigma(Z) \in \mathbb{Z}/2$ of Z counts the number of double points, which we assume without loss of generality to be disjoint from the double points of F. As usual, this is not invariant under homotopies of Z in Σ, only under regular homotopies.

(2) Let S be a compact surface, with a generic immersion of pairs $(S, \partial S) \looparrowright (M, \Sigma)$. Suppose that $w_1(\Sigma)$ is trivial on each component of ∂S, eg if Σ is orientable. Then the normal bundle of ∂S in Σ is trivial and we can pick a nowhere-vanishing section (if S is closed, this is an empty choice). Extend this section to the normal bundle of S in M (such a normal bundle exists by Corollary 2.7) such that the extension is transverse to the zero section. Then we define the *Euler number* $e(S)$ to be the number of zeros of this section modulo 2. Observe that this is analogous to how one measures the twisting of a Whitney disc with respect to the Whitney framing. For S a closed surface, this coincides with the usual definition of the Euler number.

Here is the definition of $\Theta(S)$ in the case that $w_1(\Sigma)$ is trivial on every component of ∂S.

Definition 5.12 Let $F\colon (\Sigma, \partial\Sigma) \looparrowright (M, \partial M)$ be as in Convention 1.1, with $\mu(F) = 0$. Let A be a choice of Whitney arcs pairing the double points of F. Let S be a compact surface in M with a generic immersion of pairs $(S, \partial S) \looparrowright (M, \Sigma)$ such that ∂S is transverse to A and $w_1(\Sigma)$ is trivial on every component of ∂S. Define

$$\Theta_A(S) := \mu_\Sigma(\partial S) + |\partial S \pitchfork A| + |\operatorname{Int} S \pitchfork F| + e(S) \mod 2. \tag{5-2}$$

For closed S we have $\Theta_A(S) \equiv |\operatorname{Int} S \pitchfork F| + e(S) \bmod 2$, and thus $\Theta_A(S)$ vanishes for all closed S if and only if F is characteristic in the traditional sense. In the proof of Theorem 1.6, we will only use the definition of Θ_A for bands. But the case of general surfaces will be useful for our proof that, in the cases relevant to us, Θ_A does not depend on the choice of A (see Lemma 5.16 below).

Remark 5.13 Definition 5.12 suffices in the case of orientable Σ. The reader only interested in this case may safely skip ahead to Lemma 5.16.

If a component of ∂S is orientation-reversing in Σ, then its normal bundle in Σ is nontrivial and hence we may not use it to choose a nowhere-vanishing section of the normal bundle of S on its boundary as before to define the Euler number. However, bands with such boundaries may exist in the ambient 4–manifold and must be considered. When $w_1(\Sigma)$ is nontrivial on precisely one component of ∂S, eg when ∂S is connected, we know $\lambda_\Sigma(\partial S, \partial S) = 1$. Then, by Lemma 5.11, if a band with such boundary exists, then $\operatorname{km}(F) = 0$, and there is no need to define Θ. In particular, note that this means that we need not consider the case of a Möbius band whose boundary consists of a curve on which $w_1(\Sigma)$ is nontrivial. There is one final case of relevant bands left to consider, which we do next.

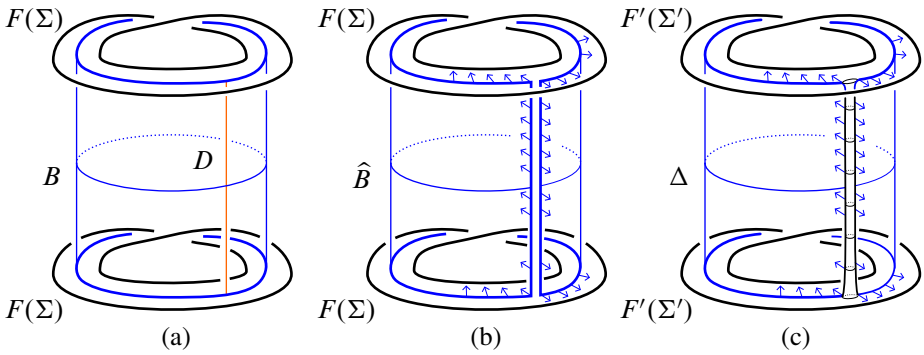


Figure 7: (a) A band B (blue) is shown for F (black). Both boundary components of B are nonorientable on Σ. A vertical arc D is shown in orange. (b) By cutting along the arc D we obtain a disc $\widehat{B}$. We show how to choose a nowhere-vanishing section of the normal bundle of $\partial\widehat{B}$ in M. (c) Adding a tube to F guided by D splits the band into a disc Δ, and changes Σ to a surface Σ'. We show how to choose a section of the normal bundle of $\partial\Delta$ in F'.

Definition 5.14 Let $F\colon(\Sigma,\partial\Sigma)\looparrowright(M,\partial M)$ be as in Convention 1.1, with $\mu(F)=0$. Let A be a choice of Whitney arcs pairing the double points of F. Let B be an annulus with a generic immersion of pairs $(B,\partial B)\looparrowright(M,\Sigma)$, such that ∂B transverse to A and $w_1(\Sigma)$ is nontrivial on both components of ∂B. Pick an embedded arc D in B connecting the two boundary components, disjoint from the self-intersections of B and the intersections of B with F. Let $\widehat{B}$ be the result of cutting B open along D. There is a canonical quotient map $\gamma\colon\widehat{B}\to B$.

Let ν_B^M denote the normal bundle of B in M. Pick a nowhere-vanishing section of $\gamma^*\nu_B^M$ on $\partial\widehat{B}$ as follows. On each part of $\partial\widehat{B}$ that maps to ∂B, use the normal bundle of ∂B in F to define the section locally. For this we require that, on ∂D, the two vectors for the two components agree up to multiplication by ±1. On the part of $\partial\widehat{B}$ that maps to D, pick a section so that, on every pair of points that map to the same point in D, the vectors agree up to multiplication by ±1. See Figure 7(b). We define

$$\Theta_A(B,D) := \mu_\Sigma(\partial B) + |\partial B \pitchfork A| + |\operatorname{Int} B \pitchfork F| + e(\widehat{B}) \mod 2. \tag{5-3}$$

Remark 5.15 An alternative definition of $\Theta_A(B,D)$ would use the arc D to add a tube to $F(\Sigma)$ in such a way that the tube intersects the band B in two parallel copies of D. More precisely, we perform an ambient surgery on $F(\Sigma)$. This requires choosing a 2–dimensional subbundle of the normal bundle of D in M — the tube itself consists of the circle bundle for this subbundle, considered within a tubular neighbourhood of D. We build the required subbundle by first choosing a section lying in the normal bundle of D in B, denoted by ν_D^B. The second section can be chosen freely. Let F' denote the immersion constructed by the tubing procedure. Observe that the domain of F', denoted by Σ', is obtained from the abstract surface Σ by adding a 1–handle. Depending on the choice of the second section above, this may be an orientation-reversing or orientation-preserving 1–handle, but this will not matter for us.

Adding the tube changes the band B to a disc Δ, by removing a thin strip neighbourhood of D. The disc Δ has boundary lying on Σ'. Observe that $\partial\Delta$ is orientation-preserving on Σ', since it is the result of banding together two orientation-reversing curves. Since no new intersection points were added, the collection A is a collection of Whitney arcs for the intersection points of F'. As a result, we may define

$$\Theta_A(B, D) := \Theta_A(\Delta),$$

where the latter is computed as in Definition 5.12. In order to see that this agrees with Definition 5.14, we need only check that the definition of the Euler number terms agree, assuming we choose the tube to be thin enough to miss any double points. For this compare Figure 7(b)–(c) to see that the sections at the boundary in both cases are the same and hence also the Euler numbers coincide. When comparing the pictures, the choice of the second section of the 2–dimensional subbundle which determines the tube corresponds to the choice of the section of $\gamma^*\nu_B^M$ on the part of $\partial\widehat{B}$ that maps to D.

Note that we did not prove that $e(\widehat{B})$ is independent of the choice of D. This will follow from the upcoming proof of Lemma 5.16(ii) below, which states that $\Theta_A(B, D)$ depends only on the homology class of B. The following lemma, whose proof is again deferred to Section 7, shows that Θ is well defined in all required cases. As a reminder, the case of orientable Σ does not require the notion of $\Theta_A(B, D)$ from Definition 5.14, so parts (ii) and (iii) of the following lemma may be skipped by anyone only interested in that case.

Lemma 5.16 *Let $F\colon (\Sigma, \partial\Sigma) \looparrowright (M, \partial M)$ be as in Convention 1.1, with $\mu(F) = 0$. Let A be a choice of Whitney arcs pairing the double points of F.*

(i) *Let S be a compact surface with a generic immersion of pairs $(S, \partial S) \looparrowright (M, \Sigma)$ such that ∂S is transverse to A and $w_1(\Sigma)$ is trivial on every component of ∂S. Then $\Theta_A(S) \in \mathbb{Z}/2$ depends only on the homology class of S in $H_2(M, \Sigma; \mathbb{Z}/2)$.*

(ii) *Let B be an annulus with a generic immersion of pairs $(B, \partial B) \looparrowright (M, \Sigma)$ such that ∂B is transverse to A and $w_1(\Sigma)$ is nontrivial on both components of ∂B. Pick an embedded arc D in B connecting the components of ∂B and disjoint from all double points. Then $\Theta_A(B, D) \in \mathbb{Z}/2$ depends only on the homology class of B in $H_2(M, \Sigma; \mathbb{Z}/2)$. In particular, $\Theta_A(B, D)$ does not depend on D, so we write $\Theta_A(B)$.*

(iii) *Let S be a surface as in* (i) *and let B be an annulus as in* (ii) *such that $[S] = [B] \in H_2(M, \Sigma; \mathbb{Z}/2)$. Then $\Theta_A(S) = \Theta_A(B) \in \mathbb{Z}/2$.*

(iv) *If $\lambda_\Sigma|_{\partial\mathcal{B}(F)} = 0$, the restriction of Θ_A to $\mathcal{B}(F)$ is independent of the choice of A, giving a well-defined map $\Theta\colon \mathcal{B}(F) \to \mathbb{Z}/2$.*

Finally, we are ready to define the required generalisation of r–characteristic maps, called *b–characteristic* maps.

Definition 5.17 Let $F\colon (\Sigma,\partial\Sigma)\looparrowright(M,\partial M)$ be as in Convention 1.1, with $\mu(F)=0$. We say F is *band characteristic* (or *b–characteristic* for short) if $\lambda_\Sigma|_{\partial\mathcal{B}(F)}=0$ and $\Theta\colon\mathcal{B}(F)\to\mathbb{Z}/2$ is trivial.

Lemma 5.18 *Every b–characteristic map is r–characteristic. Moreover, the two notions agree for unions of discs or spheres.*

Proof Let $F\colon(\Sigma,\partial\Sigma)\looparrowright(M,\partial M)$ be as in Convention 1.1, with $\mu(F)=0$. Let $R\colon\mathbb{RP}^2\to M$ be a generic immersion which is transverse to F and such that $R^*w_1(M)=0$. We apply Construction 5.10. In other words, take a small disc on $F(\Sigma)$ away from $R\pitchfork F$, and tube into the image of R. This creates a Möbius band B with boundary on Σ. Here ∂B is homotopically trivial in Σ, so $\langle w_1(\Sigma),\partial B\rangle=0$. The core C of B corresponds to $\mathbb{RP}^1$ within the original immersed $\mathbb{RP}^2$. Therefore, since $R^*w_1(M)=0$, we have that $\langle w_1(M),C\rangle=0$. So $B\in\mathcal{B}(F)$. Note that $\Theta\colon\mathcal{B}(F)\to\mathbb{Z}/2$ is well defined by Lemma 5.16(iv) since F is b–characteristic. Further, $\Theta(B)=\Theta(R)=F\cdot R+R\cdot R\in\mathbb{Z}/2$ by Lemma 5.16(i) since $[B]=[R]\in H_2(M,\Sigma;\mathbb{Z}/2)$. But this vanishes since F is b–characteristic. Hence, F is r–characteristic.

For the second sentence, suppose that Σ is a union of discs or spheres and is r–characteristic. Let $B\in\mathcal{B}(F)$ be a band. Since Σ is simply connected, the boundary of B is null-homotopic in $F(\Sigma)$. Therefore, B can be closed up using a codimension zero submanifold of Σ to either a sphere or an $\mathbb{RP}^2$ immersed in M. The resulting closed surface R again satisfies $\Theta(B)=F\cdot R+R\cdot R\in\mathbb{Z}/2$ by Lemma 5.16(i). Here $\lambda_\Sigma|_{\partial\mathcal{B}(F)}=0$ since Σ is simply connected and so $\Theta\colon\mathcal{B}(F)\to\mathbb{Z}/2$ is again well defined by Lemma 5.16(iv). Once again, since null-homotopic circles on Σ must be orientation-preserving, $\langle w_1(\Sigma),\partial B\rangle=0$ and so (5-1) implies that $R^*w_1(M)=0$. So $\Theta(B)=0$ since F is r–characteristic. Thus, F is b–characteristic. It follows that the notions of b–characteristic and r–characteristic coincide, as claimed. □

Recall that, if F^ω is b–characteristic, then Theorem 1.6 states that $\mathrm{km}(F)=t(F^\omega,\mathcal{W}^\omega)$, so we have a combinatorial description of $\mathrm{km}(F)$. Moreover, since Θ only depends on the homology class of a band in $H_2(M,\Sigma;\mathbb{Z}/2)$, we can in principle determine whether or not F is b–characteristic by computing Θ on finitely many homology classes. Having said that, as mentioned in the introduction, in practice deciding precisely which homology classes can be represented by maps of bands may be tricky.

Lemma 5.19 *Let $F\colon(\Sigma,\partial\Sigma)\looparrowright(M,\partial M)$ be as in Convention 1.1, with $\mu(F)=0$. If G is regularly homotopic to F and F is b–characteristic, then G is b–characteristic.*

Proof By definition, a regular homotopy can be decomposed into a sequence of ambient isotopies, finger moves and Whitney moves. None of these affect which classes of $H_1(\Sigma;\mathbb{Z}/2)$ bound a band. In particular, $\lambda_\Sigma|_{\partial\mathcal{B}(F)}=\lambda_\Sigma|_{\partial\mathcal{B}(G)}$. Assume that G is not b–characteristic. Then either $\lambda_\Sigma|_{\partial\mathcal{B}(G)}$ is nontrivial or there is a band in $\mathcal{B}(G)$ on which Θ is nonvanishing. In the former case, $\lambda_\Sigma|_{\partial\mathcal{B}(F)}=\lambda_\Sigma|_{\partial\mathcal{B}(G)}$ implies that F is not b–characteristic.

For the latter case, let B in $\mathcal{B}(G)$ be such that $\Theta(B) = 1$. It suffices to show that such a band still exists after an ambient isotopy, a finger move or a Whitney move. This is obvious for ambient isotopy. Recall that Θ only depends on the homology class of the band by Lemma 5.16. Hence, we can assume that the boundary of B is away from the singularity of the finger move. Then we can still consider the band B as a band for the surface after the finger move and Θ is unchanged. The argument in the case of a Whitney move is similar. We first let B undergo a homotopy to arrange that it is disjoint from the boundary of the Whitney disc W along which the Whitney move is performed. Then we can again consider the same band B for the new surface. The Whitney move leaves all terms in the definition of Θ except $|\operatorname{Int} B \pitchfork F|$ unchanged. Since the Whitney move uses two copies of the Whitney disc, the change in $|\operatorname{Int} B \pitchfork F|$ is twice $|\operatorname{Int} B \pitchfork W|$. As Θ takes values in $\mathbb{Z}/2$, $\Theta(B)$ is unchanged, as claimed. Thus, we have a band B in $\mathcal{B}(F)$ with $\Theta(B) = 1$, and so again F is not b–characteristic. We have shown the contrapositive of the desired statement. □

Remark 5.20 As a counterpoint to Lemma 5.19, there exist maps that are homotopic to each other but where one is b–characteristic and the other is not. For example, let Σ be the Klein bottle. Then an embedding $f\colon \Sigma \hookrightarrow \mathbb{R}^4$ must have normal Euler number $e(\nu f) \in \{-4, 0, 4\}$ by [Massey 1969]. It can be verified, as we do presently, that the embeddings with $e(\nu f) = 0$ are precisely those which are b–characteristic. Hence, the b–characteristic notion is not invariant under homotopy.

To see that f is b characteristic if and only if $e(\nu f) = 0$, think of $\Sigma \cong \mathbb{RP}^2 \# \mathbb{RP}^2$, with a corresponding isomorphism $H_1(\Sigma; \mathbb{Z}/2) \cong \mathbb{Z}/2 \oplus \mathbb{Z}/2$. There is a standard embedding of $\mathbb{RP}^2$ in $\mathbb{R}^4$ with normal Euler number ± 2, and there are essentially three ways to take connected sums of these embeddings, realising the three options $e(\nu f) \in \{-4, 0, 4\}$. With $e(\nu f)$ fixed, these embeddings are unique up to regular homotopy by Theorem 2.32.

We explicitly construct the standard embeddings, as follows. Take two disjoint, unlinked, unknotted Möbius bands M_1 and M_2 in $\mathbb{R}^3$, with an $\varepsilon_i \in \{\pm 1\}$ signed half-twist for $i = 1, 2$. Take the boundary connected sum $M_1 \natural M_2$ ambiently to obtain a punctured Klein bottle in $\mathbb{R}^3$ with boundary an unknot. Cap this unknot off with a standard slice disc in $\mathbb{R}^4_{\geq 0}$ to obtain a standard embedding f with normal Euler number $-2(\varepsilon_1 + \varepsilon_2)$. We do not justify the sign, which depends on conventions that are not important for us.

By Lemma 5.19, it suffices to check whether the three standard embeddings above are b–characteristic, which we do next. First one computes that $\lambda_\Sigma|_{\partial \mathcal{B}(f)} = 0$ in all three cases, as follows. We have

$$\mathcal{B}(f) \subseteq H_2(\mathbb{R}^4, \Sigma; \mathbb{Z}/2) \xrightarrow[\partial]{\cong} H_1(\Sigma; \mathbb{Z}/2) \cong \mathbb{Z}/2 \oplus \mathbb{Z}/2.$$

By (5-1), in order to have $B \in \mathcal{B}(f)$, we need $\langle w_1(\Sigma), h(\partial B)\rangle = 0$. Hence, $\partial B \in H_1(\Sigma; \mathbb{Z}/2) \cong \mathbb{Z}/2 \oplus \mathbb{Z}/2$ is either $(0, 0)$ or $(1, 1)$. To see that (x, x) for $x \in \{0, 1\}$ can be realised as the boundary of a band, pick a simple closed curve Z on Σ representing the homology class (x, x) and a generically immersed disc D bounded by Z in $\mathbb{R}^4$. Then add a tube from D to Σ to turn D into an annulus B, using Construction 5.10 (see Figure 6). Thus, $\partial \mathcal{B}(f) = \{(0, 0), (1, 1)\}$, on which λ_Σ vanishes.

Hence, whether or not the given standard embedding of Σ is b–characteristic is decided by $\Theta(B)$, where B is a band with boundary $(1,1) \in H_1(\Sigma;\mathbb{Z}/2)$. For the standard embeddings constructed above, such a band can be constructed explicitly, as follows. Take the core curves of M_1 and M_2, and connect sum them inside $M_1 \natural M_2$. This gives an unknot representing $(1,1)$, which bounds a standard slice disc D in $\mathbb{R}^4_{\leq 0}$. Construction 5.10 converts D to a band B.

Since Θ only depends on the class of a band in $H_2(M,\Sigma;\mathbb{Z}/2)$, we can use the band B from the previous paragraph. We shall compute that $\Theta(B)=0$ for this band if and only if $e(\nu f)=0$, that is, if the embedding of Σ arises from the connected sum of the standard embeddings of $\mathbb{RP}^2$ with opposite normal Euler numbers. The curve ∂B is orientation-preserving on Σ, so we use (5-2) to compute $\Theta(B)$. Most of the terms in this definition are trivial in this case, since we are working with an embedding of Σ and ∂B is itself embedded. Also D has interior disjoint from the image of f, and therefore so does B. Only the relative Euler number term remains, which can be computed from the twists in M_1 and M_2. It follows that $\Theta(B) \equiv \frac{1}{2}(\varepsilon_1+\varepsilon_2) \in \mathbb{Z}/2$, which vanishes if and only if $\varepsilon_1=-\varepsilon_2$, which in turn holds if and only if $e(\nu f)=0$.

6 Homotopy versus regular homotopy

In this short section, we describe Construction 6.1 and apply it to prove Theorem 1.12, which we used in Section 1.4 to compare homotopy and regular homotopy of maps. Note that the results in this section require that the surface Σ from Convention 1.1 is nonorientable.

If we are interested in finding an embedding in a given homotopy class, rather than a regular homotopy class, we may use the construction below to replace a given map by a homotopic map for which the invariant t is trivial. In particular, the construction is applicable in the cases from Theorem 2.32 in which there are infinitely many regular homotopy classes with $\mu(-)_1=0$ in a given homotopy class.

Construction 6.1 *Let $F=\{f_i\}_{i=1}^m\colon(\Sigma,\partial\Sigma)\looparrowright(M,\partial M)$ be as in Convention 1.1, with $\mu(F)=0$. Let $\mathcal{W}$ denote a convenient collection of Whitney discs for the double points of F. Suppose that there exists f_i with $w_1(\Sigma)|_{\ker(f_i)_\bullet}$ nontrivial.*

Let $N\subseteq\Sigma_i$ be a Möbius band with $f_i|_N$ π_1–trivial. In a small disc in N introduce four double points with the same sign by cusp homotopies and call the resulting immersion f_i'. Let F' denote the map given by $\{f_j\}_{j\neq i}\cup\{f_i'\}$. Then there is a convenient collection of Whitney discs $\mathcal{W}'$ for all the double points of F' such that

$$t(F',\mathcal{W}')\equiv t(F,\mathcal{W})+1 \mod 2.$$

While we have created four double points with the same sign, we will use in the proof that the Möbius band is nonorientable to change the sign of two of the double points, in order to then be able to find new Whitney discs.

Proof We will pair up the two new pairs of double points with new Whitney discs. Pick any pair of the four new double points and pair them by arcs in the small disc, as in Definition 2.26, such that the resulting circle is null-homotopic in M. For one of the arcs perform a connected sum in the interior with the core α of N. With this new pair of arcs, the double points have opposite sign, and by our choice of N the resulting Whitney circle bounds a Whitney disc W_1 in M with embedded boundary. By boundary twisting, arrange that W_1 is framed, and, by pushing off, ensure there is no intersection between the boundary of W_1 and the boundaries of the components of $\mathcal{W}$. For this we push the boundary arc of W_1 over the end of the boundary arc for the Whitney disc in $\mathcal{W}$. This way $t(F,\mathcal{W})$ remains unchanged. Do the same for the remaining two new double points in f_i', namely pair them by a Whitney disc W_2, which by definition is a parallel copy of W_1.

Since $\lambda_N(\alpha,\alpha)=1$, the boundaries of W_1 and W_2 intersect an odd number of times. To turn $\mathcal{W}\cup\{W_1,W_2\}$ into a convenient collection of Whitney discs we have to remove any intersections between their boundaries. For such an intersection, push the Whitney arc of W_1 over the end of the Whitney arc of W_2. This will in turn change the number of intersections between the interior of W_1 and f_i' by one mod 2; that is, $|\operatorname{Int} W_1 \pitchfork F| \equiv |\operatorname{Int} W_2 \pitchfork F| + 1 \bmod 2$. Let $\mathcal{W}'$ be the resulting collection of Whitney discs. We have

$$t(F',\mathcal{W}') \equiv t(F,\mathcal{W}) + |\operatorname{Int} W_1 \pitchfork F| + |\operatorname{Int} W_2 \pitchfork F| \equiv t(F,\mathcal{W}) + 1 \mod 2. \qquad \square$$

With the above construction in hand, we can now prove Theorem 1.12 from the introduction.

Theorem 1.12 *Let $F=\{f_i\}_{i=1}^m\colon (\Sigma,\partial\Sigma)\looparrowright(M,\partial M)$ be as in Convention 1.1 with $\mu(F)=0$. Suppose that there is at least one $f_i\in F^\omega$ with $w_1(\Sigma)|_{\ker(f_i)_\bullet}$ nontrivial. Then there exists a generic immersion F' homotopic to F with $\mu(F')=0$, and a convenient collection of Whitney discs $\mathcal{W}'$ such that $t((F')^\omega,(\mathcal{W}')^\omega)=0$. Thus, if F' has algebraically dual spheres, then $\operatorname{km}(F')=0$, and if moreover $\pi_1(M)$ is good, then F' is regularly homotopic, relative to $\partial\Sigma$, to an embedding.*

Proof By the vanishing of the intersection and self-intersection numbers, there is a convenient collection of Whitney discs W for F and therefore for F^ω. If $t(F^\omega,W^\omega)=0$, set $F'=F$. If $t(F^\omega,W^\omega)=1$, use Construction 6.1 to find a generic immersion F' homotopic to F, with $t((F')^\omega,(W')^\omega)=0$. If F' has algebraically dual spheres, then $\operatorname{km}(F')=0$ by Theorem 1.6 since either F' is not b–characteristic or $\operatorname{km}(F')=t((F')^\omega,(W')^\omega)=0$. If in addition $\pi_1(M)$ is good, apply Theorem 1.2 to see that F' is regularly homotopic, relative to $\partial\Sigma$, to an embedding. $\square$

We end this section by giving another pair of applications of Construction 6.1.

Proposition 6.2 *Let $f\colon(\Sigma,\partial\Sigma)\looparrowright(M,\partial M)$ be a generic immersion as in Convention 1.1. Assume that Σ is connected, $\mu(f)=0$ and $w_1(\Sigma)|_{\ker f_\bullet}$ is nontrivial while $f^*(w_1(M))$ is trivial. If f' is a generic immersion homotopic to f, both f and f' are b–characteristic, and $e(\nu f)-e(\nu f')=\pm 8$, then*

$$t(f') \equiv t(f) + 1 \mod 2.$$

Proof Since $f^*(w_1(M))$ is trivial, regular homotopy classes of generic immersions homotopic to f are detected by the Euler number of the normal bundle by Theorem 2.32. Further, since $e(\nu f)-e(\nu f')=\pm 8$, we may add four cusps (of the same sign) to f to obtain a map f'' which is regularly homotopic to f'.

The map f is b–characteristic by assumption, while the map f'' is b–characteristic since f' is, by Lemma 5.19. By Lemma 2.24, $\mu(f)_1 \in \mathbb{Z}/2$, so, by construction, $\mu(f)=\mu(f'')=\mu(f')=0$. So the quantities $t(f)$ and $t(f'')$ are defined and, further, $t(f'')=t(f')$. Apply Construction 6.1 to see that $t(f'') \equiv t(f)+1 \bmod 2$. □

Applying Proposition 6.2 to immersions of $\mathbb{RP}^2$ into $\mathbb{R}^4$, we obtain the following corollary, obstructing generic immersions of $\mathbb{RP}^2$ in $\mathbb{R}^4$ with $e(f) \neq \pm 2 \bmod 16$ from being regularly homotopic to an embedding and thus partially recovering the result, due to Massey [1969], that every embedding of $\mathbb{RP}^2$ in $\mathbb{R}^4$ must have Euler number ± 2. Massey stated the result for smooth embeddings, since he used the G–signature theorem. But the G–signature theorem was later extended to the topological category by [Wall 1970, Chapter 14B], so Massey's result also holds for locally flat embeddings of $\mathbb{RP}^2$ in $\mathbb{R}^4$.

Corollary 6.3 *Let $f\colon \mathbb{RP}^2 \looparrowright \mathbb{R}^4$ be a generic immersion with $\mu(f)=0$. Then $t(f)=0$ if and only if $e(\nu f)=\pm 2 \bmod 16$.*

Proof Recall that there exist embeddings $g_\pm\colon \mathbb{RP}^2 \hookrightarrow \mathbb{R}^4$ with Euler number ± 2. First we prove that $e(\nu f) \equiv 2 \bmod 4$. By Lemma 2.24, we know that $\mu(f)_1 \in \mathbb{Z}/2$. Then $\mu(g_+)=0$, and so $\mu(g_+)_1=0$. Since f is homotopic to g_+ and $\mu(f)=0$, it follows from Theorem 2.32 that $e(\nu f) \equiv e(\nu g_+) \equiv 2 \bmod 4$. (The same argument would have applied with g_-.)

Note that any generic immersion of $\mathbb{RP}^2$ into $\mathbb{R}^4$, and in particular the map f, is b–characteristic since $H_2(\mathbb{R}^4, \mathbb{RP}^2; \mathbb{Z}/2) \cong \mathbb{Z}/2$ and the nontrivial element does not satisfy condition (5-1) in Definition 5.9.

We have $t(g_\pm)=0$ since t vanishes for embeddings. Since $e(\nu f) \equiv 2 \bmod 4$, it differs from one of ± 2 by a multiple of 8. Let $k \in \mathbb{Z}$ be such that $e(\nu f)=\pm 2+8k=e(\nu g_\pm)+8k$. By Proposition 6.2,

$$t(f) \equiv t(g_\pm)+k \equiv k \mod 2.$$

Thus, $t(f)=0$ if and only if k is even, which is the case precisely when $e(\nu f)$ differs from $e(\nu g_\pm)=\pm 2$ by a multiple of 16. □

7 Proofs of statements from Section 5

In this section, we provide the proofs we skipped in Section 5. The following transfer move will be useful for arranging that algebraically cancelling intersection points occur on the same Whitney disc.

Construction 7.1 (Transfer move) *Let Σ and M be as in Convention 1.1 and let $H\colon (\Sigma, \partial\Sigma) \to (M, \partial M)$ be a generic immersion, with components $\{h_i\colon \Sigma_i \to M\}$. Assume the double points within H are paired by a convenient collection $\mathcal{W}$ of Whitney discs.*

Let W_1 and W_2 be components of $\mathcal{W}$ with $\operatorname{Int} W_1 \pitchfork H \neq \varnothing \neq \operatorname{Int} W_2 \pitchfork H$. We can perform three finger moves on H, so that the resulting generic immersion H' has six new double points, paired by three framed, embedded Whitney discs $\{V, U_1, U_2\}$, each of which has two intersections with H', such that the boundaries of $\{V, U_1, U_2\}$ are mutually disjoint and embedded. Moreover, the collection $\mathcal{W}' := \mathcal{W} \cup \{V, U_1, U_2\}$ is a convenient collection of Whitney discs for H' and we have

$$|\operatorname{Int} W_1 \pitchfork H'| = |\operatorname{Int} W_1 \pitchfork H| - 1 \quad \text{and} \quad |\operatorname{Int} W_2 \pitchfork H'| = |\operatorname{Int} W_2 \pitchfork H| - 1.$$

Proof Suppose that W_1 pairs intersections of h_a and h_b while W_2 pairs intersections of h_c and h_d, where repetition within a, b, c and d is allowed. Perform a finger move between h_a and h_c, creating two new double points paired by a corresponding framed, embedded Whitney disc V. Note that the interior of V is disjoint from the image of H. The operation depicted in Figure 8 gives a further regular homotopy, involving a finger move pushing h_e through h_a, and a finger move pushing h_f through h_c. We call the outcome of all three finger moves H'. The procedure creates six new intersections within H' compared with H. The four intersections created by the $h_e - h_a$ and $h_f - h_c$ finger moves are paired by Whitney discs U_1 and U_2. A preliminary version of these are shown in Figure 8(ii); the final versions are those arising after the boundary push-off operations indicated by the bottom panel. Overall, the move transfers an intersection of H with W_1, as well as an intersection of H with W_2, on to V, so that $|\operatorname{Int} V \pitchfork H'| = 2$. By construction, each U_i intersects H' twice. □

Now we prove Lemma 5.4, whose statement we recall.

Lemma 5.4 *Let $F\colon (\Sigma, \partial\Sigma) \looparrowright (M, \partial M)$ be as in Convention 1.1. Suppose that F admits algebraically dual spheres, and that all double points of F are paired by a convenient collection $\mathcal{W}$ of Whitney discs. Let $\mathcal{W}^\circ \subseteq \mathcal{W}$ denote the subcollection of Whitney discs for the intersections within F°, where F° is as in Definition 5.1. If $t(F^\circ, \mathcal{W}^\circ) = 0$, then $\operatorname{km}(F) = 0$.*

Proof By applying the geometric Casson lemma (Lemma 4.2) and Propositions 2.18, 2.25 and 3.1, we may arrange by a regular homotopy that F and G become geometrically dual. By definition,

$$t(F^\circ, \mathcal{W}^\circ) = \sum_{l,i} |\operatorname{Int} W_l^\circ \pitchfork f_i^\circ| = 0 \in \mathbb{Z}/2. \tag{7-1}$$

We modify the collection of Whitney discs, as follows, so that each has an even number of intersections with F°. Since the count in (7-1) is zero, the number of Whitney discs with odd intersection with F° is even, so we may pair them up (arbitrarily). For each such pair, apply Construction 7.1. This changes F by finger moves to some F', whose double points are paired by a convenient collection of Whitney discs $\mathcal{W}' := \{W_l'\}$ such that each element of $\mathcal{W}'$ has an even number of intersections with $(F')^\circ$. Note that the new Whitney discs created by the application of Construction 7.1 have been added to the collection.

For each intersection of some W_l' with $(F')^\circ$, tube W_l' into the corresponding geometrically dual sphere. Note that each sphere being tubed into is necessarily twisted, but since we tube an even number of times,

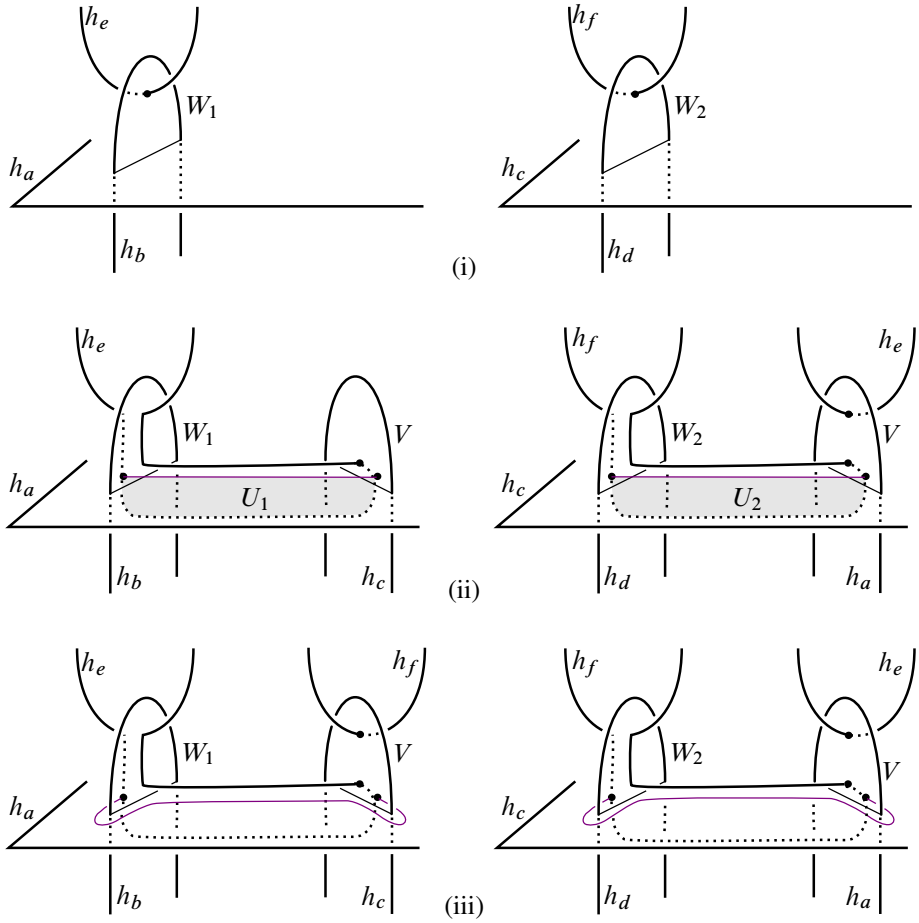


Figure 8: The transfer move. (i) Whitney discs W_1 and W_2 pairing intersections between h_a and h_b, and between h_c and h_d, respectively. (ii) A finger move between h_a and h_c creates a new pair of intersections, paired by a Whitney disc V, shown on both panels. Depicted in the left panel, an intersection between W_1 and h_e is pushed down into h_a and then one of the resulting intersection points is pushed across to V. In the right panel, we see this new intersection between V and h_e. Further, an intersection between W_2 and h_f is pushed down to h_c and one of the resulting intersection points is pushed over to V. These last three moves form a regular homotopy of H, with result H'. Each W_i has one fewer intersection with H' than with H, at the expense of creating two new intersections within H', paired by Whitney discs U_1 and U_2, both shaded grey. Additionally, V has two intersections with H'. The result of a boundary push-off operation making the Whitney arcs for the U_i (purple) disjoint from the Whitney arcs for W_i and V is shown in (iii). This operation creates two intersections of each U_i with H'.

the total change in the framing of W_l' is even. Do this for each element of $\mathcal{W}'$. The resulting family of Whitney discs may still intersect F', but not $(F')^\circ$. For each such intersection with F', again tube into the appropriate geometrically dual sphere. Now the spheres are not twisted, so the framing of the Whitney discs changes by an even number. Arrange for all the Whitney discs to be framed by adding local cusps in

the interior. We may do this because the framing coefficient of each of the Whitney discs is even. We have now produced the desired convenient collection of Whitney discs for the intersections within F', whose interiors lie in the complement of F'. This shows that $\mathrm{km}(F') = 0$, and therefore $\mathrm{km}(F) = 0$, as desired. □

Before giving the proof of Lemma 5.11, we explain the key new construction in this paper, which we already mentioned in Section 1.2. Briefly, given a band B with boundary lying on an immersed surface, a finger move along a fibre of the band produces two new double points paired by a Whitney disc arising from B.

Construction 7.2 (Band fibre finger move) *Let $F\colon (\Sigma, \partial\Sigma) \looparrowright (M, \partial M)$ be as in Convention 1.1. Suppose that $\mu(F) = 0$ and that the self-intersections of F are paired by a convenient collection $\mathcal{W} = \{W_l\}$ of Whitney discs with boundary arcs A.*

Consider $B \in \mathcal{B}(F)$ as a D^1–bundle. Then we can do a finger move along a fibre of B, with endpoints missing A, as depicted in Figure 1. We call this fibre the ***finger arc***, *and denote it by D. We assume that D misses all double points of B and all intersections between* Int B *and F.*

A finger move depends on a choice of 2–dimensional subbundle of the normal bundle to the finger arc (the proof of Lemma 7.3 will give further details). We use a subbundle that lies in the tangent bundle TF at one end of the arc, contains ν_D^B along D, and intersects TF in the line bundle ν_D^B at the other end of the arc. Here ν_D^B is the normal bundle of D in B.

Call the immersion resulting for the above finger move F'. We will check in the next paragraph that the remainder of B, ie the complement in B of a tubular neighbourhood of D, gives a Whitney disc for the new pair of double points. Make the boundary embedded and disjoint from A, by boundary push-off operations, and then boundary twist if necessary, to obtain a framed Whitney disc W_B for the new double points. Then $\mathcal{W}' := \mathcal{W} \cup \{W_B\}$ is a convenient collection of Whitney discs pairing the double points of F'. Note that both F' and W_B depend on the choice of finger arc D and the 2–dimensional subbundle of its normal bundle mentioned above.

Now, as promised, we check that W_B is a Whitney disc. The finger move creates a trivial Whitney disc and we refer to the corresponding Whitney arcs as the trivial arcs. The double points are also paired by the arcs $A_1, A_2 \subseteq \partial B$, where $A_1 \cup A_2 = \partial W_B$. The existence of the disc W_B implies that the group elements of the double points agree with respect to the arcs A_1 and A_2. It remains only to see that the double points have opposite signs with respect to the arcs A_1 and A_2 (see Definition 2.27). The case that both M and Σ are orientable is straightforward. The general case follows from the w_1 condition in the definition of a band, (5-1), *as we now check.*

First we consider the case that B is an annulus. Let $\partial_1 B$ and $\partial_2 B$ denote the two components of ∂B. Then A_1 and A_2 differ from the trivial arcs joining the new double points by $\partial_1 B$ and $\partial_2 B$, respectively. By Definition 2.27, the double points have opposite sign precisely when $\langle w_1(\Sigma), \partial B\rangle + \langle w_1(M), \partial_1 B\rangle = 0$, which matches (5-1) *since $\partial_1 B$ is homotopic in M to the core of B.*

Now suppose that B is a Möbius band. Then the union of A_1 and A_2 and the trivial pair of arcs is the circle $\partial B \subseteq \Sigma$. Moreover, the union of the image of A_1 and either one of the trivial arcs forms a circle in M homotopic to the core C of B. As before, the double points have opposite sign precisely when $\langle w_1(\Sigma), \partial B\rangle + \langle w_1(M), C\rangle = 0$, which again matches (5-1).

The following lemma explains how Construction 7.2 changes the value of t. In the proof we will also carefully explain how to make suitable choices of 2–dimensional subbundles, as required for the finger move in Construction 7.2.

Lemma 7.3 *Let $F\colon (\Sigma, \partial\Sigma) \looparrowright (M, \partial M)$ be as in Convention 1.1.*

(i) *Suppose that F' and $\mathcal{W}'$ are obtained from F and $\mathcal{W}$ by a single application of Construction 7.2 with respect to a band $B \in \mathcal{B}(F)$, where $w_1(\Sigma)$ restricted to every component of ∂B is trivial. Let A denote the Whitney arcs corresponding to $\mathcal{W}$. Then*

$$t(F', \mathcal{W}') = t(F, \mathcal{W}) + \Theta_A(B) \in \mathbb{Z}/2.$$

(ii) *Suppose that F' and $\mathcal{W}'$ are obtained from F and $\mathcal{W}$ by a single application of Construction 7.2 with respect to a band $B \in \mathcal{B}(F)$ and an arc $D \subseteq B$, where B is an annulus with $w_1(\Sigma)$ nontrivial on both boundary components and $D \subseteq B$ connects the two boundary components. Let A denote the Whitney arcs corresponding to $\mathcal{W}$. Then*

$$t(F', \mathcal{W}') = t(F, \mathcal{W}) + \Theta_A(B, D) \in \mathbb{Z}/2.$$

For the proof it will be advantageous to refrain from applying boundary twists and removing intersections involving ∂W_B, and to instead use the following alternative definition of $t(F, \mathcal{W})$, using a slightly weaker restriction on collections of Whitney discs, as in [Stong 1994]; see [Freedman and Quinn 1990, Section 10.8A].

Definition 7.4 Let $F\colon (\Sigma, \partial\Sigma) \looparrowright (M, \partial M)$ be as in Convention 1.1. A *weak* collection of Whitney discs for F is a collection $\mathcal{W}$ of Whitney discs pairing all the double points of F, with generically immersed interiors transverse to F, and with Whitney arcs whose interiors are generically immersed in $F(\Sigma)$ minus the double points of F.

In particular, compared to the definition of a convenient collection of Whitney discs (Definition 2.31), we allow the boundaries of Whitney discs to be generically immersed on Σ and to intersect one another transversely. We also allow the Whitney discs to be twisted, ie for the framing of the normal bundle restricted to the boundary to disagree with the Whitney framing. Each of the discs in a weak collection of Whitney discs admits a normal bundle. The proof is the same as for a generic immersion of pairs, but with a preliminary step that one has to first fix the normal bundle in neighbourhoods of the two double points being paired.

Definition 7.5 Given a weak collection of Whitney discs $\mathcal{W} := \{W_l\}$ for the double points of a generic immersion F as in Convention 1.1, fix an ordering on the indexing set for $\mathcal{W}$ and define

$$t_{\mathrm{alt}}(F, \mathcal{W}) := \sum_l \left(\mu_\Sigma(\partial W_l) + \sum_{k>l} |\partial W_l \pitchfork \partial W_k| + \sum_i |\mathrm{Int}\, W_l \pitchfork f_i| + e(W_l) \right) \in \mathbb{Z}/2,$$

where $e(W_l)$ is the relative Euler number of the normal bundle with respect to the Whitney framing on ∂W_l.

Note that, if $\mathcal{W}$ is a convenient collection of Whitney discs for F, then $t_{\mathrm{alt}}(F, \mathcal{W}) = t(F, \mathcal{W})$ (see Definition 1.5). In particular, since a convenient collection of Whitney discs comprises framed Whitney discs and has embedded and disjoint Whitney arcs, a majority of terms in the definition of t_{alt} vanish. The following lemma shows that t_{alt} can be used as a proxy for t in general.

Lemma 7.6 *Given a weak collection of Whitney discs $\mathcal{W} := \{W_l\}_{l=1}^n$ for the double points of a generic immersion F as in Convention 1.1, there exists a convenient collection of Whitney discs $\mathcal{W}'$ such that $t(F, \mathcal{W}') = t_{\mathrm{alt}}(F, \mathcal{W})$.*

Proof For each Whitney disc W_l with $e(W_l) \neq 0$, add boundary twists to obtain $\overline{W}_l$ with $e(\overline{W}_l) = 0$. Each boundary twist changes the Euler number by ± 1 and introduces an intersection of the Whitney disc with F. We have

$$\sum_i |\mathrm{Int}\, \overline{W}_l \pitchfork f_i| \equiv \sum_i |\mathrm{Int}\, W_l \pitchfork f_i| + e(W_l) \mod 2,$$

and also $\mu_\Sigma(\partial \overline{W}_l) = \mu_\Sigma(\partial W_l)$ and $|\partial \overline{W}_l \pitchfork \partial \overline{W}_k| = |\partial W_l \pitchfork \partial W_k|$ for each $k \neq l$.

Next we will remove intersections between Whitney arcs as well as self-intersections of Whitney arcs, at the expense of adding intersections between F and the Whitney discs. We will use the procedure described in [Powell and Ray 2021a, Section 15.2.3]. For an intersection between $\partial \overline{W}_l$ and $\partial \overline{W}_k$, where possibly $k = l$, the procedure pushes the intersection off one of the endpoints of one of the Whitney arcs of $\partial \overline{W}_l$, ie a double point of F, moving a neighbourhood of $\partial \overline{W}_k$ and creating an intersection between $\overline{W}_k$ and F. This new intersection point is created in a small neighbourhood of the double point of F chosen for the pushing-off procedure. If several Whitney arcs intersect the given Whitney arc of $\partial \overline{W}_l$, push off in order of proximity to the endpoint. This avoids extraneous intersections between Whitney arcs being created. Perform this pushing-off procedure on both arcs of $\partial \overline{W}_l$. For each of the two arcs in $\partial \overline{W}_l$, push towards one of the two double points of F paired by $\overline{W}_l$; choose these double points so that we use one double point for each arc. This ensures that the new intersections between Whitney discs and F arise in disjoint neighbourhoods in the ambient manifold.

Apply the move described in the previous paragraph to the Whitney arcs of $\{\overline{W}_l\}$ in order, beginning with $l = n$. In other words, in the i^{th} step, we push off the intersections of $\partial \overline{W}_k$ with $\partial \overline{W}_{n-i+1}$ for

$k \le n-i+1$. After the n^{th} step, we produce a convenient collection $\mathcal{W}' := \{W_l'\}$, where each W_l' is the result of applying the above procedure to $\overline{W}_l$. This yields, for each l,

$$\begin{aligned}\sum_i |\operatorname{Int} W_l' \pitchfork f_i| &\equiv \sum_i |\operatorname{Int} \overline{W}_l \pitchfork f_i| + \mu_\Sigma(\partial \overline{W}_l) + \sum_{k>l} |\partial \overline{W}_l \pitchfork \partial \overline{W}_k| \mod 2 \\ &\equiv \sum_i |\operatorname{Int} W_l \pitchfork f_i| + e(W_l) + \mu_\Sigma(\partial W_l) + \sum_{k>l} |\partial W_l \pitchfork \partial W_k| \mod 2.\end{aligned}$$

In the above expression, the term $\sum_{k>l} |\partial \overline{W}_l \pitchfork \partial \overline{W}_k|$ arises since the arcs in $\partial \overline{W}_l$ are moved, to create a new intersection point of $\overline{W}_l$ with F, precisely once for each intersection of $\partial \overline{W}_l$ with $\bigcup_{k>l} \partial \overline{W}_k$.

Sum over l to obtain that $t(F, \mathcal{W}') = t_{\text{alt}}(F, \mathcal{W}) \in \mathbb{Z}/2$, as claimed. □

Proof of Lemma 7.3 By Lemma 7.6, it will suffice to show that, in case (i),

$$t_{\text{alt}}(F', \mathcal{W}') = t_{\text{alt}}(F, \mathcal{W}) + \Theta_A(B) \in \mathbb{Z}/2,$$

and, in case (ii),

$$t_{\text{alt}}(F', \mathcal{W}') = t_{\text{alt}}(F, \mathcal{W}) + \Theta_A(B, D) \in \mathbb{Z}/2.$$

The proof splits into three cases.

Case 1 The band B is an annulus as in (i).

Recall that

$$\Theta_A(B) := \mu_\Sigma(\partial B) + |\partial B \pitchfork A| + |\operatorname{Int} B \pitchfork F| + e(B) \mod 2.$$

By the construction of F' and $\mathcal{W}'$, we have

$$t_{\text{alt}}(F', \mathcal{W}') \equiv t(F, \mathcal{W}) + \mu_\Sigma(\partial W_B) + |\partial W_B \pitchfork A| + |\operatorname{Int} W_B \pitchfork F| + e(W_B) \mod 2.$$

Every self-intersection of ∂B and each intersection $\partial B \pitchfork A$ will contribute one intersection of ∂W_B and between ∂W_B and A, respectively, while each intersection $\operatorname{Int} B \pitchfork F$ will contribute one intersection between $\operatorname{Int} W_B$ and F. Thus it remains to show that the framing $e(B)$ appearing in the definition of $\Theta_A(B)$ agrees with the framing $e(W_B)$. For this it will be helpful to pick the finger arc and the 2–dimensional subbundle for its normal bundle needed for the finger move more carefully, which we do next.

Let $\partial_i B$ denote the connected components of $\partial B \subseteq \Sigma$. Consider the following decomposition of the tangent bundle of M restricted to $\partial_i B$:

$$TM|_{\partial_i B} \cong T(\partial_i B) \oplus \nu^{\Sigma}_{\partial_i B} \oplus \nu^{B}_{\partial_i B} \oplus (\nu^{M}_{\Sigma}|_{\partial_i B} \cap \nu^{M}_{B}|_{\partial_i B}). \tag{7-2}$$

As shown in Figure 9, choose a section s of the normal bundle of B that is nonvanishing on the finger arc. In both boundary components $\partial_i B$, we assume that this section lies in $\nu^{\Sigma}_{\partial_i B}$. Now rotate the section near the top boundary component, as shown in Figure 9, middle, to obtain a section s', so that on the top component s' lies in $(\nu^{M}_{\Sigma}|_{\partial_i B} \cap \nu^{M}_{B}|_{\partial_i B})$. For the finger move, by definition, we use the 2–dimensional subbundle of ν^{M}_{D} determined by s' and $T(\partial_i B)$, as shown in Figure 9, right.

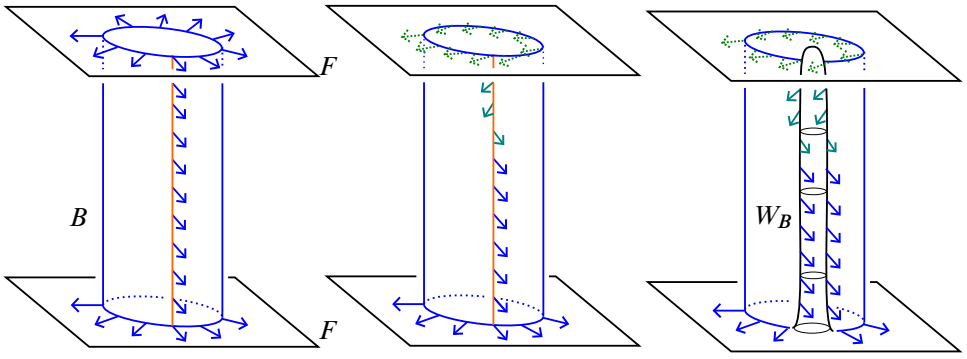


Figure 9: Left: we see a model annulus B (blue) connecting two sheets of F (black), and a finger arc $D = \{\text{pt}\} \times D^1 \subseteq S^1 \times D^1 \cong B$ (orange). We also see the surface framing on ∂B and the section s along the finger arc of the normal bundle of B in M. Recall the section is defined over all of B, but we only show it on a subset. Middle: in the top half of B, we rotate the section so that it lies in $(\nu_\Sigma^M|_{\partial_i B} \cap \nu_B^M|_{\partial_i B})$, ie in the time direction, on the top boundary component (dotted green). The modified section is called s'. Right: after performing the finger move, s' gives the Whitney framing for the new Whitney disc W_B.

Now consider the Whitney disc W_B obtained from B after performing the finger move along D using the above 2–dimensional subbundle of its normal bundle. By definition, $e(W_B)$ equals the number of zeros of $s'|_{W_B}$, since on the boundary Whitney arcs it is normal to one sheet and tangent to the other. On the other hand, the number of zeros of $s'|_{W_B}$ equals the number of zeros of s, since s' was obtained from s by a rotation, and neither section vanishes near the finger arc. Finally, by definition, $e(B)$ counts the zeros of s. Therefore we see that $e(B) = e(W_B)$.

Case 2 The band B is an annulus such that $w_1(\Sigma)$ restricted to both components of ∂B is nontrivial, as in (ii).

Assume that a finger arc D has been chosen. To define $\Theta_A(B, D)$, we pick a section s as in Definition 5.14 on $\widehat{B}$, which is by definition the result of cutting B open along D. Recall that

$$\Theta_A(B, D) := \mu_\Sigma(\partial B) + |\partial B \pitchfork A| + |\text{Int } B \pitchfork F| + e(\widehat{B}) \mod 2.$$

As in Case 1, we need only check that the term $e(\widehat{B})$ in $\Theta_A(B, D)$ agrees with the term $e(W_B)$ in t_{alt}. Again as in Case 1, we rotate the section near the top boundary component to obtain a section s', so that on the top component s' lies in $(\nu_\Sigma^M|_{\partial_i B} \cap \nu_B^M|_{\partial_i B})$. For the finger move, we use the 2–dimensional subbundle determined by s' and $T(\partial_i B)$. Note that, just like s, the section s' is not defined on all of B, but only on $\widehat{B}$; see Figure 10. Nevertheless, on points that map to the same point in D, the section s' agrees up to a sign and thus still determines a 1–dimensional subbundle. The section s' restricts to a section of the normal bundle of the Whitney disc W_B obtained from B. As in Case 1, the quantities $e(B)$ and $e(W_B)$ coincide.

Case 3 The band B is a Möbius band with $w_1(\Sigma)$ restricted to ∂B trivial, as in (i).

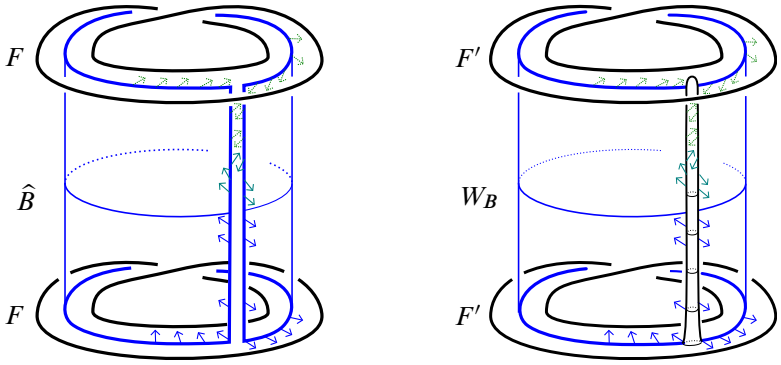
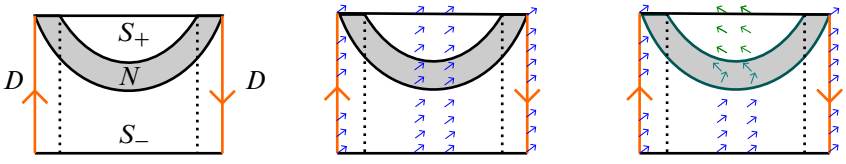


Figure 10: Left: the section s' on D (orange), and on a parallel copy of D; see Figure 7(b). Close to the top boundary the section extends into the time direction (teal and dotted green). Right: the section s' on the boundary of the new Whitney disc W_B.

As in Case 1, we only need to show that the term $e(B)$ in $\Theta_A(B)$ agrees with the term $e(W_B)$ in t_{alt}. Let D denote a properly embedded arc on B along which we wish to perform the finger move. Identify B with the quotient of the square $S := [-1,1]\times[-1,1]$ as usual, ie $(-1,x)\sim(1,-x)$ for all $x\in[-1,1]$, with D corresponding to the arc $\{-1\}\times[-1,1]\equiv\{1\}\times[-1,1]$ (see Figure 11). Pull back the normal bundle ν_B^M of B in M to S via the quotient map $\pi\colon S\to B$ and then pick a trivialisation $\pi^*(\nu_B^M)\cong S\times\mathbb{R}^2$ so that, on the horizontal boundary $H := [-1,1]\times\{-1,1\}$, we have that $\pi^*\nu^\Sigma_{\partial B}$ coincides with $H\times\mathbb{R}\times\{0\}\subseteq S\times\mathbb{R}\times\mathbb{R}$.

Then we pick a section s of ν_B^M such that $s|_{\partial B}$ lies in $\nu^\Sigma_{\partial B}$. Note that $s|_{\partial B}$ is nowhere-vanishing but the section s might have zeros. Without loss of generality we assume that any zeros of s do not lie in the strip $([-1,-1+\varepsilon]\cup[1-\varepsilon,1])\times[-1,1]$ for some $\varepsilon\in\left(0,\frac{1}{4}\right)$.

Next we modify the section s. Choose a "necklace" region, ie a subsquare N with two opposite edges coinciding with $[-1,-1+\varepsilon]\times\{1\}$ and $[1-\varepsilon,1]\times\{1\}$, and otherwise lying in the interior of $S := [-1,1]^2$.



(a) (b) (c)

Figure 11: (a) The necklace region $N\subseteq S$ is shown in grey. The dotted black lines indicate the boundary of the region where the finger move occurs. The solid black lines indicate where the band is attached to the surface F. The arc D is shown in orange. (b) The section s is shown in blue. (c) The section s' is indicated. Note that on S_-, the sections s and s' agree. On S_+, the section s' (green) is obtained by rotating s by 90 degrees. In the necklace region, the section rotates continuously (teal), interpolating between the values on S_+ and S_-. Note that the section on the arc D has not changed, but it has been modified on part of the dotted lines.

We consider the pullback of s to S, where we have a trivialisation. Modify this pullback so that it remains unchanged in the lower component S_- of $S \setminus N$ and is rotated by 90 degrees on the upper component S_+. On the region N, the section rotates continuously, interpolating between the values on S_+ and S_-. Push this forward to get a modified section s' on B. Since the modification is produced by a continuous rotation, the number of zeros of this modification agrees with the number of zeros of the original s.

Recall that we wish to perform a finger move guided by the arc $D = \{-1\} \times [-1, 1]$. Without loss of generality, we assume that the "width" of the finger move is 2ε. More precisely, to perform a finger move we need a 2–dimensional subbundle of ν_D^M. We require that this contains ν_D^B to ensure that the finger move cuts B open into a Whitney disc, as desired. Fix an identification of the total space of the normal bundle ν_D^M with $D \times \mathbb{R}^3$. We choose an embedding $\iota\colon \nu_D^M \hookrightarrow M$ restricting to the inclusion of D on $D \times \{0\}$, with the following properties:

(i) We assume the first $\mathbb{R}^1$ factor of $D \times \mathbb{R}^3$ corresponds to ν_D^B and that ι identifies $D \times \{\pm 1\} \times \{0\} \times \{0\}$ with the arcs $\{-1 + \varepsilon, 1 - \varepsilon\} \times [-1, 1]$. This is what was meant by the width of the finger move.

(ii) We also require that ι identifies $D \times \{t\} \times \{1\} \times \{0\}$ with $s'\big(\iota(D \times \{t\} \times \{0\} \times \{0\})\big)$ for $t \geq 0$, and with $s'\big(\iota(D \times \{t\} \times \{0\} \times \{0\})\big)$ rotated by 90 degrees for $t \leq -1$. (Here we also implicitly identify ν_B^M with its image in M.)

Now do the finger move using $D \times S^1 \times \{0\}$ according to this parametrisation, where S^1 is the unit circle in the $\mathbb{R}^2$ factor, along with a finger tip. The choices above imply that $s'|_{\partial W_B}$ is a Whitney framing, where W_B denotes the new Whitney disc created according to Construction 7.2. Specifically, let F' denote the result of the finger move. By our choice of the 2–dimensional subbundle for the finger move above, the section s' is normal to F' along half of ∂W_B, and tangent along the other half; see Figure 12.

As previously mentioned, we need to check that the relative Euler number $e(B)$ in $\Theta_A(B)$ agrees with the twisting number $e(W_B)$ in t_{alt}. The relative Euler number $e(B)$ is given by the number of zeros of the section s on the interior of B. As mentioned before, this coincides with the number of zeros of the section s'. Since $s'|_{\partial W_B}$ is a Whitney framing, this further coincides with the twisting number $e(W_B)$, as desired, since we assumed there are no zeros of s within the strip $([-1, -1 + \varepsilon] \cup [1 - \varepsilon, 1]) \times [-1, 1] \subseteq B$ used for the finger move. □

Next we prove Lemma 5.11. Here is the statement for the convenience of the reader.

Lemma 5.11 *Let $F\colon (\Sigma, \partial\Sigma) \looparrowright (M, \partial M)$ be as in Convention 1.1, with $\mu(F) = 0$. If the $\mathbb{Z}/2$–valued intersection form λ_Σ on $H_1(\Sigma; \mathbb{Z}/2)$ is nontrivial on $\partial\mathcal{B}(F)$, then we can change F by a regular homotopy to F' such that there are convenient collections of Whitney discs $\mathcal{W}$ and $\mathcal{W}'$ for the double points of F and F', respectively, such that $t(F, \mathcal{W}) \neq t(F', \mathcal{W}')$.*

Moreover, if F has dual spheres and the $\mathbb{Z}/2$–valued intersection form λ_{Σ^ω} on $H_1(\Sigma^\omega; \mathbb{Z}/2)$ is nontrivial on $\partial\mathcal{B}(F^\omega)$, then $\operatorname{km}(F) = 0$.

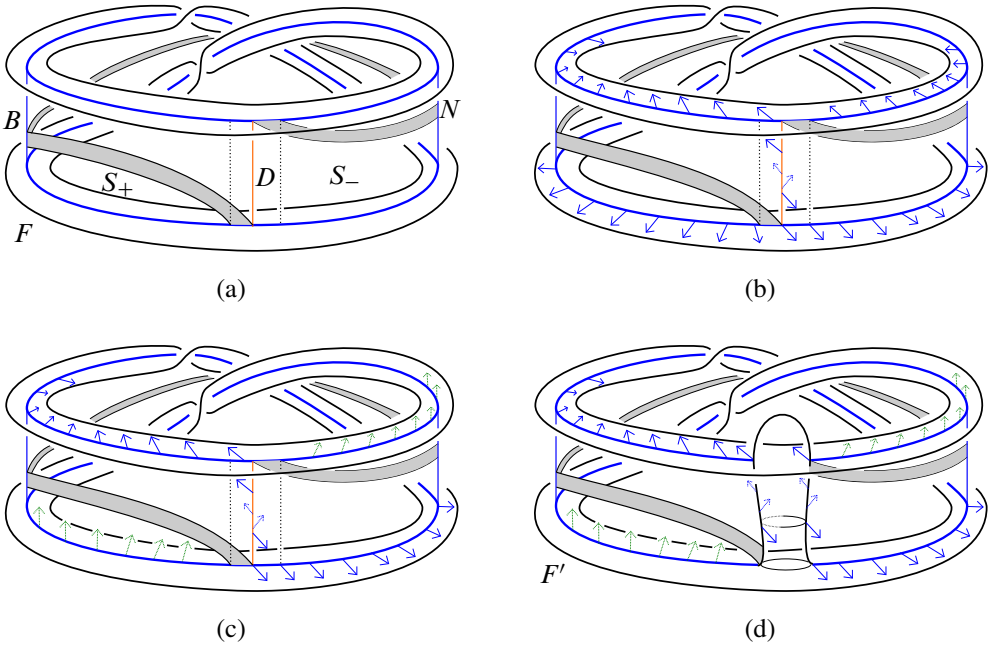


Figure 12: (a) The surface F is shown in black, and the Möbius band B in blue. Note this picture is entirely in $\mathbb{R}^3$. The necklace region N is in grey, and splits B into two components S_+ and S_-. The finger arc D is in orange, and the width of the finger move is shown with dotted lines. (b) We show the section s in blue. Note that, while there is a rotation along D, there are no zeros of s in the strip between the dotted lines. (c) The modified section s'. Note the section coincides with s on S_- and has been rotated (green) on S_+. (d) The section s' on the Whitney disc W_B formed after the band fibre finger move. By the construction of the 2–dimensional subbundle of the normal bundle of D used to guide the finger move, the section s' is tangent to F' along the right edge of the finger (corresponding to the right dotted line in (c), where the finger contains part of a Whitney arc of W_B), and s' is normal to F' along the left edge of the finger.

Proof We first prove the statement (without using dual spheres) about $t(F,\mathcal{W})$ depending on the choice of $\mathcal{W}$ under our assumption. By hypothesis, F is a generic immersion whose double points can be paired by a convenient collection $\mathcal{W}=\{W_l\}$ of Whitney discs (Corollary 2.30). By hypothesis, λ_Σ is nontrivial on $\partial\mathcal{B}(F)$, meaning that there are bands B_1 and B_2 with boundary on $F(\Sigma)$ minus double points such that $\lambda_\Sigma(\partial B_1,\partial B_2)\neq 0\in\mathbb{Z}/2$. Here it is possible that B_2 is a parallel push-off of B_1. Using B_i and Construction 7.2, perform a finger move and obtain a new framed Whitney disc, calling the resulting convenient collection of Whitney discs $\mathcal{W}_i$, for $i=1,2$, and the resulting map F_i

If $t(F_i,\mathcal{W}_i)\neq t(F,\mathcal{W})$ for some $i=1,2$, we can set $F'=F_i$ and $\mathcal{W}'=\mathcal{W}_i$. Otherwise, use Lemma 7.3 twice, for B_1 and B_2 simultaneously, and let $\mathcal{W}'$ denote the resulting convenient collection of Whitney discs for the resulting map F'. Then the change in $t(F,\mathcal{W})$ is as before, except that there is an additional contribution from the odd number of intersections between the boundary arcs for the new Whitney discs coming from B_1 and B_2. Specifically, removing these by pushing one Whitney arc off the end of the

other (as part of Construction 7.2) introduces an odd number of intersections between the Whitney discs and F. Therefore, $t(F', \mathcal{W}') \neq t(F, \mathcal{W})$, as needed.

For the second statement, apply the above argument to the subcollection $\mathcal{W}^\omega$ of $\mathcal{W}$ pairing the intersections within F^ω. It follows that we may find $\mathcal{W}^\omega$ such that $t(F^\omega, \mathcal{W}^\omega) = 0$, possibly after a regular homotopy of F^ω. Then $\mathrm{km}(F) = 0$ follows from Lemma 5.4. □

For the proof of Lemma 5.16, we will need the following Lemmas 7.7, 7.8, 7.9 and 7.10.

Lemma 7.7 *Let $F\colon (\Sigma, \partial\Sigma) \looparrowright (M, \partial M)$ be as in Convention 1.1. Every element of $H_2(M, \Sigma; \mathbb{Z}/2)$ can be represented by an immersion of some compact surface into M, with interior transverse to F, and with boundary generically immersed in $F(\Sigma)$ away from the double points.*

Proof Let $\mathcal{N}_k(M, \Sigma)$ denote the k–dimensional unoriented bordism group over (M, Σ), and let $\mathcal{N}_k$ denote the k–dimensional unoriented bordism group over a point. Using topological transversality, it suffices to show that every element of $H_2(M, \Sigma; \mathbb{Z}/2)$ can be represented by a map $(S, \partial S) \to (M, \Sigma)$ for some surface S. To show this, it suffices to see that the edge homomorphism $\mathcal{N}_2(M, \Sigma) \to H_2(M, \Sigma; \mathbb{Z}/2)$ from the Atiyah–Hirzebruch spectral sequence is onto.

Recall that the $\mathcal{N}_0$ is isomorphic to $\mathbb{Z}/2$ while the $\mathcal{N}_1$ vanishes. It follows that, in the Atiyah–Hirzebruch spectral sequence with E_2–term $H_p(M, \Sigma; \mathcal{N}_q)$ and converging to $\mathcal{N}_{p+q}(M, \Sigma)$, there is no nontrivial differential going out of $H_2(M, \Sigma; \mathcal{N}_0) \cong H_2(M, \Sigma; \mathbb{Z}/2)$; such a differential would have codomain $H_0(M, \Sigma; \mathcal{N}_1) = 0$. Thus the edge homomorphism $\mathcal{N}_2(M, \Sigma) \to H_2(M, \Sigma; \mathbb{Z}/2)$ is onto, as desired. □

Lemma 7.8 *Let $F\colon (\Sigma, \partial\Sigma) \looparrowright (M, \partial M)$ be as in Convention 1.1. Suppose that $\mu(F) = 0$ and let A be a choice of Whitney arcs pairing the double points of F. Then the function Θ_A is quadratic with respect to the $\mathbb{Z}/2$–valued intersection form λ_Σ. That is, let S and S' be compact surfaces, with generic immersions of pairs $(S, \partial S) \looparrowright (M, \Sigma)$ and $(S', \partial S') \looparrowright (M, \Sigma)$ such that ∂S and $\partial S'$ intersect A and each other transversely, and are such that $w_1(\Sigma)$ is trivial on every component of ∂S and $\partial S'$. Then we have*

$$\Theta_A(S \cup S') = \Theta_A(S) + \Theta_A(S') + \lambda_\Sigma(\partial S, \partial S').$$

Proof The term $e(S)$ in Definition 5.12 is defined componentwise and the terms $|\partial S \pitchfork A|$ and $|\mathrm{Int}\, S \pitchfork F|$ are linear in ∂S and S, respectively. Hence, the only term that is not linear in S is $\mu_F(\partial S)$. This term is also quadratic in the sense that

$$\mu_\Sigma(\partial S \cup \partial S') = \mu_\Sigma(\partial S) + \mu_\Sigma(\partial S') + \lambda_\Sigma(\partial S, \partial S'),$$

which proves the lemma. □

Lemma 7.9 *Let $F\colon (\Sigma, \partial\Sigma) \looparrowright (M, \partial M)$ be as in Convention 1.1. Suppose that $\mu(F) = 0$ and let A be a choice of Whitney arcs pairing the double points of F. Let S be a compact surface with a generic immersion of pairs $(S, \partial S) \looparrowright (M, \Sigma)$ such that ∂S is transverse to A and $w_1(\Sigma)$ is trivial*

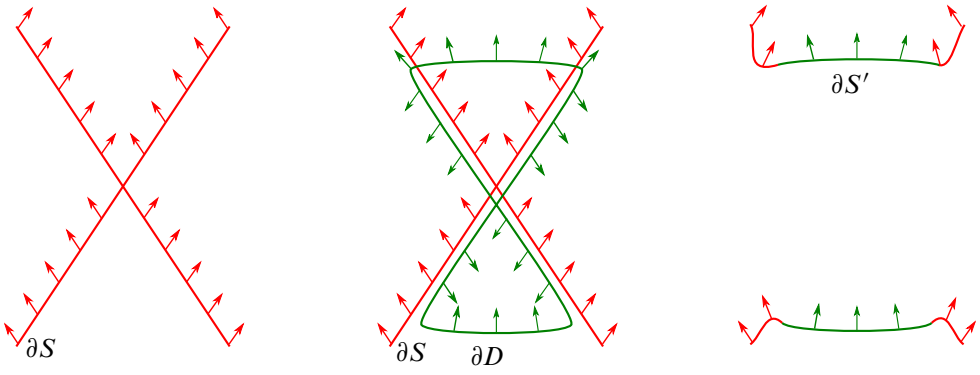


Figure 13: Adding a disc D to S to remove a self-intersection of ∂S. Left: the neighbourhood of a self-intersection of ∂S before adding the disc D. The section γ_S is shown along ∂S. Middle: the boundary of the disc D. The section γ_D is shown along ∂D. Right: after the modification; see Figure 14.

on every component of ∂S. Then there is another such surface S' with a generic immersion of pairs $(S', \partial S') \looparrowright (M, \Sigma)$ with $\partial S'$ transverse to A such that

(1) $[S] = [S'] \in H_2(M, \Sigma; \mathbb{Z}/2)$;

(2) $\Theta_A(S) = \Theta_A(S')$; *and*

(3) $\partial S'$ *is embedded in* Σ.

Proof To start, pick a section γ_S of the normal bundle ν_S^M which on ∂S is nowhere-vanishing and lies in $\nu_{\partial S}^F$ as in the definition of $\Theta_A(S)$.

The idea of the proof is to remove all intersections of ∂S by locally adding a twisted disc D as indicated in Figure 13. More precisely, we add these discs D such that the interiors are disjoint from the interior of F and the boundary is disjoint from A. Then pick a section γ_D of ν_D^M such that, along the aligned (ie parallel) parts of the boundaries, γ_D and γ_S are opposite. Glue D to S along the aligned parts of the boundaries and push this part of the boundary off F as indicated in Figure 14. Each of these local twisted discs has mod 2 Euler number 1, as can be seen from the nontrivial linking in Figure 15. Thus the resulting surface S' has embedded boundary and the mod 2 Euler number of S' differs from that of S by the number of intersections of ∂S modulo two, ie $\mu_\Sigma(\partial S)$. Since we have changed neither the

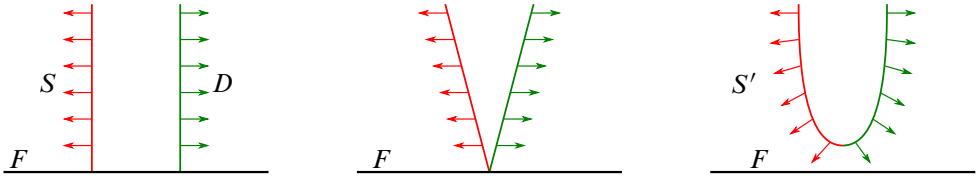


Figure 14: Glue D to S along the aligned parts of the boundaries and push this part of the boundary off F.

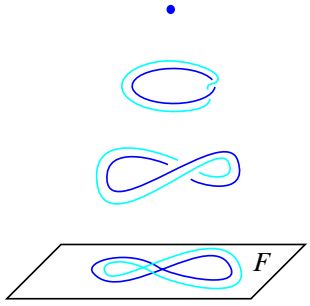


Figure 15: A twisted band with Euler number $+1$ in a movie description. Bottom: an immersed figure-eight curve (blue) is shown lying on the immersed surface $F(\Sigma)$ (black) away from the double points. A framing on the normal bundle on the boundary of the band is shown in light blue. Moving upward/forward in time, we see a simple closed curve shrinking to a point. The push-off corresponding to the framing induced by Σ is shown in light blue. For the twisted band with Euler number -1, we use the other resolution.

number of intersections of the interior with F nor the number of intersections of the boundary with A, we have $\Theta_A(S) = \Theta_A(S') \in \mathbb{Z}/2$. As the local discs are trivial in $H_2(M, \Sigma; \mathbb{Z}/2)$, we furthermore have $[S] = [S'] \in H_2(M, \Sigma; \mathbb{Z}/2)$, as needed. □

Lemma 7.10 *Let $F\colon (\Sigma, \partial\Sigma) \looparrowright (M, \partial M)$ be as in Convention 1.1. Let Z be a disjoint union of embedded circles in Σ. Let $\Sigma \mid Z$ denote Σ cut along Z, ie the completion of $\Sigma \setminus Z$ to a compact manifold with boundary. Let $\mathcal{F} = \{\Sigma_i\}$ be the connected components of $\Sigma \mid Z$ and suppose that $[Z] = 0 \in H_1(\Sigma; \mathbb{Z}/2)$. Then we can pick a subset $\mathcal{F}' \subseteq \mathcal{F}$ such that each component of Z appears exactly once as a connected component of the boundary of precisely one $\Sigma_i \in \mathcal{F}'$.*

Proof Without loss of generality, assume that Σ is connected. Considering the entire collection $\mathcal{F} = \{\Sigma_i\}$, every component of Z would appear as the boundary of precisely two of the Σ_i, since otherwise Z would be nontrivial in $H_1(\Sigma; \mathbb{Z}/2)$. To see this, note that Z can contain homologically essential curves in $H_1(\Sigma; \mathbb{Z}/2)$ provided they cancel. However, none of these can be orientation-reversing curves, since Z is embedded.

The idea of the proof is to take "half" of the components of $\mathcal{F}$. Let $x \in \Sigma \mid Z$ be an arbitrary basepoint away from Z. For each Σ_i, define $p(\Sigma_i) \in \mathbb{Z}/2$ as follows. Pick a point $y \in \operatorname{Int} \Sigma_i$ and a path w in Σ from x to y which is transverse to Z. Define $p(\Sigma_i)$ as the mod 2 intersection number of w and Z.

We show that $p(\Sigma_i)$ is independent of the choices of w and y. If w' is another path from x to y, then the concatenation $w^{-1} \cdot w'$ is a loop in Σ and we have

$$|(w^{-1} \cdot w') \pitchfork Z| = \lambda_\Sigma([w^{-1} \cdot w'], [Z]) = \lambda_\Sigma([w^{-1} \cdot w'], 0) = 0.$$

So $p(\Sigma_i)$ does not depend on the choice of w. Also, since each Σ_i is connected, $p(\Sigma_i)$ does not depend on y. To see this, let $y' \in \operatorname{Int} \Sigma_i$ and choose a path z from y to y' that lies in $\operatorname{Int} \Sigma_i$. Let w' be a path

from x to y' which is further transverse to Z. Then

$$|w \pitchfork Z| = |(w \cdot z) \pitchfork Z| = |w' \pitchfork Z|.$$

The first equation uses that $z \subseteq \Sigma_i$ and the second uses independence of the choice of w. Hence, $p(\Sigma_i) \in \mathbb{Z}/2$ is well defined, as desired.

Now let $\mathcal{F}'$ consist of all the components Σ_i for which $p(\Sigma_i) = 0$. This subset is the one we seek, since, for a fixed component Z_j of Z, the two components of $\mathcal{F}$ containing a cut-open copy of Z_j have different values of p. □

We are now ready for the proof of Lemma 5.16.

Lemma 5.16 *Let $F\colon (\Sigma, \partial\Sigma) \looparrowright (M, \partial M)$ be as in Convention 1.1, with $\mu(F) = 0$. Let A be a choice of Whitney arcs pairing the double points of F.*

(i) *Let S be a compact surface with a generic immersion of pairs $(S, \partial S) \looparrowright (M, \Sigma)$ such that ∂S is transverse to A and $w_1(\Sigma)$ is trivial on every component of ∂S. Then $\Theta_A(S) \in \mathbb{Z}/2$ depends only on the homology class of S in $H_2(M, \Sigma; \mathbb{Z}/2)$.*

(ii) *Let B be an annulus with a generic immersion of pairs $(B, \partial B) \looparrowright (M, \Sigma)$ such that ∂B is transverse to A and $w_1(\Sigma)$ is nontrivial on both components of ∂B. Pick an embedded arc D in B connecting the components of ∂B and disjoint from all double points. Then $\Theta_A(B, D) \in \mathbb{Z}/2$ depends only on the homology class of B in $H_2(M, \Sigma; \mathbb{Z}/2)$. In particular, $\Theta_A(B, D)$ does not depend on D, so we write $\Theta_A(B)$.*

(iii) *Let S be a surface as in* (i) *and let B be an annulus as in* (ii) *such that $[S] = [B] \in H_2(M, \Sigma; \mathbb{Z}/2)$. Then $\Theta_A(S) = \Theta_A(B) \in \mathbb{Z}/2$.*

(iv) *If $\lambda_\Sigma|_{\partial\mathcal{B}(F)} = 0$, the restriction of Θ_A to $\mathcal{B}(F)$ is independent of the choice of A, giving a well-defined map $\Theta\colon \mathcal{B}(F) \to \mathbb{Z}/2$.*

Proof To prove (i), assume that S and S' are immersed compact surfaces, with $w_1(\Sigma)$ trivial on each of the connected components of the boundaries, representing the same element in $H_2(M, \Sigma; \mathbb{Z}/2)$. Modulo isotopy we can assume that S and S' intersect transversely in their interiors in M, and their boundaries intersect transversely on F. In particular, their boundaries ∂S and $\partial S'$ intersect in an even number of points. Hence, $\Theta_A(S \cup S') = \Theta_A(S) + \Theta_A(S')$ by Lemma 7.8. Thus it suffices to show that $\Theta_A(S) = 0$ for a compact surface S such that $0 = [S] \in H_2(M, \Sigma; \mathbb{Z}/2)$ and $w_1(\Sigma)$ is trivial on ∂S. In particular, we know by Lemma 7.7 that every element of $H_2(M, \Sigma; \mathbb{Z}/2)$ — in particular the trivial class — can be represented by an immersed surface S.

By Lemma 7.9, we can assume that ∂S is embedded. As $0 = [S] \in H_2(M, \Sigma; \mathbb{Z}/2)$, we also have that $0 = [\partial S] \in H_1(\Sigma; \mathbb{Z}/2)$ since S maps to ∂S under the map $H_2(M, \Sigma; \mathbb{Z}/2) \to H_1(\Sigma; \mathbb{Z}/2)$. Pick a set $\mathcal{F}'$ of components of $\Sigma \setminus \partial S$ as in Lemma 7.10. Gluing the $F_i \in \mathcal{F}'$ to S along the common boundary, we obtain a closed surface N. First note that N represents the same class as S in $H_2(M, \Sigma; \mathbb{Z}/2)$ since

it only differs by a subset of $F(\Sigma)$. Hence, $0 = [N] \in H_2(M, \Sigma; \mathbb{Z}/2)$. As N is closed, it also defines an element in $H_2(M; \mathbb{Z}/2)$. Note that we have the pair sequence

$$\cdots \to H_2(\Sigma; \mathbb{Z}/2) \xrightarrow{F} H_2(M; \mathbb{Z}/2) \to H_2(M, \Sigma; \mathbb{Z}/2) \to \cdots .$$

Hence, N represents the same class in $H_2(M; \mathbb{Z}/2)$ as a subsurface Σ' of Σ. Let λ_M denote the $\mathbb{Z}/2$–valued intersection form on $H_2(M; \mathbb{Z}/2)$. By hypothesis, we have $\lambda_M(f_j, f_{j'}) = 0$ for any two connected components f_j and $f_{j'}$ of F. Thus,

$$\lambda_M([N], [F]) + \lambda_M([N], [N]) = 0. \tag{7-3}$$

We finish the proof of (i) by showing that $\Theta_A(S) = \lambda_M([N], [F]) + \lambda_M([N], [N])$.

Recall that we were able to assume that ∂S is embedded in $F(\Sigma)$ away from the double points and that $\lambda_M([N], [N]) = e(\nu N) \bmod 2$. We claim that this in turn agrees with $e(S) + \sum_{F_i \in \mathcal{F}'} e(F_i)$. Here we define $e(F_i)$ as follows. We used F to define a nowhere-vanishing section of $\nu_S^M|_{\partial S}$. Since $\nu_S^M|_{\partial S}$ is 2–dimensional, we can pick a linearly independent nonvanishing section. This can be equivalently used for the definition of $e(S)$. But this new section now can also be used to define $e(F_i)$. Combining these vector fields that are transverse to the zero section defines a vector field on the normal bundle of N, and hence computes the Euler number of the normal bundle of N. Thus we have shown

$$\lambda_M([N], [N]) = e(S) + \sum_{F_i \in \mathcal{F}'} e(F_i).$$

Now consider $\lambda_M([N], [F])$. We can use the vector field used for defining $e(F_i)$ to make N and F transverse. Then $\lambda_M([N], [F])$ is given by the sum of $|S \pitchfork F|$, $\sum_{F_i \in \mathcal{F}'} e(F_i)$ and the self-intersection points of F contained in the $F_i \in \mathcal{F}'$. As the self-intersection points of F are paired by the Whitney arcs A, we have that modulo two the number of self-intersection points of F contained inside F_i agrees with $|A \pitchfork \partial F_i|$. Since the boundary of the F_i is precisely ∂S, we have

$$\lambda_M([N], [F]) = |\operatorname{Int} S \pitchfork F| + \sum_{F_i \in \mathcal{F}'} e(F_i) + |A \pitchfork \partial S|.$$

Therefore,

$$\begin{aligned}
\lambda_M([N], [N]) + \lambda_M([N], [F]) &= e(S) + \sum_{F_i \in \mathcal{F}'} e(F_i) + |\operatorname{Int} S \pitchfork F| + \sum_{F_i \in \mathcal{F}'} e(F_i) + |A \pitchfork \partial S| \\
&= e(S) + |\operatorname{Int} S \pitchfork F| + |A \pitchfork \partial S| \\
&= \Theta_A(S) \in \mathbb{Z}/2,
\end{aligned}$$

where the last equality holds because $\mu_S(\partial S) = 0$. Combine this with (7-3) to obtain

$$\Theta_A(S) = \lambda_M([N], [N]) + \lambda_M([N], [F]) = 0.$$

This completes the proof of (i).

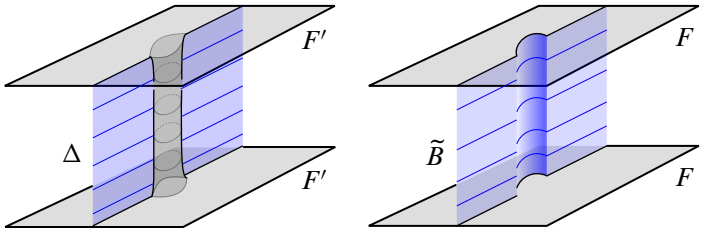


Figure 16: A strip, ie half of the tube, added to Δ.

Before proving (ii) and (iii), we introduce a general construction. Let B be an annulus as in (ii) with an embedded arc D in B connecting its two boundary components. As in Remark 5.15, add a tube to $F(\Sigma)$ along the arc D. Let d, d' denote the two discs removed from F when the tube is added. Adding the tube changes F to some F', an immersion of a surface Σ', and changes B to a disc Δ. As before, observe that $\partial\Delta$ is an orientation-preserving curve in Σ' and A is now a collection of Whitney arcs pairing the double points of F'. By construction, we see that $\Theta_A(B, D) = \Theta_A(\Delta)$.

Moreover, suppose there is either some immersed compact surface S in M as in (i) or some immersed annulus B' as in (ii), with an embedded arc D' on B' connecting its two boundary components, where possibly $B = B'$. We may choose the tube in the above construction thin enough that $\Theta_A(S)$ and $\Theta_A(B', D')$ remain unchanged. In particular, this means we assume, after a small local isotopy, that the discs d and d' do not intersect the boundaries of S and B', so both represent classes in $H_2(M, \Sigma'; \mathbb{Z}/2)$. We have the following claim.

Claim 7.11 *If $[B] = [S] \in H_2(M, \Sigma; \mathbb{Z}/2)$, then either $[\Delta] = [S] \in H_2(M, \Sigma'; \mathbb{Z}/2)$ or $[\Delta] = [S] + [d] \in H_2(M, \Sigma'; \mathbb{Z}/2)$. Similarly, if $[B] = [B'] \in H_2(M, \Sigma; \mathbb{Z}/2)$ then either $[\Delta] = [B'] \in H_2(M, \Sigma'; \mathbb{Z}/2)$ or $[\Delta] = [B'] + [d] \in H_2(M, \Sigma'; \mathbb{Z}/2)$.*

Proof The exact sequence of the triple with $\mathbb{Z}/2$ coefficients yields

$$(\mathbb{Z}/2)^2 \cong H_2(\Sigma, \Sigma \setminus (\mathring{d} \cup \mathring{d}')) \to H_2(M, \Sigma \setminus (\mathring{d} \cup \mathring{d}')) \xrightarrow{j} H_2(M, \Sigma) \to H_1(\Sigma, \Sigma \setminus (\mathring{d} \cup \mathring{d}')) = 0,$$

so j is surjective with kernel generated by the images of $[d]$ and $[d']$ from the left-hand group.

Construct a lift $\widetilde{B}$ of Δ in $H_2(M, \Sigma \setminus (\mathring{d} \cup \mathring{d}'))$ by adding a strip along the added tube to Δ, as shown in Figure 16. Since $\widetilde{B}$, S and B' are mapped by j to B, S and B' in $H_2(M, \Sigma)$, respectively, and the kernel is generated by $[d]$ and $[d']$, we see that the classes of $\widetilde{B}$, S and B' differ at most by the classes $[d]$ and $[d']$. The map $H_2(M, \Sigma \setminus (\mathring{d} \cup \mathring{d}')) \to H_2(M, \Sigma')$ identifies $[d]$ and $[d']$, so the claim follows. □

We continue now to prove (ii). Let B and B' be immersed annuli in M as in the statement of (ii). Choose arcs D in B and D' in B' connecting the boundary components of each, and assume that $[B] = [B']$. By the construction from the proof of Claim 7.11 applied twice, once to B and once to B', we find discs Δ and Δ', coming from B and B', respectively, such that $\Theta_A(B, D) = \Theta_A(\Delta)$ and $\Theta_A(B', D') = \Theta_A(\Delta')$.

From Claim 7.11, applied twice with the roles of B and B' reversed, we see that the classes $[\Delta]$ and $[\Delta']$ satisfy

$$[\Delta] = [B] \quad \text{or} \quad [\Delta] = [B] + [d],$$

and

$$[\Delta'] = [B'] \quad \text{or} \quad [\Delta'] = [B'] + [d].$$

Since also $[B] = [B']$, it follows that either $[\Delta] = [\Delta']$ or $[\Delta] = [\Delta'] + [d]$ in $H_2(M, \Sigma'; \mathbb{Z}/2)$ for Σ' the surface obtained from applying the construction (twice) to Σ.

In the first case, $[\Delta] = [\Delta']$, the proof of (ii) is completed by appealing to (i), which says that $\Theta_A(\Delta) = \Theta_A(\Delta')$, since both Δ and Δ' have $w_1(\Sigma')$ trivial on the boundary. In the second case, $[\Delta] = [\Delta'] + [d]$, we also appeal to (i), but now for the pair of surfaces Δ and $\Delta' \cup d$. So we have that $\Theta_A(\Delta) = \Theta_A(\Delta' \cup d)$. It follows directly from the definition that $\Theta_A(\Delta' \cup d) = \Theta_A(\Delta')$. This completes the proof of (ii). In particular, we have proved that $\Theta_A(B, D)$ does not depend on the choice of arc D.

The proof of (iii) is similar. Suppose we have an immersed annulus B in M as in the statement of (ii), as well as an immersed compact surface S in M as in the statement of (i). Choose an embedded arc $D \subseteq B$ connecting the boundary components. Assume that $[S] = [B] \in H_2(M, \Sigma; \mathbb{Z}/2)$. Apply the previous construction to B, yielding a disc Δ which, by Claim 7.11, satisfies either $[\Delta] = [S]$ or $[\Delta] = [S] + [d]$ in the group $H_2(M, \Sigma'; \mathbb{Z}/2)$ for the surface Σ' obtained from applying the construction to Σ. Further, we know that $\Theta_A(B, D) = \Theta_A(\Delta)$. Now, in the first case, the proof is completed by appealing to (ii), which says that $\Theta_A(\Delta) = \Theta_A(S)$. In the second case, apply (ii) to the pair Δ and $S \cup d$, to see that $\Theta_A(\Delta) = \Theta_A(S \cup d)$. It follows directly from the definition that $\Theta_A(S \cup d) = \Theta_A(S)$.

It remains to prove (iv). Let B denote an element of $\mathcal{B}(F)$. First note that only the term $|\partial B \pitchfork A|$ of $\Theta_A(B)$ depends on the Whitney arcs A. Let A' denote another collection of Whitney arcs. The quantities $\Theta_A(B)$ and $\Theta_{A'}(B)$ differ by $|\partial B \pitchfork A| + |\partial B \pitchfork A'|$, regardless of whether Σ is orientable or nonorientable.

Case 1 The collections of Whitney arcs A and A' correspond to the same choice of pairing up of the double points of F.

For each pair of double points, we can pick Whitney discs W_1 and W_2 with boundary in A and A', respectively. By adding small strips to the union of W_1 and W_2 in the neighbourhood of the double points, we can see that the difference of A and A' is the boundary of some collection of bands B'. For more details about this construction, see the upcoming proof of Theorem 1.6. Then we have

$$|\partial B \pitchfork A| + |\partial B \pitchfork A'| = |\partial B \pitchfork (A \cup A')| = |\partial B \pitchfork \partial B'| = \lambda_\Sigma(\partial B, \partial B') \mod 2,$$

which vanishes by assumption.

Case 2 The collections of Whitney arcs A and A' correspond to a different pairing up of the double points of F.

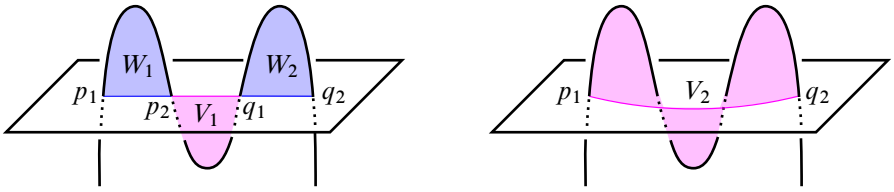


Figure 17: Left: the Whitney discs W_1, W_2 and V_1, pairing up double points as (p_1, p_2), (q_1, p_2) and (q_1, q_2), respectively. Right: the Whitney disc V_2 pairing up (p_1, q_2) is obtained as a union of W_1, W_2 and V_1, by adding small bands at the points p_2 and q_1 to resolve the singularities, and pushing the interiors of the bands into the complement of F. Compare with [Stong 1994, Figure 2].

From A we can construct Whitney arcs A'' so that A' and A'' correspond to the same pairing up of double points, as in Figure 17. Here are the details. We will define the family A'' iteratively, starting with A. Let p_1, p_2, q_1 and q_2 be double points of F. Suppose that arcs in A pair up p_1 and p_2, as well as q_1 and q_2, while arcs in A' pair up p_2 and q_1. Pick Whitney discs W_1 and W_2 with boundary in A. Let V_1 be a Whitney disc for the points p_2 and q_1 with boundary away from A. Then, as indicated in Figure 17, we may choose Whitney arcs, away from the other arcs in A, so that p_2 and q_1 are also paired by a Whitney disc V_2, obtained as a union of W_1, W_2 and V_1. Modify the family A by removing ∂W_1 and ∂W_2, and adding in ∂V_1 and ∂V_2. Comparing this new family with A', we see that we have reduced the number of mismatches in the pairing up of double points of F. Iterate this process and call the result A''.

Looking more closely at the construction in the previous paragraph, observe that, at each step, the family of arcs changes by adding in two parallel copies of the boundary of a Whitney disc V_1. Since intersection points are counted modulo 2, Θ_A and $\Theta_{A''}$ are equal. By Case 1, we know that $\Theta_{A'}$ and $\Theta_{A''}$ are equal when restricted to $\mathcal{B}(F)$. Thus, Θ_A and $\Theta_{A'}$ are equal when restricted to $\mathcal{B}(F)$, as needed. □

8 Proof of Theorems 1.6 and 1.9

First we prove Theorem 1.9 from the introduction, which shows that, for b–characteristic surfaces, $t(F, \mathcal{W}) \in \mathbb{Z}/2$ is well defined, ie independent of the Whitney discs $\mathcal{W}$. Note that the theorem has no assumption about the existence of algebraically dual spheres.

Theorem 1.9 *Let $F\colon (\Sigma, \partial\Sigma) \looparrowright (M, \partial M)$ be as in Convention 1.1 with $\mu(F) = 0$. Let $\mathcal{W}$ be a convenient collection of Whitney discs for the double points of F. Then F is b–characteristic if and only if, for every F' regularly homotopic to F and convenient collection $\mathcal{W}'$ for the double points of F', we have $t(F, \mathcal{W}) = t(F', \mathcal{W}')$.*

For b–characteristic F, we denote the resulting regular homotopy invariant by $t(F) \in \mathbb{Z}/2$. Then, if $\mathrm{km}(F) = 0$ *— eg if F is an embedding — then $t(F) = 0$.*

Proof The final sentence, that $\mathrm{km}(F) = 0$ implies $t(F) = 0$ for b–characteristic F, is an immediate consequence of the definitions.

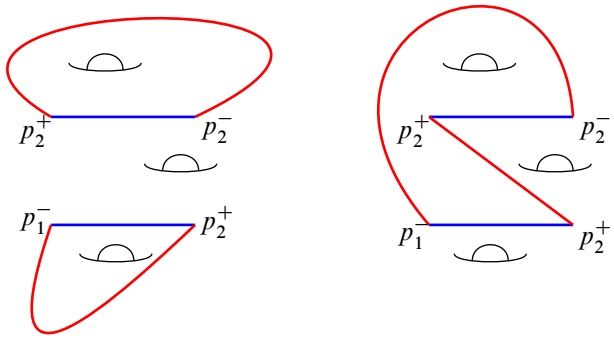


Figure 18: Within Σ we see the preimages $p_1^\pm$ and $p_2^\pm$, for the double points p_1 and p_2 of F, respectively. Blue denotes the Whitney arcs for W_1 while red denotes the Whitney arcs for the new disc V_1. On the left, the choice of sheets stays the same, while it changes on the right. Compare with [Stong 1994, Figure 3].

Now suppose that F is not b–characteristic. Then, by Lemma 5.11, we can assume that $\lambda_\Sigma|_{\partial\mathcal{B}(F)}$ is trivial, which implies that the function Θ is well defined on $\mathcal{B}(F)$. Since we assume that F is not b–characteristic, there exists $B \in \mathcal{B}(F)$ such that $\Theta(B) = 1$, so we can apply Construction 7.2 and Lemma 7.3 to find F' regularly homotopic to F, and a convenient collection of Whitney discs $\mathcal{W}'$ for the double points of F' with $t(F, \mathcal{W}) \neq t(F', \mathcal{W}')$.

If F is b–characteristic, by definition $\lambda_\Sigma|_{\partial\mathcal{B}(F)}$ is trivial and Θ is trivial on $\mathcal{B}(F)$. As indicated above, the function Θ, as well as which classes of $H_2(M, \Sigma; \mathbb{Z}/2)$ can be represented by bands, only depends on the immersion F up to regular homotopy. We need to show that $t(F, \mathcal{W})$ does not depend on the choice of pairing of the double points, the choice of Whitney arcs, nor the choice of Whitney discs; see Figure 18. Let $\mathcal{W}$ be a given initial choice of convenient collection of Whitney discs for the double points of F. Let A be the corresponding collection of Whitney arcs for the double points of F.

The remainder of the proof is similar to [Stong 1994, pages 1311–1313; Freedman and Quinn 1990, Section 10.8A]. We will work with *weak collections of framed Whitney discs* and the alternative count $t_{\text{alt}} \in \mathbb{Z}/2$, as in Definitions 7.4 and 7.5. So the boundaries of our collections of Whitney discs might not be disjointly embedded, but the Whitney discs will be framed (as can always be arranged by boundary twisting). We will show that $t_{\text{alt}}(F, \mathcal{W})$ does not depend on the choice of weak collection of Whitney discs $\mathcal{W}$, and then use that $t_{\text{alt}}(F, \mathcal{W}) = t(F, \mathcal{W})$ for $\mathcal{W}$ a convenient collection (Lemma 7.6).

Claim 8.1 *Suppose we are given a weak collection of Whitney discs $\mathcal{W}$ corresponding to some choice of pairing up of double points of F; then, for any other choice of pairing, there exists a weak collection of Whitney discs $\mathcal{V}$ for that choice such that $t_{\text{alt}}(\mathcal{V}) = t_{\text{alt}}(\mathcal{W})$.*

Proof Let p_1, p_2, q_1 and q_2 be double points of F. Suppose that, in the initial choice of data, p_1 and p_2 are paired by a Whitney disc $W_1 \in \mathcal{W}$, and q_1 and q_2 by a Whitney disc $W_2 \in \mathcal{W}$. Suppose we instead pair up p_1 and q_2 by some Whitney disc V_1. Then, as indicated in Figure 17, p_2 and q_1 are also paired by

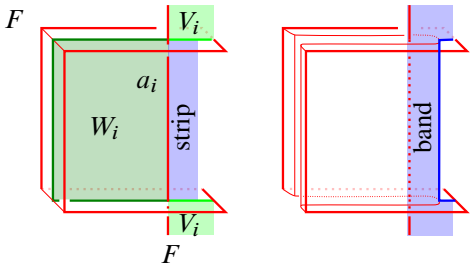


Figure 19: Left: two sheets of the surface Σ and two Whitney discs W_i and V_i between the same pair of double points. The disc W_i is assumed to be framed, embedded, have interior disjoint from F and the Whitney discs $\mathcal{U}_{i-1} \setminus \{W_i\}$, and ∂W_i disjoint from A_i. One of its Whitney arcs a_i is also labelled. The blue strip to the right of a_i is an extension of W_i beyond its boundary, which is part of the data for the Whitney move. Right: the result of the Whitney move. The strip and the disc V_i from the previous panel have formed a band B (blue).

a Whitney disc V_2, obtained as a union of W_1, W_2 and V_1. Then $(\mathcal{W} \setminus \{W_1, W_2\}) \cup \{V_1, V_2\}$ is a weak collection of framed Whitney discs. The contribution of V_1 and V_2 to $t_{\text{alt}}\big(F, (\mathcal{W} \setminus \{W_1, W_2\}) \cup \{V_1, V_2\}\big)$ counts the intersections of F with each disc W_1 and W_2 once, while it counts the intersections of F with the disc V_1 twice. Each intersection of ∂V_1 with $A \setminus (\partial W_1 \cup \partial W_2)$ can be paired with an intersection of ∂V_2 with $A \setminus (\partial W_1 \cup \partial W_2)$. Each intersection of ∂V_1 with $\partial W_1 \cup \partial W_2$ gives rise to two contributions to t_{alt}: an intersection of ∂V_2 with ∂V_1 and a self-intersection of ∂V_2. Since intersections are counted mod 2 in the definition of t_{alt}, we see that

$$t_{\text{alt}}\big(F, (\mathcal{W} \setminus \{W_1, W_2\}) \cup \{V_1, V_2\}\big) = t_{\text{alt}}(F, \mathcal{W}) \in \mathbb{Z}/2,$$

as needed. Iterate this process to complete the proof of Claim 8.1. □

Continuing with the proof of Theorem 1.9, next we check that t_{alt} is independent of the choice of Whitney discs. This includes potentially changing the Whitney arcs and the choice of sheets at each double point. Suppose we are given another weak collection of framed Whitney discs $\mathcal{V}$ for the double points of F. By applying Claim 8.1, we may assume that $\mathcal{V}$ corresponds to the same pairing of double points of F as $\mathcal{W}$. Assume the collections are indexed so that $W_l \in \mathcal{W}$ and $V_l \in \mathcal{V}$ correspond to the same pair of double points. For each i, define the weak collection of Whitney discs

$$\mathcal{U}_i := \{V_1, V_2, \dots, V_i, W_{i+1}, W_{i+2} \dots, W_N\},$$

where $\mathcal{U}_0 = \mathcal{W}$ and $\mathcal{U}_N = \mathcal{V}$. We will show that $t_{\text{alt}}(F, \mathcal{U}_{i-1}) = t_{\text{alt}}(F, \mathcal{U}_i) \in \mathbb{Z}/2$ for each i. Let A_i denote the collection of Whitney arcs for $\mathcal{U}_i$. First we prove a special case.

Claim 8.2 *Suppose the Whitney disc W_i is framed, embedded, with interior disjoint from F and the Whitney discs $\mathcal{U}_{i-1} \setminus \{W_i\}$, and with ∂W_i disjoint from A_i, other than the endpoints. Then $t_{\text{alt}}(F, \mathcal{U}_{i-1}) = t_{\text{alt}}(F, \mathcal{U}_i) \in \mathbb{Z}/2$.*

Proof A neighbourhood of W_i is depicted in Figure 19. Note that the two arcs of ∂V_i lie in A_i and thus, by hypothesis, only intersect the arcs in ∂W_i at the endpoints. As described in the figure, we wish to

perform the Whitney move using W_i pushing towards the Whitney arc a_i for W_i. Observe that the union of V_i with a strip, corresponding to the unit outward-pointing normal vector field of $a_i \subseteq \partial W_i$, is either an annulus or a Möbius band; this requires a small isotopy of V_i to ensure that the chosen vector field of a_i is compatible with the Whitney arcs of V_i, as shown in Figure 19. Denote the union of V_i and the strip by B.

We show that $B \in \mathcal{B}(F)$. For this we need to check that condition (5-1) holds. From Figure 19, right, one sees that ∂B is homotopic in Σ to the union of ∂V_i and ∂W_i. The core C of B is given by the union of a_i and either of the Whitney arcs of V_i. The Whitney arcs must induce opposite signs at the two double points, as explained in Definition 2.27. The orientation conditions in the latter definition imply that the condition in (5-1) holds, as we explain next. Let p_1 and p_2 denote the double points paired by W_i (and V_i). Let a_i and b_i denote the Whitney arcs of W_i, and let c_i and d_i denote those of V_i. Begin by fixing local orientations of M and both sheets of Σ at p_1, so that the first agrees with the one determined by the latter two. Transport the local orientations of Σ to p_2 via the Whitney arcs of W_i and form the induced local orientation of M at p_2. By Definition 2.27, this does not agree with the local orientation of M at p_2 determined by the one at p_1 by transporting along a_i. Continuing with the local orientations at p_2 determined in the previous step, transport the local orientations of Σ back to p_1, this time along the Whitney arcs of V_i. Again by Definition 2.27, the resulting induced local orientation of M at p_1 agrees with the local orientation of M transported to p_1 along c_i. In this circuit, we have constructed a new set of local orientations of M and the two sheets of Σ at p_1. Compared to the initial choice, the local orientation induced by the sheets of Σ has changed by $\langle w_1(\Sigma), a_i \cup b_i \cup c_i \cup d_i\rangle = \langle w_1(\Sigma), \partial V_i \cup \partial W_i\rangle$. On the other hand, the local orientation of M transported along $a_i \cup c_i$ has changed by $\langle w_1(M), a_i \cup c_i\rangle = \langle w_1(M), C\rangle$, where C is the core of B from above. Since the two orientations must agree, we have $\langle w_1(\Sigma), \partial V_i \cup \partial W_i\rangle = \langle w_1(M), C\rangle$, as needed.

For the band B as above, performing a finger move as in Construction 7.2 creates W_i as the standard Whitney disc, and V_i as the new Whitney disc arising from the band. Here we used the fact that ∂W_i and ∂V_i only intersect at the endpoints. Since F is b–characteristic, the disc V_i has trivial contribution to $t_{\text{alt}}(F, \mathcal{U}_i)$ by Lemma 7.3. So does W_i to $t_{\text{alt}}(F, \mathcal{U}_{i-1})$ since, by hypothesis, ∂W_i is framed, embedded, and disjoint from $A_{i-1} \setminus \{a_i, b_i\} \subseteq A_i$, and the interior of W_i is disjoint from F. Therefore, $t_{\text{alt}}(F, \mathcal{U}_{i-1}) = t_{\text{alt}}(F, \mathcal{U}_i) \in \mathbb{Z}/2$, as asserted. □

Now we prove the general case. Denote the double points paired by W_i by p_1 and p_2. By a small isotopy, assume that, other than p_1 and p_2, the arcs of ∂W_i intersect the arcs in A_i in isolated double points in the interiors. By performing a suitable finger move near p_2, split W_i into new Whitney discs W_i' and U_1, creating two new double points q_1 and q_2 in the process, paired by a standard trivial Whitney disc U_2, where U_1 satisfies the conditions of Claim 8.2. We choose both the base and tip of the finger arc to be closer to p_2 than any intersections of ∂W_i with arcs in A_i, as well as any self-intersections of ∂W_i. See Figure 20. By construction, the points p_1 and q_1 are paired by W_i', and the points q_2 and p_2 are paired by U_1. Here U_1 is framed, embedded, has interior disjoint from F and the Whitney discs

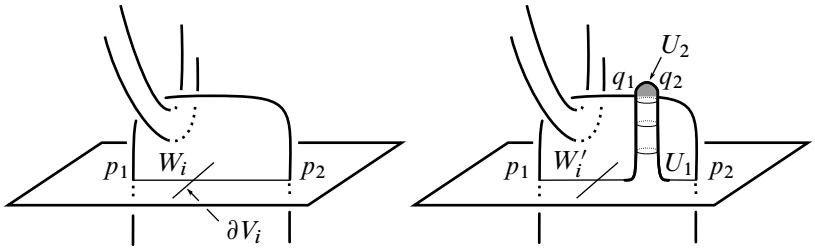


Figure 20: Splitting a Whitney disc W_i into two Whitney discs. One of the new Whitney discs, U_1, pairing p_2 and q_2, satisfies the hypotheses of Claim 8.2. The other Whitney disc W_i' intersects whatever W_i intersected. The trivial Whitney disc U_2 pairing the new double points q_1 and q_2 is shown in grey. Note that ∂W_i may intersect ∂V_i, or more generally other arcs in A_i, or itself.

$\mathcal{U}_{i-1} \setminus \{W_i\}$. In addition, ∂U_1 is disjoint from A_i, other than at p_2, and is disjoint from $\partial W_i'$. These conditions will shortly allow us to apply Claim 8.2 to U_1.

Let F' denote the result of performing the finger move above to F. Note that

$$t_{\mathrm{alt}}(F, \mathcal{U}_{i-1}) = t_{\mathrm{alt}}\big(F', (\mathcal{U}_{i-1} \setminus \{W_i\}) \cup \{W_i', U_1\}\big) \tag{8-1}$$

by construction. Let V_i' denote the Whitney disc obtained as the union of V_i, W_i' and U_2, as in Figure 17. Observe that the Whitney discs U_1 and V_i' pair the same double points, namely q_2 and p_2. Consider the two collections of Whitney discs $(\mathcal{U}_{i-1} \setminus \{W_i\}) \cup \{W_i', U_1\}$ and $(\mathcal{U}_{i-1} \setminus \{W_i\}) \cup \{W_i', V_i'\}$ for the double points of F'. The two collections differ only in that one contains the disc U_1 and the other the disc V_i'. We will apply Claim 8.2 to change between the two collections. This is permitted since U_1 is framed, embedded, has interior disjoint from F' and the Whitney discs $(\mathcal{U}_{i-1} \setminus \{W_i\}) \cup \{W_i'\}$, and ∂U_1 is disjoint, other than at the endpoints, from the Whitney arcs of $(\mathcal{U}_{i-1} \setminus \{W_i\}) \cup \{W_i', V_i'\}$, given by $A_i \cup \partial W_i' \cup \partial U_2$.

So, by Claim 8.2,

$$t_{\mathrm{alt}}\big(F', (\mathcal{U}_{i-1} \setminus \{W_i\}) \cup \{W_i', U_1\}\big) = t_{\mathrm{alt}}\big(F', (\mathcal{U}_{i-1} \setminus \{W_i\}) \cup \{W_i', V_i'\}\big). \tag{8-2}$$

By the proof of Claim 8.1 (see Figure 17),

$$t_{\mathrm{alt}}\big(F', (\mathcal{U}_{i-1} \setminus \{W_i\}) \cup \{W_i', V_i'\}\big) = t_{\mathrm{alt}}\big(F', (\mathcal{U}_{i-1} \setminus \{W_i\}) \cup \{U_2, V_i\}\big). \tag{8-3}$$

Since U_2 is trivial, we can use it to undo the Whitney move, and obtain

$$t_{\mathrm{alt}}\big(F', (\mathcal{U}_{i-1} \setminus \{W_i\}) \cup \{U_2, V_i\}\big) = t_{\mathrm{alt}}\big(F, (\mathcal{U}_{i-1} \setminus \{W_i\}) \cup \{V_i\}\big) = t_{\mathrm{alt}}(F, \mathcal{U}_i). \tag{8-4}$$

The combination of (8-1), (8-2), (8-3) and (8-4) implies $t_{\mathrm{alt}}(F, \mathcal{U}_{i-1}) = t_{\mathrm{alt}}(F, \mathcal{U}_i)$. This completes the proof that t_{alt} is independent of the choices of Whitney discs, and therefore completes the proof that t_{alt} is well defined.

Finally, by Lemma 7.6, we know that $t_{\mathrm{alt}}(F, \mathcal{W}) = t(F, \mathcal{W})$ for $\mathcal{W}$ a convenient collection, so t is well defined for convenient collections $\mathcal{W}$, as desired. □

Next we recall the statement of Theorem 1.6 for the convenience of the reader.

Theorem 1.6 *Let $F\colon (\Sigma,\partial\Sigma)\looparrowright (M,\partial M)$ be as in Convention 1.1. Suppose that $\mu(F)=0$ and that F has algebraically dual spheres. If F^ω is not b–characteristic, then $\mathrm{km}(F)=0$. If F^ω is b–characteristic, then the secondary embedding obstruction satisfies*

$$\mathrm{km}(F)=t(F^\omega,\mathcal{W}^\omega)\in\mathbb{Z}/2$$

for every convenient collection of Whitney discs $\mathcal{W}^\omega$ pairing all the double points of F^ω.

Proof First we show that, if F^ω is not b–characteristic then $\mathrm{km}(F)=0$. By Lemma 5.11, we reduce to the case that $\lambda_{\Sigma^\omega}|_{\partial\mathcal{B}(F^\omega)}$ is trivial, which implies that the function Θ is well defined on $\mathcal{B}(F^\omega)$. Since we assumed that F^ω is not b–characteristic, there exists $B\in\mathcal{B}(F^\omega)$ such that $\Theta(B)=1$, so we can apply Construction 7.2 and Lemma 7.3 to find a collection of Whitney discs $\mathcal{W}^\omega$ for the double points of F^ω with $t(F^\omega,\mathcal{W}^\omega)=0$. Then, by Lemma 5.4, we know that $\mathrm{km}(F)=0$.

By Theorem 1.9, if F^ω is b–characteristic, then $t(F^\omega,\mathcal{W}^\omega)$ is well defined, ie is independent of $\mathcal{W}^\omega$. As in Theorem 1.9, we denote the resulting invariant $t(F^\omega)$. We need to show that $\mathrm{km}(F)=t(F^\omega)$.

Recall that b–characteristic implies r–characteristic by Lemma 5.18, and also r–characteristic implies s–characteristic by Remark 5.6. Therefore Lemma 5.4 applies, which says that, if $t(F^\omega)=t(F^\omega,\mathcal{W}^\omega)=0$, then $\mathrm{km}(F)=0$. On the other hand, if $\mathrm{km}(F)=0$, then after a regular homotopy the double points of F can be paired up by a convenient collection of Whitney discs with interiors disjoint from F. Using these Whitney discs to calculate $t(F^\omega)$, and regular homotopy invariance of t from Theorem 1.9, it follows that $t(F^\omega)=0$.

Thus we have shown that, for F^ω b–characteristic and F with algebraically dual spheres, $\mathrm{km}(F)=0$ if and only if $t(F^\omega)=0$ or, equivalently, $\mathrm{km}(F)=t(F^\omega)\in\mathbb{Z}/2$, as desired. □

9 Examples and applications

Proposition 9.1 *Let $F\colon (\Sigma,\partial\Sigma)\looparrowright (M,\partial M)$ be as in Convention 1.1 and assume that $\mu(F)=0$. If there are two orientation-preserving immersed loops in Σ that intersect transversely in an odd number of points and are null-homotopic in M, then F is not b–characteristic.*

Proof The two immersed loops in Σ from the assumption bound immersed discs in M. These discs give classes in $\mathcal{B}(F)$ by Construction 5.10 and, by assumption, $\lambda_\Sigma|_{\mathcal{B}(F)}$ is nontrivial. It follows by definition that F is not b–characteristic. □

This applies to every simply connected target M whenever Σ has a component of positive genus. Proposition 9.1 also implies Corollaries 1.7 and 1.8, whose statements we recall, as follows.

Corollary 1.7 *If M is a simply connected 4–manifold and Σ is a connected, oriented surface with positive genus, then any generic immersion $F\colon (\Sigma,\partial\Sigma)\looparrowright(M,\partial M)$ with vanishing self-intersection number is not b–characteristic. Thus, if F has an algebraically dual sphere, then $\mathrm{km}(F)=0$, and, since $\pi_1(M)$ is good, the map F is regularly homotopic, relative to $\partial\Sigma$, to an embedding.*

Proof As $\pi_1(M)$ is trivial and Σ has positive genus, F is not b–characteristic by Proposition 9.1. By Theorem 1.6, it follows that $\mathrm{km}(F)=0$ if F has an algebraically dual sphere. In this case F is regularly homotopic, relative to $\partial\Sigma$, to an embedding, by Theorem 1.2. The theorem applies because $\pi_1(M)$ is good. □

Corollary 1.8 *Let $F\colon(\Sigma,\partial\Sigma)\looparrowright(M,\partial M)$ be as in Convention 1.1, with $\mu(F)=0$ and Σ connected. If F' is obtained from F by an ambient connected sum with an embedding $S^1\times S^1\hookrightarrow S^4$, then F' is not b–characteristic. Thus, if F has an algebraically dual sphere, then $\mathrm{km}(F')=0$, and, if $\pi_1(M)$ is good, then F' is regularly homotopic, relative to $\partial\Sigma$, to an embedding.*

Proof Since F' is obtained from F by an ambient connected sum with an embedding $S^1\times S^1\hookrightarrow S^4$, we can apply Proposition 9.1 to see that F' is not b–characteristic. By Theorem 1.6, it follows that $\mathrm{km}(F')=0$ if F has an algebraically dual sphere, as this sphere remains algebraically dual to F'. If in addition $\pi_1(M)$ is good, then, by Theorem 1.2, F is regularly homotopic, relative to $\partial\Sigma$, to an embedding. □

Example 9.2 To illustrate the difference between r–characteristic and b–characteristic surfaces, we give an example of a surface that is r–characteristic but not b–characteristic. Consider any r–characteristic immersed sphere with trivial self-intersection number. Add a single trivial tube to obtain an immersed torus. As this will not change the intersection number with any closed surface, the new torus is still r–characteristic. But it fails to be b–characteristic by Corollary 1.8.

Example 9.3 We explain next why our methods allow us to obtain embeddings where [Freedman and Quinn 1990, Theorem 10.5A(1)] would not produce them (see the discussion directly following Theorem 1.2).

Let $f\colon S^2\looparrowright M$ be a generic immersion in a 4–manifold with $\pi_1(M)$ good, equipped with an algebraically dual sphere and with $\mathrm{km}(f)=1$, for example a sphere representing a generator of $H_2(*\mathbb{C}\mathbb{P}^2)$. Other such spheres may be constructed as in [Kasprowski et al. 2021a, Theorem 2]. Let T be a generic immersion of a torus produced by adding a trivial tube to f, ie by taking the ambient connected sum of f with the standard embedding $S^1\times S^1\hookrightarrow S^4$. Then by Corollary 1.8 we see that $\mathrm{km}(T)=0$. Thus T is regularly homotopic to an embedding since $\pi_1(M)$ is good. Fix a 1–skeleton Σ_0 for $S^1\times S^1$. Then T is not regularly homotopic to an embedding relative to Σ_0, since the Kervaire–Milnor invariant for f restricted to the 2–cell(s) $(S^1\times S^1)\setminus\nu\Sigma_0$ considered as a map to $M\setminus T(\nu\Sigma_0)$ equals $\mathrm{km}(f)=1$.

We emphasise that this holds for every choice of 1–skeleton $\Sigma_0\subseteq S^1\times S^1$. In order to apply the strategy of [Freedman and Quinn 1990, Theorem 10.5A(1)] to find an embedding, one needs to first make a

judicious choice of finger moves. But, without our theory, there is no clear strategy for finding these finger moves. To obtain an embedding obstruction in this way, matters are worse, since one would need to compute the Kervaire–Milnor invariant of the 2–skeleton for every choice of finger moves and for every choice of 1–skeleton.

Example 9.4 We construct an immersed torus with nontrivial Kervaire–Milnor invariant. In contrast to Proposition 9.1, the torus in this example is not π_1–trivial. Consider an immersion f_1 of a 2–sphere in $*\mathbb{CP}^2$ representing a generator of $H_2(*\mathbb{CP}^2;\mathbb{Z})$ with trivial self-intersection number. Let $K\colon S^1 \hookrightarrow S^3$ be an arbitrary knot and consider the embedding of a torus given by the product $f_2 := K \times \mathrm{Id}\colon S^1 \times S^1 \hookrightarrow S^3 \times S^1$. Let F denote the interior connected sum $f_1 \# f_2\colon S^1 \times S^1 \looparrowright W := *\mathbb{CP}^2 \# (S^3 \times S^1)$.

First we claim that F is b–characteristic. To see this, we start by computing $H_2(W, S^1 \times S^1; \mathbb{Z}/2)$ using the long exact sequence of the pair with $\mathbb{Z}/2$ coefficients:

$$\begin{array}{ccccccccc} H_2(S^1 \times S^1) & \twoheadrightarrow & H_2(W) & \xrightarrow{0} & H_2(W, S^1 \times S^1) & \longrightarrow & H_1(S^1 \times S^1) & \twoheadrightarrow & H_1(W) \\ & & & & & & \wr\| & & \wr\| \\ & & & & & & \mathbb{Z}/2 \oplus \mathbb{Z}/2 & & \mathbb{Z}/2 \end{array}$$

Therefore, $H_2(W, S^1 \times S^1; \mathbb{Z}/2) \cong \mathbb{Z}/2$ is generated by $S \times \{p\}$, where $S \subseteq S^3$ is a Seifert surface for the knot $K(S^1)$ and $p \in S^1$. The intersection form of $S^1 \times S^1$ restricted to ∂S is trivial. Since Θ is well defined on homology classes we can compute it using S. But S has interior disjoint from the image of F, embedded boundary and trivial relative Euler number, so $\Theta(S) = 0$. If follows that Θ vanishes on all of $H_2(W, S^1 \times S^1; \mathbb{Z}/2)$, in particular it vanishes on the subset $\mathcal{B}(F)$. Thus F is b–characteristic, as claimed.

Observe that $\mathrm{km}(f_1) = 1$ inside $*\mathbb{CP}^2$ (see eg [Freedman and Quinn 1990, Section 10.8]). We can pick a convenient collection of Whitney discs for f_1 in $*\mathbb{CP}^2$. Since f_2 is an embedding, these constitute a convenient collection of Whitney discs for F. It follows that $\mathrm{km}(F) = \mathrm{km}(f_1) = 1$. Note that the choice of knot K was irrelevant, since, for any two choices, the resulting immersions F are regularly homotopic and hence have equal Kervaire–Milnor invariant.

Example 9.5 In the previous example we constructed a generically immersed torus in $*\mathbb{CP}^2 \# (S^1 \times S^3)$ with nontrivial Kervaire–Milnor invariant. In particular, this torus is not homotopic to an embedding (see Section 1.4). Now we show that, in contrast to this, every map f from a closed surface Σ to $S^1 \times S^3$ is homotopic to an embedding. Note that these classes do not have algebraically dual spheres since $\pi_2(S^1 \times S^3) = 0$. The surfaces in the regular homotopy class with $\mu(f)_1 = 0$ are either not b–characteristic or $t(f^\omega)$ vanishes.

Since the projection $S^1 \times S^3 \to S^1$ is 3–connected, the induced map $[\Sigma, S^1 \times S^3] \to [\Sigma, S^1]$ is bijective. In particular, the homotopy class of a map $f\colon \Sigma \to S^1 \times S^3$ is determined by the induced map on fundamental groups.

We first consider the case that Σ is connected. Since $\pi_1(S^1 \times S^3) \cong \mathbb{Z}$, we can find a generating set for $\pi_1(\Sigma)$ such that at most one generator is nontrivial in $\pi_1(S^1 \times S^3)$. Thus there exists a decomposition $\Sigma = H \# \Sigma'$, where H is either a sphere, a torus or a Klein bottle, with respect to which f can be written as an internal connected sum $T \# f'$, where T is a map on H and f' is π_1–trivial. In particular, f' is homotopic to an embedding inside a ball $D^4 \subseteq S^1 \times S^3$. It remains to show that T is homotopic to an embedding, which will show that the connected sum is homotopic to an embedding.

If H is a sphere, we are done. If H is a torus, let $i : S^1 \times S^1 \hookrightarrow S^3$ be an embedding. For each $k \in \mathbb{Z}$, define the embedding $h'_k : S^1 \times S^1 \to S^1 \times (S^1 \times S^1)$ by $(s,t) \mapsto (s^k, (s,t))$. Let $h_k := (\mathrm{Id} \times i) \circ h'_k$. There exists some k and some identification of H with $S^1 \times S^1$ such that T and h_k induce the same map on fundamental groups and thus are homotopic. If H is a Klein bottle, let $p : H \to S^1$ be a fibre bundle with fibre S^1. For each $k \in \mathbb{Z}$, there exists an immersion $i : H \looparrowright S^3$ such that $h_k(x) := (p(x)^k, i(x))$ is an embedding $H \hookrightarrow S^1 \times S^3$. As before, there exists some k such that T and h_k are homotopic.

The above embeddings can be realised as embeddings into $S^1 \times D^3 \subseteq S^1 \times S^3$. The argument generalises to disconnected surfaces by picking disjoint copies of $S^1 \times D^3$ in $S^1 \times S^3$ for each connected component.

Next we prove Proposition 1.10 and Corollary 1.11, which we restate for the convenience of the reader.

Proposition 1.10 *Let M_1 and M_2 be oriented 4–manifolds. Let $F_1 : (\Sigma_1, \partial\Sigma_1) \looparrowright (M_1, \partial M_1)$ and $F_2 : (\Sigma_2, \partial\Sigma_2) \looparrowright (M_2, \partial M_2)$ be generic immersions of connected, compact, oriented surfaces, each with vanishing self-intersection number. If F_i is b–characteristic for each i, then both the disjoint union*

$$F_1 \sqcup F_2 : \Sigma_1 \sqcup \Sigma_2 \looparrowright M_1 \# M_2$$

and any interior connected sum

$$F_1 \# F_2 : \Sigma_1 \# \Sigma_2 \looparrowright M_1 \# M_2$$

are b–characteristic, and satisfy

$$t(F_1 \sqcup F_2) = t(F_1 \# F_2) = t(F_1) + t(F_2).$$

Proof The vanishing of the self-intersection number of F_i is witnessed by a convenient collection of Whitney discs $\mathcal{W}_i$ in M_i for each i. The union $\mathcal{W}_1 \sqcup \mathcal{W}_2$, now considered in $M_1 \# M_2$, shows that the intersection and self-intersection numbers of $F_1 \sqcup F_2$, as well as for $F_1 \# F_2$, vanish in $M_1 \# M_2$. Since the union $\mathcal{W}_1 \sqcup \mathcal{W}_2$ pairs all the double points of $F_1 \sqcup F_2$ (resp. $F_1 \# F_2$), and since components of $\mathcal{W}_i$ cannot intersect $F_j(\Sigma_j)$ for all $i \neq j$, the claimed relationship $t(F_1 \sqcup F_2) = t(F_1 \# F_2) = t(F_1) + t(F_2)$ holds as long as $F_1 \sqcup F_2$ and $F \# F_2$ are b–characteristic.

As a preliminary step, note that neither F_i has a framed dual sphere in M_i, since otherwise it would not be s–characteristic, and therefore not b–characteristic by Lemma 5.18. As a result, $\Sigma_i^{\circ} = \Sigma_i$ for $i = 1, 2$.

Next we consider the immersion $F_1 \sqcup F_2 : (\Sigma_1 \sqcup \Sigma_2, \partial\Sigma_1 \sqcup \partial\Sigma_2) \looparrowright M_1 \# M_2$. Let $S \subseteq M_1 \# M_2$ denote a connected sum 3–sphere. Consider a band $[B] \in H_2(M_1 \# M_2, \Sigma_1 \sqcup \Sigma_2)$. By (topological) transversality, we can assume that B is immersed, the double points of B are disjoint from S, and the

intersection $B \cap S$ corresponds to an embedded 1–manifold in the interior of the domain of B, since $\partial B \subseteq \Sigma_1 \sqcup \Sigma_2 \subseteq (M_1 \# M_2) \setminus S$. The image of this 1–manifold in S is embedded and bounds a collection of immersed (perhaps intersecting) discs in S. Surger B using two copies each of these discs to produce $B_1 \subseteq M_1$ and $B_2 \subseteq M_2$, where each B_i is an immersed collection of surfaces with $\partial B_i \subseteq \Sigma_i$.

Each component of B_i can be replaced by a band as follows. Recall that since each M_i and Σ_i is oriented, there is no condition on Stiefel–Whitney classes for bands, and we need only arrange that each component is either a Möbius band or an annulus. By considering the Euler characteristic, we see that each component is homeomorphic to either a sphere, an $\mathbb{RP}^2$, a disc, a Möbius band or an annulus. Then use the tubing procedure from Construction 5.10 to replace each sphere, $\mathbb{RP}^2$ or disc component by a band. More precisely, choose a small disc on Σ_1 or Σ_2, as appropriate, away from all Whitney arcs and double points, and tube into the disc, sphere or $\mathbb{RP}^2$.

Since each F_i is b–characteristic, $\lambda_{\Sigma_1}|_{\partial \mathcal{B}(F_i)}$ is trivial for each i. Therefore, $\lambda_{\Sigma_1 \sqcup \Sigma_2}$ is trivial on $\partial B = \partial B_1 \cup \partial B_2$. It follows by Lemma 5.16(iv) that $\Theta\colon \mathcal{B}(F_1 \sqcup F_2) \to \mathbb{Z}/2$ is well defined. By Lemma 7.8, Θ extends to a linear map $\langle \mathcal{B}(F_1 \sqcup F_2) \rangle \to \mathbb{Z}/2$ on the subspace

$$\langle \mathcal{B}(F_1 \sqcup F_2) \rangle \subseteq H_2(M_1 \# M_2, \Sigma_1 \sqcup \Sigma_2; \mathbb{Z}/2)$$

generated by the bands. Then, since $[B_1 \cup B_2] = [B] \in H_2(M_1 \# M_2, \Sigma_1 \sqcup \Sigma_2; \mathbb{Z}/2)$, we see that $\Theta(B) = \Theta(B_1) + \Theta(B_2)$.

For each i, the value of $\Theta(B_i)$ does not depend on whether the ambient manifold is M_i or $M_1 \# M_2$, since B_i does not intersect $F_j(\Sigma_j)$ for all $i \neq j$ (see Definition 5.12). Since each F_i is b–characteristic, $\Theta(B) = \Theta(B_1) + \Theta(B_2) = 0 + 0 = 0 \in \mathbb{Z}/2$. This completes the proof that $F_1 \sqcup F_2$ is b–characteristic.

Now we consider the connected sum $F_1 \# F_2$. Let $B \in H_2(M_1 \# M_2, \Sigma_1 \# \Sigma_2)$ be a band. As above, we assume that the intersection $B \cap S$ corresponds to an embedded 1–manifold in the domain of B. Unlike above, this may include embedded arcs with endpoints on the boundary. These endpoints are mapped to the intersection $(F_1 \# F_2)(\Sigma_1 \# \Sigma_2) \cap S$. By connecting the endpoints with arcs on $(F_1 \# F_2)(\Sigma_1 \# \Sigma_2) \cap S$, we again get a collection of closed circles in S, which bound an immersed collection of discs in S. Surger using these discs as before to produce collections $B_i \subseteq M_i$. Once again, each component of B_i is homeomorphic to either a sphere, an $\mathbb{RP}^2$, a disc, a Möbius band or an annulus. By Construction 5.10 applied to the sphere, $\mathbb{RP}^2$ and disc components, we may arrange that each component is a band. The argument of the previous paragraph now applies to show that $F_1 \# F_2$ is b–characteristic. □

Corollary 1.11 *For any g, there exists a smooth, closed 4–manifold M_g, a closed, connected, oriented surface Σ_g of genus g, and a smooth, b–characteristic, generic immersion $F\colon \Sigma_g \looparrowright M_g$ with $t(F) \neq 0$ and therefore* $\mathrm{km}(F) \neq 0$.

Proof By the same proof as in Example 9.4, for any knot K the product $T := K \times \mathrm{Id}\colon S^1 \times S^1 \to S^3 \times S^1$ is an embedded b–characteristic torus. Since T is an embedding, $t(T) = 0$. A computation using the intersection form shows that a generic immersion $S\colon S^2 \to \mathbb{CP}^2$ representing three times a generator of

$H_2(\mathbb{CP}^2;\mathbb{Z})$ is s–characteristic. Since $\pi_1(\mathbb{CP}^2)$ has no 2–torsion, the map S is also r–characteristic and thus b–characteristic by Lemma 5.18. We will show that $\mathrm{km}(S)=1$. This was the original example of Kervaire and Milnor [1961]. To see that $\mathrm{km}(S)=1$, represent S in the following way. Take a cuspidal cubic, which is a smooth embedding of a 2–sphere away from a single singular point. In a neighbourhood of the singular point we see a cone on the trefoil. Replace a neighbourhood of the singular point with an immersed disc Δ in D^4 with boundary the trefoil, and two double points that are paired by a framed Whitney disc that intersects Δ once. Alternatively, we can compute $t(S)$ as $\frac{1}{8}(\sigma(\mathbb{CP}^2)-S\cdot S)=\frac{1}{8}(1-9)\equiv 1 \bmod 2$; see Section 3. This gives us the case $g=0$. Next, by Proposition 1.10, for every $g\in\mathbb{N}$,

$$S\,\#\,\#^g\,T\colon \Sigma_g \to \mathbb{CP}^2\,\#\,\#^g(S^3\times S^1)$$

is a b–characteristic generic immersion of a closed surface of genus g with nontrivial t, and therefore $\mathrm{km}\big(S\,\#\,\#^g\,T\big)\neq 0$. In particular, $S\,\#\,\#^g\,T$ is not regularly homotopic to an embedding. Note these examples are smooth, but have no algebraically dual sphere. We could replace $(\mathbb{CP}^2,S)$ with $\big(\mathbb{CP}^2\,\#\,\#^8\,\overline{\mathbb{CP}^2},S'\big)$, where S' is a generic immersion representing the class $(3,1,\dots,1)\in\mathbb{Z}^9\cong H_2\big(\mathbb{CP}^2\,\#\,\#^8\,\overline{\mathbb{CP}^2}\big)$, to obtain an example with an algebraically dual sphere and $\mathrm{km}(S')=\frac{1}{8}(-7-1)\equiv 1 \bmod 2$. □

Remark 9.6 Let M denote the infinite connected sum $\mathbb{CP}^2\#\#^\infty(S^3\times S^1)$. The proof of Corollary 1.11, along with the formula from Proposition 1.10, shows that, for every g, there exists a smooth generic immersion $F\colon \Sigma_g \looparrowright M$ with $t(F)\neq 0$ and therefore $\mathrm{km}(F)\neq 0$. The following proposition shows that, if there is such a compact 4–manifold M and such an F, then the 4–manifolds must have nonabelian fundamental group. In other words, if there is an immersed surface in a 4–manifold with abelian fundamental group with nontrivial km, then we give a bound on the complexity of that surface.

Proposition 9.7 *Let M be a compact 4–manifold such that $\pi_1(M)$ is abelian with n generators. Let $F\colon \Sigma \looparrowright M$ be a b–characteristic generic immersion, where Σ is a closed, connected surface. Then the Euler characteristic satisfies $\chi(\Sigma)\geq -2n$.*

Proof Suppose that $\chi(\Sigma)<-2n$. Note that Σ can be written as a connected sum of a genus g orientable surface Σ' for some $g>n$ with zero, one or two copies of $\mathbb{RP}^2$. There exists a surjection $\mathbb{Z}^n \twoheadrightarrow \pi_1(M)$. Then the induced map $H_1(\Sigma')\to H_1(M)\cong\pi_1(M)$ admits a lift $H_1(\Sigma')\to\mathbb{Z}^n$, which has kernel of rank at least $2g-n>g$. So there exist closed curves γ_1 and γ_2 in $\Sigma'\setminus \mathring{D}^2\subseteq\Sigma$ that are null-homotopic in M and $\lambda_\Sigma(\gamma_1,\gamma_2)\equiv 1 \bmod 2$. It follows that F is not b–characteristic. □

Next we prove our corollaries on knot theory from Section 1.5.

Corollary 1.15 *For every knot $K\subseteq S^3$,*

(1) *$g_M(K)=0$ for every simply connected 4–manifold M not homeomorphic to one of S^4, $\mathbb{CP}^2$ or $*\mathbb{CP}^2$;*

(2) $g_{\mathbb{CP}^2}(K) \leq 1$ *and* $g_{\mathbb{CP}^2}\big(\#^3 T(2,3)\big) = 1$; *and*

(3) $g_{*\mathbb{CP}^2}(K) \leq 1$ *and* $g_{*\mathbb{CP}^2}\big(\#^2 T(2,3)\big) = 1$.

Proof Let $K \subseteq S^3$ be an arbitrary knot and let M be an arbitrary closed, simply connected 4–manifold. Let Δ' be a generically immersed disc bounded by K in a collar $S^3 \times [0,1]$ of ∂M°. Since M is simply connected, every class in $H_2(M;\mathbb{Z}) \cong \pi_2(M)$ is represented by a generically immersed sphere. By assumption, M is not homeomorphic to S^4 and thus $H_2(M;\mathbb{Z})$ is nontrivial. Since M is closed, every primitive class $\alpha \in H_2(M;\mathbb{Z})$ has an algebraic dual $\beta \in H_2(M;\mathbb{Z})$, ie $\lambda(\alpha,\beta) = 1$. Represent α and β by generically immersed spheres, and tube the interior of Δ' into β to obtain Δ. Add local cusps to arrange $\mu(\Delta) = 0$.

First we prove (1). In this case we claim that in the construction of Δ we can choose the primitive class α to satisfy $\lambda(\alpha,\alpha) \in 2\mathbb{Z}$, as we explain presently. Then the disc Δ constructed above is not r–characteristic, since $\Delta \cdot \alpha \not\equiv \alpha \cdot \alpha \bmod 2$ (see Remark 5.6). By Theorem 5.7, this implies that $\mathrm{km}(\Delta) = 0$. Since the disc Δ has the algebraically dual sphere α and $\pi_1(M) = 1$ is good, by Theorem 1.2, Δ is homotopic rel boundary to an embedding. To see the claim regarding α, note that, when $M \ncong S^4, \mathbb{CP}^2, *\mathbb{CP}^2$, the group $H_2(M;\mathbb{Z})$ has rank at least 2 by the classification of closed, simply connected 4–manifolds up to homeomorphism. Then $H_2(M;\mathbb{Z})$ has a summand isomorphic to $\mathbb{Z} \oplus \mathbb{Z}$, so the classes x, y and $x+y$, for the generators x and y of the $\mathbb{Z}$ factors, are primitive, and at least one of $\lambda(x,x)$, $\lambda(y,y)$ or $\lambda(x+y,x+y)$ is even.

In (2) and (3), we have $M = \mathbb{CP}^2$ or $*\mathbb{CP}^2$. The only primitive classes are $\pm[\mathbb{CP}^1]$, so we choose $\alpha = \beta = [\mathbb{CP}^1]$ in the construction of the first paragraph. We construct the disc Δ as before, but are no longer able to conclude that it is r–characteristic. However, by Corollary 1.8, we know that the connected sum of Δ with an unknotted torus is homotopic to an embedding. This completes the proof of the first parts of (2) and (3).

Now we prove that $g_{\mathbb{CP}^2}\big(\#^3 T(2,3)\big) = 1$. Let $K := \#^3 T(2,3)$. Let $g^d_{\mathbb{CP}^2}(K)$ denote the minimal genus of a surface bounded by K in $(\mathbb{CP}^2)^\circ$ in the homology class $d \in \mathbb{Z} \cong H_2(\mathbb{CP}^2;\mathbb{Z})$. First we consider $d = \pm 1$, where the class is b–characteristic (or equivalently, s–characteristic; see Lemma 5.18). As before, construct the disc $\Delta' \subseteq S^3 \times [0,1]$, and tube into $\mathbb{CP}^1$ to obtain the disc Δ. We assume that Δ' has trivial self-intersection number, so $1 = \mathrm{Arf}(K) = t(\Delta')$ by [Matsumoto 1978; Freedman and Kirby 1978; Conant et al. 2014, Lemma 10]. Since $\mathbb{CP}^1$ is embedded disjointly from Δ', $t(\Delta) = 1$. Thus, by Theorem 1.9, Δ is not homotopic to an embedding and so $g^{\pm 1}_{\mathbb{CP}^2}(K) \neq 0$.

Next let $\sigma_d(K) := \sigma_K(e^{\pi i(d-1)/d})$, where σ_K denotes the Levine–Tristram signature function of K. By [Gilmer 1981; Viro 1975], for even d,

$$2g^d_{\mathbb{CP}^2}(K) + 1 \geq \left|\tfrac{1}{2}d^2 - 1 - \sigma(K)\right|,$$

while, if d is divisible by an odd prime p, then

$$2g^d_{\mathbb{CP}^2}(K)+1 \geq \left|\frac{p^2-1}{2p^2}d^2-1-\sigma_d(K)\right|.$$

In our case, $\sigma(K)=\sigma_d(K)=-6$ for all d, and so $g^d_{\mathbb{CP}^2}(K)\geq 1$ for all $d\neq\pm1$. This completes the argument that $g_{\mathbb{CP}^2}(K)=1$.

Finally we show that $g_{*\mathbb{CP}^2}\big(\#^2 T(2,3)\big)=1$. Write $K:=\#^2 T(2,3)$ and let $g^d_{*\mathbb{CP}^2}(K)$ denote the minimal genus of a surface bounded by K in $(*\mathbb{CP}^2)^\circ$ in the homology class $d\in H_2(*\mathbb{CP}^2;\mathbb{Z})$. For $d=\pm1$, modify the argument above for the case of $\mathbb{CP}^2$, using that tubing into a sphere representing a generator of $H_2(*\mathbb{CP}^2;\mathbb{Z})$ to obtain a disc Δ' adds 1 to the t count, and so $1=1+\mathrm{Arf}(K)=t(\Delta')$. Therefore, again by Theorem 1.9, $g^{\pm1}_{*\mathbb{CP}^2}(K)\neq 0$. Next, for $*\mathbb{CP}^2$, the same inequalities from [Gilmer 1981; Viro 1975] hold, and $\sigma(K)=\sigma_d(K)=-4$ for all d. Therefore, applying the inequalities, we see that $g^d_{*\mathbb{CP}^2}(K)\geq 1$ for all d. It follows that $g_{*\mathbb{CP}^2}(K)=1$, as asserted. □

Corollary 1.16 *For any knot $K\subseteq S^3$, $g^{\mathrm{sh}}_{\pm1}(K)=\mathrm{Arf}(K)\in\{0,1\}$.*

Proof A generator of $H_2(X_{\pm1}(K);\mathbb{Z})$ can be represented by a generically immersed sphere F which is b–characteristic (or equivalently s–characteristic; see Lemma 5.18), has trivial $\mu(F)$, and has an algebraically dual sphere. We also recall from [Matsumoto 1978; Freedman and Kirby 1978; Conant et al. 2014, Lemma 10] that $\mathrm{Arf}(K)$ coincides with the count $t(F)$. Then, by Theorems 1.6 and 1.9, the sphere F is homotopic to an embedding if and only if $\mathrm{Arf}(K)=0$. We have an embedded torus representative for both generators by Corollary 1.8. □

References

[Aït Nouh 2009] **M Aït Nouh**, *Genera and degrees of torus knots in* $\mathbb{C}\mathrm{P}^2$, J. Knot Theory Ramifications 18 (2009) 1299–1312 MR Zbl

[Behrens et al. 2021] **S Behrens**, **B Kalmár**, **M H Kim**, **M Powell**, **A Ray** (editors), *The disc embedding theorem*, Oxford Univ. Press (2021) MR Zbl

[Casson 1986] **A J Casson**, *Three lectures on new-infinite constructions in 4–dimensional manifolds*, from "À la recherche de la topologie perdue" (L Guillou, A Marin, editors), Progr. Math. 62, Birkhäuser, Boston, MA (1986) 201–244 MR Zbl

[Cochran and Ray 2016] **T D Cochran**, **A Ray**, *Shake slice and shake concordant knots*, J. Topol. 9 (2016) 861–888 MR Zbl

[Cochran et al. 2003] **T D Cochran**, **K E Orr**, **P Teichner**, *Knot concordance, Whitney towers and L^2–signatures*, Ann. of Math. 157 (2003) 433–519 MR Zbl

[Conant et al. 2012a] **J Conant**, **R Schneiderman**, **P Teichner**, *Universal quadratic forms and Whitney tower intersection invariants*, from "Proceedings of the Freedman Fest" (R Kirby, V Krushkal, Z Wang, editors), Geom. Topol. Monogr. 18, Geom. Topol. Publ., Coventry (2012) 35–60 MR Zbl

[Conant et al. 2012b] **J Conant**, **R Schneiderman**, **P Teichner**, *Whitney tower concordance of classical links*, Geom. Topol. 16 (2012) 1419–1479 MR Zbl

[Conant et al. 2014] **J Conant**, **R Schneiderman**, **P Teichner**, *Milnor invariants and twisted Whitney towers*, J. Topol. 7 (2014) 187–224 MR Zbl

[Connelly 1971] **R Connelly**, *A new proof of Brown's collaring theorem*, Proc. Amer. Math. Soc. 27 (1971) 180–182 MR Zbl

[Feller et al. 2021] **P Feller**, **A N Miller**, **M Nagel**, **P Orson**, **M Powell**, **A Ray**, *Embedding spheres in knot traces*, Compos. Math. 157 (2021) 2242–2279 MR Zbl

[Freedman 1982] **M H Freedman**, *The topology of four-dimensional manifolds*, J. Differential Geom. 17 (1982) 357–453 MR Zbl

[Freedman and Kirby 1978] **M Freedman**, **R Kirby**, *A geometric proof of Rochlin's theorem*, from "Algebraic and geometric topology, II" (R J Milgram, editor), Proc. Sympos. Pure Math. 32, Amer. Math. Soc., Providence, RI (1978) 85–97 MR Zbl

[Freedman and Quinn 1990] **M H Freedman**, **F Quinn**, *Topology of* 4*–manifolds*, Princeton Math. Ser. 39, Princeton Univ. Press (1990) MR Zbl

[Freedman and Teichner 1995] **M H Freedman**, **P Teichner**, 4*–manifold topology, I: Subexponential groups*, Invent. Math. 122 (1995) 509–529 MR Zbl

[Gilmer 1981] **P M Gilmer**, *Configurations of surfaces in* 4*–manifolds*, Trans. Amer. Math. Soc. 264 (1981) 353–380 MR Zbl

[Golubitsky and Guillemin 1973] **M Golubitsky**, **V Guillemin**, *Stable mappings and their singularities*, Graduate Texts in Math. 14, Springer (1973) MR Zbl

[Guillou and Marin 1980] **L Guillou**, **A Marin**, *Une extension d'un théorème de Rohlin sur la signature*, from "Seminar on real algebraic geometry" (J J Risler, editor), Publ. Math. Univ. Paris VII 9, Univ. Paris VII (1980) 69–80 MR Zbl

[Hambleton and Kreck 1993] **I Hambleton**, **M Kreck**, *Cancellation of hyperbolic forms and topological four-manifolds*, J. Reine Angew. Math. 443 (1993) 21–47 MR Zbl

[Hirsch 1976] **M W Hirsch**, *Differential topology*, Graduate Texts in Math. 33, Springer (1976) MR Zbl

[Kasprowski et al. 2017] **D Kasprowski**, **M Land**, **M Powell**, **P Teichner**, *Stable classification of* 4*–manifolds with* 3*–manifold fundamental groups*, J. Topol. 10 (2017) 827–881 MR Zbl

[Kasprowski et al. 2020] **D Kasprowski**, **M Powell**, **P Teichner**, *Algebraic criteria for stable diffeomorphism of spin* 4*–manifolds*, preprint (2020) arXiv 2006.06127

[Kasprowski et al. 2021a] **D Kasprowski**, **P Lambert-Cole**, **M Land**, **A G Lecuona**, *Topologically flat embedded* 2*–spheres in specific simply connected* 4*–manifolds*, from "2019–20 MATRIX annals" (D R Wood, J de Gier, C E Praeger, T Tao, editors), MATRIX Book Ser. 4, Springer (2021) 111–116 MR Zbl

[Kasprowski et al. 2021b] **D Kasprowski**, **M Powell**, **P Teichner**, *The Kervaire–Milnor invariant in the stable classification of spin* 4*–manifolds*, preprint (2021) arXiv 2105.12153

[Kervaire and Milnor 1961] **M A Kervaire**, **J W Milnor**, *On* 2*–spheres in* 4*–manifolds*, Proc. Nat. Acad. Sci. U.S.A. 47 (1961) 1651–1657 MR Zbl

[Kim et al. 2021] **M H Kim**, **P Orson**, **J Park**, **A Ray**, *Good groups*, from "The disc embedding theorem" (S Behrens, B Kalmár, M H Kim, M Powell, A Ray, editors), Oxford Univ. Press (2021) 273–282 MR Zbl

[Kirby and Taylor 2001] **R C Kirby**, **L R Taylor**, *A survey of* 4*–manifolds through the eyes of surgery*, from "Surveys on surgery theory, II" (S Cappell, A Ranicki, J Rosenberg, editors), Ann. of Math. Stud. 149, Princeton Univ. Press (2001) 387–421 MR Zbl

[Krushkal and Quinn 2000] **V S Krushkal**, **F Quinn**, *Subexponential groups in* 4*–manifold topology*, Geom. Topol. 4 (2000) 407–430 MR Zbl

[Lee and Wilczyński 1990] **R Lee**, **D M Wilczyński**, *Locally flat* 2*–spheres in simply connected* 4*–manifolds*, Comment. Math. Helv. 65 (1990) 388–412 MR Zbl

[Lee and Wilczyński 1997] **R Lee**, **D M Wilczyński**, *Representing homology classes by locally flat surfaces of minimum genus*, Amer. J. Math. 119 (1997) 1119–1137 MR Zbl

[Lisca and Matić 1998] **P Lisca**, **G Matić**, *Stein* 4*–manifolds with boundary and contact structures*, Topology Appl. 88 (1998) 55–66 MR Zbl

[Manolescu et al. 2024] **C Manolescu**, **M Marengon**, **L Piccirillo**, *Relative genus bounds in indefinite four-manifolds*, Math. Ann. (online publication January 2024)

[Marengon et al. 2024] **M Marengon**, **A N Miller**, **A Ray**, **A I Stipsicz**, *A note on surfaces in* $\mathbb{CP}^2$ *and* $\mathbb{CP}^2\#\mathbb{CP}^2$, Proc. Amer. Math. Soc. Ser. B 11 (2024) 187–199 MR Zbl

[Massey 1969] **W S Massey**, *Proof of a conjecture of Whitney*, Pacific J. Math. 31 (1969) 143–156 MR Zbl

[Matsumoto 1978] **Y Matsumoto**, *Secondary intersectional properties of* 4*–manifolds and Whitney's trick*, from "Algebraic and geometric topology, II" (R J Milgram, editor), Proc. Sympos. Pure Math. 32, Amer. Math. Soc., Providence, RI (1978) 99–107 MR Zbl

[Matsumoto 1986] **Y Matsumoto**, *An elementary proof of Rochlin's signature theorem and its extension by Guillou and Marin*, from "À la recherche de la topologie perdue" (L Guillou, A Marin, editors), Progr. Math. 62, Birkhäuser, Boston, MA (1986) 119–139 MR Zbl

[Norman 1969] **R A Norman**, *Dehn's lemma for certain* 4*–manifolds*, Invent. Math. 7 (1969) 143–147 MR Zbl

[Pichelmeyer 2020] **J Pichelmeyer**, *Genera of knots in the complex projective plane*, J. Knot Theory Ramifications 29 (2020) art. id. 2050081 MR Zbl

[Powell and Ray 2021a] **M Powell**, **A Ray**, *Basic geometric constructions*, from "The disc embedding theorem" (S Behrens, B Kalmár, M H Kim, M Powell, A Ray, editors), Oxford Univ. Press (2021) 217–226 MR

[Powell and Ray 2021b] **M Powell**, **A Ray**, *Intersection numbers and the statement of the disc embedding theorem*, from "The disc embedding theorem" (S Behrens, B Kalmár, M H Kim, M Powell, A Ray, editors), Oxford Univ. Press (2021) 155–170 MR Zbl

[Powell et al. 2020] **M Powell**, **A Ray**, **P Teichner**, *The* 4*–dimensional disc embedding theorem and dual spheres*, preprint (2020) arXiv 2006.05209

[Quinn 1982] **F Quinn**, *Ends of maps, III: Dimensions* 4 *and* 5, J. Differential Geom. 17 (1982) 503–521 MR Zbl

[Rokhlin 1952] **V A Rokhlin**, *New results in the theory of four-dimensional manifolds*, Doklady Akad. Nauk SSSR 84 (1952) 221–224 MR Zbl In Russian

[Rokhlin 1972] **V A Rokhlin**, *Proof of Gudkov's hypothesis*, Funct. Anal. Appl. 6 (1972) 136–138 Zbl

[Schneiderman 2003] **R Schneiderman**, *Algebraic linking numbers of knots in* 3*–manifolds*, Algebr. Geom. Topol. 3 (2003) 921–968 MR Zbl

[Schneiderman and Teichner 2001] **R Schneiderman**, **P Teichner**, *Higher order intersection numbers of* 2*–spheres in* 4*–manifolds*, Algebr. Geom. Topol. 1 (2001) 1–29 MR Zbl

[Stong 1994] **R Stong**, *Existence of π_1–negligible embeddings in 4–manifolds: a correction to Theorem 10.5 of* [Freedman and Quinn 1990], Proc. Amer. Math. Soc. 120 (1994) 1309–1314 MR Zbl

[Tristram 1969] **A G Tristram**, *Some cobordism invariants for links*, Proc. Cambridge Philos. Soc. 66 (1969) 251–264 MR Zbl

[Viro 1975] **O Y Viro**, *Positioning in codimension 2, and the boundary*, Uspehi Mat. Nauk 30 (1975) 231–232 MR Zbl In Russian

[Wall 1970] **C T C Wall**, *Surgery on compact manifolds*, Lond. Math. Soc. Monogr. 1, Academic, London (1970) MR Zbl

[Yasuhara 1991] **A Yasuhara**, *$(2,15)$–torus knot is not slice in $\mathbb{C}P^2$*, Proc. Japan Acad. Ser. A Math. Sci. 67 (1991) 353–355 MR Zbl

[Yasuhara 1992] **A Yasuhara**, *On slice knots in the complex projective plane*, Rev. Mat. Univ. Complut. Madrid 5 (1992) 255–276 MR Zbl

School of Mathematical Sciences, University of Southampton
Southampton, United Kingdom

School of Mathematics and Statistics, University of Glasgow
Glasgow, United Kingdom

Max Planck Institute for Mathematics
Bonn, Germany

Max Planck Institute for Mathematics
Bonn, Germany

d.kasprowski@soton.ac.uk, mark.powell@glasgow.ac.uk, aruray@mpim-bonn.mpg.de, teichner@mpim-bonn.mpg.de

Proposed: Ciprian Manolescu
Seconded: Anna Wienhard, Mladen Bestvina

Received: 4 October 2022
Revised: 28 February 2023

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.

The normal language used by the journal is English. Articles written in other languages are acceptable, provided your chosen editor is comfortable with the language and you supply an additional English version of the abstract.

Preparing your article for Geometry & Topology

At the time of submission you need only supply a PDF file. Once accepted for publication, the paper must be supplied in LaTeX, preferably using the journal's class file. More information on preparing articles in LaTeX for publication in GT is available on the GT website.

`arXiv` papers

If your paper has previously been deposited on the `arXiv`, we will need its `arXiv` number at acceptance time. This allows us to deposit the DOI of the published version on the paper's `arXiv` page.

References

Bibliographical references should be listed alphabetically at the end of the paper. All references in the bibliography should be cited at least once in the text. Use of BibTeX is preferred but not required. Any bibliographical citation style may be used, but will be converted to the house style (see a current issue for examples).

Figures

Figures, whether prepared electronically or hand-drawn, must be of publication quality. Fuzzy or sloppily drawn figures will not be accepted. For labeling figure elements consider the pinlabel LaTeX package, but other methods are fine if the result is editable. If you're not sure whether your figures are acceptable, check with production by sending an email to graphics@msp.org.

Proofs

Page proofs will be made available to authors (or to the designated corresponding author) in PDF format. Failure to acknowledge the receipt of proofs or to return corrections within the requested deadline may cause publication to be postponed.

GEOMETRY & TOPOLOGY

Volume 28 Issue 5 (pages 1995–2482) 2024